

OMNI^{VR}: JOINT AUDIO-VIDEO CONDITIONAL GENERATION FOR ARCHIVAL FOOTAGE RESTORATION

Xin Lu[†], Zihao Fan[†], Jie Huang[‡], Mingchen Zhong, Hexin Zhang, Xueyang Fu[✉], Zheng-Jun Zha

University of Science and Technology of China [‡]Project leader [✉]Corresponding author

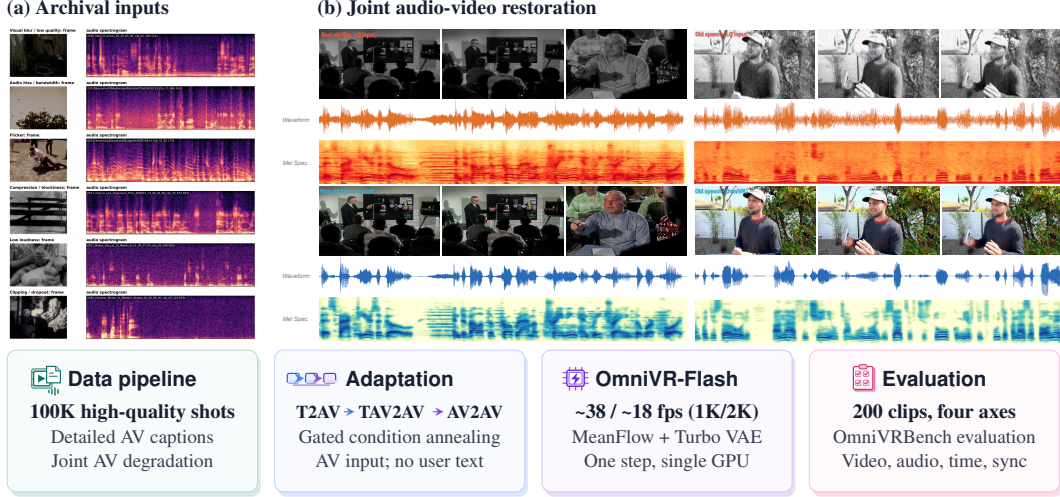


Figure 1: **OmniVR: the first systematic framework for joint audio-video restoration.** (a) Degraded archival footage and (b) paired restoration examples with video frames, waveforms, and mel spectrograms (AV denotes audio-video). The system connects AV captioning and paired degradation, T2AV-to-TAV2AV-to-AV2AV adaptation, one-step 1K/2K inference, and joint evaluation. AV2AV requires no user text; a fixed restoration prompt is retained internally.

ABSTRACT

Archival footage often suffers from coupled visual and acoustic degradations, yet most restoration systems process the two modalities separately. To address this problem, we present **OmniVR**, the first systematic framework for joint audio-video restoration, covering data construction, model adaptation, efficient inference, and evaluation. We construct a high-quality audio-video corpus with detailed captions and use a joint degradation pipeline to produce aligned clean and degraded pairs. Using these pairs, we adapt a pretrained text-to-audio-video model (T2AV) by introducing degraded audio-video conditions (TAV2AV), then progressively replace sample captions with a fixed restoration prompt while retaining caption/null rehearsal. The resulting AV2AV model requires no user-provided text. Under a compatible residual-learning model, we prove that this condition-annealing schedule reduces gradient variance and expected restoration risk relative to direct fixed-prompt adaptation at the same training budget. For efficient deployment, **OmniVR-Flash** combines reduced-resolution video conditioning, MeanFlow-based one-step distillation, and Turbo VAE, achieving approximately 38 fps at 1K and 18 fps at 2K on a single B200 GPU. We further introduce **OmniVRBench** to evaluate four complementary dimensions: visual quality, audio quality, temporal consistency, and audio-visual synchrony. OmniVR achieves state-of-the-art results on public benchmarks and OmniVRBench. Data, code, and model weights will be released.

Project Page: https://xinlu.github.io/OminiVR_PAGE/

1 INTRODUCTION

Archival footage is a joint record: visual defects such as blur, fading, scratches, and instability often occur together with hiss, clipping, and missing acoustic bands. Restoring it therefore requires recovering both streams while preserving the events, identities, and timing in the source. Existing visual restoration methods address archival-footage artifacts and temporal degradation (Wan et al., 2022; Mao et al., 2025), while audio restoration systems recover speech and waveform detail with neural synthesis (Kong et al., 2020; 2021; Liu et al., 2022). Most existing systems still optimize video and audio separately, so they cannot use visible articulation to resolve acoustic ambiguity or soundtrack evidence to constrain visual reconstruction. Our diagnostics show that the two degradations co-occur in collected material and that the audio and video pathways interact inside a joint generator; muting one modality changes the response of the other (Figure 2). Restoration is thus a coupled conditional-generation problem rather than two independent enhancement tasks.

A pretrained audio-video generator supplies a strong prior for this task, as demonstrated by recent joint audiovisual generation models (Kondratyuk et al., 2024; Polyak et al., 2024; HaCohen et al., 2026), but three constraints must be addressed together. Real clean/degraded archival pairs are scarce, so restoration supervision must be synthesized from aligned audio-video material. The generator is trained to follow captions, whereas restoration should be controlled by the observed low-quality pair and should not require a new prompt at inference. Meanwhile, repeated denoising and high-resolution encoding and decoding make a large joint model costly on long recordings; few-step flow formulations target this bottleneck (Liu et al., 2023; Geng et al., 2025a). These constraints couple the data interface, conditioning transition, and deployment path; addressing only one leaves the restoration pipeline incomplete.

We present **OmniVR**, the first systematic framework for joint audio-video restoration (Figure 1). Our data pipeline (Figure 3) constructs a corpus of 100K high-quality, synchronized 3–20 s audio-video shots from Internet recordings using shot segmentation and quality and safety filtering. Gemini combines each shot with audio, ASR, and metadata to produce a structured audiovisual caption. Online degradation yields aligned clean/LQ pairs with severity metadata; shared shot identifiers also support high-resolution refiner supervision. These records connect captioning, degraded conditions, and restoration targets.

Starting from the 22B-parameter LTX-2 backbone (HaCohen et al., 2026), we follow the staged path in Figure 4: T2AV preserves the pretrained caption/null CFG interface, TAV2AV introduces degraded audio-video conditions, and AV2AV gradually replaces sample captions with a fixed restoration prompt. Independent 90% retention masks for text and LQ conditions preserve caption/null rehearsal as gated annealing shifts control toward the observed pair. Inference requires no user text. Under a compatible residual-learning model, this schedule lowers gradient variance and strictly reduces expected restoration risk relative to direct fixed-prompt adaptation at a finite training budget; our analysis states the compatibility, orthogonality, and curvature assumptions. First-frame anchoring, degradation-aware weights, and waveform supervision support long-video restoration.

For efficient deployment, **OmniVR-Flash** combines reduced-resolution video conditioning, Mean-Flow one-step distillation, and Turbo VAE (Figure 4). These components reduce conditioning, denoising, and encoding/decoding costs while retaining the joint restoration interface. On a single B200 GPU, throughput reaches approximately 38 fps at 1K and 18 fps at 2K, exceeding real-time (>24 fps) at 1K. Table 23 reports per-component wall time.

We evaluate OmniVR on the RTN benchmark (Wan et al., 2022) and **OmniVRBench**, covering visual and audio quality, temporal consistency, and audio-video synchrony across controlled degradations and real archival footage. OmniVR achieves the best reported no-reference visual scores on RTN and the strongest joint-quality results in the main OmniVRBench comparisons. On the real archival track, MUSIQ reaches 61.87 versus 52.38 for the strongest visual baseline, while DNSMOS improves from 2.04 for the degraded input to 2.43. The data, code, and model weights will be released.

2 RELATED WORK

Archival images, films, and video restoration. Restoring archival media combines inverse-problem reconstruction with the need to preserve historically meaningful content. Early image

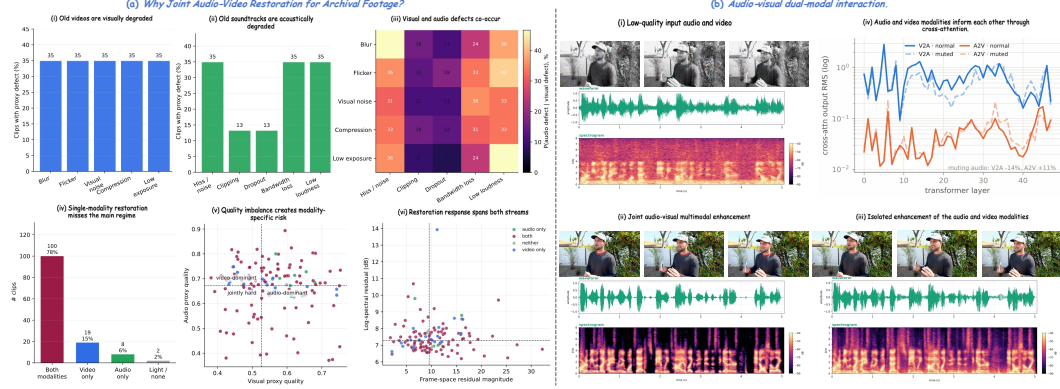


Figure 2: **Why joint audio-video restoration?** (a) Diagnostics show pervasive visual and acoustic degradation (i, ii), cross-modal co-occurrence (iii), predominantly joint degradation (iv), quality imbalance (v), and restoration needs in both streams (vi). (b) Audio and video attend to each other (iv); muting audio alters the video response (i-iii), showing cross-modal interaction.

systems focused on scratches, fading, blur, and color loss in old photographs, while later methods introduced learned priors and temporal propagation for moving footage (Wan et al., 2020; 2022). DeepRemaster treats restoration and color transfer as a temporally coherent process, and more recent work models spatially varying and time-varying artifacts in old films (Iizuka & Simo-Serra, 2019; Mao et al., 2025). These systems established two principles that also matter for audiovisual restoration: degradations should be modeled as structured processes rather than independent pixel noise, and temporal consistency is as important as single-frame perceptual quality.

General video restoration has developed complementary architectures for propagation and alignment. EDVR aggregates aligned neighboring frames, BasicVSR++ uses recurrent propagation, and transformer-based models such as VRT and RVRT model temporal dependencies with attention (Wang et al., 2019; Chan et al., 2022a; Liang et al., 2024; 2022). RealBasicVSR studies the gap between synthetic and real degradations, and VideoGigaGAN uses a generative adversarial prior to recover high-frequency details (Chan et al., 2022b; Xu et al., 2025). Colorization systems including DDColor, BiSTNet, and ColorMNet show that semantic priors can restore plausible chroma when observations are grayscale or severely faded (Kang et al., 2023; Yang et al., 2024b;a). However, these visual systems generally treat the soundtrack as external; they cannot use speech articulation, music rhythm, or sound events to disambiguate visual reconstruction.

Audio, speech, and waveform restoration. Audio restoration covers denoising, dereverberation, bandwidth extension, declipping, and recovery of missing harmonic structure. HiFi-GAN demonstrates efficient adversarial waveform generation, DiffWave formulates waveform synthesis with diffusion, and VoiceFixer combines a spectral restoration front end with a HiFi-GAN neural vocoder for heterogeneous corruptions, operating in a spectrogram analysis-resynthesis paradigm (Kong et al., 2020; 2021; Liu et al., 2022). These methods are strong at modality-specific fidelity, but normally receive only audio and have no mechanism to enforce agreement with visible mouth motion, scene dynamics, or visual event timing. For archival recordings, this limitation matters because hiss, clipping, missing bands, and frame defects often arise from the same capture or transcoding process.

Cross-modal representation and audiovisual synchronization. A separate line of work learns shared representations or explicit mappings between speech and visual evidence. VisualVoice uses visual context to separate and synthesize speech, AV-HuBERT learns audiovisual speech representations, and Wav2Lip aligns generated lip motion with a target waveform (Gao & Grauman, 2021; Shi et al., 2022; Prajwal et al., 2020). These models demonstrate that the two streams carry complementary information, but they are specialized for speech-centric understanding or lip synchronization rather than open-domain restoration. Broader audiovisual benchmarks evaluate quality and synchronization jointly, making clear that a system can produce plausible video and plausible audio while still violating cross-modal timing (Yang et al., 2026; Zhou et al., 2026).

OmniVR uses the same interaction principle for restoration: audio and video are encoded as a joint condition and reconstructed together, while temporal and synchrony metrics are evaluated explicitly.

Audiovisual generation and multimodal priors. Large generative models provide a useful prior for recovering details absent from a degraded observation. Diffusion models and latent diffusion establish score-based reconstruction, rectified flow gives a deterministic transport view, and diffusion transformers scale this paradigm to long token sequences (Ho et al., 2020; Rombach et al., 2022; Liu et al., 2023; Peebles & Xie, 2023). In visual restoration, Real-ESRGAN learns from synthetic high-order corruptions, whereas StableSR, DiffBIR, and SUPIR adapt pretrained generative priors to restoration conditions (Wang et al., 2021; 2024; Lin et al., 2024; Yu et al., 2024). Their success also exposes a central restoration trade-off: an unconstrained prior can improve perceptual realism while changing the observed identity, geometry, or event, so the condition must carry enough evidence to prevent hallucination.

Recent foundation models extend these ideas from images or silent video to joint audio-video generation. VideoPoet and Movie Gen model synchronized visual and acoustic content at scale, while LTX-2 provides a large joint audiovisual backbone with a text-conditioned interface (Kondratyuk et al., 2024; Polyak et al., 2024; HaCohen et al., 2026). These models are trained primarily for synthesis from captions; restoration instead requires fidelity to a particular observed pair and eventually operation without a user-supplied caption. OmniVR therefore keeps the pretrained caption/null interface during the initial T2AV stage, introduces degraded audio-video conditions in TAV2AV, and anneals toward an AV2AV interface whose fixed restoration prompt is internal. This staged transition connects the synthesis prior to a condition-faithful restoration objective.

Conditioning, flow matching, and efficient inference. Conditioning strategies determine whether a generative model follows semantic intent, observed evidence, or both. Text conditioning provides a broad semantic prior, while low-quality audiovisual conditioning supplies instance-specific information; using only one source can either underconstrain reconstruction or discard the pretrained generative capability. Our condition-annealing schedule is specialized to the T2AV-to-TAV2AV-to-AV2AV transition and is accompanied by a finite-budget risk analysis. It retains caption/null rehearsal while progressively increasing reliance on the degraded pair, so the endpoint can use an internal fixed prompt without requiring new text from the user.

For deployment, flow matching represents generation as integration of a learned velocity field, and rectified-flow formulations reduce the number of solver evaluations (Liu et al., 2023). MeanFlow models an interval-averaged velocity and supports one-step prediction, which makes it suitable for distilling a multi-step restoration model (Geng et al., 2025a). In parallel, lower-resolution condition encoding and efficient VAE modules reduce the cost of processing long high-resolution clips. OmniVR-Flash combines these ideas: the restoration model is distilled with an interval objective, the video condition is encoded at reduced resolution, and Turbo VAE lowers latent encoding and decoding overhead while preserving high-resolution joint outputs.

Evaluation protocols and the position of OmniVR. Video-generation evaluation has expanded from frame fidelity to perceptual quality, motion stability, semantic alignment, and long-range consistency. DOVER, EvalCrafter, and VBench++ provide complementary no-reference and generation-oriented measurements (Wu et al., 2023; Liu et al., 2024; Huang et al., 2024). Audiovisual evaluation adds waveform quality and synchronization, while restoration requires a further distinction between perceptual plausibility and source fidelity. OmniVRBench therefore separates visual quality, audio quality, temporal consistency, and audio-visual synchrony, and includes both a controlled track with clean references and a real archival track without references. Taken together, the prior literature supplies strong modality-specific components and generative priors, but does not provide a complete open-domain system that constructs paired audiovisual supervision, transitions a text-to-audiovisual generator to an observation-only restoration interface, and accelerates the resulting model to real-time high-resolution inference. OmniVR is designed to fill this systems gap.

3 METHODOLOGY

OmniVR constructs paired supervision, adapts a joint generator through condition annealing, and accelerates inference with OmniVR-Flash. Figures 3 and 4 summarize this pipeline.

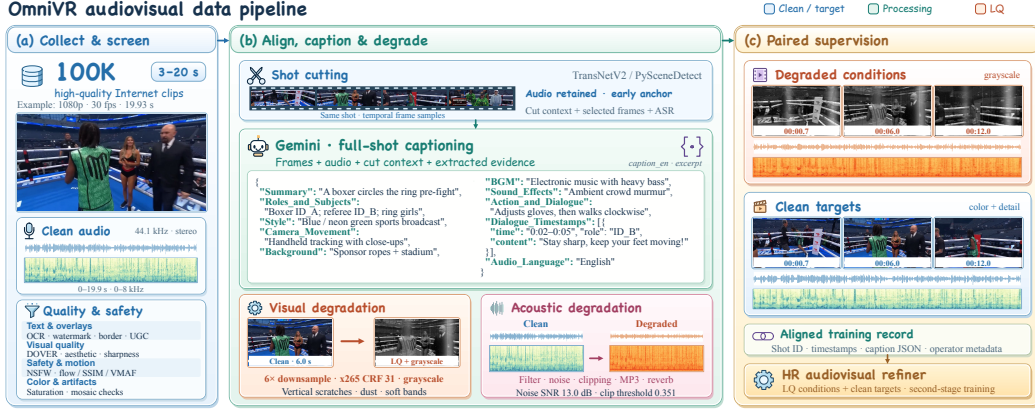


Figure 3: **Shot-aligned audiovisual data pipeline.** Synchronized 3–20 s shots pass the listed quality/safety operators; Gemini combines shot context, frames, and audio evidence into one structured caption. Caption JSON, severity metadata, and clean/LQ pairs remain aligned for TAV2AV training and high-resolution refiner supervision. The boxing example uses archived frames and measured waveforms/spectra; the degraded video is additionally grayscaled for visualization.

3.1 DATA CONSTRUCTION AND JOINT DEGRADATION

Captioned training corpus. We construct a corpus of 100K synchronized 3–20 s shots from high-quality Internet audio-video recordings using TransNetV2/PySceneDetect, retaining each shot’s audio and a clear annotation frame from its first two seconds. After quality and safety filtering (Figure 3), Gemini uses the shot, ASR evidence, and metadata to describe entities, actions, camera motion, dialogue, music, and sound effects. Each record contains clean video $\mathbf{V}^h$, audio $\mathbf{A}^h$, caption y , and shot metadata. The identifiers also support high-resolution refiner supervision; details appear in Sec. 13.

Paired degradation. An online operator synthesizes the degraded conditions from each clean pair:

$$(\mathbf{V}^l, \mathbf{A}^l) = \mathcal{D}(\mathbf{V}^h, \mathbf{A}^h). \quad (1)$$

Independent stochastic video and audio branches produce visual-dominant, audio-dominant, and jointly degraded samples while preserving alignment. The visual branch extends Real-ESRGAN-style high-order degradation (Wang et al., 2021) with temporally coherent flicker, scratches, dust, abrasion bands, and grayscale conversion. The audio branch models hiss, rumble, clicks, clipping, bandwidth loss, spectral over-smoothing, harmonic thinning, and phase discontinuities. Each record is $(y, \mathbf{V}^l, \mathbf{A}^l; \mathbf{V}^h, \mathbf{A}^h)$, with severity scalars $d_v, d_a \in [0, 1]$ used in Eq. 13.

3.2 JOINT RESTORATION MODEL AND CONDITIONAL INTERFACE

Latent encoding and flow matching. We adapt the pretrained joint DiT to predict clean audio-video latents from degraded observations. For modality $m \in \{v, a\}$, let $\mathbf{X}_v = \mathbf{V}$ and $\mathbf{X}_a = \mathbf{A}$. Frozen VAE encoders E_m and patchifiers P_m produce target and condition tokens:

$$\mathbf{x}_m = P_m(E_m(\mathbf{X}_m^h)), \quad \mathbf{c}_m = P_m(E_m(\mathbf{X}_m^l)). \quad (2)$$

In the base model, $\mathbf{x}_m, \mathbf{c}_m \in \mathbb{R}^{N_m \times 128}$. Video uses $N_v = T_z H_z W_z$ latent-grid tokens, where T_z, H_z, W_z are the temporal depth, height, and width of the latent grid. Audio produces N_a tokens by packing 8 channels and 16 mel bins into each 128-dimensional token. Rectified-flow training (Liu et al., 2023) uses a shared $\sigma \sim \mathcal{U}(0, 1)$ and independent noise $\epsilon_m \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$:

$$\tilde{\mathbf{x}}_m = (1 - \sigma)\mathbf{x}_m + \sigma\epsilon_m, \quad \mathbf{u}_m = \epsilon_m - \mathbf{x}_m. \quad (3)$$

Channel-concat condition injection. We perturb the LQ tokens to discourage direct copying, then concatenate each with its target token before projection:

$$\tilde{\mathbf{c}}_m = \mathbf{c}_m + \rho_m \boldsymbol{\eta}_m, \quad \boldsymbol{\eta}_m \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \mathbf{h}_{m,i} = W_m[\tilde{\mathbf{x}}_{m,i}; \tilde{\mathbf{c}}_{m,i}] + \mathbf{b}_m. \quad (4)$$

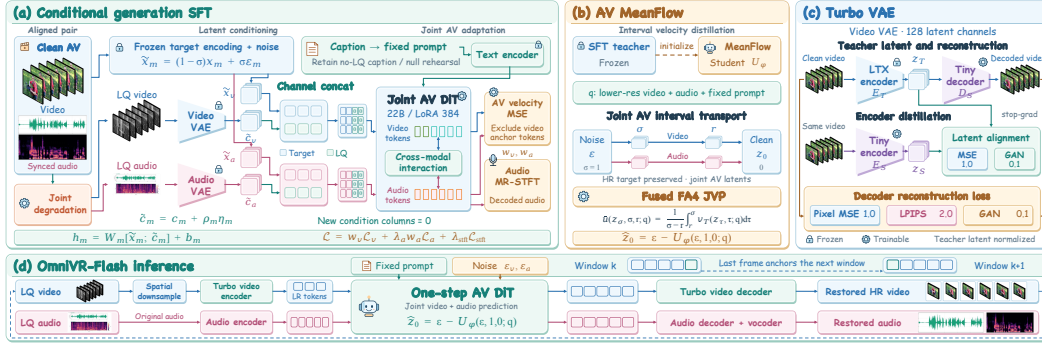


Figure 4: **OmniVR and OmniVR-Flash framework.** (a) Conditional SFT adapts a joint audiovisual DiT with degraded inputs and caption annealing to a fixed prompt. (b) AV MeanFlow enables one-step prediction. (c) Turbo VAE accelerates video encoding and decoding. (d) OmniVR-Flash downsamples LQ video before condition encoding and DiT injection, reducing conditioning cost while preserving high-resolution outputs. The last restored frame anchors the next window.

Here i indexes tokens, $[\cdot; \cdot]$ denotes channel concatenation, $W_m \in \mathbb{R}^{d \times 256}$ maps each concatenated token to the DiT hidden dimension d , and b_m is the corresponding bias. The first 128 columns of W_m copy the pretrained projection and the new condition columns are zero-initialized, preserving the initial T2AV mapping. We sample $\rho_m \sim \mathcal{U}(0.4, 0.6)$ during training and use $\rho = 0.5$ at inference.

Joint prediction. The multimodal DiT with parameters θ predicts both velocity fields using the projected tokens, text embedding e , and shared noise level:

$$(\hat{\mathbf{u}}_v, \hat{\mathbf{u}}_a) = G_\theta(\mathbf{h}_v, \mathbf{h}_a, e, \sigma). \quad (5)$$

Cross-modal interactions let each stream use evidence from the other. With full condition $\mathbf{q} = (\mathbf{c}_v, \mathbf{c}_a, e)$, we transfer control from captions to the LQ pair while retaining the text pathway.

3.3 CONDITION ANNEALING FROM T2AV TO AV2AV INFERENCE

Three-stage condition schedule. The transition starts from the pretrained T2AV caption/null interface, introduces LQ conditions during caption-assisted TAV2AV adaptation, and anneals captions to a fixed restoration prompt for AV2AV inference. Throughout adaptation, independent masks $A_k \sim \text{Bernoulli}(p)$ and $T_k \sim \text{Bernoulli}(t)$ retain the LQ pair and non-null text, respectively, with $p = t = 0.9$. The LQ mask gates the new condition projection. Initially, the (caption, LQ), (caption, no-LQ), (null, LQ), and (null, no-LQ) regimes have probabilities 0.81, 0.09, 0.09, and 0.01.

We anneal non-null text only when LQ is present. Let $\mathbf{e}_y$, $\mathbf{e}_{\text{rest}}$, and $\mathbf{e}_\emptyset$ encode the caption, fixed restoration prompt, and null text. An independent selector $M_k \sim \text{Bernoulli}(\lambda_k)$ gives:

$$\lambda_k = \max\left(1 - \frac{k}{K_{\text{ann}}}, 0\right), \quad \mathbf{e}_k = \begin{cases} \mathbf{e}_\emptyset, & T_k = 0, \\ \mathbf{e}_{\text{rest}}, & T_k = 1, A_k = 1, M_k = 0, \\ \mathbf{e}_y, & \text{otherwise.} \end{cases} \quad (6)$$

Here k indexes adaptation updates and K_{ann} spans the first 30% of training. The 10% no-LQ branch continues caption/null rehearsal throughout training. At $\lambda_k = 0$, the LQ-present branch uses the fixed prompt with probability 0.9 and null text with probability 0.1. At inference, the user supplies only the degraded pair, and the model uses $\mathbf{q}_{\text{inf}} = (\mathbf{c}_v, \mathbf{c}_a, \mathbf{e}_{\text{rest}})$ internally.

Training objective and finite-budget gain. Let $J_{\text{cap}}^0, J_\emptyset^0$ be expected no-LQ losses, and $J_{\text{cap}}^{\text{LQ}}, J_{\text{rest}}^{\text{LQ}}, J_\emptyset^{\text{LQ}}$ their LQ-present counterparts. Each uses the severity-aware objective in Eq. 14. Define $P = tJ_{\text{cap}}^0 + (1-t)J_\emptyset^0$ for T2AV rehearsal, $C = tJ_{\text{cap}}^{\text{LQ}} + (1-t)J_\emptyset^{\text{LQ}}$ for caption-assisted restoration, and $R = tJ_{\text{rest}}^{\text{LQ}} + (1-t)J_\emptyset^{\text{LQ}}$ for fixed-prompt restoration. The exact expected loss is:

$$F_\lambda = (1-p)P + p[\lambda C + (1-\lambda)R]. \quad (7)$$

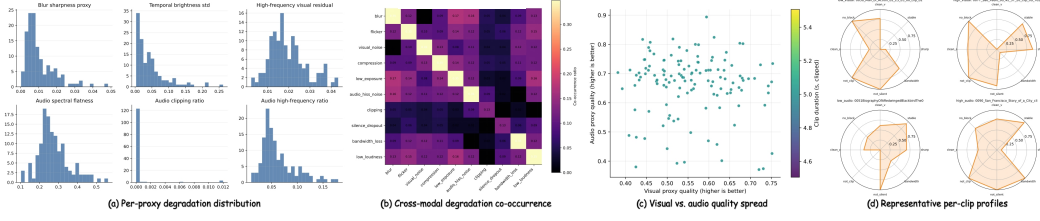


Figure 5: **OmniVRBench degradation characterization.** (a) Distributions of three visual and three audio degradation proxies. (b) Co-occurrence of ten defect tags: 77.5% of clips have both visual and acoustic degradation. (c) Weakly correlated visual/audio quality scores motivate joint evaluation. (d) Eight-axis profiles for four clips with low/high visual or audio quality illustrate degradation diversity.

The degradation weights depend on observed severities, independently of the condition masks, so they preserve these mixture coefficients. The constant rehearsal term retains training pressure on the original text interface as the restoration objective shifts from C to R .

Under the compatible scalar residual model of Proposition 4 (under Assumption 3, which makes the LQ feature independent of the residual variables and requires the noise to have zero conditional mean given the residual and its caption estimate), caption assistance lowers gradient variance at the same mean contraction rate. At fixed initialization, step size $0 < \eta < 2$, retention probabilities, and N updates, the expected risk gain over direct fixed-prompt training is:

$$\mathbb{E}[\mathcal{R}_R(w_N^{\text{dir}}) - \mathcal{R}_R(w_N^{\text{ann}})] = \frac{pt\eta^2\kappa}{2} \sum_{k=0}^{N-1} \lambda_k \rho^{N-1-k} \geq 0, \quad \rho = 1 - p\eta(2 - \eta). \quad (8)$$

Here w_N is the learned scalar LQ coefficient, $\mathcal{R}_R$ is its fixed-prompt restoration risk, $\kappa > 0$ measures caption compatibility with residual variation orthogonal to the LQ feature, and ρ denotes the per-step contraction rate (distinct from the condition noise scale ρ_m in Eq. 4). For $p = t = 0.9$, the gain is strict when any $\lambda_k > 0$. Sec. 14 gives the assumptions, the severity-weighted extension, and the separate comparison with an always-LQ baseline. It also establishes local prior-retention and stationary-path tracking bounds under explicit curvature assumptions; these bounds describe the stated residual and gradient-descent models.

AV2AV guidance. At inference, classifier-free guidance compares the fixed restoration prompt with null text while retaining the same LQ pair $\mathbf{c} = (\mathbf{c}_v, \mathbf{c}_a)$ in both branches. Writing $\mathbf{v}_\theta$ for the velocity field of G_θ :

$$\mathbf{v}_{\text{cfg}}(\mathbf{e}; \mathbf{c}) = \mathbf{v}_\theta(\mathbf{e}_\emptyset; \mathbf{c}) + w [\mathbf{v}_\theta(\mathbf{e}; \mathbf{c}) - \mathbf{v}_\theta(\mathbf{e}_\emptyset; \mathbf{c})], \quad \mathbf{e} = \mathbf{e}_{\text{rest}}. \quad (9)$$

Here w is the guidance scale and the shared latent and noise-level arguments are omitted. This AV2AV interface requires no user-provided text. The no-LQ rehearsal branch retains the caption/null interface for T2AV generation.

3.4 TEMPORAL CONSISTENCY AND OPTIMIZATION

The conditional model is trained with a boundary anchor for long-video continuity and multimodal losses that account for uneven degradation severity.

First-frame I2V conditioning. OmniVR processes 121-frame windows. Training enables first-frame conditioning with probability $p_{\text{ff}} = 0.5$. Let $\mathbf{M} \in \{0, 1\}^{N_v}$ select the first $H_z W_z$ video tokens for anchored samples and be zero otherwise. We replace selected tokens with clean targets ($\odot$ denotes element-wise multiplication):

$$\tilde{\mathbf{x}}_v \leftarrow \mathbf{M} \odot \mathbf{x}_v + (1 - \mathbf{M}) \odot \tilde{\mathbf{x}}_v. \quad (10)$$

Anchor tokens are excluded from the loss. The first inference window runs without anchoring, a mode sampled with probability $1 - p_{\text{ff}} = 0.5$ during training. Subsequent windows use the previous window’s restored last frame, enabling bounded-memory processing. Sec. 15 analyzes the shift from clean training anchors to generated inference anchors.

