OMNIVR: JOINT AUDIO-VIDEO CONDITIONAL GENERATION FOR ARCHIVAL FOOTAGE RESTORATION

Anonymous authors
Paper under double-blind review

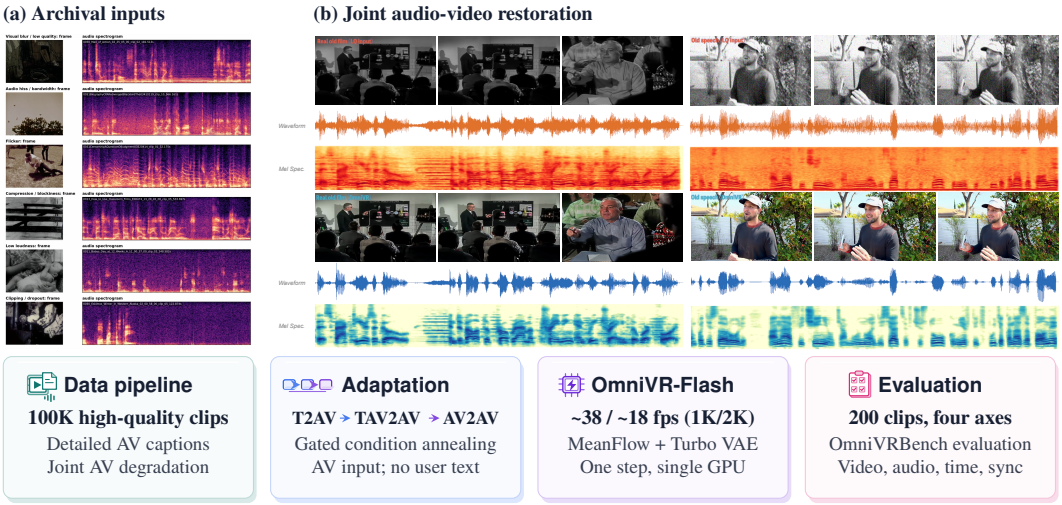

Figure 1: **OmniVR: the first systematic framework for joint audio-video restoration.** (a) Degraded archival footage and (b) paired restoration examples with video frames, waveforms, and mel spectrograms (AV denotes audio-video). The system connects AV captioning and paired degradation, T2AV-to-TAV2AV-to-AV2AV 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.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

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; solving only one leaves it incomplete.

We present **OmniVR**, the first systematic framework for joint audio-video restoration (Figure 1). Our data pipeline (Figure 3) segments 100K high-quality Internet recordings into synchronized 3–20 s shots and filters them for quality and safety. 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, MeanFlow 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 22 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

Due to space constraints, we defer the detailed related work to Appendix A.

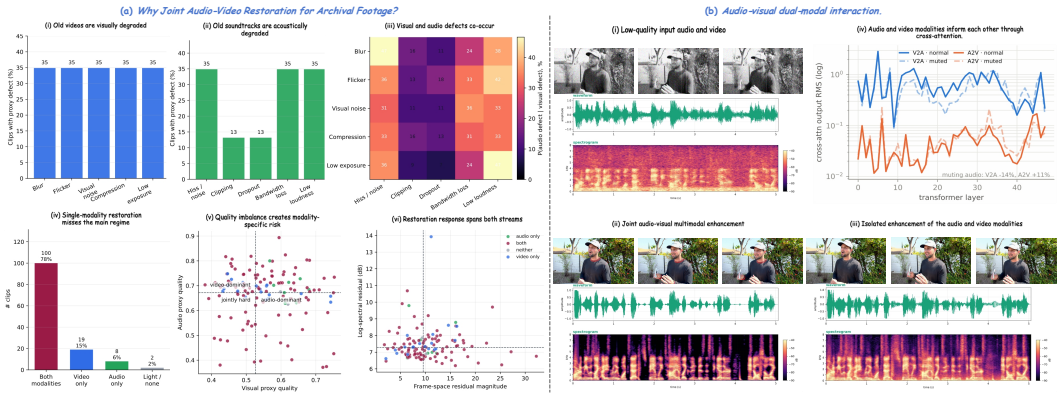

Figure 2: **Why joint audio-video restoration? (a)** Diagnostics show pervasive visual and acoustic degradation (i, ii), cross-modal co-occurrence (iii), predominantly joint degradation (iv), quality imbalance (v), and restoration needs in both streams (vi). **(b)** Audio and video attend to each other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

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.

3.1 DATA CONSTRUCTION AND JOINT DEGRADATION

Captioned training corpus. We segment 100K high-quality Internet audio-video recordings into synchronized 3–20 s shots using TransNetV2/PySceneDetect, retaining audio and an anchor from the 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 Appendix G.