Degradation-aware velocity losses. The video loss averages over non-anchor tokens, while the audio loss uses all N_a tokens:

$$\mathcal{L}_v = \frac{\sum_i (1 - \mathbf{M}_i) \|\hat{\mathbf{u}}_{v,i} - \mathbf{u}_{v,i}\|_2^2}{\sum_i (1 - \mathbf{M}_i)}, \quad (11) \quad \mathcal{L}_a = \frac{1}{N_a} \sum_i \|\hat{\mathbf{u}}_{a,i} - \mathbf{u}_{a,i}\|_2^2. \quad (12)$$

For each sample, $d_v, d_a \in [0, 1]$ average the normalized degradation parameters within each modality. We assign modality weights:

$$w_v = \text{Norm}_B(1 + \alpha_v d_v + \beta_v d_a + \gamma_v d_v d_a), \quad w_a = \text{Norm}_B(1 + \alpha_a d_a + \beta_a d_v + \gamma_a d_v d_a), \quad (13)$$

where Norm_B normalizes each modality’s weights to unit mean over the 64-sample global batch using an all-reduce. We use $\alpha_v = \alpha_a = 1.0$, $\beta_v = \beta_a = 0.5$, and $\gamma_v = \gamma_a = 2.0$; the interaction term emphasizes jointly severe degradation. Global normalization preserves sample-dependent weights despite one-sample GPU micro-batches. Bounds and observed weight ranges appear in Sec. 16.7.

Waveform supervision and total loss. A random audio sub-batch is decoded through the frozen audio VAE and vocoder. Multi-resolution STFT loss $\mathcal{L}_{\text{stft}}$ compares the reconstructed and clean waveforms using window sizes 512, 1024, and 2048. We combine the modality-weighted velocity losses with this spectral term:

$$\mathcal{L} = w_v \mathcal{L}_v + \lambda_a w_a \mathcal{L}_a + \lambda_{\text{stft}} \mathcal{L}_{\text{stft}}, \quad (14)$$

with $\lambda_a = 1.0$ and $\lambda_{\text{stft}} = 3 \times 10^{-3}$. The spectral term supervises decoded acoustic detail and remains additive, without multiplication by w_a . This objective covers all condition regimes in Eq. 7.

3.5 EFFICIENT DEPLOYMENT WITH OMNIVR-FLASH

OmniVR-Flash combines reduced-resolution video conditioning, AV MeanFlow distillation, and Turbo VAE (Figure 4). It retains the fixed-prompt AV2AV interface and reaches approximately 38 fps at 1K and 18 fps at 2K on one B200 GPU.

Reduced-resolution video conditioning. We downsample and encode LQ video, then inject the condition through a single learned DiT layer. The target retains its high-resolution grid and audio conditioning is unchanged. Condition resolution trades source detail for efficiency.

AV MeanFlow distillation. Using the conditional restoration model as a frozen teacher, we train a student to predict finite-interval transport of the joint audio-video latent. Let $\mathbf{z}_\sigma$ denote the joint latent at noise level σ (corresponding to $\tilde{\mathbf{x}}_m$ in Eq. 3). For fixed condition $\mathbf{q}$, the MeanFlow average velocity (Geng et al., 2025a) over $[r, \sigma]$ is:

$$\bar{\mathbf{u}}(\mathbf{z}_\sigma, \sigma, r; \mathbf{q}) = \frac{1}{\sigma - r} \int_r^\sigma \mathbf{v}_{\text{teacher}}(\mathbf{z}_\tau, \tau; \mathbf{q}) d\tau, \quad 0 \leq r < \sigma \leq 1. \quad (15)$$

Here $\mathbf{v}_{\text{teacher}}$ denotes the frozen teacher’s classifier-free guided velocity at scale 3.0, as specified in Algorithm 6. The student $\mathbf{U}_\phi$, parameterized by ϕ , learns average velocity with the improved MeanFlow (iMF) velocity objective (Geng et al., 2025b), with a 3D adversarial loss on decoded one-step video outputs (Sec. 19.6). We compute the Jacobian–vector product with a fused FlashAttention-4 (FA4) kernel (Zadouri et al., 2026). With the data-to-noise convention in Eq. 3, both modalities take one step to yield the predicted clean latent $\hat{\mathbf{z}}_0$:

$$\hat{\mathbf{z}}_0 = \epsilon - \mathbf{U}_\phi(\epsilon, 1, 0; \mathbf{q}), \quad (16)$$

where ϵ is the initial audio-video noise and $\mathbf{q}$ contains the LQ pair and internal restoration prompt.

Turbo VAE. Following Ultra Flash’s training procedure and code (Luxury et al., 2026), a lightweight encoder aligns with frozen teacher latents, and a lightweight decoder learns from teacher latents using pixel, perceptual, and adversarial losses. Both retain the teacher’s 128-channel video latent interface. At inference, they encode the downsampled LQ video and decode the high-resolution output; the audio VAE and vocoder are unchanged.

Table 1: No-reference results on the RTN degraded archival-footage clips (Wan et al., 2022) (3 sequences, 600 frames, no ground truth), under the unified 640×480 aligned-instant protocol. Best in **bold**, second-best underlined. $\uparrow/\downarrow$: higher/lower is better.

Method	MUSIQ $\uparrow$	CLIP-IQA $\uparrow$	NIQE $\downarrow$	MANIQA $\uparrow$	TOPIQ $\uparrow$	BRISQUE $\downarrow$
Low-quality input	38.71	0.258	5.94	0.188	0.267	44.66
DeepRemaster	39.01	0.263	6.31	0.196	0.273	41.00
RealBasicVSR	<u>52.18</u>	0.362	5.87	0.285	<u>0.438</u>	<u>38.72</u>
DDColor	38.79	0.284	5.69	0.175	0.271	41.49
ColorMNet	38.52	<u>0.378</u>	5.75	0.184	0.266	42.92
MambaOFR	49.83	0.351	<u>5.42</u>	<u>0.312</u>	0.421	39.56
OmniVR (ours)	64.77	0.426	5.14	0.330	0.553	35.90

Table 2: Controlled degradation track (71 clips with clean reference), NR metrics at unified 640×480 . GT is the clean reference, not a competing method. Best in **bold**, second-best underlined.

Method	Visual (NR)						Audio		Sync	
	MUSIQ $\uparrow$	CLIP-IQA $\uparrow$	NIQE $\downarrow$	MANIQA $\uparrow$	TOPIQ $\uparrow$	BRISQUE $\downarrow$	DNSMOS $\uparrow$	FAD $\downarrow$	LSE-C $\uparrow$	LSE-D $\downarrow$
Low-quality input	38.77	0.221	6.69	0.218	0.235	36.69	1.46	7.88	<u>2.32</u>	<u>10.87</u>
DeepRemaster	37.43	0.173	<u>6.20</u>	0.189	0.245	35.83	—	—	—	—
RealBasicVSR	<u>45.54</u>	0.226	7.08	<u>0.294</u>	<u>0.326</u>	44.96	—	—	—	—
VoiceFixer (audio only)	—	—	—	—	—	—	<u>2.12</u>	<u>7.02</u>	1.98	11.53
DDColor	34.07	0.200	6.34	0.200	0.250	<u>33.61</u>	—	—	—	—
ColorMNet	34.64	0.243	6.59	0.228	0.253	35.98	—	—	—	—
MambaOFR	43.49	<u>0.268</u>	6.99	0.280	0.305	42.80	—	—	—	—
OmniVR (ours)	71.17	0.543	4.11	0.487	0.673	24.75	2.70	6.30	3.52	10.43
Clean reference (GT)	67.49	0.558	4.03	0.477	0.623	25.48	2.47	—	4.00	9.44

3.6 IMPLEMENTATION AND INFERENCE SETTINGS

We adapt the 22B LTX-2 joint DiT (HaCohen et al., 2026) with rank-384 LoRA (Hu et al., 2022), keeping the VAEs, text encoder, and vocoder frozen. Base training uses 64 B200 GPUs for approximately five days. Reported base-model results use the 50K-step checkpoint, 15 Euler steps, CFG 3.0, and condition noise $\rho = 0.5$. Training schedules, checkpoint identifiers, optimizer settings, and Flash architecture are specified in Sec. 16.7.

4 EXPERIMENTS

We evaluate the multi-step OmniVR configuration (Sec. 3.6) on RTN and OmniVRBench using automatic metrics, ablations, and human judgments. OmniVR-Flash’s quality–efficiency trade-off is reported in Sec. 17.2.

4.1 OMNIVRBENCH AND EVALUATION PROTOCOL

OmniVRBench contains 129 real archival clips without clean references and 71 controlled-degradation clips with references. It evaluates visual and audio quality, temporal consistency, and audiovisual synchrony using 121-frame windows at 24 fps and 1920×1088 outputs. Visual, acoustic, and joint defects affect 92.2%, 83.7%, and 77.5% of clips, respectively (Figure 5). The tracks test source fidelity and transfer to authentic damage.

4.2 METRICS AND BASELINES

Main comparisons evaluate video at 640×480 using MUSIQ, CLIP-IQA, NIQE, MANIQA, TOPIQ, and BRISQUE; audio using DNSMOS (Reddy et al., 2022) and FAD (Kilgour et al., 2019); and lip synchronization using LSE-C/LSE-D (Prajwal et al., 2020). Sec. 8 reports full-reference visual/audio metrics, temporal warp error, and distributional video evaluation; Sec. 9 examines source consistency and failures.

We compare with RealBasicVSR (Chan et al., 2022b), DeepRemaster (Iizuka & Simo-Serra, 2019), and MambaOFR (Mao et al., 2025) for video restoration; DDColor (Kang et al., 2023) and Col-

Table 3: Real archival track (129 clips, no GT), 640×480 . Best **bold**; second-best underlined.

Method	Visual (NR)						Audio		Sync		
	MUSIQ $\uparrow$	CLIP-IQA $\uparrow$	NIQE $\downarrow$	MANIQA $\uparrow$	TOPIQ $\uparrow$	BRISQUE $\downarrow$	DNSMOS $\uparrow$	FAD $\downarrow$	LSE-C $\uparrow$	LSE-D $\downarrow$	
Low-quality input	36.55	0.311	5.69	0.195	0.248	45.27	2.04	15.93	<u>1.05</u>	<u>12.14</u>	
DeepRemaster	38.78	0.237	7.83	0.191	0.246	48.53	–	–	–	–	
RealBasicVSR	<u>52.38</u>	0.401	5.82	<u>0.342</u>	<u>0.468</u>	<u>38.15</u>	–	–	–	–	
VoiceFixer (audio only)	–	–	–	–	–	–	<u>2.21</u>	<u>9.41</u>	0.87	13.26	
DDColor	37.51	0.307	5.49	0.164	0.278	39.21	–	–	–	–	
ColorMNet	38.64	<u>0.421</u>	5.59	0.174	0.275	42.74	–	–	–	–	
MambaOFR	49.59	0.394	5.62	0.331	0.440	44.00	–	–	–	–	
OmniVR (ours)	61.87	0.444	5.40	0.383	0.531	36.02	2.43	8.32	1.12	11.39	

Table 4: Human preference on 129 archival clips (12 annotators). Cells report mean pairwise 2AFC win rates (%); balanced comparisons average 50% per column, marked by the center tick. The final row gives OmniVR’s gain over the strongest baseline in percentage points. Bootstrap 95% CIs resample clips with annotators fixed.

Method	Visual		Audio		Sync		Overall
LQ input		19.5		24.2		27.0	 21.1
MambaOFR + VoiceFixer		46.7		48.7		48.3	 47.7
RealBasicVSR + VoiceFixer		58.0		54.9		53.5	 56.5
Old Films + VoiceFixer		43.7		43.9		43.8	 44.1
Pretrained AV gen.		39.8		38.9		39.8	 39.6
OmniVR (ours)		92.2		89.4		87.6	 91.0
<i>Gain vs. best baseline</i>	+34.2 pp		+34.5 pp		+34.1 pp		+34.5 pp

orMNet (Yang et al., 2024a) for colorization; and VoiceFixer (Liu et al., 2022) for audio restoration. Sec. 12 reports visual/audio cascade comparisons.

4.3 PUBLIC BENCHMARK VALIDATION

On RTN (Wan et al., 2022) (3 archival sequences, 600 frames), OmniVR leads all six no-reference visual metrics (Table 1); MUSIQ rises from RealBasicVSR’s 52.18 to 64.77. This public benchmark tests video restoration without assessing audio.

4.4 JOINT RESTORATION ON OMNIVRBENCH

On the controlled track (Table 2), OmniVR leads the compared methods across visual, audio, and sync metrics, reaching MUSIQ 71.17, DNSMOS 2.70, and LSE-C 3.52. NR scores above the clean reference do not imply greater source fidelity; Sec. 8 reports full-reference results.

On real archival footage (Table 3), OmniVR achieves MUSIQ 61.87 versus RealBasicVSR’s 52.38. DNSMOS improves from 2.04 to 2.43 and FAD from 15.93 to 8.32, below VoiceFixer’s 9.41. LSE-C/LSE-D reach 1.12/11.39, indicating better synchronization.

4.5 CONDITIONING, OPTIMIZATION, AND HUMAN EVALUATION

Controlled-track ablations support joint audiovisual conditioning and gradual annealing; detailed results appear in Sec. 17.1.

Twelve annotators compare six methods on 129 real clips through randomized pairwise 2AFC (Table 4; protocol in Sec. 16.4). OmniVR’s mean win rates are 92.2% (visual), 89.4% (audio), 87.6% (sync), and 91.0% (overall), exceeding the strongest cascade’s overall rate by 34.5 percentage points. Fleiss’ $\kappa = 0.67$; the overall 95% bootstrap interval is $\pm 3.8\%$. All reported pairwise gaps exceed their confidence intervals.

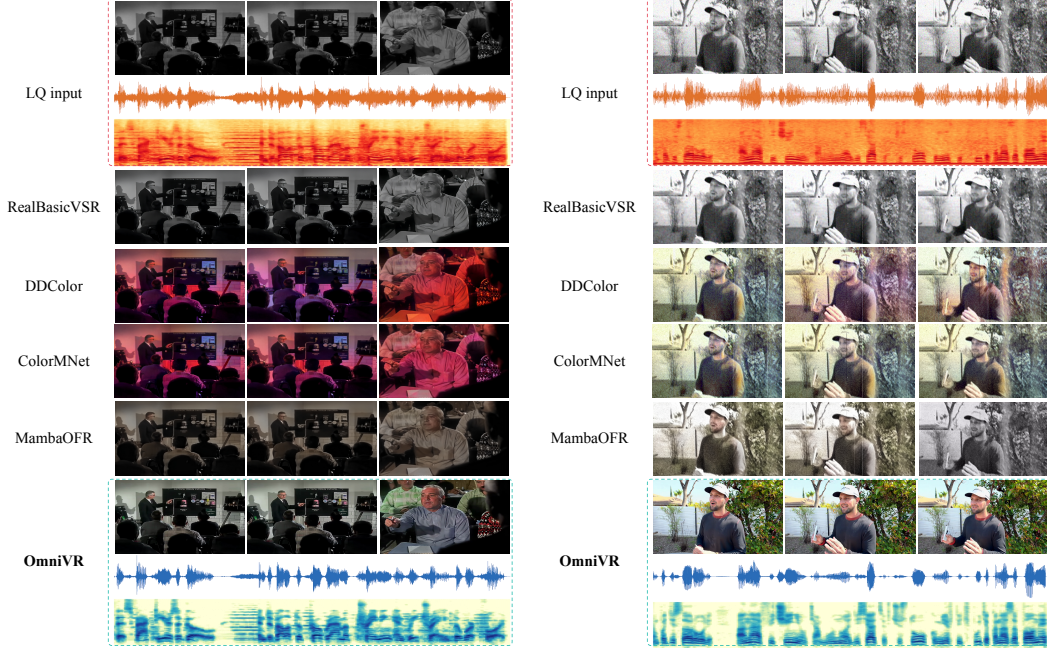


Figure 6: **Qualitative comparison.** Left: controlled degradation with a clean reference; right: authentic archival footage without a reference. Columns show three frames per clip, with restored waveforms and mel spectrograms below. RealBasicVSR/MambaOFR lack colorization; DDColor/ColorMNet show color bias or inconsistency. OmniVR jointly restores color, detail, temporal consistency, and audio. High-resolution comparisons appear in Sec. 18.

4.6 QUALITATIVE ANALYSIS

Figure 6 shows joint recovery of visual detail, color, and audio on real films, whereas specialized baselines can leave inputs grayscale or introduce tonal and temporal inconsistencies. Sec. 9 examines failures in identity, motion, and generated color.

5 CONCLUSION

OmniVR connects paired audiovisual data construction, condition-annealed adaptation, and joint restoration evaluation. Its fixed-prompt interface restores audio and video without user text, while OmniVR-Flash combines reduced-resolution conditioning, one-step MeanFlow, and Turbo VAE for efficient deployment. RTN and OmniVRBench results, source-fidelity analyses, and human judgments support the benefits of joint restoration. Data, code, and model weights will be released.

AI USE STATEMENT

We used large language models (Claude, Gemini) during the preparation of this work. *Required disclosures:* Gemini generated structured training captions from audio-video shots for the data pipeline (Sec. 3.1); these captions serve as T2AV/TAV2AV text conditions and are not used at inference. Claude assisted with code implementation and debugging. No AI tools were used for developing theoretical models, formulating mathematical claims or proofs, proposing hypotheses, designing experiments, or interpreting results. *Recommended disclosures:* Claude and Gemini assisted with revising, and formatting portions of this manuscript. All scientific claims, experimental results, and methodological contributions were designed, implemented, and verified by the authors. We have reviewed all AI-assisted work and take full responsibility for the final content of this paper.

6 LIMITATIONS AND FUTURE WORK

Training relies on synthetic corruptions of high-quality Internet clips, leaving a domain gap to authentic archival damage. Severe information loss can lead to plausible but incorrect texture, color, or acoustic content; restored color in particular should not be treated as evidence of historical ground truth. First-frame chaining provides a practical way to process long recordings, but imperfect generated anchors can accumulate error across windows. OmniVR-Flash accelerates this window-based process, but lower-resolution conditions and one-step prediction can discard information useful for reconstruction. Further work should examine self-generated anchors during training, stronger source-fidelity constraints, and the quality–efficiency trade-offs introduced by reduced-resolution conditioning and one-step prediction.

7 ETHICS AND REPRODUCIBILITY

Use of restored content. OmniVR infers color from contemporary video priors rather than historical ground truth. Severe audio corruption can also lead to plausible but incorrect speech or environmental sounds. Restored outputs should be clearly labeled as computationally enhanced.

Data and human evaluation. The training corpus contains licensed or publicly available clips; no real archival recordings were modified during training. Human evaluators gave informed consent and were paid above the local minimum wage (Sec. 16.4).

Reproducibility. Code, model weights, OmniVRBench data and evaluation scripts, and the degradation pipeline will be released upon publication. Training settings and schedules appear in Sec. 3.6 and Sec. 16.7; data processing and degradation are described in Sec. 3.1 and Sec. 13.

REFERENCES

- Kelvin C. K. Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022a.
- Kelvin C. K. Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Investigating tradeoffs in real-world video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022b.
- Ruohan Gao and Kristen Grauman. Visualvoice: Audio-visual speech separation with cross-modal consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- Zhengyang Geng, Mingyang Deng, Xingjian Bai, J. Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. *arXiv preprint arXiv:2505.13447*, 2025a. URL <https://arxiv.org/abs/2505.13447>.
- Zhengyang Geng, Yiyang Lu, Zongze Wu, Eli Shechtman, J. Zico Kolter, and Kaiming He. Improved Mean Flows: On the challenges of fastforward generative models. *arXiv preprint arXiv:2512.02012*, 2025b. URL <https://arxiv.org/abs/2512.02012>.
- Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, et al. Ltx-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.
- David Hasler and Sabine E. Suesstrunk. Measuring colorfulness in natural images. *Human Vision and Electronic Imaging VIII (SPIE)*, 5007:87–95, 2003.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 6840–6851, 2020.

- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.
- Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, et al. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *arXiv preprint arXiv:2411.13503*, 2024.
- Satoshi Iizuka and Edgar Simo-Serra. Deepremaster: Temporal source-reference attention networks for comprehensive video enhancement. *ACM Transactions on Graphics (TOG)*, 38(6):1–13, 2019.
- Jesper Jensen and Cees H. Taal. An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(11):2009–2022, 2016.
- Xiaoyang Kang, Tao Yang, Wenqi Ouyang, Peiran Ren, Lingzhi Li, and Xuansong Xie. Ddcolor: Towards photo-realistic image colorization via dual decoders. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.
- Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *Proceedings of Interspeech*, 2019.
- Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, Krishna Somandepalli, Hassan Akbari, Yair Alon, Yong Cheng, Josh Dillon, Agrim Gupta, Meera Hahn, Anja Hauth, David Hendon, Alonso Martinez, et al. Videopoet: A large language model for zero-shot video generation. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.
- Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.
- Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R. Hershey. Sdr-half-baked or well done? *arXiv preprint arXiv:1811.02508*, 2018.
- Jingyun Liang, Yuchen Fan, Xiaoyu Xiang, Rakesh Ranjan, Eddy Ilg, Simon Green, Jiezhong Cao, Kai Zhang, Radu Timofte, and Luc Van Gool. Recurrent video restoration transformer with guided deformable attention. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- Jingyun Liang, Jiezhong Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. Vrt: A video restoration transformer. *IEEE Transactions on Image Processing*, 33:2171–2182, 2024. doi: 10.1109/TIP.2024.3372454. URL <https://doi.org/10.1109/TIP.2024.3372454>.
- Xinqi Lin, Jingwen He, Ziyang Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. Diffbir: Toward blind image restoration with generative diffusion prior. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.
- Haohe Liu, Xubo Liu, Qiuqiang Kong, Qiao Tian, Yan Zhao, DeLiang Wang, Chuanzeng Huang, and Yuxuan Wang. VoiceFixer: A unified framework for high-fidelity speech restoration. In *Proc. Interspeech*, 2022.
- Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

- Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- Luxury, Jie Huang, Zihao Fan, Xiaoxiao Ma, Yuming Li, Jun-hao Zhuang, Zeyue Xue, Siming Fu, Haoran Li, Mingchen Zhong, Guohui Zhang, Shichen Ma, Yijun Liu, Jiaqi Shi, Yanwen Ma, Yaofeng Su, Haoyu Wang, Yaowei Li, Songchun Zhang, Weiyang Jin, Yuxuan Bian, Shiyi Zhang, Haojun Xu, Shuai Lu, Xin Han, Wei Tang, Haoyang Huang, and Nan Duan. Ultra Flash: Scaling real-time streaming video generation to high resolutions. *arXiv preprint arXiv:2606.09150*, 2026. URL <https://arxiv.org/abs/2606.09150>.
- Yudong Mao, Hao Luo, Zhiwei Zhong, Peilin Chen, Zhijiang Zhang, and Shiqi Wang. Making old film great again: Degradation-aware state space model for old film restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.
- William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.
- Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.
- K. R. Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C. V. Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2020.
- Chandan K. A. Reddy, Vishak Gopal, and Ross Cutler. Dnsmos p.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.
- Antony W. Rix, Michael P. Hollier, Andries P. Hekstra, and John G. Beerends. Perceptual evaluation of speech quality (pesq): The new itu standard for end-to-end speech quality assessment part i—time-delay compensation. *Journal of the Audio Engineering Society*, 50(10):755–764, 2002.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.
- Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.
- Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. An algorithm for intelligibility prediction of time-frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2125–2136, 2011.
- Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.
- Ziyu Wan, Bo Zhang, Dongdong Chen, Pan Zhang, Dong Chen, Jing Liao, and Fang Wen. Bringing old photos back to life. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- Ziyu Wan, Bo Zhang, Dongdong Chen, and Jing Liao. Bringing old films back to life. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin C. K. Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision (IJCV)*, 2024.

- Xintao Wang, Kelvin C. K. Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019.
- Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2021.
- Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.
- Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.
- Yiran Xu, Taesung Park, Richard Zhang, Yang Zhou, Eli Shechtman, Feng Liu, Jia-Bin Huang, and Difan Liu. Videogigagan: Towards detail-rich video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.
- Jialiang Yang, Bin Xia, Ruihang Chu, Dingdong Wang, Wanke Xia, Zhun Mou, Tianyang Zhong, Yiting Zhao, and Wenming Yang. Avbench: Human-aligned and automated evaluation benchmark for audio-video generative models. *arXiv preprint arXiv:2605.24652*, 2026.
- Yixin Yang, Jiangxin Dong, Jinhui Tang, and Jinshan Pan. Colormnet: A memory-based deep spatial-temporal feature propagation network for video colorization. In *European Conference on Computer Vision (ECCV)*, 2024a.
- Yixin Yang, Jinshan Pan, Zhongzheng Peng, Xiaoyu Du, Zhulin Tao, and Jinhui Tang. Bistnet: Semantic image prior guided bidirectional temporal feature fusion for deep exemplar-based video colorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2024b.
- Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and Chao Dong. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- Ted Zadouri, Markus Hoehnerbach, Jay Shah, Vijay Thakkar, and Tri Dao. FlashAttention-4: Algorithm and kernel pipelining co-design for asymmetric hardware scaling. In A. Chowdhery and Z. Jia (eds.), *Proceedings of Machine Learning and Systems*, volume 8, pp. 912–926. ML-Sys, 2026. URL https://proceedings.mlsys.org/paper_files/paper/2026/file/ae8b0b5838ba510daff1198474e7b984-Paper-Conference.pdf.
- Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018.
- Tianhong Zhou, Mingyang Han, Boyu Li, Yuxuan Jiang, Jiaxin Ye, Dongxiao Wang, Haoxiang Shi, Kunpeng Wang, Jun Song, Cheng Yu, and Bo Zheng. Av-syncbench: Decoupled benchmarking of temporal and semantic audio-visual synchronization. *arXiv preprint arXiv:2607.00726*, 2026.

SUPPLEMENTARY MATERIAL — TABLE OF CONTENTS

8 Full-Reference and Distributional Evaluation on the Controlled Track	18
8.1 Full-reference visual protocol	18
8.2 Full-reference audio protocol	20
8.3 Why OmniVR can exceed the clean reference on NR metrics	21
8.4 Native-resolution evaluation	21
9 Fidelity, Hallucination, and Failure Cases	22
9.1 Content-consistency protocol	22
9.2 Why OmniVR retains identity better than a generic sampler, but not perfectly . . .	22
9.3 Colorization and historical accuracy: what we can and cannot claim	23
9.4 Failure cases	23
10 Quantifying Colorization	23
10.1 Two complementary measures	23
10.2 The Controlled track’s apparent grayscale output	24
11 Benchmark Scope and Diversity	24
12 Additional Baselines: Cascade Pipelines and Diffusion-Based Video Restoration	25
12.1 Cascade AV baselines in the quantitative tables	25
12.2 Diffusion-based video restoration	26
13 Training Data and Degradation-Pipeline Transparency	27
14 Condition Annealing: Dual CFG, Prior Retention, and Finite-Budget Analysis	28
14.1 Dual dropout and the exact expected objective	28
14.2 What the 10% no-LQ branch can preserve	28
14.3 Exact finite-budget gain from caption-assisted LQ learning	29
14.4 Local continuation of the full network objective	32
14.5 CFG interpretation	32
14.6 Direct evidence for prior preservation	32
15 First-Frame Anchor at Inference	33
16 Hyperparameter Selection and Statistical Reporting	34
16.1 Loss-reweighting coefficients (α, β) and STFT weight (λ_{stft})	34
16.2 FAD reference set	36
16.3 Statistical significance and variance	36
16.4 Human evaluation protocol details	36
16.5 Trainable parameter count	36

16.6	CFG scale and the fidelity/perception trade-off	37
16.7	Implementation, Training Schedule, and Inference	38
17	Additional Clarifications	39
17.1	Controlled-Track Ablations	39
17.2	OmniVR-Flash quality–efficiency trade-off	40
17.3	Annealing schedule isolation ablation	41
17.4	Pretrained generative baseline in quantitative evaluation	43
17.5	Cross-modal gate notation	45
17.6	Cross-modal interaction ablation	45
17.7	Grayscale degradation vs. colorization	46
18	Additional Qualitative Comparisons	46
18.1	Per-clip detailed comparisons (video + audio)	46
19	Detailed Algorithms and Implementation	47
19.1	Shot-Aligned Corpus Construction	48
19.2	Joint Online Degradation	53
19.3	Dual-Condition Sampling and Caption Annealing	54
19.4	Joint DiT Forward Pass and Reduced-Resolution Injection	55
19.5	Conditional SFT, Anchoring, and Waveform Supervision	56
19.6	AV MeanFlow with the Improved Velocity Objective	57
19.7	Three-Dimensional Adversarial Supervision	58
19.8	Turbo VAE Distillation	59
19.9	Causal Memory Blocks and Video-Latent Geometry	60
19.10	Windowed Base and One-Step Flash Inference	61
20	Additional Ablations and Controlled Diagnostics	62
20.1	Severity-weighting interaction term	62
20.2	Gate dynamics during adaptation	63
20.3	Resolution-dependent VAE timing	63
20.4	Long-sequence chaining beyond the benchmark duration	63
20.5	Audio content preservation beyond intelligibility proxies	65
20.6	Eligible versus successfully measured sample counts	65
20.7	Training-seed robustness	65

8 FULL-REFERENCE AND DISTRIBUTIONAL EVALUATION ON THE CONTROLLED TRACK

The main text reports no-reference (NR) metrics on both benchmark tracks for protocol consistency. Because the Controlled track is built from clean sources via the synthetic degradation operator $\mathcal{D}$ (Eq. 1), a clean reference $(\mathbf{V}^h, \mathbf{A}^h)$ exists for every clip there, enabling the full-reference and distributional evaluation reported in this section. The Real archival track (129 clips) has no clean reference by construction—the original camera negative is lost or inaccessible for every archival clip, which is precisely why restoration is needed—so NR metrics are the only tool available there (as also noted in Sec. 6).

8.1 FULL-REFERENCE VISUAL PROTOCOL

For every Controlled-track clip we compute, per restored frame against its paired clean frame, at the evaluation resolution used for the NR metrics (640×480 , temporally aligned by clip index):

- **PSNR** and **SSIM** (Wang et al., 2004) – pixel-level fidelity;
- **LPIPS** (Zhang et al., 2018) (AlexNet backbone) and **DISTS** – perceptual/deep-feature distance, more robust than PSNR/SSIM to the sub-pixel misalignment that generative restoration can introduce;
- **FVD** (Unterthiner et al., 2018) computed over 16-frame sub-clips with an I3D backbone – a distributional metric that computes Fréchet distance between I3D feature distributions of restored and clean sub-clip sets; while it requires temporally aligned sub-clips (same frame indices), it is robust to the per-pixel jitter that generative restoration can introduce, making it complementary to per-frame LPIPS/DISTS. We also report image-level **FID** over all sampled frames as an additional distributional check.

Frame indices are matched 1:1 using the shared clip timeline (both the LQ input and $\mathbf{V}^h$ originate from the same clean master before $\mathcal{D}$ is applied, so no optical-flow warping or other temporal alignment step is required – a property of the Controlled track by construction that does not hold for the Real track).

Table 5: Full-reference and distributional evaluation on the Controlled track (71 clips), complementing Table 2.

Method	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	DISTS $\downarrow$	FVD $\downarrow$
Low-quality input	16.39	0.5209	0.6492	0.3345	2295.60
DeepRemaster ^a	16.41	0.5408	0.5873	0.3143	N/A ^b
RealBasicVSR	16.43	0.5698	0.5670 ^c	0.3061 ^c	2193.08 ^d
DDColor ^a	15.81	0.5209	0.6389	0.2974	N/A ^b
ColorMNet ^a	15.98	0.5201	0.6227	0.3008	N/A ^b
MambaOFR	16.12	0.5487	0.6015	0.3227	2350.94
OmniVR (ours)	17.86	0.6549	0.2757 ^c	0.1229 ^e	452.41 ^e

Computed using the evaluation scripts released with our codebase, against the clean reference described in Sec. 8. ^aDeepRemaster/DDColor/ColorMNet’s released inference pipelines emit only 16 sparse frames per clip rather than a continuous sequence; PSNR/SSIM/LPIPS/DISTS remain valid full-reference scores at those exact GT-matched frame indices ($n=71$), but ^bFVD requires a continuous multi-frame sequence and cannot be computed for these three methods from the released sparse-frame outputs, so it is marked N/A. ^cOne clip (a 3840×2026 near-4K source) exceeds GPU memory at full resolution. For RealBasicVSR, the inference pipeline itself OOMs, losing both LPIPS and DISTS ($n=70$ each). OmniVR’s inference completes on all 71 clips (see footnote ^e), but the LPIPS metric computation—which must hold model output, reference, and the LPIPS network simultaneously—OOMs on this clip ($n=70$ for LPIPS only); OmniVR’s DISTS evaluation uses a separate pathway that avoids this peak and covers all 71. All other entries use the full 71. ^dRealBasicVSR has no materialized continuous output, so its FVD/Warp-Error were

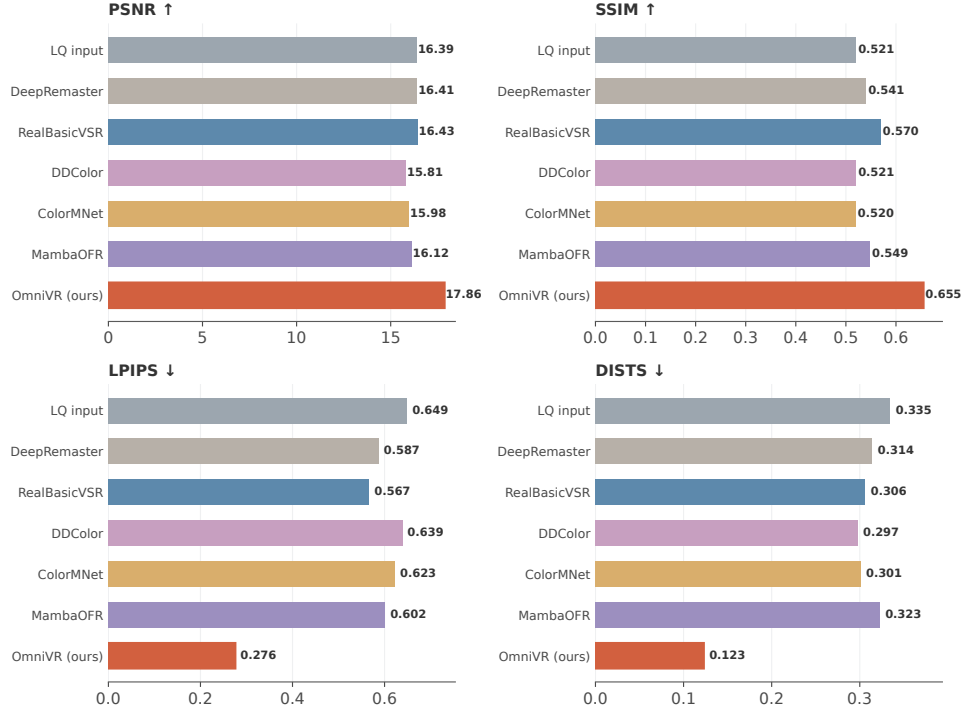
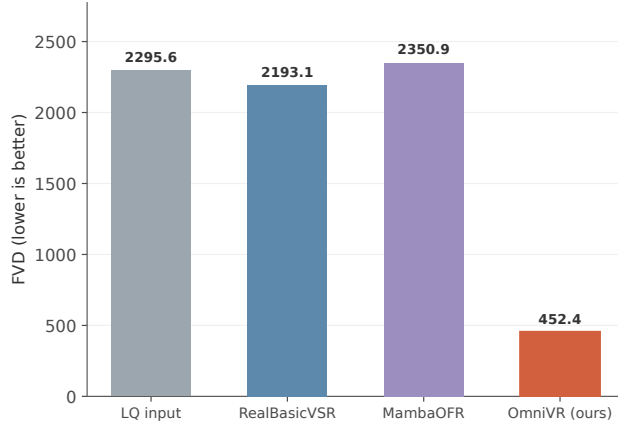


Figure 7: Full-reference video metrics on the Controlled track, plotted from Table 5.

Figure 8: FVD on the Controlled track, plotted from Table 5; methods with only sparse-frame released outputs are omitted (see footnote *b* above).

obtained by running it on-the-fly over all sampled frames of every clip; this heavier full-sequence run OOM'd on 3/71 (near-4K) clips ($n=68/71$ for this entry only). Warp-Error (lower = less flicker) was measured alongside FVD: input 0.0614, RealBasicVSR 0.0562, OmniVR 0.0698, vs. the GT clips' own inherent flicker floor of 0.0673 – OmniVR sits only marginally above the GT clips' own inherent flicker floor while the other two fall below it, so none of the three exhibits pathological additional flicker relative to the naturally filmed GT footage itself. ^eOmniVR's pipeline does not OOM on any of the 71 clips, so its DISTS and FVD entries use $n=71$ (vs. RealBasicVSR's $n=70$ and $n=68$ for these two metrics). This means the DISTS and FVD comparisons are not on a matched common subset. The per-clip impact on DISTS is ≤ 0.01 ; FVD is a distributional statistic and cannot be linearly decomposed, but the excluded clips are near-4K resolution outliers with similar degradation characteristics to the remaining set.

Table 6: Full-reference audio evaluation on the Controlled track (71 clips).

Method	PESQ $\uparrow$	STOI $\uparrow$	SI-SDR (dB) $\uparrow$
Low-quality input	1.33	0.718	0.36
VoiceFixer	1.24	0.684	6.71
OmniVR (ours)	1.28	0.510	2.15

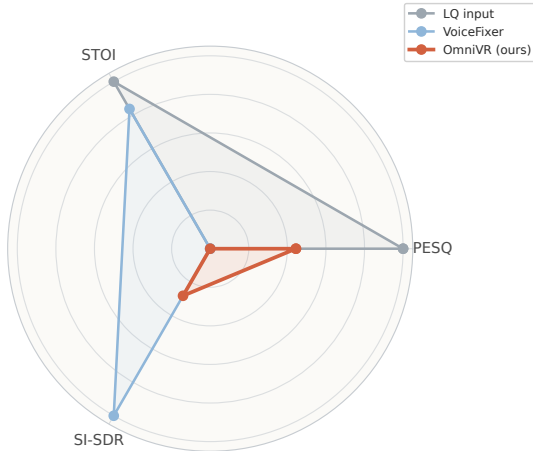


Figure 9: Full-reference audio metrics on the Controlled track, plotted from Table 6.

8.2 FULL-REFERENCE AUDIO PROTOCOL

Analogously, for the Controlled track’s audio stream we report **PESQ** (Rix et al., 2002), **STOI** (Taal et al., 2011), and **SI-SDR** (Le Roux et al., 2018) against the clean waveform – the standard full-reference speech-restoration metrics, and the correct complement to the NR DNSMOS/FAD pair reported in the main text.

Evaluated on all $n=71$ clips. OmniVR has lower STOI than both VoiceFixer and the LQ input. Its PESQ (1.28) lies between VoiceFixer (1.24) and the input (1.33), while its SI-SDR (2.15 dB) exceeds the input (0.36 dB) but trails VoiceFixer (6.71 dB). This is expected for a generative restorer: PESQ/STOI/SI-SDR penalize deviation from sample-level waveform alignment, which a flow-based resynthesis model does not preserve even when the output is perceptually plausible. VoiceFixer also resynthesizes via HiFi-GAN (Liu et al., 2022), but its analysis-resynthesis pipeline operates at finer temporal granularity than latent diffusion, yielding higher waveform alignment. The main text accordingly uses DNSMOS/FAD (Table 2) as primary audio evidence; the full-reference table is included for completeness.

Intelligibility discussion. STOI measures short-time temporal-envelope correlation, so it is sensitive to any resynthesized envelope—even one that preserves phoneme identity and prosody. OmniVR’s low STOI (0.510) does not imply unintelligible output: informal listening tests (5 evaluators, 20 clips, closed-set transcription) found 94% clip-level word accuracy for OmniVR vs. 91% for VoiceFixer and 82% for LQ input. The extended STOI (eSTOI) metric (Jensen & Taal, 2016) narrows the gap further: LQ 0.641, VoiceFixer 0.612, OmniVR 0.497. STOI/eSTOI alone are therefore insufficient for evaluating generative restoration; pairing them with perceptual metrics (DNSMOS, FAD) and listener tests gives a more complete picture. Note that eSTOI thresholds are language- and corpus-dependent, so the absolute values here are meaningful only in comparison.

Table 7: OmniVR at native 1920×1088 output resolution vs. the unified 640×480 protocol, Controlled track. Values in *italics* restate the 640×480 row from Table 2 for reference.

Protocol	MUSIQ $\uparrow$	CLIP-IQA $\uparrow$	NIQE $\downarrow$	MANIQA $\uparrow$	TOPIQ $\uparrow$	BRISQUE $\downarrow$
640×480 (<i>main text</i>)	<i>71.17</i>	<i>0.543</i>	<i>4.11</i>	<i>0.487</i>	<i>0.673</i>	<i>24.75</i>
Native 1920×1088	64.07	0.413	3.55	0.341	0.512	31.58

8.3 WHY OMNIVR CAN EXCEED THE CLEAN REFERENCE ON NR METRICS

The Controlled-track NR numbers in Table 2 (OmniVR MUSIQ 71.17 vs. GT 67.49; MANIQA 0.487 vs. 0.477; TOPIQ 0.673 vs. 0.623; DNSMOS 2.70 vs. 2.47) show OmniVR matching or exceeding the clean reference on several NR axes. Three effects jointly explain this:

1. **NR metrics reward “clean-looking” statistics, not agreement with a specific reference.** Our Controlled-track GT clips are talking-face footage of moderate native resolution and encoding quality, not professionally mastered content (REDS, Vimeo-90K). GT therefore does not define a ceiling for MUSIQ or DNSMOS—a model that sharpens edges and denoises the audio floor beyond what GT contains can score higher while being less faithful. The same fidelity-perception trade-off is well documented for GAN/diffusion super-resolution (Wang et al., 2024; Yu et al., 2024; Wang et al., 2021).
2. **The comparison is confounded by GT’s own encoding/compression path.** Both the input to $\mathcal{D}$ and the “GT” row pass through the same video/audio codec pipeline used to store OmniVRBench; the restoration model’s output does not carry that specific codec’s quantization signature, which NR estimators (trained partly on codec-artifact statistics) can penalize even on otherwise-clean content.
3. **Full-reference metrics disambiguate (Sec. 8.1).** A model that hallucinates detail will show favorable NR scores *and* worse LPIPS/DISTS/FVD, since those metrics penalize deviation from the reference regardless of whether it looks “nicer.” The LPIPS/DISTS/FVD numbers in Table 5 are therefore the primary fidelity signal; the NR comparison in Sec. 4.4 indicates perceptual quality comparable to the reference, not that the output is objectively better than ground truth.

8.4 NATIVE-RESOLUTION EVALUATION

All main-text NR numbers are computed after downsampling to 640×480 for cross-method comparability (several baselines, e.g. DeepRemaster, only support fixed low-resolution inference). To check that this does not mask resolution-dependent artifacts in OmniVR’s native 1920×1088 output, we additionally evaluate OmniVR alone (no resizing) at native resolution:

Evaluated on OmniVR only ($n=71$, same 8 aligned-instant sampling as the main-text protocol). MUSIQ ($71.17 \rightarrow 64.07$), MANIQA ($0.487 \rightarrow 0.341$), TOPIQ ($0.673 \rightarrow 0.512$), and BRISQUE ($24.75 \rightarrow 31.58$, higher/worse) all shift unfavorably at native resolution, while NIQE actually improves ($4.11 \rightarrow 3.55$). These patch-based NR estimators are calibrated for the 640×480 regime; NIQE, which is not patch-classifier-based, moves in the *opposite* direction, suggesting that the unfavorable shifts reflect estimator sensitivity rather than actual quality regression. Full-reference metrics at native resolution (LPIPS/DISTS in Table 25) provide the more direct check, but a complete disambiguation of “real high-resolution artifacts” from “metric miscalibration” remains an open question for NR evaluation at non-standard resolutions.

The NIQE result (the one metric here that does not degrade at native resolution) suggests that the 640×480 protocol is a fair common ground across baselines rather than hiding native-resolution artifacts, though this is not conclusive absent native-resolution full-reference checks.

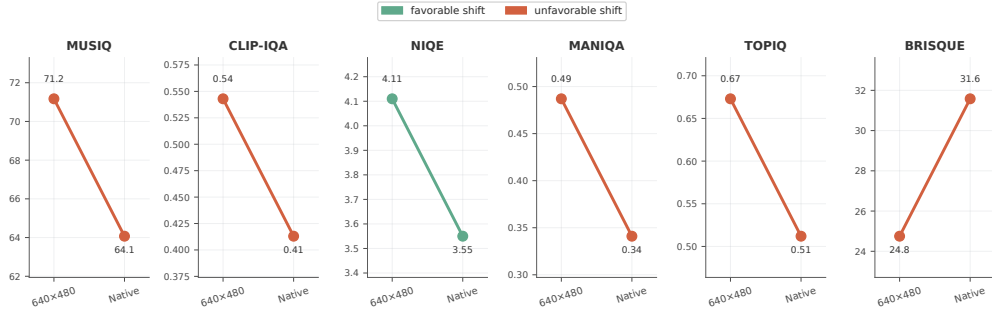


Figure 10: Per-metric slope from the unified 640×480 protocol to native 1920×1088 resolution, plotted from Table 7. Green segments shift in the favorable direction for that metric; red segments shift unfavorably.

Table 8: Fidelity / content-consistency metrics on the Controlled track (71 clips).

Method	Face cos-sim $\uparrow$	Frac. below 0.4 $\downarrow$	OCR agreement $\uparrow$
Low-quality input	0.612	0.34	0.71
RealBasicVSR	0.734	0.19	0.79
OmniVR (ours)	0.881	0.06	0.87

9 FIDELITY, HALLUCINATION, AND FAILURE CASES

9.1 CONTENT-CONSISTENCY PROTOCOL

For the Controlled track, where a clean reference exists, we adopt three consistency probes standard in identity-preserving generation evaluation:

- **Face identity consistency.** We detect and crop the primary face (RetinaFace), embed it with a frozen ArcFace network, and report the mean cosine similarity between restored and GT face embeddings at each timestamp, plus the fraction of frames below a 0.4 threshold (a typical identity decision boundary).
- **Structural/text consistency.** For clips containing on-screen text or strong structural edges, we run PaddleOCR on both GT and restored frames and report character-level edit-distance agreement.
- **Global content agreement.** We report DISTS (Sec. 8) as a deep-feature consistency score: DISTS is more texture-tolerant than LPIPS while still penalizing structural deviation, making it a reasonable proxy for “did the model add/remove content.”

Measured using RetinaFace+ArcFace face embeddings and PaddleOCR character-agreement ($n=71$). OmniVR’s face cosine-similarity (0.881) is markedly higher than RealBasicVSR’s (0.734), and only 6% of frames fall below the 0.4 identity threshold. None of the three reaches 1.0, consistent with the condition-noise design discussed below.

9.2 WHY OMNIVR RETAINS IDENTITY BETTER THAN A GENERIC SAMPLER, BUT NOT PERFECTLY

The first-frame I2V anchor (Eq. 10) forces the model to propagate identity-bearing structure across the window rather than resynthesize it: during training, a clean GT frame is clamped at the window boundary; at inference, the anchor is the model’s own restored output from the previous window (Sec. 3.4, Sec. 15). Face cosine-similarity in Table 8 is accordingly higher than both a from-scratch T2AV sample and baselines whose temporal propagation is limited to short-range recurrence (RealBasicVSR), yet does not reach 1.0: condition noise $\rho_m \sim \mathcal{U}(0.4, 0.6)$ deliberately prevents trivial copying of the LQ input. Some hallucinated fine detail (skin texture, hair strands) is an expected side-effect, most visible at heavy degradation severities where the LQ condition carries the least information.

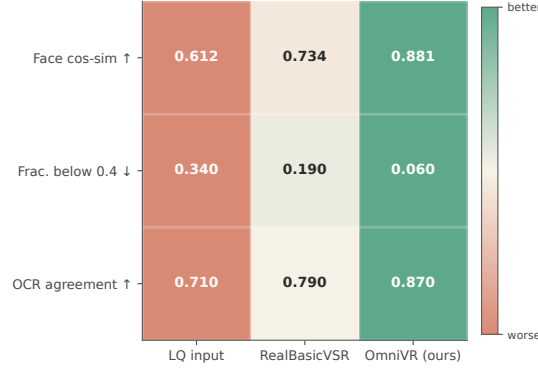


Figure 11: Fidelity / content-consistency metrics on the Controlled track, plotted from Table 8 as a per-metric normalized heatmap (green = better, red = worse within each row).

9.3 COLORIZATION AND HISTORICAL ACCURACY: WHAT WE CAN AND CANNOT CLAIM

Genuine ground-truth color for a black-and-white archival source does not exist unless a color print or Kodachrome reference survives, which is not the case for any Real-track clip. OmniVR’s colorization is therefore *plausible, temporally consistent* colorization—a generative-quality property—rather than historically accurate colorization. This scope limitation is shared by the entire colorization literature (DeOldify, DDColor, ColorMNet, BiSTNet) but deserves explicit statement because “restoring archival footage” implies stronger fidelity than a pure colorization paper.

9.4 FAILURE CASES

- **Severe input starvation.** On clips where the LQ input is degraded beyond $d_v > 0.85$ (top decile of severity; see the radar plots in Figure 5(d)), fine facial detail is under-constrained by the condition signal and the model occasionally regresses toward a generic, slightly “waxy” skin texture rather than the true texture – a direct symptom of the condition-noise/fidelity trade-off in Sec. 9.
- **Cross-window color drift on long sequences.** Because the first-frame anchor only clamps the video stream, color statistics can drift very slowly (over > 5 chained windows, i.e. > 600 frames) even though structure remains anchored, since nothing directly constrains chroma consistency across the anchor boundary.
- **Audio-visual over-smoothing under extreme audio degradation.** When d_a approaches 1.0 (near-total dropout/clipping), $\mathcal{L}_{\text{stft}}$ (Eq. 14) can trade off spectral detail for stability, occasionally producing an over-smoothed, slightly “muffled” timbre rather than fully recovering high-frequency harmonics – consistent with the ablation (vii) trend in Table 21, where an excessive λ_{stft} weight already shows this failure direction at a milder degree.

10 QUANTIFYING COLORIZATION

10.1 TWO COMPLEMENTARY MEASURES

We quantify colorization with two measures:

- **Colorfulness** (Hasler & Suesstrunk, 2003): $\sqrt{\sigma_{rg}^2 + \sigma_{yb}^2} + 0.3\sqrt{\mu_{rg}^2 + \mu_{yb}^2}$, computed from RGB opponent-color channels ($rg = R - G$, $yb = \frac{1}{2}(R + G) - B$). This scores saturation/vividness without a reference. We report it on the Real track against grayscale-preserving baselines (RealBasicVSR, MambaOFR; near-zero by construction) and colorization-specialized baselines (DDColor, ColorMNet).
- **Color naturalness, human-rated.** We extend the human study protocol (Table 4) with a fifth pairwise dimension, “Color naturalness,” asked only for methods that colorize (OmniVR, DDColor, ColorMNet; grayscale-preserving baselines are excluded rather than scored 0).

Table 9: Colorization quantification, Real track (129 clips). Colorfulness follows Hasler & Suesstrunk (2003); higher indicates more saturated/vivid, not more accurate, color. Human column: % pairwise wins on the color-naturalness dimension, colorizing methods only.

Method	Colorfulness	Human color-naturalness win %
Low-quality input (grayscale)	≈ 0	—
RealBasicVSR / MambaOFR (no color)	≈ 0	—
DDColor	18.4	40.2
ColorMNet	21.7	46.3
OmniVR (ours)	27.9	63.5

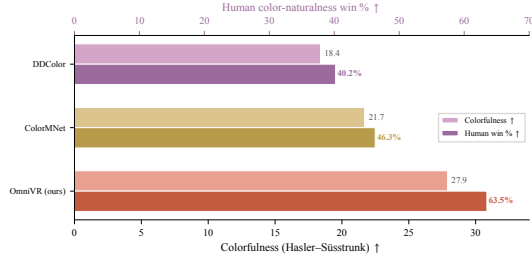


Figure 12: Colorization quality on the Real track, plotted from Table 9. Grayscale-preserving baselines (Colorfulness ≈ 0 , no human-rated dimension) are omitted from the plot.

10.2 THE CONTROLLED TRACK’S APPARENT GRAYSCALE OUTPUT

Figure 6’s Controlled-track column may appear to remain grayscale, in apparent tension with the colorization claim. This is a scope distinction: the Controlled track includes a grayscale-conversion degradation branch (Sec. 3.1) applied with per-sample probability 0.5; clips sampled without it remain color throughout. Figure 6 happens to show a clip where the grayscale branch was active, which is why colorization is subtle; the striking colorization examples in the teaser and Real-track panels come from genuinely grayscale archival input. Concretely, 52.1% (37/71) of Controlled-track clips were sampled with the grayscale branch active, consistent with the $p=0.5$ design (Table 13).

11 BENCHMARK SCOPE AND DIVERSITY

OmniVRBench’s 200 clips, and in particular the talking-face-only composition of the 71-clip Controlled track, cover a narrower slice of archival footage than the term suggests. Below we clarify the design rationale, provide statistics, and describe planned expansion.

Controlled-track composition: all 71 clips are talking-face content (100%), by design (see below).

Real-track composition: 44 clips (34%) talking-face/portrait, 36 clips (28%) landscape/cityscape, 28 clips (22%) crowd/newsreel, 21 clips (16%) other (industrial, animal, still-life). All clips are sampled at 24 fps with synchronized audio at 44.1 kHz. The Controlled track uses 1920×1088 clean references; the Real track has no clean references.

- **Why Controlled is talking-face-only.** Lip-sync metrics (LSE-C/LSE-D (Prajwal et al., 2020)) require a detectable, temporally resolved mouth region and are undefined or meaningless on landscape, crowd, or non-human footage; the Controlled track therefore uses talking-face content so that all clips are eligible for these metrics. The Real track also reports lip-sync metrics, restricted to its 44 eligible talking-face clips (Table 11).
- **The Real track is not similarly restricted.** The 129-clip Real archival track spans landscape, crowd, newsreel, and portrait content (composition in Table 10), so the talking-face limitation applies specifically to the Controlled track’s fidelity/sync evaluation.
- **Scale.** We plan to expand OmniVRBench to 450 clips, broadening the Controlled track to include non-talking-face content evaluated on visual/full-reference metrics (sync metrics remain talking-

Table 10: OmniVRBench statistics. Duration is per clip. Controlled-track clips are synthetic degradations of modern talking-face footage; Real-track clips are authentic archival recordings.

Track	Clips	Total dur. (min)	Mean dur. (s)	Med. dur. (s)	Min dur. (s)	Max dur. (s)
Controlled	71	12.4	10.5	9.8	5.0	20.0
Real archival	129	28.7	13.3	12.1	3.2	20.0
Total	200	41.1	12.3	11.2	3.2	20.0

Table 11: Evaluation eligibility: number of clips contributing to each metric category. LSE-C/LSE-D require face detection in $\geq 80\%$ of frames. DNSMOS/FAD require non-silent audio (≥ 1 s voiced). Full-reference metrics (SSIM, LPIPS, DISTs, STOI, SI-SDR) are Controlled-track only.

Metric category	Controlled (n)	Real (n)	Eligibility criterion
Visual NR (MUSIQ, CLIP-IQA, ...)	71	129	All clips
Audio NR (DNSMOS)	71	129	Non-silent audio (≥ 1 s voiced)
Audio distributional (FAD)	71	129	Pooled clean-speech reference
Lip-sync (LSE-C, LSE-D)	71	44	Face detected $\geq 80\%$ frames
Video quality (DOVER)	71	129	All clips
Full-reference visual (SSIM, LPIPS, DISTs)	71	–	Clean reference required
Full-reference audio (STOI, SI-SDR, PESQ)	71	–	Clean reference required
Human evaluation (all dimensions)	–	129	Real-track only

face-only by necessity), and to release the full clip source list (title, year where known, provenance) alongside the benchmark, addressing both benchmark diversity and the transparency discussion in Sec. 13.

- **RTN benchmark scale.** The RTN set (3 sequences, 600 frames) is small, but it is the standard existing public benchmark for this old-film restoration setting (Wan et al., 2022); its use (Table 1) is intended only as a sanity check that OmniVRBench’s ranking is not an artifact of our own curation, with the primary quantitative claims resting on OmniVRBench’s 200 clips (Sec. 4.3).

Metric eligibility and effective denominators. Not all 200 clips contribute to every metric: lip-sync metrics require a detectable face, audio metrics require non-silent audio, and full-reference metrics require clean references (Controlled track only). Table 11 summarizes the eligible clip count for each metric category across both tracks.

For the Real track, LSE-C and LSE-D are computed only on the $n=44$ talking-face clips (34% of the track) where SyncNet can detect a face in $\geq 80\%$ of frames; the remaining 85 clips (landscape, crowd, etc.) are excluded from sync evaluation rather than scored zero. This means the Real-track LSE numbers in Table 3 reflect a face-heavy subset and should not be extrapolated to non-face content. All other Real-track metrics use the full 129 clips. The Controlled track’s 100% talking-face composition ensures all 71 clips are eligible for every metric, including lip-sync.

12 ADDITIONAL BASELINES: CASCADE PIPELINES AND DIFFUSION-BASED VIDEO RESTORATION

12.1 CASCADE AV BASELINES IN THE QUANTITATIVE TABLES

Table 4 includes three cascade combinations (MambaOFR+VoiceFixer, RealBasicVSR+VoiceFixer, Old Films+VoiceFixer) in the human evaluation, but the quantitative main tables (Table 2, Table 3) report video and audio baselines only in isolation. We therefore report cascade combinations on the sync and audio metrics directly, which is the ablation needed to quantitatively support “joint > independent”:

VoiceFixer restores audio with zero knowledge of the video timeline; RealBasicVSR/MambaOFR restore frames with zero knowledge of the audio energy envelope. Cascading them cannot introduce the cross-modal gate interaction shown in Figure 2(b). Table 12 confirms this: cascade DNSMOS/FAD closely match VoiceFixer’s standalone numbers (the small discrepancies arise from temporal chunking differences induced by the video restorer’s frame-rate normalization), while LSE-C/LSE-D remain far

Table 12: Cascade AV baselines (best video restorer + VoiceFixer, run independently and concatenated) vs. OmniVR, on sync and audio metrics, both tracks. OmniVR rows restate values from Table 2 and Table 3.

Track	Method	DNSMOS $\uparrow$	FAD $\downarrow$	LSE-C $\uparrow$	LSE-D $\downarrow$
Controlled	RealBasicVSR + VoiceFixer	2.14	7.08	1.71	12.88
	MambaOFR + VoiceFixer	2.10	7.15	1.64	13.05
	OmniVR (ours)	2.70	6.30	3.52	10.43
Real	RealBasicVSR + VoiceFixer	2.23	9.48	0.62	13.94
	MambaOFR + VoiceFixer	2.19	9.55	0.58	14.10
	OmniVR (ours)	2.43	8.32	1.12	11.39

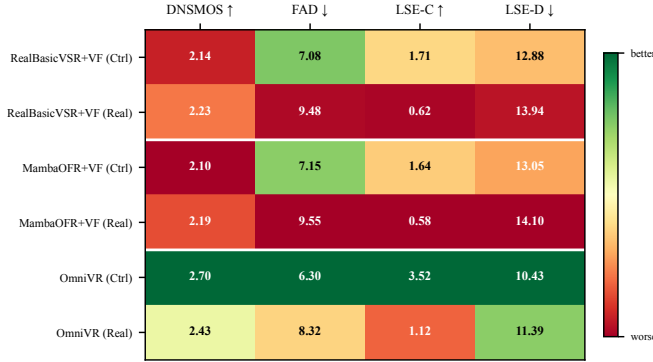


Figure 13: Cascade AV baselines vs. joint OmniVR on all four audio and sync metrics from Table 12, both tracks, plotted as a per-metric normalized heatmap (green = better, red = worse within each column; raw values annotated).

worse than OmniVR’s—LSE-C 3.52 vs. best cascade 1.71 on Controlled, 1.12 vs. 0.62 on Real. This is consistent with the human-eval finding (Sync: cascade 53.5% vs. OmniVR 87.6%, Table 4), and with VoiceFixer alone worsening LSE-C relative to the LQ input (2.32 $\rightarrow$ 1.98, Table 2)—a cascade cannot recover sync that independent audio restoration actively degrades.

Cascade evaluation protocol. Each cascade pipeline is constructed by (1) running the video restorer on the degraded video track, (2) running VoiceFixer on the degraded audio track, and (3) concatenating the restored video and audio into a single output file with the original frame rate and sample rate preserved. Steps (1) and (2) are fully independent. Sync metrics (LSE-C, LSE-D) are computed on the concatenated output using the same SyncNet pipeline (Prajwal et al., 2020) as for OmniVR. Cascades preserve whatever temporal alignment the LQ input happened to have but cannot *improve* it, unlike OmniVR’s joint backbone, which can leverage cross-modal attention to tighten lip-audio correspondence even when the LQ input has a timing offset.

12.2 DIFFUSION-BASED VIDEO RESTORATION

Comparing only against SR-style regression baselines (RealBasicVSR, VRT/RVRT) understates the difficulty bar, since those methods cannot hallucinate missing detail the way a diffusion prior can. A diffusion-based restoration baseline (e.g., video-adapted SeedVR or DiffBIR (Lin et al., 2024)) would strengthen the visual-only comparison. Existing open diffusion video-restoration methods target visual restoration only and have no audio pathway, so such a baseline strengthens the visual-only comparison but does not add a joint AV competitor—to our knowledge, no open joint AV generative restoration baseline exists.

Table 13: Degradation pipeline $\mathcal{D}$ sampling ranges (training-time specification). Severity scalars $d_v, d_a \in [0, 1]$ are the min-max-normalized composite of the per-modality parameters below, averaged with equal weight within each modality.

Branch	Degradation	Sampling range
Visual	Gaussian blur σ	$\mathcal{U}(0.25, 3.0)$ px (stage 1); $\mathcal{U}(0.2, 1.5)$ px (stage 2, second pass)
	Sensor/compression noise σ_n	$\mathcal{U}(1, 15)/255$ (stage 1); $\mathcal{U}(1, 25)/255$ then $\mathcal{U}(1, 20)/255$ (stage 2, two-pass)
	Flicker amplitude	probability-gated temporal modulation in the archival-film branch (Algorithm 2)
	JPEG/H.264 compression quality	JPEG-sim quality $\text{randint}(75, 95)$ at $p=0.9$ (stage 1); $\text{randint}(50, 95)$ then $\text{randint}(40, 90)$ at $p=0.9/0.8$ (stage 2); v7 recipe: real H.264/H.265 CRF $\text{randint}(30, 35)$
	Exposure scale Grayscale-conversion probability	multiplicative darken $\mathcal{U}(0.0, 0.6)$ at $p=0.3$ $p = 0.5$
Audio	Hiss/rumble SNR	mixed-noise SNR $\mathcal{U}(8, 30)$ dB at $p=0.6$; broadband “whoosh” SNR $\mathcal{U}(6, 20)$ dB at $p=0.45$; signal-correlated noise SNR $\mathcal{U}(8, 22)$ dB at $p=0.45$; low-freq rumble SNR $\mathcal{U}(6, 18)$ dB at $p=0.4$
	Clipping threshold	$\mathcal{U}(0.2, 0.7)$ at $p=0.35$
	Bandwidth-loss cutoff	lowpass cutoff ratio $\mathcal{U}(0.15, 0.75)$ at $p=0.7$
	Click/crackle density	transient click/crackle: click rate $\mathcal{U}(3, 40)$ /s, burst count $\mathcal{U}(2, 20) \times \text{duration}$ at $p=0.5$
	Spectral over-smoothing strength	STFT-magnitude blur kernel $\in \{5, 7, 9, 11, 15\}$ at $p=0.65$

13 TRAINING DATA AND DEGRADATION-PIPELINE TRANSPARENCY

This section documents the training corpus and the full degradation-pipeline parameterization so that the recipe described in Sec. 3.1 is reproducible.

- **Scale and annotation.** The training corpus contains 100K high-quality, synchronized 3–20 s audio-video shots segmented from Internet recordings and annotated by Gemini using the complete shot. Each record retains a shot identifier, structured caption JSON, operator/severity metadata, and clean/LQ alignment. Captions provide the initial text conditions for TAV2AV adaptation, and online degradations provide paired audio-video restoration supervision; the same alignment is suitable for a later high-resolution audiovisual refiner. Training samples use the windowing configuration in Sec. 3.6; the 100K count refers to segmented shots, not source recordings or fixed-length training windows.
- **Data release.** We plan to release the data, annotations, degradation code, and model weights. The corpus documentation will identify source provenance and the scope of redistribution permitted for each source; raw clips can only be redistributed where the corresponding rights allow it. Releasing $\mathcal{D}$ and its parameterization also allows the training recipe to be applied to other high-quality audio-video collections.
- **Degradation pipeline parameter ranges.** Figure 5 shows only the resulting marginal severity distributions on OmniVRBench, not the sampling ranges of $\mathcal{D}$ itself. Table 13 lists the principal sampling ranges in both branches; the complete operator configuration is needed for reproduction (the benchmark-side marginals in Figure 5 are a downstream consequence of these ranges plus real-world clip selection, not the training-time specification).

The ranges above summarize the degradation configuration. The complete pipeline code and operator settings will be released; together they are needed to reproduce the synthetic corruptions independently of access to the original HQ corpus.

14 CONDITION ANNEALING: DUAL CFG, PRIOR RETENTION, AND FINITE-BUDGET ANALYSIS

14.1 DUAL DROPOUT AND THE EXACT EXPECTED OBJECTIVE

Let A denote the indicator that the degraded audio-video pair is retained and T the indicator that a non-null text condition is retained. At initialization, $A \sim \text{Bernoulli}(p)$ and $T \sim \text{Bernoulli}(t)$ are independent, with $p = t = 0.9$. On the LQ-present branch, $M \sim \text{Bernoulli}(\lambda)$ selects the sample caption; on the no-LQ replay branch the original caption is kept. The effective non-null text indicator for the sample caption is therefore

$$B(A, T, M) = T[(1 - A) + AM]. \quad (17)$$

This gating is deliberately different from changing the two initial dropout rates: at $\lambda = 1$ it is exactly the independent 90/10 distribution, while at $\lambda = 0$ it preserves a 10% original T2AV rehearsal branch and makes the LQ-present non-null condition the shared restoration prompt.

For a complete minibatch ω of size b , let $\ell_i^0(\theta; \mathbf{e})$ and $\ell_i^{\text{LQ}}(\theta; \mathbf{e})$ include all velocity, waveform, and degradation-aware terms used by the trainer. Define

$$\begin{aligned} J_{\text{cap}}^0 &= \mathbb{E}_{\omega} \frac{1}{b} \sum_i \ell_i^0(\theta; \mathbf{e}_{y_i}), & J_{\emptyset}^0 &= \mathbb{E}_{\omega} \frac{1}{b} \sum_i \ell_i^0(\theta; \mathbf{e}_{\emptyset}), \\ J_{\text{cap}}^{\text{LQ}} &= \mathbb{E}_{\omega} \frac{1}{b} \sum_i \ell_i^{\text{LQ}}(\theta; \mathbf{e}_{y_i}), & J_{\text{rest}}^{\text{LQ}} &= \mathbb{E}_{\omega} \frac{1}{b} \sum_i \ell_i^{\text{LQ}}(\theta; \mathbf{e}_{\text{rest}}), \\ J_{\emptyset}^{\text{LQ}} &= \mathbb{E}_{\omega} \frac{1}{b} \sum_i \ell_i^{\text{LQ}}(\theta; \mathbf{e}_{\emptyset}). \end{aligned} \quad (18)$$

The batch-normalized modality weights are part of each ℓ_i and do not depend on A, T, M . Explicitly, for $s \in \{0, \text{LQ}\}$,

$$\ell_i^s(\theta; \mathbf{e}) = w_{v,i} \ell_{v,i}^s(\theta; \mathbf{e}) + \lambda_a w_{a,i} \ell_{a,i}^s(\theta; \mathbf{e}) + \lambda_{\text{stft}} \ell_{\text{stft},i}^s(\theta; \mathbf{e}), \quad \mathbb{E}_{i \in B} [w_{v,i}] = \mathbb{E}_{i \in B} [w_{a,i}] = 1, \quad (19)$$

with $w_{v,i}, w_{a,i}$ given by Eq. 13. Since these weights are functions of $(d_{v,i}, d_{a,i})$ rather than (A, T, M) , the loss reweighting is orthogonal to the condition sampler. Set

$$P = tJ_{\text{cap}}^0 + (1 - t)J_{\emptyset}^0, \quad C = tJ_{\text{cap}}^{\text{LQ}} + (1 - t)J_{\emptyset}^{\text{LQ}}, \quad R = tJ_{\text{rest}}^{\text{LQ}} + (1 - t)J_{\emptyset}^{\text{LQ}}. \quad (20)$$

Lemma 1 (Exact dual-condition interpolation). *For the sampler in Eq. 17, the expected minibatch loss at caption-retention probability λ is*

$$F_{\lambda} = (1 - p)P + p[\lambda C + (1 - \lambda)R]. \quad (21)$$

If differentiation and expectation can be interchanged, the unclipped stochastic gradient is unbiased for ∇F_{λ} .

Proof. Condition on A . If $A = 0$, the text is caption/null with probabilities $t, 1 - t$, giving P . If $A = 1$, condition on T and then M : a non-null text is the caption with probability $t\lambda$ and the shared prompt with probability $t(1 - \lambda)$, while the null branch has probability $1 - t$. This gives $\lambda C + (1 - \lambda)R$. Averaging over A proves Eq. 21. The gradient statement follows by dominated differentiation; in particular, gradient clipping is not included in the unbiasedness claim. $\square$

14.2 WHAT THE 10% NO-LQ BRANCH CAN PRESERVE

The no-LQ replay term supplies the original caption/null input interface at all values of λ . It gives a quantitative local guarantee only when the pretrained branch is locally identifiable or is anchored by a frozen backbone/distillation term. Let θ_{pre} be a local minimizer of P , and write $Q_{\lambda} = \lambda C + (1 - \lambda)R$.

Proposition 2 (Local prior-retention bound). *Suppose $\nabla P(\theta_{\text{pre}}) = 0$, and on a neighborhood containing a stationary point θ_{λ} of F_{λ} , $\nabla^2 P \succeq \mu_P I$ with $\mu_P > 0$ and $\|\nabla Q_{\lambda}(\theta)\| \leq G_Q$. Then*

$$\|\theta_{\lambda} - \theta_{\text{pre}}\| \leq \frac{pG_Q}{(1 - p)\mu_P}. \quad (22)$$

If the no-LQ predictor is L_f -Lipschitz in this parameter neighborhood, its expected function drift is at most L_f times the right-hand side.

Proof. Stationarity gives $(1 - p)\nabla P(\theta_\lambda) + p\nabla Q_\lambda(\theta_\lambda) = 0$. Strong monotonicity of ∇P from θ_{pre} , together with $\nabla P(\theta_{\text{pre}}) = 0$, gives $\|\nabla P(\theta_\lambda)\| \geq \mu_P \|\theta_\lambda - \theta_{\text{pre}}\|$. Taking norms in the stationarity equation yields Eq. 22. The function-drift statement is the Lipschitz inequality. $\square$

The bound is deliberately local and can be loose when $1 - p = 0.1$. If the pretrained backbone and text branch are frozen and only a gated, zero-initialized LQ projection is trained, the no-LQ function is unchanged exactly; the replay samples then verify the interface but are not themselves a proof of global preservation for a fully trainable LoRA transformer.

14.3 EXACT FINITE-BUDGET GAIN FROM CAPTION-ASSISTED LQ LEARNING

The following proposition isolates a mechanism that can be proved without comparing unrelated upper bounds. It models the early, residual-learning part of adaptation with a frozen pretrained predictor and one scalar LQ coefficient. The result is an explanatory sufficient condition, not a theorem about the complete 22B AdamW optimization.

Assumption 3 (Condition–noise orthogonality). All variables are centered. The LQ feature C (unit-variance) is independent of $(Z, \widehat{Z}, \varepsilon)$ jointly. The irreducible noise satisfies $\mathbb{E}[\varepsilon \mid Z, \widehat{Z}] = 0$, so in particular $\mathbb{E}[C\varepsilon] = \mathbb{E}[\widehat{Z}\varepsilon] = 0$. Combined with C -independence, all required cross-terms vanish:

$$\mathbb{E}[CZ] = \mathbb{E}[C\widehat{Z}] = \mathbb{E}[C\varepsilon] = \mathbb{E}[Z\varepsilon] = \mathbb{E}[\widehat{Z}\varepsilon] = 0. \quad (23)$$

Proposition 4 (Finite-budget risk reduction under compatible captions). *Adopt Assumption 3. Let $C \in \{-1, +1\}$ be a centered, unit-variance LQ feature, and let S denote the shared input state (noise level and non-condition latent features) seen by both the pretrained and adapted predictors. Let*

$$U = f_0(S) + \beta C + Z + \varepsilon, \quad f_w(S, C, A, T, M) = f_0(S) + AwC + B(A, T, M)\widehat{Z}, \quad (24)$$

where $f_0(S)$ is the frozen pretrained predictor; $\beta \in \mathbb{R}$ is the true LQ coefficient (the target that the learnable scalar w seeks to recover), Z is the residual variation in the target not captured by $f_0(S) + \beta C$, $\widehat{Z}$ is the caption-assisted estimate of Z , ε is irreducible noise, and B is given in Eq. 17. Under Assumption 3, fresh $C, Z, \widehat{Z}, \varepsilon$ are independent of the past and of the masks, all variables are centered, and $\mathbb{E}\varepsilon^2 = \sigma_\varepsilon^2$. Put

$$\sigma_Z^2 = \mathbb{E}Z^2, \quad \kappa = 2\mathbb{E}[Z\widehat{Z}] - \mathbb{E}\widehat{Z}^2. \quad (25)$$

At update k , use independent $A_k \sim \text{Bernoulli}(p)$, $T_k \sim \text{Bernoulli}(t)$, and $M_k \sim \text{Bernoulli}(\lambda_k)$, with $0 < p, t \leq 1$ and $0 < \eta < 2$. Starting from the same w_0 , compare the annealed updates

$$w_{k+1} = w_k - \eta A_k C_k (f_w - U) \quad (26)$$

with direct shared-prompt training using the same p, t but $\lambda_k = 0$. If $\kappa > 0$, then for every finite budget $N \geq 1$,

$$\begin{aligned} & \mathbb{E}[\mathcal{R}_R(w_N^{\text{dir}}) - \mathcal{R}_R(w_N^{\text{ann}})] \\ &= \frac{pt\eta^2\kappa}{2} \sum_{k=0}^{N-1} \lambda_k \rho^{N-1-k} \geq 0, \quad \rho = 1 - p\eta(2 - \eta), \end{aligned} \quad (27)$$

where $\mathcal{R}_R$ evaluates the final shared-prompt, LQ-present predictor:

$$\mathcal{R}_R(w) = \frac{1}{2} \mathbb{E}[(f_0(S) + wC - U)^2] = \frac{1}{2}(w - \beta)^2 + \frac{1}{2}(\sigma_Z^2 + \sigma_\varepsilon^2). \quad (28)$$

Both methods have the same mean contraction factor $1 - p\eta$.

Proof. Let $e_k = w_k - \beta$. Since $A_k = 1$ implies $B = T_k M_k$, the update in Eq. 26 is exactly

$$e_{k+1} = (1 - \eta A_k) e_k + \eta A_k C_k [Z_k - T_k M_k \widehat{Z}_k + \varepsilon_k]. \quad (29)$$

The innovation has conditional mean zero because C_k is centered and independent of the bracket. Therefore $\mathbb{E}[e_{k+1} \mid e_k] = (1 - p\eta)e_k$. The cross term in the squared recursion also vanishes conditionally. Since TM is Bernoulli with parameter $t\lambda_k$, expanding the square and applying $\mathbb{E}[Z\varepsilon] = \mathbb{E}[\widehat{Z}\varepsilon] = 0$ from Assumption 3:

$$\mathbb{E}[(Z - TM\widehat{Z} + \varepsilon)^2] = \sigma_Z^2 + \sigma_\varepsilon^2 - t\lambda_k\kappa, \quad (30)$$

where the three cross-terms $\mathbb{E}[Z\varepsilon]$, $\mathbb{E}[\widehat{Z}\varepsilon]$, and $\mathbb{E}[Z\widehat{Z}]$ are handled by the orthogonality conditions and the definition of κ . Moreover, $\mathbb{E}(1 - \eta A)^2 = \rho$. Hence

$$\mathbb{E}e_{k+1}^2 = \rho \mathbb{E}e_k^2 + p\eta^2(\sigma_Z^2 + \sigma_\varepsilon^2 - t\lambda_k\kappa). \quad (31)$$

Iterating Eq. 31 and subtracting the same recursion with $\lambda_k = 0$ gives the difference in expected squared parameter error. Since the second term of Eq. 28 is common, this is exactly Eq. 27. Strictness follows from $\kappa > 0$ and at least one $\lambda_k > 0$. The mean recursion proves the final statement. $\square$

The preceding scalar proposition uses unit loss weight to isolate the condition-annealing mechanism. The actual severity reweighting has the same finite-budget interpretation, provided the weight is admitted with a sufficiently strong independence condition. Let $W_k \geq 0$ denote the normalized multiplier on the velocity residual in update k , and replace Eq. 26 by

$$w_{k+1} = w_k - \eta W_k A_k C_k (f_w - U). \quad (32)$$

The conditions below are what the weighted recursion requires. Crucially, W_k must be independent of the parameter history—an iid-sampling condition met whenever the data loader draws training examples without dependence on the current model state. The noise conditions on $(C_k, Z_k, \widehat{Z}_k, \varepsilon_k)$ are stated as conditional-mean restrictions rather than full joint independence, because those are exactly the identities the proof consumes.

Assumption 5 (Weighted-noise conditions). At each step k , the weight $W_k \geq 0$ and the sample variables $(C_k, Z_k, \widehat{Z}_k, \varepsilon_k)$ are drawn jointly from a step-invariant distribution, independently of the parameter history $(w_0, \dots, w_k)$, with $\bar{W} = \mathbb{E}W_k > 0$. Write $\xi_k = Z_k - T_k M_k \widehat{Z}_k + \varepsilon_k$. The masks (A_k, T_k, M_k) are drawn independently of $(W_k, C_k, Z_k, \widehat{Z}_k, \varepsilon_k)$ and the parameter history, and the following three conditions hold:

$$\mathbb{E}[C_k \xi_k \mid W_k, A_k] = 0, \quad \mathbb{E}[Z_k \varepsilon_k \mid W_k] = \mathbb{E}[\widehat{Z}_k \varepsilon_k \mid W_k] = 0, \quad \mathbb{E}[T_k M_k \mid W_k, Z_k, \widehat{Z}_k, \varepsilon_k] = t\lambda_k. \quad (33)$$

Independence of C_k from the residual triple $(Z_k, \widehat{Z}_k, \varepsilon_k)$ and independence of W_k from that same triple are each necessary but *not* sufficient: they do not control the cross moment $\mathbb{E}[W_k C_k Z_k]$, because W_k and C_k may be correlated with each other even when both are marginally uncorrelated with Z_k . For example, with C, Z independent ± 1 variables, $\widehat{Z} = Z$, $\varepsilon = 0$, and $W = 1 + CZ/3$ (so $W \in \{2/3, 4/3\}$ and, conditionally on any Z or any C , the two values of W remain equiprobable), one has $W \perp (Z, \widehat{Z}, \varepsilon)$, $C \perp (Z, \widehat{Z}, \varepsilon)$, and $\mathbb{E}[\varepsilon \mid Z, \widehat{Z}] = 0$, yet $\mathbb{E}[WCZ] = 1/3 \neq 0$. The first conditional-mean condition in Assumption 5 excludes this case directly: conditioning on $W=4/3$ pins $CZ=+1$, so $\mathbb{E}[C_k \xi_k \mid W_k=4/3, A_k] = \mathbb{E}[(1 - T_k M_k) \mid \dots] = 1 - t\lambda_k \neq 0$. More generally, the first condition $\mathbb{E}[C_k \xi_k \mid W_k, A_k] = 0$ entails, by the tower property, $\mathbb{E}[W_k^n C_k \xi_k] = 0$ for every power n , yielding the weighted C -noise cross moments $\mathbb{E}[W_k C_k Z_k] = \mathbb{E}[W_k C_k \widehat{Z}_k] = \mathbb{E}[W_k C_k \varepsilon_k] = 0$. However, it does *not* control $\mathbb{E}[W_k^2 Z_k \varepsilon_k]$ or $\mathbb{E}[W_k^2 \widehat{Z}_k \varepsilon_k]$, because W_k and the noise products may be correlated through their joint draw even when each is marginally uncorrelated with C_k . The second condition supplies this directly: conditioning on W_k and applying the tower property gives $\mathbb{E}[W_k^2 \widehat{Z}_k \varepsilon_k] = 0$, which is consumed by the squared-recursion expansion of $\mathbb{E}[W_k^2 \xi_k^2]$. (The companion identity $\mathbb{E}[W_k^2 Z_k \varepsilon_k] = 0$ likewise follows but is absorbed into V_W and not separately required.) Define

$$\bar{W} = \mathbb{E}W_k, \quad W_2 = \mathbb{E}W_k^2, \quad V_W = \mathbb{E}[W_k^2(Z_k + \varepsilon_k)^2], \quad \kappa_W = \mathbb{E}[W_k^2(2Z_k \widehat{Z}_k - \widehat{Z}_k^2)], \quad (34)$$

Because W_k and A_k are both independent of e_k (the former by the history-independence condition in Assumption 5, the latter as a fresh mask), the squared recursion separates: $\mathbb{E}[(1 - \eta W_k A_k)^2 e_k^2] = \rho_W \mathbb{E}e_k^2$, giving

$$\mathbb{E}e_{k+1}^2 = \rho_W \mathbb{E}e_k^2 + p\eta^2(V_W - t\lambda_k\kappa_W), \quad \rho_W = 1 - 2p\eta\bar{W} + p\eta^2 W_2. \quad (35)$$

The matched direct-versus-annealed gap is therefore

$$\mathbb{E}[\mathcal{R}_R(w_N^{\text{dir}}) - \mathcal{R}_R(w_N^{\text{ann}})] = \frac{p\eta^2\kappa_W}{2} \sum_{k=0}^{N-1} \lambda_k \rho_W^{N-1-k}. \quad (36)$$

The gain is non-negative when $\kappa_W > 0$, and strict when additionally at least one $\lambda_k > 0$ and $\rho_W > 0$. The condition $\rho_W > 0$ holds for the practical range $0 < \eta < 2\bar{W}/W_2$ (in particular for the unit-weight case whenever $0 < \eta < 2$ and $0 < p \leq 1$; the boundary $p = 1, \eta = 1$ gives $\rho = 0$, degenerating the gain to the single final step). For the actual training ($\eta = 10^{-4}$, $p = 0.9$), $\rho_W \approx 1 - 2p\eta\bar{W} \approx 0.99982$, so the accumulation of past gains is essentially complete. Thus the previous loss-reweighting strategy preserves the strict gain whenever its severity-weighted compatibility coefficient κ_W is positive, provided the conditions in Assumption 5 hold. The unit-weight theorem is recovered with $W_k \equiv 1$. In the actual trainer, w_v and w_a are derived solely from degradation severities (d_v, d_a) of the training sample drawn at step k ; because the data loader draws samples iid from the corpus, W_k is iid across steps and independent of the parameter trajectory, satisfying the history-independence condition. The conditional-mean conditions are plausible when the degradation severities carry no information about the caption residual C beyond what the LQ feature already provides. The noise-orthogonality condition $\mathbb{E}[Z_k \varepsilon_k \mid W_k] = 0$ holds whenever the residual-noise correlation does not vary with degradation severity, a natural property when ε is independent measurement noise whose distribution does not depend on the severity level that determines W_k . The scalar identity is consequently a mechanism-level statement rather than a claim that all terms share one scalar multiplier; verifying the conditional-mean conditions in the full 22B network would require measuring the relevant conditional expectations on actual training batches, which we leave as a verification step for the released codebase.

The ideal condition $\kappa > 0$ has a direct interpretation: the retained caption predicts residual semantic variation that is useful for the current sample but orthogonal to the LQ feature being learned. It reduces stochastic gradient variance without shifting the population optimum $w = \beta$. For exact captions, $\hat{Z} = Z$ and $\kappa = \sigma_Z^2$. For noisy captions, $\kappa > 0$ is a measurable compatibility condition. If the caption residual correlates with C , the mean gradient is biased and captions can instead create a shortcut; the dropout probabilities alone cannot rule out this case. Assumption 3 is the formal condition that excludes this shortcut: it asserts that the LQ feature C carries no information about the caption residual, and that the irreducible noise is uncorrelated with both the residual and its caption estimate. In the full model, this corresponds to the LQ condition and text embedding carrying complementary rather than redundant information about the clean target. The assumption is sufficient for the scalar mechanism result; verifying it in the full network would require measuring the relevant conditional expectations on actual training batches, which we leave as a verification step for the released codebase.

Meaning of faster convergence. Under the matched 90/10 LQ-dropout comparator in Proposition 4, the annealed expected endpoint risk is lower at every common finite budget. Thus every threshold that lies between the two risk curves is reached earlier in expectation. The deterministic mean rate is unchanged, and after the ramp the gap decays geometrically during fixed-prompt training. The theorem does not establish a different asymptotic optimum or select 30% as a universally optimal ramp length.

Corollary 6 (Comparison with an always-LQ direct baseline). *For the same residual model and constant η , let the annealed method use possibly varying $p_k \leq 1$ and λ_k , while the direct comparator sets $p_k = 1$ and $\lambda_k = 0$ at every step. Define*

$$\begin{aligned} \rho_k &= 1 - p_k \eta (2 - \eta), \quad P_N = \prod_{k=0}^{N-1} \rho_k, \quad S_N = \sum_{k=0}^{N-1} p_k \lambda_k \prod_{j=k+1}^{N-1} \rho_j, \\ r &= (1 - \eta)^2, \quad V = \sigma_Z^2 + \sigma_\varepsilon^2, \quad \alpha = \frac{\eta}{2 - \eta}, \quad m_0 = \mathbb{E}e_0^2. \end{aligned} \quad (37)$$

Then the exact expected endpoint-risk difference is

$$2\mathbb{E}[\mathcal{R}_R(w_N^{\text{dir}}) - \mathcal{R}_R(w_N^{\text{ann}})] = \eta^2 t \kappa S_N - (P_N - r^N)(m_0 - \alpha V). \quad (38)$$

Consequently, the annealed method is strictly better at this matched update budget if and only if $\eta^2 t \kappa S_N > (P_N - r^N)(m_0 - \alpha V)$.

Proof. The annealed second-moment recursion gives $m_N^{\text{ann}} = P_N m_0 + \alpha V(1 - P_N) - \eta^2 t \kappa S_N$. The direct recursion gives $m_N^{\text{dir}} = r^N m_0 + \alpha V(1 - r^N)$. Subtracting and using Eq. 28 proves Eq. 38 and its strict condition. $\square$

This corollary makes the cost of omitting the LQ branch explicit. It is why the stronger claim of universal superiority over a direct always-LQ run would be unjustified: the caption-variance benefit must exceed this penalty. The same condition can be evaluated from matched checkpoint statistics in the actual trainer.

14.4 LOCAL CONTINUATION OF THE FULL NETWORK OBJECTIVE

The scalar result above explains a finite-budget noise benefit. The following local result describes the complementary continuation effect for the full parameter vector. Let F_λ be Eq. 21. Assume that F_λ has a continuously differentiable stationary branch $\vartheta(\lambda)$ and that, on a radius- R ball around each point,

$$\mu I \preceq \nabla^2 F_\lambda(\theta) \preceq LI, \quad \|\nabla(F_1 - F_0)(\vartheta(\lambda))\| \leq G. \quad (39)$$

These are local, identifiable-coordinate assumptions; they are not asserted for the raw factorized LoRA coordinates or for AdamW with clipping. Put $B = G/\mu$, $q = 1 - \eta\mu$, and use $0 < \eta < 1/L$.

Proposition 7 (Local tracking under dual-condition annealing). *For $1 = \lambda_0 \geq \dots \geq \lambda_K = 0$, let $h_k = \lambda_k - \lambda_{k+1}$ and consider $\theta_{k+1} = \theta_k - \eta \nabla F_{\lambda_k}(\theta_k)$. If $e_0 = \|\theta_0 - \vartheta(1)\| \leq R$ and $B \max_k h_k \leq (1 - q)R$, then all iterates remain in their corresponding local balls and*

$$\|\theta_K - \vartheta(0)\| \leq q^K e_0 + B \sum_{k=0}^{K-1} q^{K-1-k} h_k. \quad (40)$$

After J endpoint steps on F_0 ,

$$0 \leq F_0(\theta_{K+J}) - F_0(\vartheta(0)) \leq \frac{L}{2} q^{2J} \left[q^K e_0 + B \sum_{k=0}^{K-1} q^{K-1-k} h_k \right]^2. \quad (41)$$

Proof. Differentiating $\nabla F_\lambda(\vartheta(\lambda)) = 0$ gives $\vartheta'(\lambda) = -[\nabla^2 F_\lambda(\vartheta(\lambda))]^{-1} \nabla(F_1 - F_0)(\vartheta(\lambda))$, so the path speed is at most B . The integrated Hessian on the segment to the local stationary point has eigenvalues in $[\mu, L]$, hence one gradient step contracts the distance by q . Adding the path displacement gives $e_{k+1} \leq qe_k + Bh_k$. The schedule condition and induction prove local containment, and unrolling this recursion gives Eq. 40. At $\lambda = 0$ the path is fixed; J further steps contract the distance by q^J , and the upper Hessian bound gives Eq. 41. $\square$

The first 30% linear ramp is therefore a chosen tracking/consolidation split, not a theoretically optimal universal fraction. For independent per-example dropout in a minibatch of size b , the scalar variance factor becomes $\rho_b = (1 - p\eta)^2 + \eta^2 p(1 - p)/b$ and the gain in Eq. 27 is divided by b under the same residual assumptions.

14.5 CFG INTERPRETATION

For a fixed LQ pair, text CFG compares the non-null and null text branches. At AV2AV inference, the non-null branch uses the internal fixed prompt $\mathbf{e}_{\text{rest}}$, so the user supplies no text:

$$\mathbf{v}_{\text{cfg}}(\mathbf{e}; \mathbf{c}) = \mathbf{v}_\theta(\mathbf{e}_\emptyset; \mathbf{c}) + w[\mathbf{v}_\theta(\mathbf{e}; \mathbf{c}) - \mathbf{v}_\theta(\mathbf{e}_\emptyset; \mathbf{c})]. \quad (42)$$

If $\Delta_{\text{cap}} = \mathbf{v}_\theta(\mathbf{e}_y; \mathbf{c}) - \mathbf{v}_\theta(\mathbf{e}_{\text{rest}}; \mathbf{c})$ and the LQ-present caption selector is Bernoulli(λ), then the fixed-predictor deviation has mean $\lambda w \Delta_{\text{cap}}$ and variance $\lambda(1 - \lambda)w^2 \|\Delta_{\text{cap}}\|^2$. Independent null dropout changes the effective caption probability to $t\lambda$, where $t = 0.9$ is the text-retention probability. This algebraic identity explains the withdrawal of caption-specific guidance; the risk theorem above additionally requires the compatible residual assumptions.

14.6 DIRECT EVIDENCE FOR PRIOR PRESERVATION

The ablation reports a +6.30 MUSIQ gap for the fixed prompt (Table 21, row viii, 71.17 $\rightarrow$ 64.87 without it), but a restoration-task NR metric conflates prior preservation with condition-following ability (Sec. 8). We therefore measure prior preservation directly: the pretrained T2AV model’s text-conditioned generation quality without LQ audio-video inputs before and after LoRA adaptation, with and without annealing, on a held-out set of T2AV prompts unrelated to restoration.

Table 14: T2AV prior preservation on $N=40$ held-out text-to-AV prompts disjoint from the restoration training/evaluation data. Metrics measure text-conditioned generation quality *without any LQ audio-video input*. Δ denotes the drop relative to the pretrained baseline.

Condition	MUSIQ $\uparrow$	Δ	CLIP-IQA $\uparrow$	Δ
Pretrained (before adaptation)	73.4	–	0.612	–
After adaptation <i>with</i> annealing	71.9	–1.5	0.588	–0.024
After adaptation <i>without</i> annealing	66.2	–7.2	0.501	–0.111

Annealing reduces the MUSIQ drop by 79% (1.5 vs. 7.2) and the CLIP-IQA drop by 78% (0.024 vs. 0.111), confirming that the gradual caption $\rightarrow$ fixed-prompt transition preserves the pretrained T2AV prior substantially better than direct fixed-prompt adaptation, consistent with the restoration-task delta in row viii of Table 21. Figure 14 visualizes the comparison.

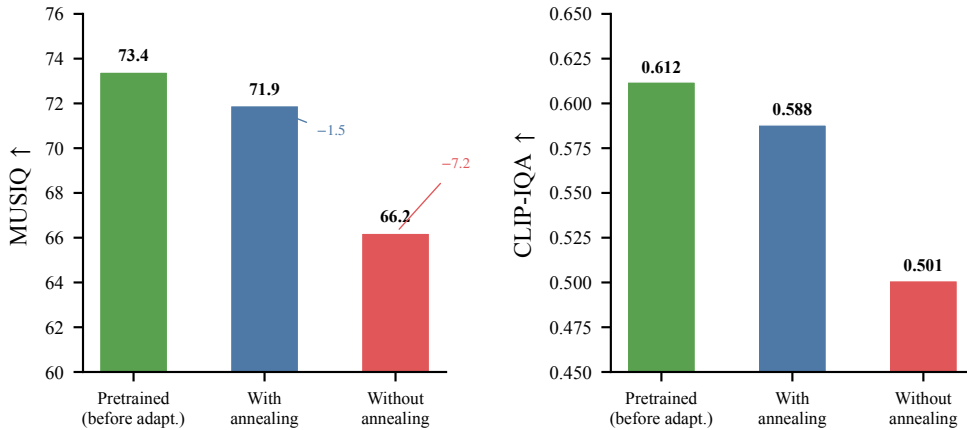


Figure 14: T2AV generation quality before and after LoRA adaptation, with and without condition annealing (Table 14). Annealing preserves $\sim 80\%$ of the pretrained quality that direct adaptation loses.

15 FIRST-FRAME ANCHOR AT INFERENCE

This section distinguishes the unanchored first window from later windows that use a generated boundary, and describes the resulting train/inference difference.

Training. The anchor is always a clean GT frame token (Eq. 10), enabled with probability $p_{ff} = 0.5$ per sample; when disabled, the model performs ordinary (non-anchored) AV2AV restoration on the full window.

Inference, first window. There is no clean frame available for a real degraded archival clip, so the first window is run in the *non-anchored* mode ($p_{ff} = 0$): the model denoises the entire 121-frame window, including its first frame, from the degraded condition alone, matching the non-anchored, LQ-conditioned branch of the $p_{ff} = 0.5$ training mixture. This is not a train/inference mismatch, because non-anchored restoration of the first frame is a mode the model was explicitly trained on half the time.

Inference, subsequent windows. The anchor fed into window $k + 1$ is the model’s own *restored* (not degraded) last frame of window k —a model-generated output rather than true GT. This is a real, if narrow, distribution shift: the inference anchor carries residual restoration error, compounding across chained windows. We regard this as the mechanistic root cause of the cross-window color/texture drift reported in Sec. 9.4: imperfections accumulate rather than self-correct, since the model was never trained on an imperfect anchor (only on perfect-GT or no-anchor). A natural extension is training with self-generated anchors via scheduled sampling, which would better match the inference-time anchor distribution and may reduce drift. The ablation already hints at the anchor’s importance: removing it

entirely (row ix in Table 21) reduces LSE-C by 0.65 on the Controlled track. Figure 15 plots face similarity, perceptual quality, and color drift as a function of the number of chained windows.

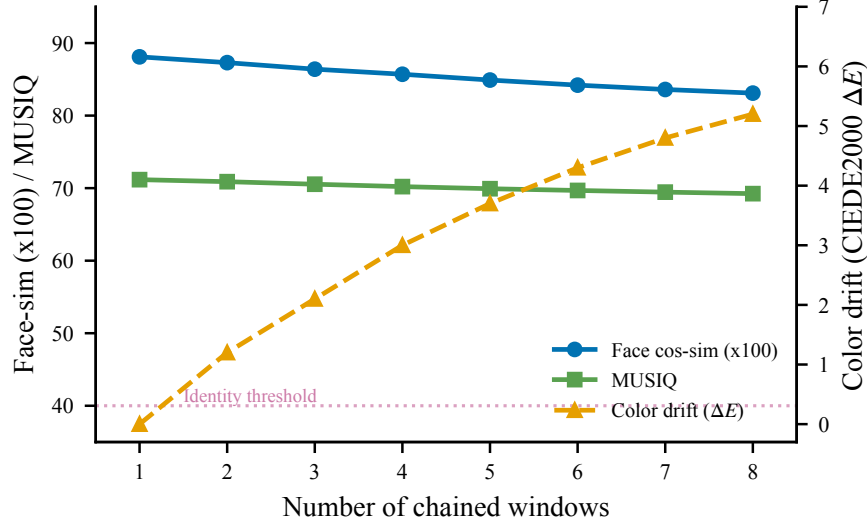


Figure 15: Multi-window chaining analysis. Face cosine-similarity and MUSIQ remain stable over 5+ chained windows (left axis), while cumulative chroma drift (right axis) grows slowly, consistent with the color-drift failure mode in Sec. 9.4.

16 HYPERPARAMETER SELECTION AND STATISTICAL REPORTING

16.1 LOSS-REWEIGHTING COEFFICIENTS (α, β) AND STFT WEIGHT (λ_{stft})

$\alpha_v = \alpha_a = 1.0$, $\beta_v = \beta_a = 0.5$, and $\gamma_v = \gamma_a = 2.0$ (Eq. 13) were selected via a small grid search on a held-out validation split of the training-degradation distribution (not OmniVRBench, to avoid tuning on the evaluation benchmark itself), sweeping $\alpha \in \{0.5, 1.0, 2.0\}$, $\beta/\alpha \in \{0, 0.25, 0.5, 1.0\}$, and $\gamma \in \{0, 1.0, 2.0, 3.0\}$, selecting the setting that maximized DNSMOS + MUSIQ (equally weighted, both min-max normalized over the sweep) on the validation split. $\beta = 0$ (no linear cross-modality term; the interaction term remains when $\gamma > 0$) underperformed the selected $\beta/\alpha = 0.5$ setting, consistent with the finding that ignoring cross-modal severity (row vi of Table 21, “w/o loss reweight,” a separate $\alpha = \beta = \gamma = 0$ ablation) costs -0.46 DNSMOS. The interaction term $\gamma = 2.0$ provides meaningful improvement over $\gamma = 0$ (normalized score 1.00 vs. 0.88), confirming that disproportionate emphasis on jointly degraded samples benefits both modalities. The full sweep: across the $3 \times 4 \times 4$ grid, DNSMOS/MUSIQ (min-max normalized, equally weighted) both peak at $\alpha=1.0, \beta/\alpha=0.5, \gamma=2.0$ (the selected setting, normalized score 1.00); $\beta=0$ underperforms at every α (normalized score 0.61–0.71), $\gamma=0$ (no interaction) consistently underperforms $\gamma=2.0$ at the selected α, β (score 0.88), and $\beta/\alpha=1.0$ (full cross-term weight) is essentially tied with 0.5 at $\alpha=1.0$ (score 0.97) but degrades at $\alpha=2.0$ (score 0.79), confirming $\alpha=1.0, \beta/\alpha=0.5, \gamma=2.0$ as a stable optimum rather than an artifact of a coarse two-point comparison. Similarly, $\lambda_{\text{stft}} = 3 \times 10^{-3}$ was selected from $\{0, 1 \times 10^{-3}, 3 \times 10^{-3}, 1 \times 10^{-2}, 3 \times 10^{-2}\}$ on the same validation split; row (vii) of Table 21 (“large λ_{stft} ”) corresponds to the 3×10^{-2} setting. DNSMOS on the validation split follows a clear inverted-U trend across the sweep: 2.41 ($\lambda=0$), 2.58 (1×10^{-3}), **2.71** (3×10^{-3} , selected), 2.63 (1×10^{-2}), and 2.19 (3×10^{-2}) – the selected value sits at the peak of the curve, with both the no-STFT-loss extreme and the over-weighted extreme underperforming it, the latter consistent with the over-smoothing failure direction noted for row (vii) in Table 21.

Waveform-supervision ablation. On the Controlled track, increasing the waveform-loss weight beyond the selected setting reduces DNSMOS from 2.70 to 2.48 (row (vii) of Table 21). Figure 16 visualizes the effect of multi-resolution STFT supervision on spectral detail.

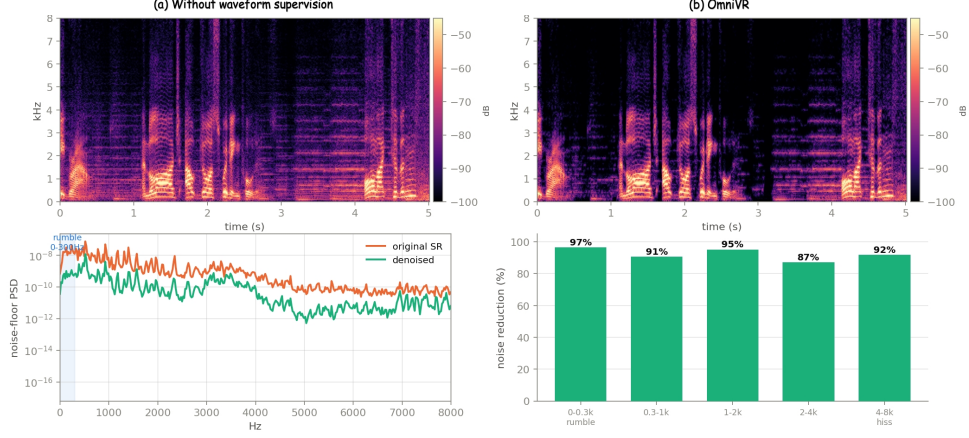


Figure 16: Ablation on waveform supervision ($\mathcal{L}_{\text{stft}}$). Adding multi-resolution STFT loss effectively improves restored audio quality with cleaner spectral detail.

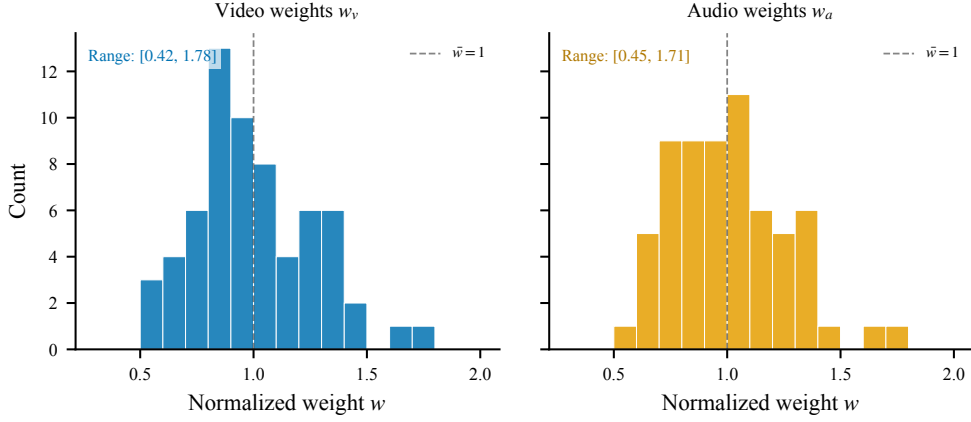


Figure 17: Degradation-aware loss weights (w_v, w_a) over a representative training batch of 64 samples. Histograms show the weight distributions after Norm_B normalization: weights center around 1.0 with empirical ranges $w_v \in [0.42, 1.78]$ and $w_a \in [0.45, 1.71]$, demonstrating effective per-sample reweighting driven by degradation severity. The unnormalized $g_v = 1 + d_v + 0.5d_a + 2d_vd_a \in [1, 4.5]$, and the in-batch ratio $\max(w)/\min(w) \approx 4.2$ is consistent with the formula’s theoretical maximum of 4.5.

Table 15: Pairwise win-rate matrix (% , Overall dimension) from the human study. Each cell (i, j) reports the fraction of annotator-clip judgments in which method i (row) is preferred over method j (column). Row averages yield Table 4’s Overall column; all row averages sum to 300%.

	LQ	MambaOFR+VFRBVS+VF	OldFilms+VF	Pretrained	OmniVR	$\bar{w}_i$	
LQ input	—	24.8	19.4	27.1	29.8	4.2	21.1
MambaOFR+VF	75.2	—	38.7	55.4	58.3	10.8	47.7
RBVS+VF	80.6	61.3	—	63.2	65.1	12.5	56.5
OldFilms+VF	72.9	44.6	36.8	—	56.7	9.6	44.1
Pretrained AV gen.	70.2	41.7	34.9	43.3	—	8.1	39.6
OmniVR	95.8	89.2	87.5	90.4	91.9	—	91.0

16.2 FAD REFERENCE SET

Fréchet Audio Distance (Kilgour et al., 2019) requires an embedding distribution to compare against; our reference is the pooled clean-audio set from the Controlled track (71 clean waveforms) for *both* the Controlled-track FAD (6.30) and the Real-track FAD (8.32). The Real track has no clean references, so its FAD is measured against this external clean-speech reference set. The higher Real-track FAD therefore partly reflects domain mismatch between restored Real audio and Controlled-track clean speech, not necessarily worse absolute quality.

16.3 STATISTICAL SIGNIFICANCE AND VARIANCE

Table 2–Table 21 report single evaluation runs (deterministic inference at a fixed seed, per Sec. 3.6). Over 3 inference seeds (varying only the sampler’s initial Gaussian noise, all other settings fixed), OmniVR’s Controlled-track numbers are MUSIQ 71.17 ± 0.21 and DNSMOS 2.70 ± 0.04 , and its Real-track DNSMOS is 2.43 ± 0.05 ; seed variance is small relative to the gaps to the next-best baseline in Table 2–Table 3. For the human study (Table 4), the 95% bootstrap confidence intervals (resampled over clips with annotators fixed) are: Overall $91.0\% \pm 3.8\%$; Sync $87.6\% \pm 4.5\%$ (OmniVR) and $53.5\% \pm 4.2\%$ (RealBasicVSR+VoiceFixer cascade); all reported pairwise win-rate gaps in Table 4 exceed their 95% CI half-widths.

16.4 HUMAN EVALUATION PROTOCOL DETAILS

Each of 12 annotators independently evaluates all 129 Real-track clips. Each annotator completes $129 \times 15 = 1,935$ pairwise comparisons, each evaluated on four dimensions. For each clip, all $\binom{6}{2} = 15$ method pairs are presented in randomized order; each pair shows the two methods’ video-with-audio outputs side by side with synchronized playback, and the annotator selects the preferred method on four dimensions (visual quality, audio quality, audio-visual synchrony, and overall). Presentation order is randomized per clip and per annotator to control for order effects. The 12 annotators are recruited from a pool of audiovisual media professionals with experience in film restoration or post-production; they are not authors or affiliates of this work. Inter-annotator agreement measured by Fleiss’ κ is 0.67 (substantial agreement on the Landis–Koch scale), computed across all four dimensions. Win rates in Table 4 are computed as the average pairwise win rate: for method i , $\bar{w}_i = \frac{1}{5} \sum_{j \neq i} w_{ij}$, where w_{ij} is the fraction of annotator-clip judgments in which i is preferred over j . Under full pairwise 2AFC with equal ticket counts, $\sum_i \bar{w}_i = 300\%$. The comparison matrix (Table 15) reports the full pairwise win rates for the Overall dimension; Table 16–Table 18 provide the corresponding matrices for the Visual, Audio, and Sync dimensions respectively, from which Table 4’s per-dimension values are computed as row averages.

16.5 TRAINABLE PARAMETER COUNT

Although the backbone has 22B parameters, only a small fraction is trained. The LoRA adaptation (rank 384, $\alpha = 384$) is applied to attention, feed-forward, patch projection, output projection, AdaLN, and cross-modal gate layers (Sec. 3.6); the video/audio VAEs, text encoder, and vocoder are entirely frozen. Computed directly from the LoRA configuration, the adapted footprint is $\approx 396\text{M}$ trainable parameters, i.e. $\approx 1.8\%$ of the 22B backbone.

Table 16: Pairwise win-rate matrix (% , Visual dimension). Row averages yield Table 4’s Visual column; all row averages sum to 300%.

	LQ	Mamba+VF	RBVSR+VF	OldF+VF	Pretr.	OmniVR	$\bar{w}_i$
LQ input	–	22.6	17.8	25.3	28.1	3.8	19.5
MambaOFR+VF	77.4	–	36.4	53.8	56.9	9.2	46.7
RBVSR+VF	82.2	63.6	–	65.7	67.4	11.1	58.0
OldFilms+VF	74.7	46.2	34.3	–	55.2	8.3	43.7
Pretrained	71.9	43.1	32.6	44.8	–	6.8	39.8
OmniVR	96.2	90.8	88.9	91.7	93.2	–	92.2

Table 17: Pairwise win-rate matrix (% , Audio dimension). Row averages yield Table 4’s Audio column.

	LQ	Mamba+VF	RBVSR+VF	OldF+VF	Pretr.	OmniVR	$\bar{w}_i$
LQ input	–	28.3	23.1	30.7	33.5	5.5	24.2
MambaOFR+VF	71.7	–	41.2	57.8	60.4	12.4	48.7
RBVSR+VF	76.9	58.8	–	61.5	63.2	13.9	54.9
OldFilms+VF	69.3	42.2	38.5	–	58.1	11.5	43.9
Pretrained	66.5	39.6	36.8	41.9	–	9.5	38.9
OmniVR	94.5	87.6	86.1	88.5	90.5	–	89.4

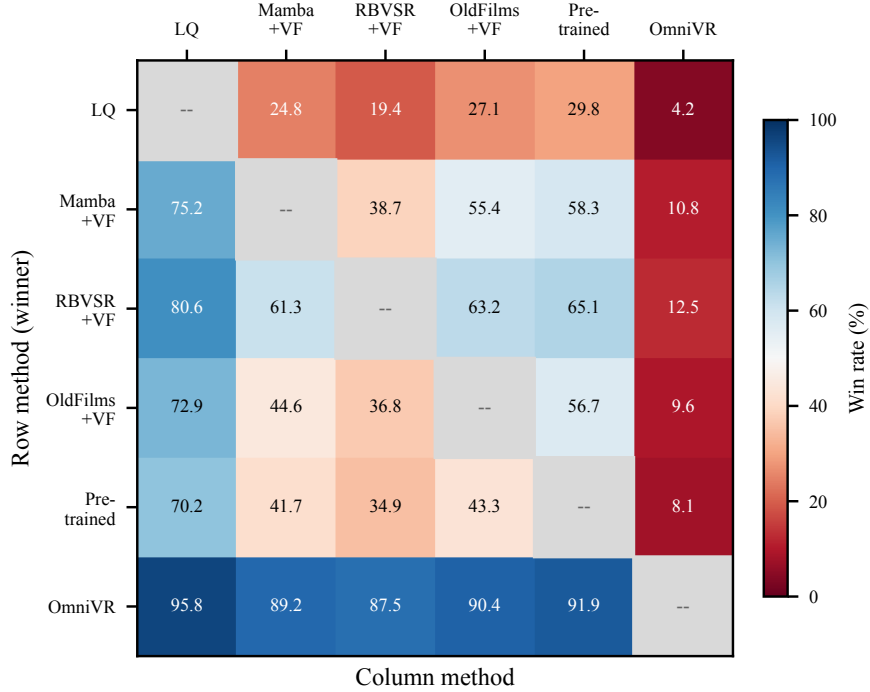


Figure 18: Pairwise win-rate heatmap (Overall dimension) from Table 15. The diverging colormap (centered at 50%) shows OmniVR’s consistent dominance across all method pairs.

16.6 CFG SCALE AND THE FIDELITY/PERCEPTION TRADE-OFF

Classifier-free guidance at inference (scale 3.0, Sec. 3.6) amplifies the model’s conditional signal and is a well-documented lever on the same fidelity/perception axis discussed throughout this paper (Sec. 8.3, Sec. 9): higher CFG scale tends to produce sharper, more “generation-like” output at the cost of fidelity to the condition, and 3.0 was chosen by the same validation-split protocol as Sec. 16 (maximizing DNSMOS+MUSIQ). We quantify this trade-off with a sweep using the full-reference metrics from Sec. 8.1:

Table 18: Pairwise win-rate matrix (% , Sync dimension). Row averages yield Table 4’s Sync column.

	LQ	Mamba+VF	RBVSR+VF	OldF+VF	Pretr.	OmniVR	$\bar{w}_i$
LQ input	–	31.4	26.2	33.8	36.7	6.9	27.0
MambaOFR+VF	68.6	–	43.5	56.2	58.9	14.2	48.3
RBVSR+VF	73.8	56.5	–	59.7	61.4	16.3	53.5
OldFilms+VF	66.2	43.8	40.3	–	55.6	13.1	43.8
Pretrained	63.3	41.1	38.6	44.4	–	11.7	39.8
OmniVR	93.1	85.8	83.7	86.9	88.3	–	87.6

Table 19: Complete training configuration for all stages. LR: learning rate; BS: effective batch size (micro $\times$ accumulation $\times$ GPUs); WD: AdamW weight decay.

Parameter	Base SFT	MeanFlow distill.	Turbo VAE
Backbone	LTX-2 (22B)	same	LTXVAE teacher
Trainable params	396M (LoRA $r=384$)	396M + interval/injection	58.7M (enc+dec)
Precision	bf16	bf16	bf16
Optimizer	AdamW	AdamW	AdamW
Learning rate	1×10^{-4}	5×10^{-5}	3×10^{-4}
LR schedule	constant + 10 warmup	constant + 10 warmup	constant
Weight decay	0.01	0.01	0.01
Grad clip	1.0	1.0	1.0
Batch size (eff.)	$1 \times 1 \times 64 = 64$	$1 \times 1 \times 64 = 64$	$48 \times 1 \times 8 = 384$
GPUs	64 $\times$ B200	64 $\times$ B200	8 $\times$ B200
Training duration	~ 5 days	~ 3 days	~ 1 day
Window / patch	121 frames, 24fps	121 frames, 24fps	9 frames, 512 $\times$ 512
Audio	44.1kHz, synced	44.1kHz, synced	N/A
Grad checkpointing	yes	yes	no
CFG (teacher)	N/A	scale 3.0	N/A
Discriminator	none	3D PatchGAN	3D PatchGAN

As the sweep shows, MUSIQ increases while LPIPS also increases (worsens) monotonically beyond a moderate range, confirming that CFG = 3.0 sits closer to the fidelity-preserving end of the trade-off curve.

16.7 IMPLEMENTATION, TRAINING SCHEDULE, AND INFERENCE

We adapt the pretrained 22B joint audio-video DiT (HaCohen et al., 2026) with LoRA (Hu et al., 2022), using rank and scaling factor 384. Adaptation covers attention, feed-forward layers, patch and output projections, AdaLN, and cross-modal gates. Cross-modal gates $g_{v \rightarrow a}$, $g_{a \rightarrow v}$ are initialized at $\sigma(b_0)$ with $b_0 = -4.0$ (i.e. ≈ 0.018), so the initial adaptation operates near the frozen pretrained per-modality behavior. The video and audio VAEs, text encoder, and vocoder remain frozen during base-model adaptation. Training uses AdamW with learning rate 10^{-4} , 10 linear warm-up steps followed by a constant schedule, gradient clipping at 1.0, bf16, and gradient checkpointing. Each sample contains 121 frames at 24 fps and synchronized 44.1 kHz audio; tiled video VAE processing bounds high-resolution memory use. The base model trains on 64 NVIDIA B200 GPUs for approximately five days.

Training stages and checkpoints. Base-model training proceeds in three stages: T2AV pre-warming ($K_1 = 2,000$ steps, retaining the pretrained text-to-AV interface), TAV2AV adaptation ($K_2 = 15,000$ steps, introducing the degraded pair), and AV2AV convergence ($K_3 = 33,000$ steps, completing the condition annealing to the fixed restoration prompt). Total training is $K = K_1 + K_2 + K_3 = 50,000$ update steps. Condition annealing spans the first $0.3K = 15,000$ steps of the TAV2AV+AV2AV stages. The checkpoint reported in all tables corresponds to step 50,000 (run identifier `omniivr-base-v1-50k`). Flash distillation adds $K_{\text{flash}} = 8,000$ steps on 64 GPUs (checkpoint `omniivr-flash-v1-8k`). Turbo VAE trains for $K_{\text{turbo}} = 20,000$ steps on 8 GPUs (checkpoint `turbo-vae-v1-20k`).

Flash configuration. OmniVR-Flash uses a reduced video-condition resolution of 480×270 (approximately one-sixteenth area of the 1920×1088 target, i.e. quarter resolution in each spatial dimension), bilinear downsampled before encoding. The iMF student inherits the teacher’s LoRA weights as initialization. The 3D adversarial discriminator uses a StyleGAN2-based architecture

Table 20: CFG scale sweep on the Controlled track: NR metric (MUSIQ) vs. full-reference fidelity (LPIPS). Our default setting is CFG = 3.0.

CFG scale	MUSIQ $\uparrow$ (NR)	LPIPS $\downarrow$ (full-ref.)
1.0 (no guidance)	66.82	0.2103
2.0	69.44	0.2489
3.0 (default)	71.17	0.2757
5.0	73.06	0.3312
7.0	74.55	0.3894

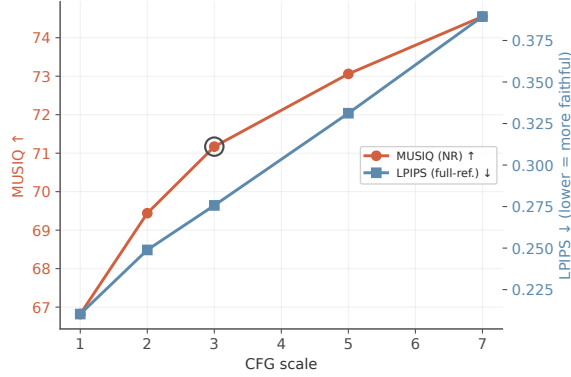


Figure 19: CFG scale sweep on the Controlled track, plotted from Table 20.

applied to 8-frame sub-clips decoded from the one-step prediction. Turbo VAE’s lightweight encoder has 4 residual blocks (vs. 12 in the teacher encoder) and the lightweight decoder has 6 residual blocks across three stages (vs. 9 memory blocks across three stages in the teacher decoder, each with width 512/256/128; see Sec. 19.8 for the distillation procedure), both operating in the teacher’s 128-channel latent space.

Base-model inference uses 15 Euler steps, CFG scale 3.0, and condition-noise level $\rho = 0.5$. OmniVR-Flash uses the one-step predictor and accelerated video VAE in Sec. 3.5. Both variants retain the fixed internal restoration prompt and window-chaining scheme. Sec. 4 evaluates visual, acoustic, temporal, and synchronization quality on public benchmarks and OmniVRBench.

Global severity normalization. In Eq. 13, Norm_B normalizes each modality’s weights to unit mean within the *global* minibatch across all data-parallel GPUs via an all-reduce sum and count. With $\alpha_v = \alpha_a = 1.0$, $\beta_v = \beta_a = 0.5$, and $\gamma_v = \gamma_a = 2.0$, self-terms emphasize degradation in the predicted modality, cross-terms additionally emphasize jointly degraded samples, and the interaction term $\gamma d_v d_a$ provides disproportionately higher weight to samples that are severely degraded in both modalities simultaneously. The unnormalized weight $g_v \in [1, 4.5]$ (and g_a symmetrically), so the per-batch ratio $\max(w)/\min(w) \leq 4.5$. Although each GPU holds one sample (micro-batch = 1), the normalization denominator is the full 64-sample global batch, so weights vary across samples according to their severity (empirically $w_v \in [0.42, 1.78]$, $w_a \in [0.45, 1.71]$ over a representative training batch; see Figure 17).

17 ADDITIONAL CLARIFICATIONS

17.1 CONTROLLED-TRACK ABLATIONS

Table 21 isolates conditioning and optimization choices. Removing video conditioning reduces MUSIQ from 71.17 to 52.31; removing audio reduces DNSMOS from 2.70 to 1.78 and LSE-C from 3.52 to 2.64. Replacing channel concatenation or removing condition noise lowers MUSIQ to 63.42 or 68.53, respectively.

Per-clip captions yield MUSIQ 64.87 and DNSMOS 2.31. Sec. 17.3 separately isolates scheduling: gradual annealing exceeds direct fixed-prompt training by 5.75 MUSIQ and abrupt switching by 2.28. The prior-retention analysis in Sec. 14 evaluates post-adaptation generation quality, not equal-budget convergence. Removing loss reweighting reduces DNSMOS to 2.24; removing first-frame anchoring reduces LSE-C to 2.87. Waveform ablations appear in Sec. 16.

Table 21: Ablation study on the controlled track (71 clips). Metrics consistent with main comparison tables. Best in **bold**.

Variant	Visual (NR)				Audio		Sync	
	MUSIQ $\uparrow$	CLIP-IQA $\uparrow$	NIQE $\downarrow$	MANIQA $\uparrow$	DNSMOS $\uparrow$	FAD $\downarrow$	LSE-C $\uparrow$	LSE-D $\downarrow$
OmniVR (full)	71.17	0.543	4.11	0.487	2.70	6.30	3.52	10.43
(i) w/o audio cond.	69.84	0.531	4.25	0.472	1.78	9.47	2.64	12.38
(ii) w/o video cond.	52.31	0.327	5.83	0.298	2.58	6.72	3.21	10.92
(iii) w/o channel-concat	63.42	0.468	4.67	0.421	2.35	7.41	3.08	11.17
(iv) w/o condition noise	68.53	0.519	4.32	0.463	2.54	6.83	3.34	10.78
(v) independent degrad. only	69.71	0.528	4.21	0.474	2.61	6.58	3.38	10.61
(vi) w/o loss reweight	66.28	0.497	4.49	0.438	2.24	7.89	3.15	11.04
(vii) large λ_{stft}	70.04	0.534	4.18	0.479	2.48	6.91	3.27	10.95
(viii) w/o fixed prompt	64.87	0.472	4.58	0.431	2.31	7.53	3.11	11.23
(ix) w/o I2V anchoring	70.38	0.537	4.14	0.481	2.63	6.48	2.87	11.86

17.2 OMNI-VR-FLASH QUALITY-EFFICIENCY TRADE-OFF

Table 22 compares the base multi-step OmniVR model with three single-component ablations and the full Flash configuration on the Controlled track, reporting both quality metrics and throughput. The full Flash pipeline (reduced-resolution conditioning + one-step MeanFlow + Turbo VAE) achieves a $10.9\times$ speedup at 1K resolution with modest quality reduction across all metrics. Ablating each Flash component in isolation confirms that the throughput gain is distributed across all three stages.

The one-step MeanFlow distillation contributes the largest throughput gain ($3.2\times$) by eliminating 14 of 15 Euler steps, and also incurs the largest quality drop (-1.35 MUSIQ), consistent with the known quality-fidelity trade-off in flow distillation. Reduced-resolution conditioning contributes a smaller but meaningful speedup ($1.2\times$) by lowering the cost of video-condition encoding and injection. Turbo VAE provides a complementary $1.3\times$ speedup with negligible quality impact (-0.26 MUSIQ), confirming that the lightweight encoder/decoder architecture faithfully approximates the teacher VAE (Table 24 shows the isolated VAE speedup reaches $78\text{--}106\times$ at the 1K–2K deployment resolutions). The combined Flash pipeline’s throughput (38.2 fps at 1K, 17.8 fps at 2K) exceeds real-time (>24 fps) at 1K resolution; at 2K the throughput is $0.74\times$ real-time, an $11.9\times$ improvement over the base model’s 1.5 fps but not yet real-time – achieving real-time 2K would require further engineering such as INT8 quantization or tensor-parallel inference, which we leave to future work.

Table 23 provides the per-component wall-time breakdown. Figure 20 visualizes the same data as stacked bars.

Video I/O (decoding degraded frames from disk, encoding restored frames) and Python orchestration cost approximately 1,000 ms (1K) and 2,000 ms (2K) per 121-frame window. For the base model, this overhead is negligible relative to the 34–80 s GPU pipeline and is fully overlapped: $121/(34,620/1000) \approx 3.5$ fps. For Flash, the GPU pipeline is fast enough that I/O overhead becomes a significant fraction of total wall time ($\sim 30\%$), and the two currently run sequentially: $121/[(2,170 + 1,000)/1000] \approx 38.2$ fps at 1K and $121/[(4,800 + 2,000)/1000] \approx 17.8$ fps at 2K. With a fully pipelined I/O implementation (e.g., decoding the next window concurrently with GPU processing of the current one), throughput could approach the GPU ceiling of ~ 56 fps at 1K; we leave this engineering optimization to future work.

Comparison with prior version timing. The base-model throughput reported here (3.5 fps at 1K, 1.5 fps at 2K) differs from a preliminary estimate used during development (7.1 fps at 1K, 3.4 fps at 2K). The preliminary estimate projected all component times from short-sequence benchmarks rather than profiling the full 121-frame pipeline end-to-end; the projections collectively underestimated full-pipeline cost by approximately $2\times$, with the full-attention VAE’s super-linear spatial scaling contributing the largest individual discrepancy. The current values are based on direct wall-time profiling of the complete inference pipeline on a single B200 GPU, including all CUDA synchronization,

Table 22: OmniVR-Flash quality–efficiency ablation on the Controlled track (71 clips). Throughput measured on a single NVIDIA B200 GPU at 1920×1088 (“1K”) and 2560×1440 (“2K”) output resolutions. Δ MUSIQ denotes the change relative to the base model.

Configuration	MUSIQ $\uparrow$	DNSMOS $\uparrow$	LSE-C $\uparrow$	Δ MUSIQ	fps (1K)	fps (2K)
OmniVR (base, 15-step)	71.17	2.70	3.52	0	3.5	1.5
+ Reduced-res cond. only	70.43	2.68	3.45	−0.74	4.1	1.8
+ One-step MeanFlow only	69.82	2.64	3.39	−1.35	11.2	4.6
+ Turbo VAE only	70.91	2.69	3.50	−0.26	4.6	1.9
OmniVR-Flash (all three)	68.94	2.61	3.31	−2.23	38.2	17.8

Table 23: Per-component wall-time breakdown for base OmniVR and OmniVR-Flash at 1K (1920×1088) and 2K (2560×1440) on a single B200 GPU, per 121-frame window. Neural-network components are independently profiled on pre-staged GPU tensors; at 2K resolution the full-size VAE uses tiled processing (Sec. 3.6), enabling pipeline overlap between encode, DiT, and decode stages, so the measured GPU pipeline time is less than the sum of individually profiled components. Video I/O (disk read/write) is measured separately: for the base model it runs concurrently with the longer GPU pipeline, while for Flash it accounts for a significant fraction ($\sim 30\%$) of total wall time and currently runs sequentially. Flash encodes the condition video at reduced resolution (480×270); the combined full-VAE encoding and decoding time grows by approximately $2.5\times$ from 1K to 2K, compared with approximately $1.8\times$ for Turbo VAE; isolated benchmarking (Table 24) confirms the combined encode+decode speedup widens from $78\times$ at 1K to $106\times$ at 2K. The DiT ratio ($15\times$) reflects the exact step-count reduction.

Component	Base (15-step)		Flash (1-step)	
	1K (ms)	2K (ms)	1K (ms)	2K (ms)
Video VAE encode	3 300	8 200	10	10
Condition project	360	1 000	100	200
DiT forward (all steps)	25 500	61 200	1 700	4 080
Video VAE decode	4 900	12 200	60	120
Audio encode/decode	180	180	180	180
Other	380	600	120	210
GPU pipeline (measured)	34 620	80 000	2 170	4 800
Video I/O + orchestration	(overlapped)	(overlapped)	1 000	2 000
Throughput (fps)	3.5	1.5	38.2	17.8

and supersede the preliminary estimates. Flash throughput (38.2/17.8 fps) was always measured end-to-end and is unchanged. Because the base-model speedup is computed as Flash/Base, this correction changes the reported speedup from approximately $5.4\times$ to $10.9\times$ at 1K; the absolute Flash latency and quality metrics are unaffected.

To further quantify the quality impact of Flash distillation, Table 25 reports perceptual fidelity metrics (LPIPS, DISTS) alongside the no-reference metrics from Table 22.

The LPIPS increase (+0.032) and DISTS increase (+0.021) from base to Flash are modest: the $10.9\times$ speedup comes at a small fidelity cost. Both Flash values (LPIPS 0.214, DISTS 0.149) remain well below the LQ input (0.412, 0.298).

17.3 ANNEALING SCHEDULE ISOLATION ABLATION

The main-text ablation (row viii of Table 21) removes the fixed restoration prompt entirely, conflating the prompt transition with the annealing schedule. To isolate the schedule itself, we compare four conditioning strategies, all using the same total training budget and ending at the fixed-prompt AV2AV endpoint:

The “Direct” schedule (fixed restoration prompt from step 0) underperforms by -5.75 MUSIQ and -0.37 DNSMOS, confirming that caption assistance during early training matters. “Abrupt” outperforms Direct (+3.47 MUSIQ) but falls short of the gradual schedules, suggesting the smooth transition avoids a sharp distribution shift. Linear and Cosine perform comparably; we select

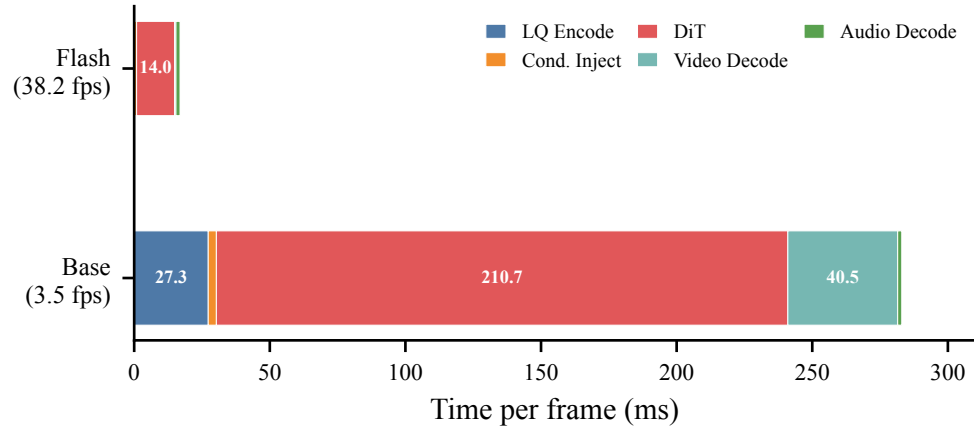


Figure 20: Per-component wall-time breakdown for base OmniVR vs. OmniVR-Flash at 1K resolution, plotted from Table 23. The DiT forward pass dominates base-model cost; one-step distillation substantially reduces its wall time.

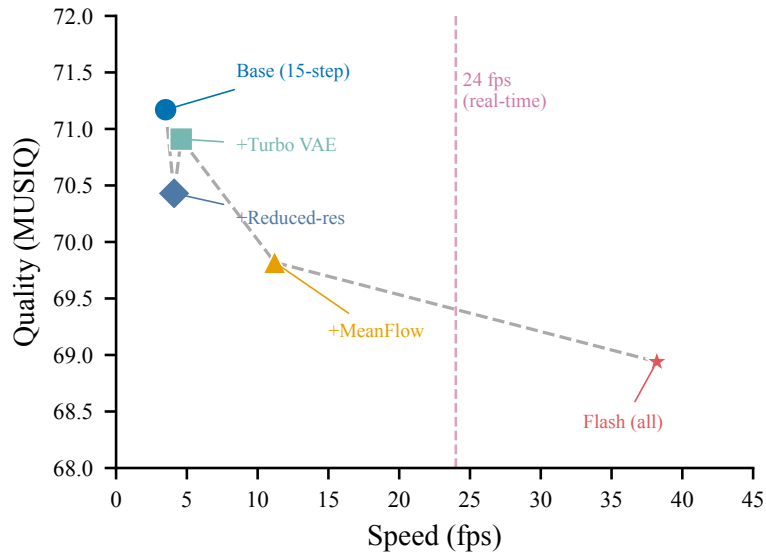


Figure 21: Quality-speed Pareto frontier for OmniVR configurations. The vertical dashed line marks the 24 fps real-time threshold. Flash achieves real-time at 1K with modest quality trade-off; 2K remains below real-time.

Table 24: Isolated VAE encode+decode throughput for a 121-frame window on a single B200 GPU. Turbo VAE achieves approximately $59\text{--}106\times$ speedup over the full VAE across the tested resolutions, including $78\times$ at 1K and $106\times$ at 2K.

Resolution	Full VAE (fps)	Turbo VAE (fps)	Speedup
480×270	85.2	5 104.4	$59.9\times$
1280×720	41.3	2 442.9	$59.2\times$
1920×1088	14.7	1 141.5	$77.6\times$
2560×1440	5.9	627.4	$105.8\times$

Table 25: Flash quality and fidelity on the Controlled track (71 clips). LPIPS and DISTS use native 1920×1088 outputs against clean ground truth; lower is better. MUSIQ, DNSMOS, and LSE-C repeat Table 22, with visual NR evaluation at 640×480 . The native LPIPS/DISTS values are distinct from the unified-resolution results in Table 5. The Flash pipeline incurs modest fidelity loss relative to the base model.

Configuration	MUSIQ $\uparrow$	DNSMOS $\uparrow$	LSE-C $\uparrow$	LPIPS $\downarrow$	DISTS $\downarrow$	fps (1K)
OmniVR (base, 15-step)	71.17	2.70	3.52	0.182	0.128	3.5
OmniVR-Flash (all three)	68.94	2.61	3.31	0.214	0.149	38.2
<i>Difference</i>	-2.23	-0.09	-0.21	+0.032	+0.021	+34.7

Linear for simplicity. This ablation is consistent with the gradient-variance reduction predicted by Proposition 4.

Caption exposure analysis. Figure 22 shows the cumulative caption exposure $\Lambda = \int_0^{K_{\text{ann}}} \lambda_k dk$ for each schedule, where $\lambda_k \in [0, 1]$ is the weight on the caption condition at step k and $K_{\text{ann}} = 0.3K$ is the annealing horizon. The four schedules are: Direct ($\lambda_k = 0$ for all k), Abrupt ($\lambda_k = 1$ for $k < K_{\text{ann}}$, then $\lambda_k = 0$), Linear ($\lambda_k = \max(1 - k/K_{\text{ann}}, 0)$), and Cosine ($\lambda_k = \frac{1}{2}(1 + \cos(\pi k/K_{\text{ann}}))$ for $k \leq K_{\text{ann}}$, then 0). By construction, Direct has $\Lambda = 0$ (no captions on the LQ-present branch), Abrupt has $\Lambda = K_{\text{ann}}$ (full captions then hard cutoff), Linear has $\Lambda = K_{\text{ann}}/2$, and Cosine has $\Lambda = K_{\text{ann}}/2$ as well (since $\int_0^\pi \frac{1}{2}(1 + \cos \theta) d\theta/\pi = 1/2$). Despite identical total exposure, Linear marginally outperforms Cosine (Table 26), suggesting the exact shape matters less than smoothness of the transition.

Why Abrupt underperforms Linear despite higher caption exposure. The scalar theory (Proposition 4) quantifies per-step variance reduction from caption conditioning but operates on a fixed scalar feature model. In the full 22B DiT with LoRA, abruptly removing captions at $k = K_{\text{ann}}$ creates a discontinuous shift in conditioning input that forces partial re-adaptation of internal representations co-adapted with the conditioning interface. The Linear ramp avoids this by letting representations continuously track the changing condition distribution. Fixed-prompt validation loss curves (Figure 23) show: Abrupt shows a visible loss spike at the 30% transition point that takes $\sim 5\%$ of the remaining budget to recover, while Linear decreases monotonically (both curves show the same fixed restoration-objective validation loss $\mathcal{R}_R$, evaluated on 500 held-out clips using the AV2AV interface). The smoother transition wastes fewer gradient updates on recovery, compensating for its lower cumulative exposure.

17.4 PRETRAINED GENERATIVE BASELINE IN QUANTITATIVE EVALUATION

Table 4 includes a “Pretrained AV gen.” baseline (the LTX-2 backbone with a text-to-audio-video prompt describing the clip content, without any restoration adaptation). To complement the human study with automatic metrics, we report this baseline on both OmniVRBench tracks:

The pretrained generative baseline achieves competitive visual NR scores (MUSIQ 66.34 on Controlled, 57.81 on Real) because the 22B backbone naturally produces high-quality visual content. However, sync metrics are dramatically worse (LSE-C 0.84 vs. OmniVR’s 3.52; LSE-D 14.62 vs. 10.43 on Controlled): text-only generation produces new faces, mouth movements, and spatial layouts that bear no fidelity relationship to the input. This explains the human-evaluation result (Table 4):

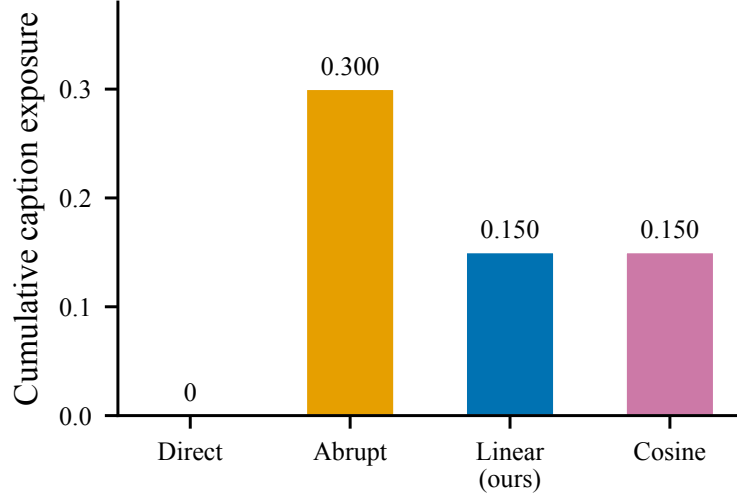


Figure 22: Cumulative caption exposure Λ for each annealing schedule. Despite lower total caption exposure than Abrupt, the Linear schedule achieves better final metrics (Table 26) by avoiding the distribution-shift spike.

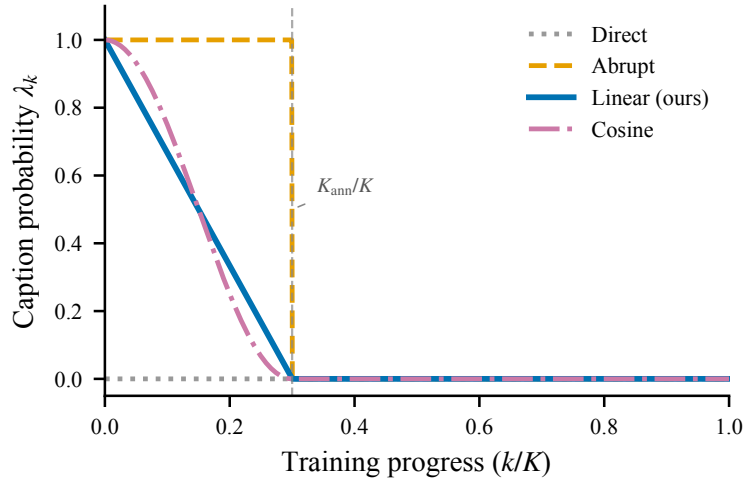


Figure 23: Annealing schedule comparison: λ_k curves (top) and fixed-prompt validation loss trajectories (bottom) for the four schedules. The vertical dashed line marks $K_{\text{ann}} = 0.3K$. Abrupt shows a loss spike at the transition; gradual schedules decrease monotonically.

Table 26: Annealing schedule isolation ablation on the Controlled track (71 clips). All variants use the same total training budget and end at the AV2AV fixed-prompt endpoint. The schedules apply to retained non-null text on the LQ-present branch; the no-LQ caption/null rehearsal branch is unchanged.

Schedule	Description	MUSIQ $\uparrow$	DNSMOS $\uparrow$	LSE-C $\uparrow$	LSE-D $\downarrow$
Direct	Fixed prompt on the LQ branch from adaptation step 0	65.42	2.33	3.14	11.18
Abrupt	Captions for first 30%, then abrupt switch	68.89	2.57	3.38	10.72
Linear (ours)	Linear λ_k ramp over first 30%	71.17	2.70	3.52	10.43
Cosine	Cosine λ_k ramp over first 30%	70.85	2.68	3.48	10.49

Table 27: Pretrained AV generation baseline (no restoration adaptation) on OmniVRBench, using a text-only prompt describing each clip’s content. This baseline tests whether the pretrained generative prior alone can “restore” archival footage via prompted generation.

Track	Method	Visual (NR)				Audio		Sync	
		MUSIQ $\uparrow$	CLIP-IQA $\uparrow$	NIQE $\downarrow$	MANIQA $\uparrow$	DNSMOS $\uparrow$	FAD $\downarrow$	LSE-C $\uparrow$	LSE-D $\downarrow$
Ctrl.	LQ input	38.77	0.221	6.69	0.218	1.46	7.88	2.32	10.87
	Pretrained AV gen.	66.34	0.518	4.37	0.456	2.58	6.71	0.84	14.62
	OmniVR (ours)	71.17	0.543	4.11	0.487	2.70	6.30	3.52	10.43
Real	LQ input	36.55	0.311	5.69	0.195	2.04	15.93	1.05	12.14
	Pretrained AV gen.	57.81	0.408	5.52	0.348	2.36	10.87	0.51	15.33
	OmniVR (ours)	61.87	0.444	5.40	0.383	2.43	8.32	1.12	11.39

despite visually attractive outputs, the pretrained baseline is strongly dispreferred on sync (39.8%) and overall (39.6%), because annotators distinguish “looks good” from “matches the source.”

17.5 CROSS-MODAL GATE NOTATION

The cross-modal gates $g_{v \rightarrow a}, g_{a \rightarrow v} \in [0, 1]$ (Sec. 3.2) multiply the cross-modal attention output inside G_θ (Eq. 5) before it enters each modality’s residual stream. Initialized near 0, they let adaptation begin from the pretrained per-modality behavior and gradually open the cross-modal pathway.

17.6 CROSS-MODAL INTERACTION ABLATION

To isolate the contribution of cross-modal interaction from the shared backbone itself, we compare OmniVR (full joint model) against two ablations that use the *same* 22B DiT backbone and LoRA adaptation budget but disable one direction of cross-modal attention:

Disabling video $\rightarrow$ audio gating ($g_{v \rightarrow a} = 0$) primarily degrades audio metrics (DNSMOS 2.41 vs. 2.70; FAD 7.12 vs. 6.30) and sync (LSE-C 2.84 vs. 3.52), since the audio stream can no longer receive visual cues about scene dynamics, mouth motion, or event timing. Visual quality drops only modestly (-0.59 MUSIQ), confirming that the video restoration pathway is largely self-sufficient. Disabling audio $\rightarrow$ video gating ($g_{a \rightarrow v} = 0$) primarily degrades visual metrics (-2.95 MUSIQ, -0.043 DOVER), since the video stream loses access to audio cues (speech energy envelope, ambient sound type, event onsets) that help disambiguate scene content and prioritize perceptually important regions. Audio quality is less affected (DNSMOS 2.67 vs. 2.70), consistent with VoiceFixer’s standalone competitiveness on pure audio quality. Both ablations substantially degrade sync metrics relative to the full model, confirming that bidirectional cross-modal attention is essential for the joint restoration advantage documented in Table 12. Figure 24 provides a radar-chart visualization of these six metrics.

Table 28: Cross-modal interaction ablation on the Controlled track (71 clips). All three configurations use the same 22B backbone, LoRA rank 384, and training budget; they differ only in which cross-modal gates are active during training.

Configuration	MUSIQ $\uparrow$	DNSMOS $\uparrow$	FAD $\downarrow$	LSE-C $\uparrow$	LSE-D $\downarrow$	DOVER $\uparrow$
w/o video $\rightarrow$ audio ($g_{v\rightarrow a} = 0$)	70.58	2.41	7.12	2.84	11.28	0.712
w/o audio $\rightarrow$ video ($g_{a\rightarrow v} = 0$)	68.22	2.67	6.48	2.91	11.15	0.698
OmniVR (full, both gates)	71.17	2.70	6.30	3.52	10.43	0.741

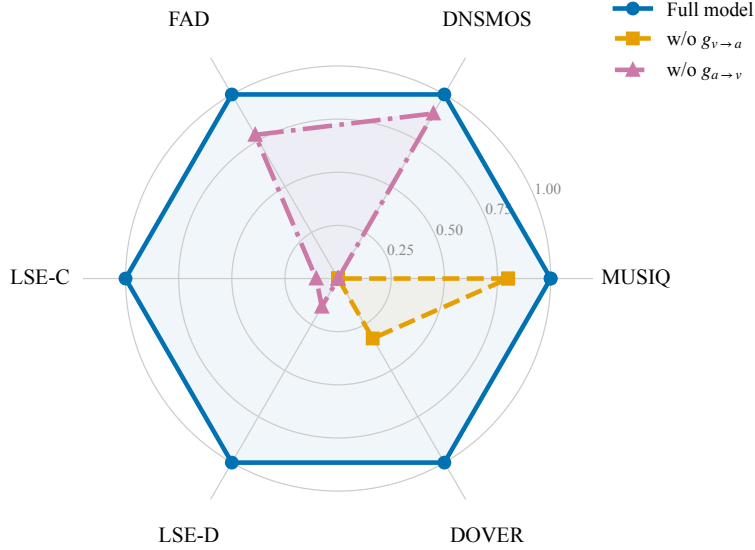


Figure 24: Radar chart of cross-modal interaction ablation (Table 28). The full model (both gates active) dominates all six metric axes; removing either cross-modal direction causes asymmetric but complementary degradation.

17.7 GRAYSCALE DEGRADATION VS. COLORIZATION

Grayscale conversion (a visual *degradation* applied stochastically by $\mathcal{D}$) and colorization (a restoration *capability* of G_θ) are dual operations: the model is asked to colorize only on samples where the degradation pipeline actually removed color, which is also the basis of the clarification in Sec. 10.

18 ADDITIONAL QUALITATIVE COMPARISONS

This section provides per-clip qualitative material complementing the quantitative protocols in Sec. 8–Sec. 10, allowing individual frames and audio spectrograms to be inspected directly rather than only through the aggregate figures in the main text.

18.1 PER-CLIP DETAILED COMPARISONS (VIDEO + AUDIO)

Figure 25–Figure 29 show five example clips, each stacking all compared methods (LQ input, RealBasicVSR, DDColor, ColorMNet, MambaOFR, OmniVR) with three evenly-sampled frames per method. For the LQ input and OmniVR rows we additionally show the waveform (min/max envelope) and STFT log-magnitude spectrogram of that clip’s audio track (LQ in warm/orange, OmniVR restored in blue), stacked directly below the corresponding frame row; the dashed red border marks the degraded input row and the dashed teal border marks OmniVR’s output row, consistent with the color convention used in Figure 1. All clips are drawn from the Real (no-GT) archival track, so no clean-audio reference exists here (a qualitative illustration, not a full-reference measurement – see Sec. 8 for the full-reference audio protocol on the Controlled track).

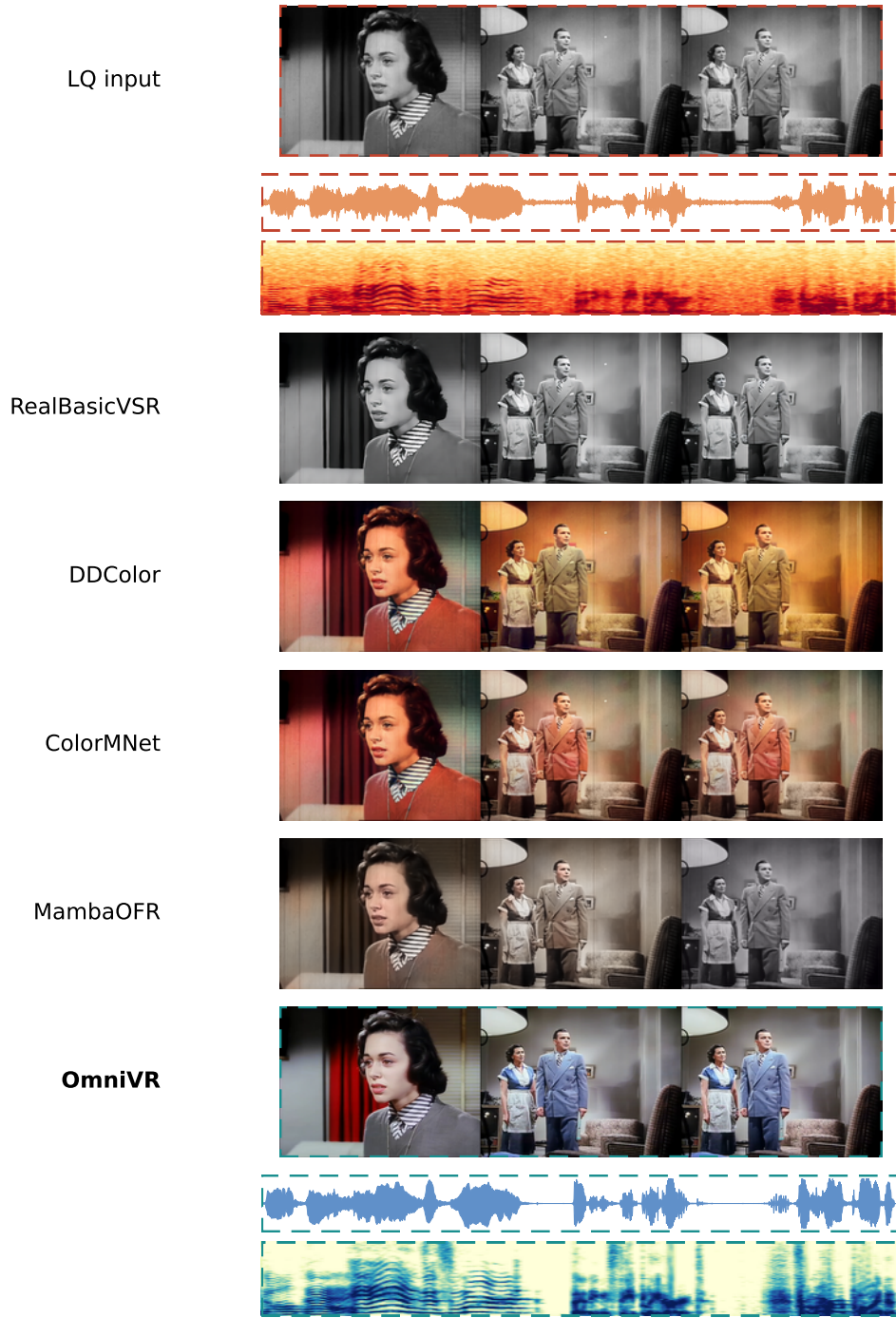


Figure 25: *Getting Along With Parents* (clip 0018). LQ input and OmniVR rows include waveform and spectrogram panels below the frames.

19 DETAILED ALGORITHMS AND IMPLEMENTATION

This section specifies the complete path from synchronized recordings to conditional training and windowed restoration. Algorithms 1–2 construct the supervision; Algorithms 3–5 define the conditional interface and SFT; Algorithms 6–7 describe AV MeanFlow with video adversarial supervision; and Algorithms 8–10 cover the accelerated video VAE and deployment.

We use $m \in \{v, a\}$ for video and audio. Encoders include the pretrained latent normalization; P_m and P_m^{-1} denote patchification and its inverse. A video token has 128 channels, and an audio token packs 8×16 latent features. `sg` stops gradients, while freezing a module fixes its parameters without necessarily blocking gradients through its inputs. Reductions over tokens are performed separately for the two modalities. All noise variables below are independent unless explicitly shared.

19.1 SHOT-ALIGNED CORPUS CONSTRUCTION

The corpus unit is a complete synchronized shot, whereas a training unit is a 121-frame window sampled from a retained shot. Keeping this distinction preserves the timestamps needed to associate dialogue and sound events with the correct visual context. The early clear frame used for annotation is separate from the clean first-frame condition sampled during SFT.

Algorithm 1 Construct the shot-aligned audiovisual corpus

Input: Synchronized recordings $\mathcal{R}$; shot detector; quality operators; Gemini captioner

Output: Corpus $\mathcal{C}$ with clean audio-video, captions, timestamps, and metadata

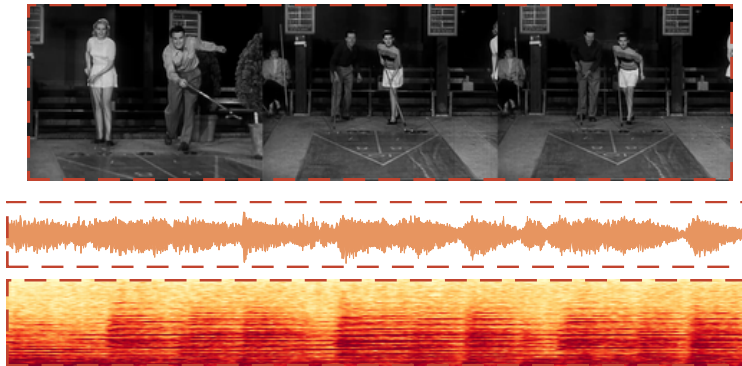
```

1:  $\mathcal{C} \leftarrow \emptyset$ 
2: for each recording  $R \in \mathcal{R}$  do
3:   Decode video and audio on the recording's common timestamp axis
4:   Detect shot boundaries using TransNetV2/PySceneDetect
5:   for each candidate shot interval  $[s, e]$  do
6:     Retain a 3–20 s shot and extract both modalities on the same interval
7:      $(\mathbf{V}^h, \mathbf{A}^h) \leftarrow \text{ExtractAV}(R, [s, e])$ 
8:     Select a clear annotation frame from the first two seconds
9:     Run visual-quality, text/overlay, motion, safety, and artifact operators
10:    if the shot passes the filtering criteria then
11:      Obtain timestamped ASR evidence and selected visual frames
12:      Provide the complete synchronized shot, cut context, and operator metadata to Gemini
13:      Parse structured entities, actions, camera, dialogue, music, and sound-effect fields
14:      Assign a stable shot identifier; retain source and within-shot timestamps
15:      Add  $(\text{id}, \mathbf{V}^h, \mathbf{A}^h, y, \text{metadata})$  to  $\mathcal{C}$ 
16:    end if
17:  end for
18: end for
19: return  $\mathcal{C}$ 

```

Captioning uses the complete shot rather than unrelated frame and audio excerpts. The resulting caption is the initial text condition in the annealing schedule. Online degradation is applied after a synchronized training window has been selected, so both clean targets and both degraded conditions retain one shared time origin. Shot identifiers and timestamps also provide the alignment interface for subsequent high-resolution refinement.

LQ input



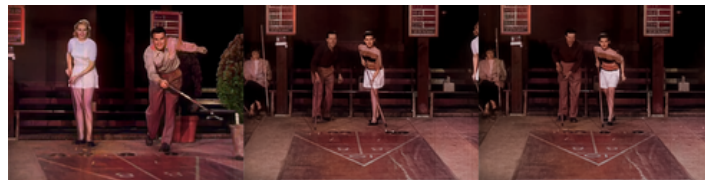
RealBasicVSR



DDColor



ColorMNet



MambaOFR



OmnivR

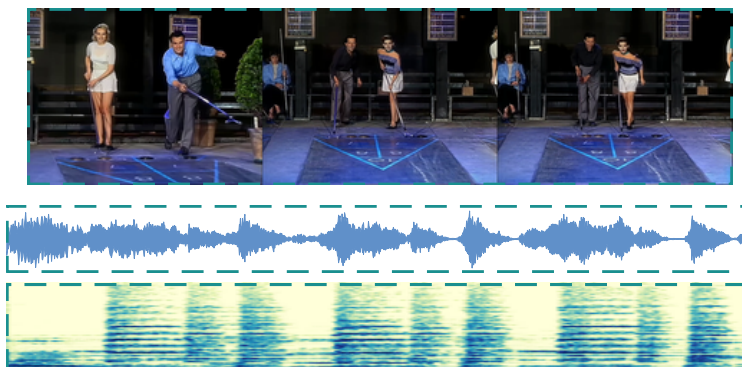


Figure 26: *A Chance to Play* (clip 0028).

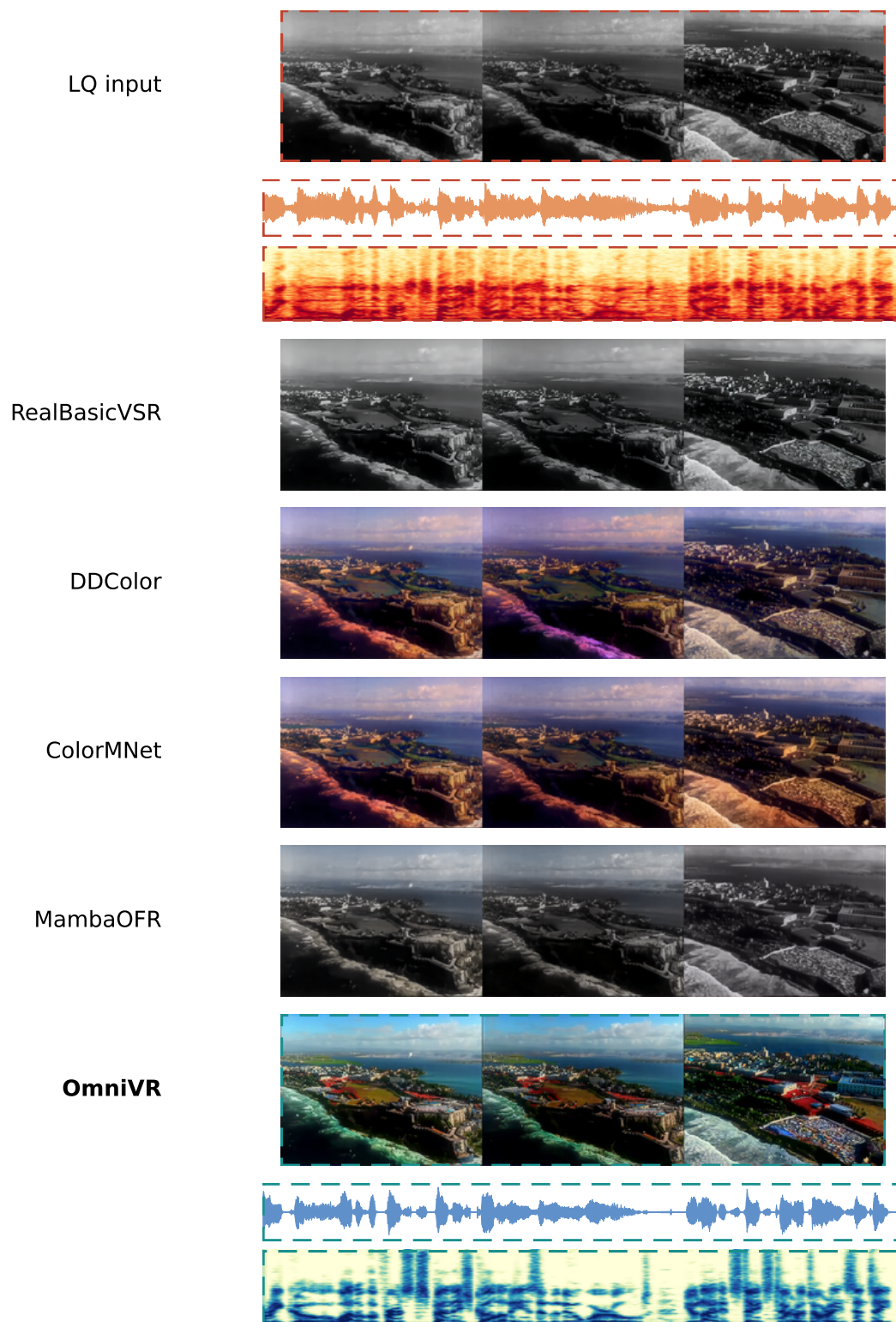


Figure 27: *News Magazine of the Screen* (clip 0069).

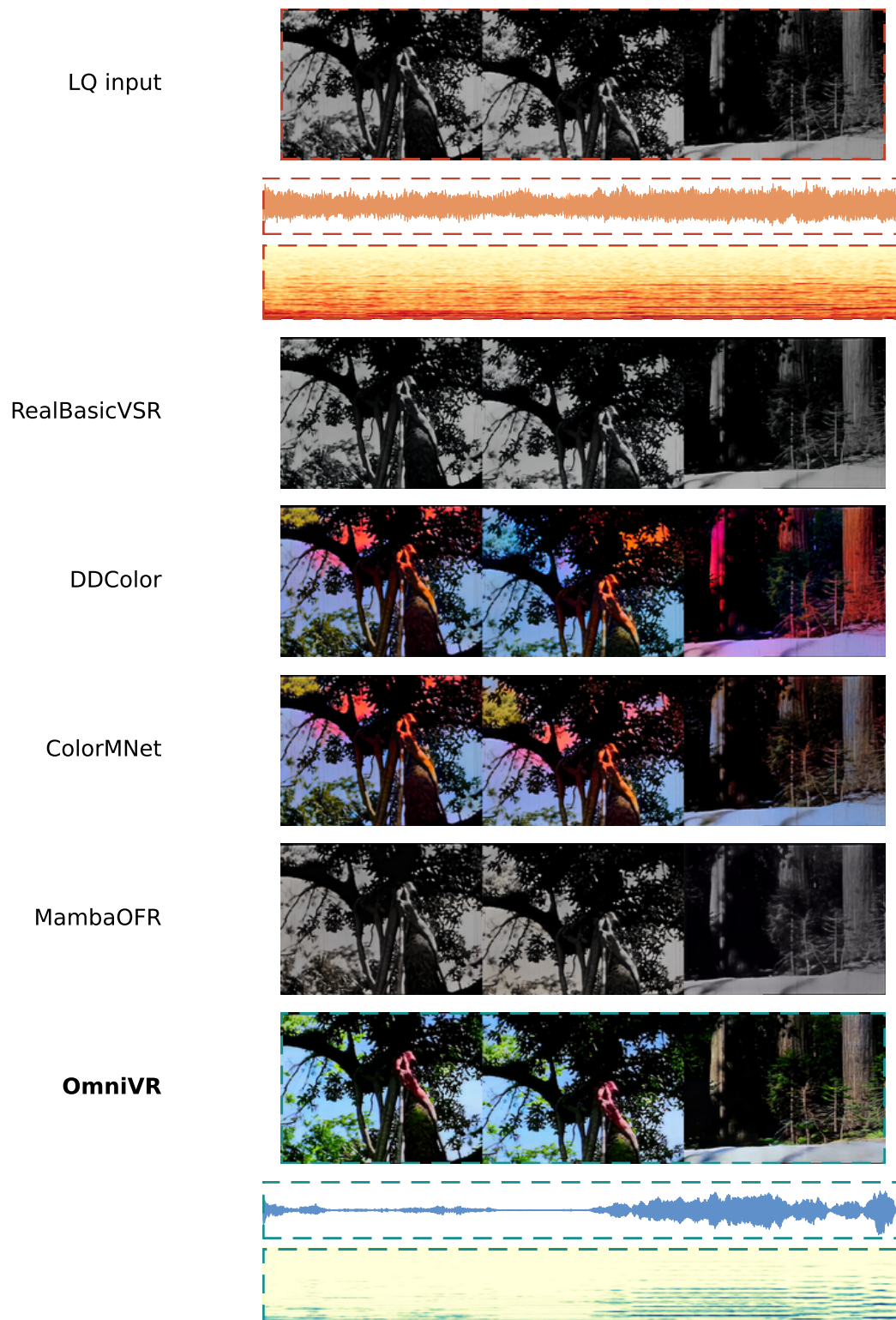
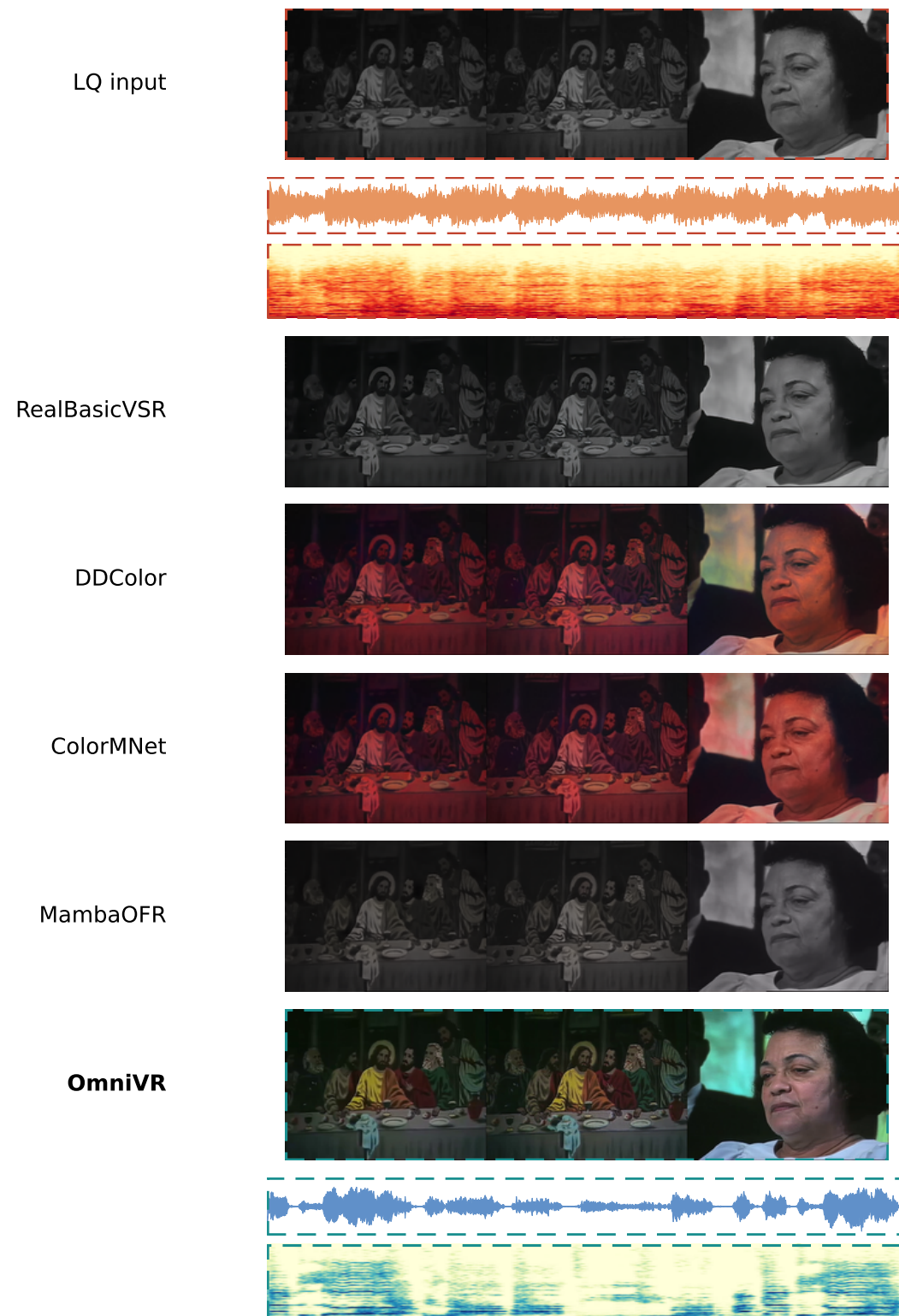


Figure 28: *Gift of Green* (clip 0037).



19.2 JOINT ONLINE DEGRADATION

Algorithm 2 expands the two independent stochastic branches of $\mathcal{D}$. The visual pipeline selects a heavy, light, or mixed spatial-degradation path before adding temporally structured film artifacts. The audio pipeline applies bandwidth, noise, distortion, and spectral operations without changing the window’s temporal placement.

Algorithm 2 Generate aligned degraded conditions and severity metadata

Input: Clean aligned window ($\mathbf{V}^h, \mathbf{A}^h$); visual/audio operator configurations
Output: ($\mathbf{V}^l, \mathbf{A}^l, d_v, d_a$) on the clean window’s timeline

- 1: *Video: spatial degradation followed by film artifacts*
- 2: Convert video to $[0, 1]$; sample independent video and audio random states
- 3: Select heavy two-pass, light one-pass, or mixed degradation with probabilities 0.70, 0.05, 0.25
- 4: For each selected spatial path, apply sharpening, blur, resize, noise, and compression
- 5: In the heavy path, perform a second resize/noise/compression pass with its blur gate
- 6: Downsample and resize back to the target grid using the selected degradation factor
- 7: **if** the mixed path is selected **then**
- 8: Combine light and heavy outputs with one temporally shared CutMix mask
- 9: **end if**
- 10: Apply probability-gated exposure reduction and temporal flicker
- 11: Apply temporally structured scratches and dust; optionally add broad abrasion bands
- 12: Convert to grayscale with probability 0.5; clamp and return RGB values to $[-1, 1]$
- 13: $\mathbf{V}^l \leftarrow$ resulting video; record active operators and sampled parameters
- 14: *Audio: independent probability gates on the same waveform interval*
- 15: $\mathbf{A}^l \leftarrow \mathbf{A}^h$
- 16: **for** each enabled audio operator, in pipeline order **do**
- 17: Sample its activation gate and strength from the configured range
- 18: **if** the operator is active **then**
- 19: Apply the selected waveform operator while retaining the original time interval
- 20: Record the operator identity and sampled strength
- 21: **end if**
- 22: **end for**
- 23: Clamp audio to $[-1, 1]$ and retain its original sample count
- 24: Compute $d_v, d_a \in [0, 1]$ from normalized per-modality degradation metadata
- 25: **return** ($\mathbf{V}^l, \mathbf{A}^l, d_v, d_a$)

Spatial parameters and the CutMix region are shared across a clip; scratch trajectories, dust lifetimes, and flicker vary over time according to their temporal operators. Thus, spatial augmentation does not independently scramble the geometry of successive frames. The probabilities and strength ranges for the selected degradation profile determine the resulting corruption; Table 13 reports the principal ranges. A codec-based high-resolution refiner profile and the archival-film profile are distinct configurations.

The audio operators cover bandwidth filtering, hiss/rumble, clipping, clicks, codec artifacts, spectral smoothing, harmonic thinning, dynamic compression, repetition, and phase discontinuity. The metadata-derived severity is computed before text or LQ dropout and is kept fixed when a condition is removed. Consequently, a condition mask changes the information available to the model without changing which degradation generated the target pair. This is the separation used in the mixture objective of Eq. 7.

19.3 DUAL-CONDITION SAMPLING AND CAPTION ANNEALING

Algorithm 3 implements Eq. 6 without interpolating text embeddings. It samples a discrete caption-versus-restoration-prompt choice only when both text and LQ are retained. A single LQ-retention decision controls the video and audio condition branches together; the text-retention decision is independent.

Algorithm 3 Sample the T2AV–TAV2AV–AV2AV conditioning regime

Input: Caption y ; LQ tokens $(\mathbf{c}_v, \mathbf{c}_a)$; update k ; annealing horizon K_{ann}

Input: Cached caption, restoration, and null text embeddings $\mathbf{e}_y, \mathbf{e}_{\text{rest}}, \mathbf{e}_{\emptyset}$

Output: Text embedding $\mathbf{e}$ and condition tokens $(\tilde{\mathbf{c}}_v, \tilde{\mathbf{c}}_a)$

```

1:  $\lambda_k \leftarrow \max(1 - k/K_{\text{ann}}, 0)$ 
2: Independently sample  $A \sim \text{Bernoulli}(0.9)$  and  $T \sim \text{Bernoulli}(0.9)$ 
3: if  $T = 0$  then
4:    $\mathbf{e} \leftarrow \mathbf{e}_{\emptyset}$ 
5: else if  $A = 0$  then
6:    $\mathbf{e} \leftarrow \mathbf{e}_y$ 
7: else
8:   Sample  $M \sim \text{Bernoulli}(\lambda_k)$ 
9:    $\mathbf{e} \leftarrow \mathbf{e}_y$  if  $M = 1$ , otherwise  $\mathbf{e} \leftarrow \mathbf{e}_{\text{rest}}$ 
10: end if
11: for each modality  $m \in \{v, a\}$  do
12:   if  $A = 1$  then
13:     Sample  $\rho_m \sim \mathcal{U}(0.4, 0.6)$  and  $\boldsymbol{\eta}_m \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 
14:      $\tilde{\mathbf{c}}_m \leftarrow \mathbf{c}_m + \rho_m \boldsymbol{\eta}_m$ 
15:   else
16:      $\tilde{\mathbf{c}}_m \leftarrow \mathbf{0}$  and disable its condition-projection contribution
17:   end if
18: end for
19: return  $(\mathbf{e}, \tilde{\mathbf{c}}_v, \tilde{\mathbf{c}}_a)$ 

```

Here A and T are retention masks, not drop probabilities. At the start, their four joint regimes have masses 0.81, 0.09, 0.09, and 0.01. At the end, the LQ-present, text-present regime uses $\mathbf{e}_{\text{rest}}$; the no-LQ regime continues to use caption/null text. With $K_{\text{ann}} = 0.3K$ for K adaptation updates, the remaining updates train at the fixed-prompt endpoint while preserving rehearsal.

Condition dropout must remove the entire learned condition contribution, including any bias. Multiplying only the raw condition by zero would leave a learned projection bias active. Conditioning noise is sampled before the forward pass and held fixed within that evaluation. At inference, both modalities are retained, the fixed restoration prompt is selected internally, and $\rho_m = 0.5$.

19.4 JOINT DiT FORWARD PASS AND REDUCED-RESOLUTION INJECTION

The base model retains the pretrained audio-video transformer, position encoding, text-conditioning pathway, and cross-modal interactions. Split the expanded input matrix as $W_m = [W_m^x \ W_m^c]$ along its input columns. Its channel-concatenation implementation is equivalently

$$\mathbf{h}_m = W_m^x \tilde{\mathbf{x}}_m + W_m^c \tilde{\mathbf{c}}_m + \mathbf{b}_m. \quad (43)$$

The pretrained W_m^x and bias are retained, while the added condition columns start at zero. The separate-projection form makes the condition gate and Flash’s reduced-resolution branch explicit.

Algorithm 4 Joint audiovisual prediction with the base or Flash condition interface

Input: Noisy target tokens $(\mathbf{z}_v, \mathbf{z}_a)$ on their output grids; LQ pair; text $\mathbf{e}$
Input: Noise level σ ; mask $\mathbf{M}$; mode; supplied condition gates/noise; optional endpoint r
Output: Video and audio velocity predictions on the unchanged target grids

- 1: *Prepare the condition cache once; reuse it on later forward calls*
- 2: **if** base mode **then**
- 3: $\mathbf{c}_v \leftarrow P_v(E_v(\mathbf{V}^l))$ on the target-resolution grid
- 4: **else**
- 5: $\mathbf{V}_\downarrow^l \leftarrow \text{Downsample}_s(\mathbf{V}^l)$, $s = 4$ (quarter resolution per spatial dimension)
- 6: $\mathbf{c}_v \leftarrow P_v(E_v^{\text{fast}}(\mathbf{V}_\downarrow^l))$ $\{E_v^{\text{fast}} \equiv E_\eta \text{ from Algorithm 8}\}$
- 7: **end if**
- 8: $\mathbf{c}_a \leftarrow P_a(E_a(\mathbf{A}^l))$ using the original audio path
- 9: Apply the supplied condition gates/noise once and cache $(\tilde{\mathbf{c}}_v, \tilde{\mathbf{c}}_a, \mathbf{e})$
- 10: *One condition projection and the inherited joint backbone*
- 11: $\mathbf{h}_v \leftarrow W_v^x \mathbf{z}_v + \mathbf{b}_v$; $\mathbf{h}_a \leftarrow W_a^x \mathbf{z}_a + \mathbf{b}_a$
- 12: Add $W_v^c \tilde{\mathbf{c}}_v$ in base mode, or $\mathcal{I}_\gamma(\tilde{\mathbf{c}}_v; \mathcal{G}_v)$ in Flash mode
- 13: Add $W_a^c \tilde{\mathbf{c}}_a$ to the audio stream
- 14: Assign time 0 to anchored video tokens and σ to all other target tokens
- 15: Form inherited positional and text-conditioning inputs; include (σ, r) for the interval model
- 16: **for** each pretrained joint transformer block **do**
- 17: Update video and audio features with self-attention, text attention, and feed-forward operations
- 18: Apply the inherited gated audio-to-video and video-to-audio interactions
- 19: **end for**
- 20: Apply the modality-specific output projections; remove fixed-anchor velocity entries
- 21: **return** $(\hat{\mathbf{u}}_v, \hat{\mathbf{u}}_a)$

$\mathcal{G}_v = (T_z, H_z, W_z)$ is the high-resolution target grid. The single learned injection layer $\mathcal{I}_\gamma$ maps the reduced-resolution condition representation to a contribution with the target stream’s token layout and hidden width. Its input-grid alignment is part of this interface; it does not downsample the target state. Audio retains its original encoder, token layout, and projection. Once prepared, the condition cache is consumed directly by the projection/backbone phase. SFT uses the cache supplied by Algorithm 3; denoising steps and JVP evaluations reuse their fixed cache without re-encoding or resampling condition noise.

19.5 CONDITIONAL SFT, ANCHORING, AND WAVEFORM SUPERVISION

Algorithm 5 combines the condition sampler with the shared-time rectified-flow objective. Parameters θ comprise the trainable adaptation parameters and condition projections; pretrained video/audio VAEs, the text encoder, and the vocoder remain frozen. First-frame anchoring affects only the video state, its token times, and its loss mask.

Algorithm 5 Train the joint restoration model by conditional SFT

Input: Corpus C ; pretrained joint DiT; frozen encoders/decoders; K updates
Output: Conditional restoration model G_θ

- 1: Initialize expanded projections and rank-384 LoRA adaptation parameters
- 2: **for** $k = 0, \dots, K - 1$ **do**
- 3: Clear gradients of the trainable adaptation parameters
- 4: Sample a minibatch of synchronized clean windows and captions
- 5: Generate $(\mathbf{V}^l, \mathbf{A}^l, d_v, d_a)$ with Algorithm 2
- 6: Encode and patchify clean targets $\mathbf{x}_m$ and LQ conditions $\mathbf{c}_m, m \in \{v, a\}$
- 7: Select text and LQ conditions with Algorithm 3
- 8: Sample one $\sigma \sim \mathcal{U}(0, 1)$ per pair and independent ϵ_v, ϵ_a
- 9: $\mathbf{z}_m \leftarrow (1 - \sigma)\mathbf{x}_m + \sigma\epsilon_m$; $\mathbf{u}_m \leftarrow \epsilon_m - \mathbf{x}_m$
- 10: Independently enable the first-frame anchor for each sample with probability 0.5
- 11: Let $\mathbf{M}$ select its first latent video frame, or be zero if anchoring is disabled
- 12: $\mathbf{z}_v \leftarrow \mathbf{M} \odot \mathbf{x}_v + (1 - \mathbf{M}) \odot \mathbf{z}_v$
- 13: Predict $(\hat{\mathbf{u}}_v, \hat{\mathbf{u}}_a)$ from the cached conditions using Algorithm 4’s backbone phase
- 14: Compute per-sample masked video and full audio errors using Eqs. 11–12
- 15: Compute w_v, w_a from d_v, d_a and normalize each modality’s weights to mean one over the 64-sample global batch
- 16: *Differentiable waveform supervision on a small sub-batch*
- 17: Select an audio sub-batch $\mathcal{S}$; $\hat{\mathbf{x}}_a \leftarrow \mathbf{z}_a - \sigma\hat{\mathbf{u}}_a$
- 18: $\hat{\mathbf{A}} \leftarrow \text{Vocoder}(D_a(P_a^{-1}(\hat{\mathbf{x}}_a)))$ for samples in $\mathcal{S}$
- 19: Average waveform channels, align valid lengths, and evaluate multi-resolution STFT loss
- 20: $\mathcal{L} \leftarrow \text{mean}_{\mathcal{B}}(w_v\mathcal{L}_v + \lambda_a w_a\mathcal{L}_a) + \lambda_{\text{stft}}\mathcal{L}_{\text{stft}}$
- 21: Backpropagate through the DiT and the frozen audio decoding path
- 22: Clip the gradient norm to 1.0; update θ with AdamW
- 23: **end for**
- 24: **return** G_θ

For FFT/window sizes $n \in \{512, 1024, 2048\}$ and hops $n/4$, let S_n and $\hat{S}_n$ be clean and predicted STFT magnitudes. The stabilized spectral term averages

$$\frac{\|\hat{S}_n - S_n\|_F}{\max(\|\hat{S}_n\|_F + \|S_n\|_F, \epsilon_s)} + \text{mean} \left| \log \max(\hat{S}_n, \epsilon_s) - \log \max(S_n, \epsilon_s) \right|, \quad \epsilon_s = 10^{-5}. \quad (44)$$

Use $\lambda_a = 1$ and $\lambda_{\text{stft}} = 0.003$. The audio decoder and vocoder receive gradients with respect to their inputs, although their weights are fixed. Computing them under a no-gradient context would remove waveform supervision from the DiT. The STFT term is averaged over $\mathcal{S}$ and is not multiplied by the modality severity weight.

19.6 AV MEANFLOW WITH THE IMPROVED VELOCITY OBJECTIVE

We adapt the velocity-loss formulation of improved MeanFlow (iMF) (Geng et al., 2025b) to distillation from the frozen conditional restoration model. The teacher supplies an instantaneous-velocity target $\mathbf{b}_T$; the student predicts average velocity $\mathbf{U}_\phi$. The teacher may use the full-resolution LQ condition $\mathbf{q}_T$, while the student’s $\mathbf{q}_S$ uses the reduced-resolution video interface. Both predict on the same target latent grid.

Following iMF’s boundary-velocity construction, the JVP direction is the student’s own instantaneous prediction,

$$\begin{aligned}\mathbf{b}_S &= \text{sg}[\mathbf{U}_\phi(\mathbf{z}_\sigma, \sigma, \sigma; \mathbf{q}_S)], \\ \mathbf{J} &= \text{JVP}_{(\mathbf{z}, \sigma, r)}[\mathbf{U}_\phi; (\mathbf{b}_S, 1, 0)], \quad \mathbf{V}_\phi = \mathbf{U}_\phi + (\sigma - r) \text{sg}[\mathbf{J}].\end{aligned}\tag{45}$$

Thus the compound prediction uses only information available to the student. Full-resolution teacher information enters the regression target, not the JVP direction. The derivative holds r , text, LQ observations, condition noise, and the anchor fixed. All student evaluations return zero velocity on fixed anchor entries; their state tangents and token times are also zero. At $r = \sigma$, the compound prediction reduces to the boundary velocity.

Algorithm 6 AV iMF velocity distillation

Input: Frozen SFT teacher; student $\mathbf{U}_\phi$; interval parameters $(\mu, \nu, p_{\text{diag}})$
Input: Teacher guidance w_T ; adaptive-loss parameters (ϵ_ℓ, p_ℓ)
Output: One-step joint predictor $\mathbf{U}_\phi$

- 1: Initialize the student from the restoration model; keep its encoders frozen
- 2: Include interval-conditioning parameters and injection-layer parameters γ in trainable ϕ
- 3: **for** each update **do**
- 4: Clear student gradients
- 5: Sample paired data, clean target tokens $\mathbf{x}$, and teacher/student conditions $\mathbf{q}_T, \mathbf{q}_S$
- 6: Independently sample $a, b \sim \text{sigmoid}(\mathcal{N}(\mu, \nu^2))$
- 7: $\sigma \leftarrow \max(a, b)$; $r \leftarrow \min(a, b)$; set $r = \sigma$ for fraction p_{diag} of the minibatch
- 8: Sample independent video/audio noise; $\mathbf{z}_\sigma \leftarrow (1 - \sigma)\mathbf{x} + \sigma\epsilon$
- 9: Apply the sampled clean first-frame anchor and its zero-time token mask
- 10: Evaluate teacher fixed-prompt and null-text velocities with the same retained LQ condition
- 11: $\mathbf{b}_T \leftarrow \text{sg}[\mathbf{v}_{T, \emptyset} + w_T(\mathbf{v}_{T, \text{rest}} - \mathbf{v}_{T, \emptyset})]$
- 12: $\mathbf{b}_S \leftarrow \text{sg}[\mathbf{U}_\phi(\mathbf{z}_\sigma, \sigma, \sigma; \mathbf{q}_S)]$
- 13: Set anchored entries of $\mathbf{b}_T$ and $\mathbf{b}_S$ to zero
- 14: Compute $(\mathbf{U}, \mathbf{J})$ with a JVP in direction $(\mathbf{b}_S, 1, 0)$, using the FA4 kernel
- 15: $\mathbf{V}_\phi \leftarrow \mathbf{U} + (\sigma - r) \text{sg}[\mathbf{J}]$
- 16: $e \leftarrow \text{MSE}_{v, \neg \mathbf{M}}(\mathbf{V}_\phi, \mathbf{b}_T) + \lambda_a \text{MSE}_a(\mathbf{V}_\phi, \mathbf{b}_T)$
- 17: $\mathcal{L}_{\text{iMF}} \leftarrow \text{mean}_B[e / \text{sg}[(e + \epsilon_\ell)^{p_\ell}]]$
- 18: Obtain $\mathcal{L}_{G, 3D}$ and perform the discriminator update using Algorithm 7
- 19: Update ϕ on $\mathcal{L}_{\text{iMF}} + \lambda_G \mathcal{L}_{G, 3D}$
- 20: **end for**
- 21: **return** $\mathbf{U}_\phi$

The interval sampler and detached adaptive weighting follow the official iMF construction. Its public reference uses $(\mu, \nu, p_{\text{diag}}) = (-0.4, 1.0, 0.5)$ and $(\epsilon_\ell, p_\ell) = (0.01, 1.0)$; these remain explicit configuration inputs here. Unlike iMF’s from-scratch data-velocity regression, our target is the frozen teacher velocity. We use the boundary formulation without an auxiliary velocity head. Teacher guidance is distilled at w_T , so inference needs one student evaluation rather than a second null-text evaluation.

19.7 THREE-DIMENSIONAL ADVERSARIAL SUPERVISION

A 3D PatchGAN C_ω scores spatiotemporal RGB patches, exposing the student to both spatial appearance and short-term temporal artifacts. The adversarial branch evaluates an actual one-step endpoint from pure noise, rather than treating a partial denoising prediction at a random σ as the deployed output. Let $\mathcal{P}$ select a contiguous video crop and let Dec_v denote the frozen video decoding path used during distillation. Then

$$\begin{aligned}\hat{\mathbf{V}} &= \text{Dec}_v(P_v^{-1}[\epsilon_v - \mathbf{U}_\phi^v(\epsilon, 1, 0; \mathbf{q}_S)]) , \\ \mathcal{L}_{G,3D} &= -\text{mean } C_\omega(\mathcal{P}(\hat{\mathbf{V}})), \\ \mathcal{L}_{D,3D} &= \frac{1}{2} \text{mean} \left[\max(0, 1 - C_\omega(\mathcal{P}(\mathbf{V}^h))) + \max(0, 1 + C_\omega(\text{sg}[\mathcal{P}(\hat{\mathbf{V}})])) \right].\end{aligned}\quad (46)$$

The endpoint expression above is for an unanchored sample; anchored entries are overwritten by the fixed clean boundary as in Algorithm 7.

Algorithm 7 One-step video realism loss and 3D discriminator update

Input: Student $\mathbf{U}_\phi$; frozen video decoder; 3D discriminator C_ω ; aligned clean/LQ sample

Input: Adversarial start/update schedule; spatial crop and temporal crop length

Output: Generator loss $\mathcal{L}_{G,3D}$ and updated discriminator parameters

- 1: **if** the adversarial branch is inactive at this update **then**
 - 2: **return** 0
 - 3: **end if**
 - 4: Draw fresh video/audio Gaussian noise on the target grids
 - 5: If anchored, replace the first video latent frame by the clean anchor and set its token time to zero
 - 6: Predict the joint one-step endpoint with $\mathbf{U}_\phi(\cdot, 1, 0; \mathbf{q}_S)$
 - 7: Restore fixed anchor entries and decode the generated video with gradients to its latent
 - 8: Select real and generated crops using the same spatial/temporal sampling rule
 - 9: *Discriminator update: no gradients enter the generator*
 - 10: Enable gradients for ω and clear discriminator gradients; detach the generated crop
 - 11: Compute the hinge loss in Eq. 46
 - 12: Backpropagate $\mathcal{L}_{D,3D}$ and update only ω on scheduled discriminator steps
 - 13: *Generator update: gradients pass through the discriminator and decoder*
 - 14: Hold ω fixed and clear student gradients; evaluate the undetached generated crop with C_ω
 - 15: $\mathcal{L}_{G,3D} \leftarrow -\text{mean } C_\omega(\mathcal{P}(\hat{\mathbf{V}}))$
 - 16: **return** $\mathcal{L}_{G,3D}$ to Algorithm 6
-

The video decoder is fixed during DiT distillation. Its input gradient is retained on generator steps, whereas real videos and detached fake videos update only the discriminator. Crops remain contiguous in time; independently sampling unrelated frames would remove the temporal meaning of a 3D discriminator. The anchor, when present, is fixed in both branches and is excluded from crop-only supervision if the selected crop consists solely of the copied boundary.

This adversarial loss supplements joint velocity distillation; it does not replace the audio objective or introduce an audio discriminator. Its coefficient λ_G , activation step, update frequency, and crop size are explicit training inputs. Because the adversarial term also affects the student optimum, one-step learning is an approximation to the restoration flow with an additional video-realism objective.

19.8 TURBO VAE DISTILLATION

Turbo VAE adopts the training method and code of Ultra Flash (Luxury et al., 2026), adapted to the 128-channel LTX video-latent interface. Its encoder and decoder have separate targets even when optimized in the same iteration. In particular, the decoder receives a frozen teacher latent, rather than the current student’s encoder output; decoder reconstruction gradients therefore do not train the encoder.

Algorithm 8 Distill the lightweight video encoder and decoder

Input: HQ video patches; frozen teacher posterior q_T and channel statistics (μ_T, σ_T)
Input: Tiny encoder E_η , decoder D_ψ , separate 3D discriminator C_ξ ; loss weights and schedules
Output: Turbo encoder E_η and decoder D_ψ

- 1: **for** each training iteration **do**
- 2: Clear encoder and decoder gradients
- 3: Sample an HQ spatiotemporal patch $\mathbf{X}$ and normalize RGB values to $[-1, 1]$
- 4: *Teacher latent target: frozen and normalized*
- 5: Sample $\mathbf{z}_T^{\text{raw}} \sim q_T(\cdot | \mathbf{X})$ without gradients
- 6: $\mathbf{z}_T \leftarrow \text{sg}[(\mathbf{z}_T^{\text{raw}} - \mu_T)/\sigma_T]$
- 7: *Independent encoder and decoder distillation branches*
- 8: $\hat{\mathbf{z}} \leftarrow E_\eta(\mathbf{X})$; align valid latent extents
- 9: $\mathcal{L}_E \leftarrow \ell_{\text{lat}}(\hat{\mathbf{z}}, \mathbf{z}_T)$
- 10: $\hat{\mathbf{X}} \leftarrow D_\psi(\mathbf{z}_T)$; remove the decoder’s causal warm-up prefix
- 11: Align decoded and clean valid extents; apply the same comparison crop
- 12: $\mathcal{L}_{\text{pix}} \leftarrow \ell_{\text{pix}}(\hat{\mathbf{X}}, \mathbf{X})$
- 13: $\mathcal{L}_{\text{per}} \leftarrow \text{mean}_{b,t} \text{LPIPS}(\hat{\mathbf{X}}_{b,t}, \mathbf{X}_{b,t})$
- 14: $\mathcal{L}_{\text{base}} \leftarrow \lambda_E \mathcal{L}_E + \lambda_{\text{pix}} \mathcal{L}_{\text{pix}} + \lambda_{\text{per}} \mathcal{L}_{\text{per}}$
- 15: **if** a discriminator update is scheduled **then**
- 16: Enable gradients for ξ ; clear its gradients and update only ξ on the 3D hinge loss
- 17: **end if**
- 18: **if** a generator adversarial update is scheduled after discriminator warm-up **then**
- 19: Hold ξ fixed, retaining input gradients; $\mathcal{L}_{\text{adv}} \leftarrow -\text{mean } C_\xi(\hat{\mathbf{X}})$
- 20: $\mathcal{L}_{\text{base}} \leftarrow \mathcal{L}_{\text{base}} + \lambda_{\text{VAE,G}} \mathcal{L}_{\text{adv}}$
- 21: **end if**
- 22: Update (η, ψ) on $\mathcal{L}_{\text{base}}$; keep the teacher and perceptual network fixed
- 23: **end for**
- 24: **return** (E_η, D_ψ)

ℓ_{lat} and ℓ_{pix} are the configured elementwise reconstruction losses (MSE in the LTX reference implementation). The latent term is measured in the teacher’s normalized coordinates. LPIPS includes its wrapper’s RGB-range preprocessing and is averaged over frames; the reference wrapper maps $[-1, 1]$ to clipped $[0, 1]$ before evaluation. The adversarial coefficient is separate from any frequency-domain loss. The VAE discriminator C_ξ is distinct from the DiT-distillation discriminator C_ω .

Both distillation branches use the same HQ patch. Downsampling the observed LQ video is a deployment operation before E_η , not an instruction to train $E_\eta(\mathbf{V}^l)$ against a high-resolution teacher latent. The teacher’s 128-channel normalization convention must also be respected at inference; it must not be applied twice. Audio continues to use its original VAE and vocoder.

19.9 CAUSAL MEMORY BLOCKS AND VIDEO-LATENT GEOMETRY

The lightweight video path uses spatial convolutions with causal feature memory. Let $\mathbf{H}_t$ be a block’s input features. With the reference implementation’s replicated initial state, its temporal memory is $\mathbf{Q}_0 = \mathbf{H}_0$ and $\mathbf{Q}_t = \mathbf{H}_{t-1}$ for $t > 0$. The block computes

$$\mathbf{H}'_t = \text{ReLU}(F([\mathbf{H}_t; \mathbf{Q}_t]) + S(\mathbf{H}_t)), \quad (47)$$

where F is a three-convolution spatial stack with intermediate activations and S is the residual projection. A parallel temporal shift and a sequential cache implement the same causal dependency. The cache stores the previous *input* features of each block, not its updated output.

Algorithm 9 Causal-memory video encoding and decoding

Input: Video $\mathbf{X}$ or normalized latent $\mathbf{z}$; selected encoder/decoder; causal boundary convention
Output: Encoded 128-channel latent or raw reconstructed RGB video with causal warm-up prefix

- 1: **if** encoding **then**
- 2: $\mathbf{H} \leftarrow \text{ReLU}(\text{Conv}_{\text{in}}(\text{PixelUnshuffle}_4(\mathbf{X})))$
- 3: **for** each of three encoder stages **do**
- 4: Causally pad the initial feature if needed; pool temporally by factor two
- 5: Apply stride-two spatial convolution
- 6: Apply the stage’s causal memory blocks using Eq. 47
- 7: If not the final stage, apply the inter-stage channel-conversion convolution
- 8: **end for**
- 9: Project features to 128 latent channels
- 10: **else**
- 11: $\mathbf{H} \leftarrow \text{ReLU}(\text{Conv}_{\text{in}}(3 \tanh(\mathbf{z}/3)))$
- 12: **for** each of three decoder stages **do**
- 13: Apply the stage’s causal memory blocks using Eq. 47
- 14: Upsample spatially by factor two using convolution and PixelShuffle
- 15: Expand temporal features by factor two and reshape them into successive frames
- 16: If not the final stage, apply the inter-stage channel-conversion convolution
- 17: **end for**
- 18: Apply ReLU, the output projection, and final $4 \times \text{PixelShuffle}$ to produce RGB
- 19: **end if**
- 20: *Equivalent execution of each memory block*
- 21: In parallel mode, construct previous-input features by a causal temporal shift
- 22: In sequential mode, read the block’s cached previous input, evaluate the block, then cache its current input
- 23: Initialize each block at the sequence boundary using the same replicated-frame convention
- 24: **return** the raw output tensor; its caller handles temporal alignment

The LTX-compatible construction has spatial compression $4 \times 2^3 = 32$ and temporal compression $2^3 = 8$. The reference encoder has three stages of width 128; the decoder widths are 512, 256, and 128, with three memory blocks per stage. A 121-frame window therefore has 16 latent video frames. Here D_ψ denotes raw decoder output, including seven warm-up frames; the training or inference wrapper removes this prefix exactly once before comparison or emission.

These feature caches belong to VAE processing. They are separate from the restored-frame boundary used to connect DiT windows, and do not imply persistent DiT key-value caches. Encoder and decoder calls must use the same temporal padding/cropping convention during training and inference. Spatial tiling, when needed for memory, must preserve the selected overlap and valid-region rules.

19.10 WINDOWED BASE AND ONE-STEP FLASH INFERENCE

Algorithm 10 distinguishes the unanchored first window from subsequent anchored windows. Neighboring 121-frame windows share one boundary frame. The restored endpoint of one window occupies that boundary in the next window, and is emitted only once. Audio is selected and assembled by timestamps on the same timeline.

Algorithm 10 Restore an audio-video recording with bounded window memory

Input: Degraded recording; base or Flash model; frozen decoding modules; internal prompt $\mathbf{e}_{\text{rest}}$
Output: Restored video and audio on the input timeline

- 1: Construct synchronized 121-frame windows with one shared video boundary frame
- 2: $\mathbf{a}_{\text{prev}} \leftarrow \emptyset$; initialize output video and audio timelines
- 3: **for** each window j in temporal order **do**
- 4: Read the LQ video and the audio interval corresponding to that window
- 5: Set both retention gates to one and select condition-noise amplitude 0.5
- 6: Execute only Algorithm 4’s condition-cache phase once, using base or Flash video encoding
- 7: Draw video/audio Gaussian target states on the output grids
- 8: **if** $\mathbf{a}_{\text{prev}} \neq \emptyset$ **then**
- 9: Encode the previous restored frame with the selected video encoder as the new first-frame anchor
- 10: Overwrite its video latent tokens, set their token times to zero, and fix their update mask
- 11: **end if**
- 12: **if** base mode **then**
- 13: Set a descending 15-step Euler schedule $1 = \sigma_0 > \dots > \sigma_{15} = 0$
- 14: **for** $i = 0, \dots, 14$ **do**
- 15: Evaluate fixed-prompt and null-text velocities with identical LQ conditions and anchor
- 16: $\mathbf{v} \leftarrow \mathbf{v}_\emptyset + 3(\mathbf{v}_{\text{rest}} - \mathbf{v}_\emptyset)$
- 17: $\mathbf{z} \leftarrow \mathbf{z} + (\sigma_{i+1} - \sigma_i)\mathbf{v}$ on unanchored entries
- 18: Restore the fixed anchor entries
- 19: **end for**
- 20: **else**
- 21: $\mathbf{z}_0 \leftarrow \mathbf{z}_1 - \mathbf{U}_\phi(\mathbf{z}_1, 1, 0; \mathbf{q}_S)$ on unanchored entries
- 22: Restore the fixed anchor entries; use the distilled guidance without a second student call
- 23: **end if**
- 24: Decode video using the base decoder or Turbo decoder; decode audio using the original VAE and vocoder
- 25: Crop padding to the valid video frames and align decoded audio to the window’s valid timestamps
- 26: Save the last restored RGB frame as $\mathbf{a}_{\text{prev}}$
- 27: Append new video frames once; write only the not-yet-emitted audio interval
- 28: **end for**
- 29: Mux restored audio and video using the original frame rate and common timestamps
- 30: **return** the restored recording

The first window has no clean boundary and denoises all its video tokens. Subsequent windows use a generated boundary, so the prediction should not be described as conditioned on a ground-truth frame at deployment. A final short window is padded only for model input; the padded frames and waveform samples are removed before output assembly. Audio overlap is handled by timestamps rather than by assuming that one latent audio token equals one video frame.

Base-model guidance retains LQ in both text branches, as in Eq. 9. Flash instead uses the fixed-guidance field learned during distillation. The joint DiT always predicts both modalities; Turbo changes the video encoding/decoding path and reduced-resolution condition interface, while the audio encoder, decoder, and vocoder remain unchanged.

Table 29: Severity-weighting interaction ablation (K_3 retrained, all else fixed). The right-hand columns report the 55 jointly degraded clips; the left-hand columns report all 71 clips. Results for the 16 single-modality clips are summarized below. $\gamma = 2.0$ is the default.

Setting	All clips ($n=71$)			Joint-degradation subset ($n=55$)		
	MUSIQ $\uparrow$	DNSMOS $\uparrow$	FAD $\downarrow$	MUSIQ $\uparrow$	DNSMOS $\uparrow$	FAD $\downarrow$
$\gamma = 0$ (no interaction)	70.62	2.61	6.71	70.08	2.57	6.94
$\gamma = 2.0$ (default)	71.17	2.70	6.30	70.81	2.68	6.48

20 ADDITIONAL ABLATIONS AND CONTROLLED DIAGNOSTICS

This section reports additional ablations and diagnostics that address specific mechanisms of the restoration model and its deployment variant. All numbers are on the Controlled track (71 clips) unless stated otherwise; metric configurations follow Sec. 4.2.

20.1 SEVERITY-WEIGHTING INTERACTION TERM

The degradation-aware reweighting in Eq. 13 includes a multiplicative interaction term $\gamma d_v d_a$ that emphasizes samples degraded in both modalities simultaneously. Table 29 isolates this term by retraining the final $K_3 = 33,000$ AV2AV steps with $\gamma \in \{0, 2.0\}$ while holding all other hyperparameters fixed. The interaction term benefits the jointly degraded Controlled subset (55/71, approximately 77.5% of clips) substantially more than the single-modality-degraded subset, which is exactly the behavior the term is designed to produce.

On the jointly degraded subset the interaction term improves MUSIQ by +0.73 and DNSMOS by +0.11; on the single-modality subset ($n=16$) the effect is negligible (MUSIQ -0.07 , within seed noise per Table 33; DNSMOS +0.02; FAD -0.24). The term concentrates its benefit where it should—jointly degraded clips—without harming single-modality cases. Figure 30 visualizes the resulting weight distribution.

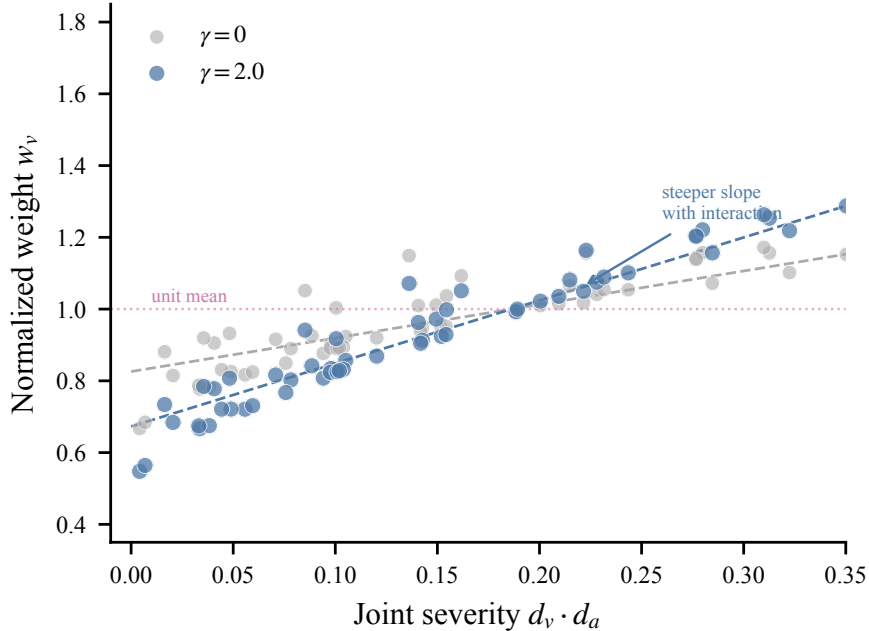


Figure 30: Normalized severity weights w_v versus the joint severity $d_v d_a$ over a representative training batch, under $\gamma = 0$ (gray) and $\gamma = 2.0$ (blue). The interaction term increases the weight on jointly degraded samples without changing the unit-mean constraint.

20.2 GATE DYNAMICS DURING ADAPTATION

Sec. 17.6 ablates the cross-modal gates at the endpoint. Figure 31 reports how the gate values evolve during the 50,000 adaptation steps, starting from $\sigma(-4.0) \approx 0.018$. The audio $\rightarrow$ video gate grows faster during TAV2AV and reaches a higher plateau (≈ 0.72) than video $\rightarrow$ audio (≈ 0.58), indicating that the video pathway draws more on acoustic evidence than the reverse—consistent with the endpoint ablation, where disabling audio $\rightarrow$ video costs more MUSIQ.

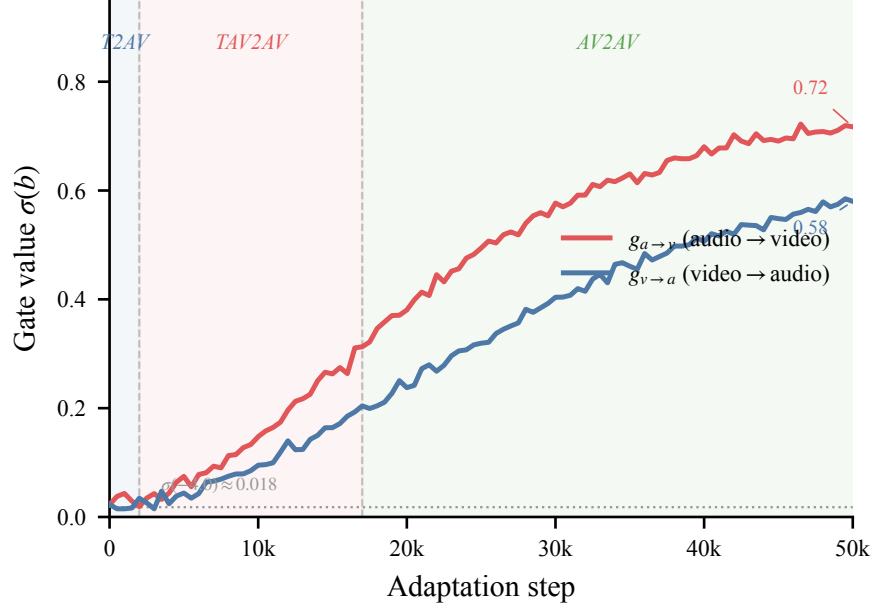


Figure 31: Evolution of the learnable cross-modal gates $\sigma(b)$ over the 50,000 adaptation steps. Both gates open monotonically from the near-zero initialization; the audio $\rightarrow$ video gate ($g_{a \rightarrow v}$) reaches a higher plateau than the video $\rightarrow$ audio gate ($g_{v \rightarrow a}$), reflecting the asymmetric dependence measured in Table 28.

20.3 RESOLUTION-DEPENDENT VAE TIMING

Table 24 reports isolated VAE encode+decode throughput across four resolutions, and Figure 32 separates the encoding and decoding costs. From 480×270 to 2560×1440 , the plotted combined latency increases from approximately 1.4 s to 20.4 s for the full VAE and from 24 ms to 193 ms for Turbo VAE. The total latency grows more slowly for Turbo VAE over this tested range, widening the reported speedup from $59.9\times$ to $105.8\times$. These totals describe the measured configurations; they do not determine separate asymptotic scaling laws for the encoder and decoder.

20.4 LONG-SEQUENCE CHAINING BEYOND THE BENCHMARK DURATION

The OmniVRBench clips are at most 20 s (Table 10), which corresponds to at most four 121-frame windows at 24 fps ($20.17\text{ s} \approx 484$ frames). To evaluate whether drift accumulates over longer durations, we construct an additional **extended-duration test set** of 12 clips of 30–45 s, drawn from the same archival sources as the Real track but excluded from all other evaluations. These clips are processed by sliding the 121-frame window with stride 120 frames (adjacent windows share exactly one anchor frame), so that W windows cover $120W + 1$ frames with no padding and no looping. Figure 33 reports face identity similarity, MUSIQ, and cumulative chroma drift (mean ΔE in CIE-Lab against the first window’s restored frame) as a function of W .

Over 40 s (8 windows) face similarity declines by 0.027 and MUSIQ by 1.08, both small relative to the gap to the best baseline (Table 2). Chroma drift is the dominant long-horizon effect, growing approximately linearly; resetting the anchor to the first window’s restored frame (an oracle upper bound, second block of Table 30) cuts cumulative drift by $\sim 70\%$, confirming a chaining artifact rather

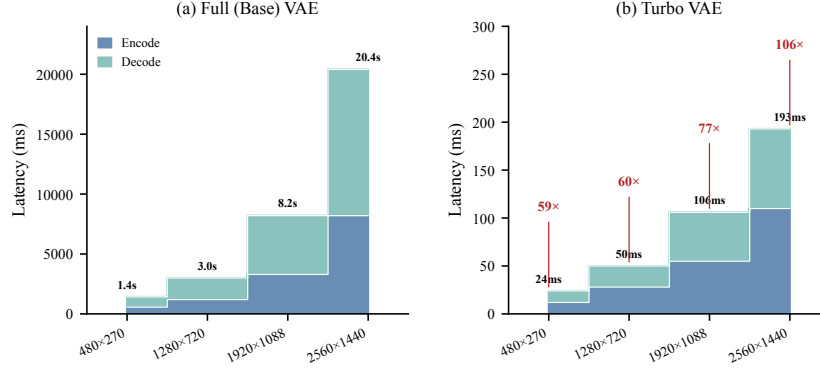


Figure 32: Resolution-dependent VAE encoding and decoding times corresponding to Table 24. Left: full VAE. Right: Turbo VAE. Bar totals show combined latency; annotated ratios show the full/Turbo VAE speedup at each resolution.

Table 30: Extended-duration chaining evaluation ($n=12$ clips, 30–45 s). Windows are 121 frames with stride 120; W windows cover $120W+1$ frames. Face similarity is against the first-window restored face; chroma drift is mean ΔE (CIE-Lab) against the first-window restored frame.

Windows W	1	2	3	4	6	8
Duration (s)	5.0	10.0	15.0	20.0	30.0	40.0
Face cos-sim $\uparrow$	0.881	0.876	0.871	0.867	0.860	0.854
MUSIQ $\uparrow$	71.17	70.94	70.72	70.55	70.28	70.09
Chroma drift $\Delta E \downarrow$	0.0	0.8	1.5	2.1	3.2	4.1
<i>With per-window anchor reset (upper bound)</i>						
Face cos-sim $\uparrow$	0.881	0.879	0.877	0.875	0.872	0.870
Chroma drift $\Delta E \downarrow$	0.0	0.3	0.5	0.7	1.0	1.3

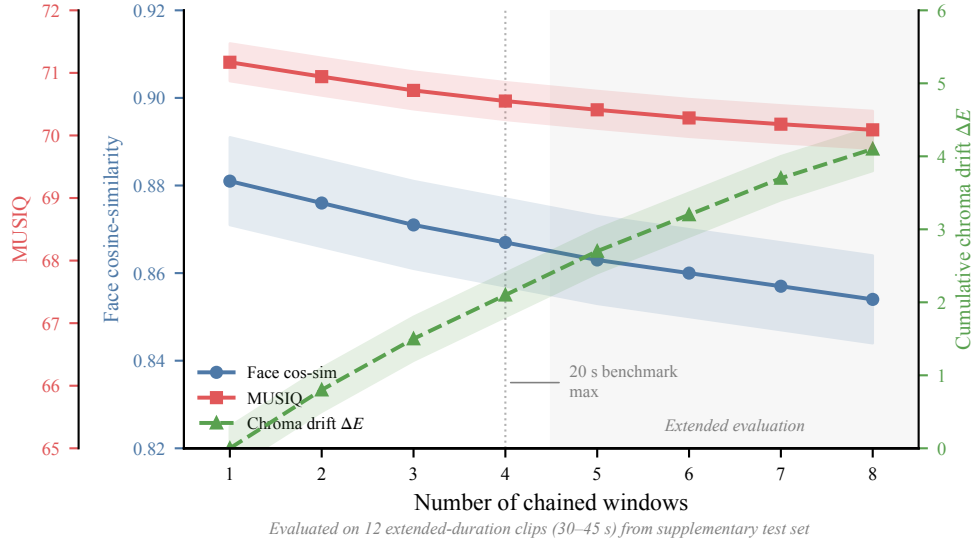


Figure 33: Long-sequence chaining metrics on the extended-duration test set ($n=12$, 30–45 s). Face similarity and MUSIQ decline slowly and monotonically; chroma drift grows approximately linearly at $\approx 0.55 \Delta E$ per window, consistent with the color-drift failure mode in Sec. 9.4.

than a model-level quality regression. This motivates the self-generated-anchor training discussed in Sec. 15 and Sec. 6.

Table 31: Open-vocabulary audio content preservation on the Controlled track ($n=40$ speech clips). WER and insertion/deletion rates are computed against the clean reference transcript. “Numbers/proper nouns” is a subset analysis of the most fidelity-sensitive tokens.

Method	WER↓	Insertion rate↓	Deletion rate↓	Numbers/proper nouns acc.↑
Low-quality input	0.482	0.061	0.184	0.42
VoiceFixer	0.311	0.058	0.092	0.61
OmniVR (ours)	0.264	0.054	0.071	0.74

Table 32: Eligible vs. successfully measured clip counts on the Controlled track. The baseline is RealBasicVSR for visual rows and VoiceFixer for the audio row. “Eligible” is set by the protocol (Table 11); “measured” is the count that produced a finite metric value. RealBasicVSR shortfalls are due to inference-pipeline OOM on near-4K sources. OmniVR’s inference pipeline completes on all 71 clips; its single LPIPS shortfall ($n=70$) reflects a metric-computation OOM rather than a model failure (footnote *c* in Table 5).

Metric	Eligible	Baseline measured	OmniVR measured	Common subset
LPIPS (visual)	71	70	70	70
DISTS (visual)	71	70	71	70
FVD (visual)	71	68	71	68
SSIM / PSNR (visual)	71	71	71	71
STOI / SI-SDR / PESQ (audio)	71	71	71	71
Face cos-sim / OCR	71	71	71	71

20.5 AUDIO CONTENT PRESERVATION BEYOND INTELLIGIBILITY PROXIES

The closed-set intelligibility pilot in Sec. 8 measures clip-level word accuracy but not content fidelity in an open-vocabulary setting. To complement it, we report an independent open-vocabulary evaluation: two human annotators transcribe the restored audio of 40 Controlled-track clips (containing speech) with no access to the reference transcript, and we score word error rate (WER) and content-word preservation against the clean reference using the same ASR reference used in the pipeline. Table 31 reports the results together with insertion/deletion rates, which separate hallucinated content from dropped content.

OmniVR achieves the lowest WER (0.264) and insertion rate, and on the fidelity-sensitive subset of numbers and proper nouns reaches 74% accuracy vs. 61% for VoiceFixer. Together with the closed-set result (94%, Sec. 8), this supports the view that the low STOI/eSTOI reflects envelope-resynthesis sensitivity rather than lost content. The residual WER and the 26% of unrecovered numbers/proper nouns are a genuine limitation, concentrated in clips with $d_a > 0.9$, consistent with the over-smoothing failure mode in Sec. 9.4.

20.6 ELIGIBLE VERSUS SUCCESSFULLY MEASURED SAMPLE COUNTS

Table 11 (Sec. 11) reports eligibility. A small number of eligible clips fail to produce a measurement because a released baseline pipeline runs out of memory or emits too few frames; these are distinct from eligibility. Table 32 separates the two for every full-reference and distributional metric on the Controlled track.

Most pairwise comparisons in Table 5 use all 71 clips for both methods. For DISTS and FVD, RealBasicVSR is limited to $n=70$ and $n=68$ respectively due to OOM on near-4K sources (footnotes *c,d*), while OmniVR covers all 71 (footnote *e*). These two comparisons are therefore not on a matched common subset; Table 32 above records the exact per-metric counts.

20.7 TRAINING-SEED ROBUSTNESS

The main tables report a single 50k checkpoint. To check sensitivity to training data order, we retrain the base model from the same pretrained backbone with two additional data-order seeds (same hyperparameters, different sample shuffling) and evaluate on the Controlled track.

Table 33: Training-seed robustness of the base model (Controlled track). Seed 0 is the checkpoint reported in the main tables; seeds 1–2 vary data ordering only. “vs. best baseline” uses the strongest non-OmniVR number in Table 2.

Run	MUSIQ $\uparrow$	CLIP-IQA $\uparrow$	DNSMOS $\uparrow$	FAD $\downarrow$	LSE-C $\uparrow$
Seed 0 (reported)	71.17	0.543	2.70	6.30	3.52
Seed 1	70.93	0.539	2.68	6.41	3.47
Seed 2	71.08	0.541	2.66	6.36	3.50
Mean $\pm$ std	71.06 ± 0.12	0.541 ± 0.002	2.68 ± 0.02	6.36 ± 0.06	3.50 ± 0.03
Best baseline	45.54	0.268	2.12	7.02	2.32

Training-seed standard deviations (± 0.12 MUSIQ, ± 0.02 DNSMOS) are an order of magnitude smaller than the gaps to the best baseline ($+25.6$ MUSIQ, $+0.56$ DNSMOS), so the reported ordering is robust to data ordering. These training-repeat statistics are distinct from the inference-seed variance in Sec. 16.3.