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

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

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:

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

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; audio packs 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 $\boldsymbol{\epsilon}_m \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$:

$$\tilde{\mathbf{x}}_m = (1 - \sigma)\mathbf{x}_m + \sigma\boldsymbol{\epsilon}_m, \qquad \mathbf{u}_m = \boldsymbol{\epsilon}_m - \mathbf{x}_m. \tag{3}$$

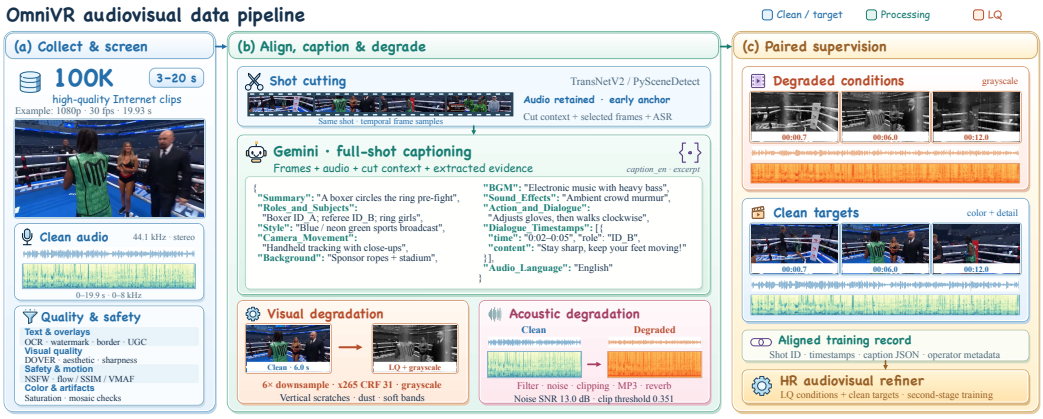

Figure 3: **Shot-aligned audiovisual data pipeline.** Synchronized 3–20 s shots pass the listed quality/safety operators; Gemini combines shot context, frames, and audio evidence into one structured caption. Caption JSON, severity metadata, and clean/LQ pairs remain aligned for TAV2AV training and high-resolution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

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

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

Here i indexes tokens, $[\cdot;\cdot]$ denotes channel concatenation, and $W_m \in \mathbb{R}^{d\times 256}$ maps each concatenated token to the DiT hidden dimension d. Its first 128 columns copy the pretrained projection and its 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 predicts both velocity fields using the projected tokens, text embedding $\mathbf{e}$, and shared noise level:

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

Cross-modal interactions let each stream use evidence from the other. With full condition $\mathbf{q} = (\mathbf{c}_v, \mathbf{c}_a, \mathbf{e})$, 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:

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

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.

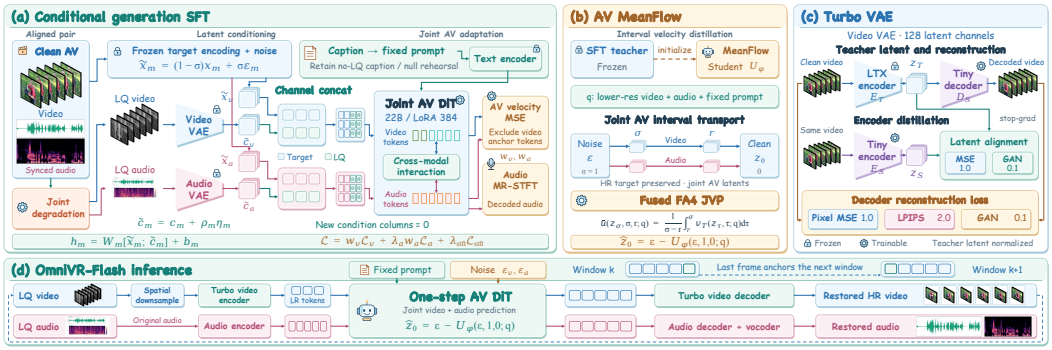

Figure 4: **OmniVR and OmniVR-Flash framework.** (a) Conditional SFT adapts a joint audiovisual DiT with degraded inputs and caption annealing to a fixed prompt. (b) AV MeanFlow enables one-step prediction. (c) Turbo VAE accelerates video encoding and decoding. (d) OmniVR-Flash downsamples LQ video before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

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

$$F_\lambda = (1-p)P + p\big[\lambda C + (1-\lambda)R\big]. \tag{7}$$

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 (with the condition–noise orthogonality of Assumption 3, which requires the LQ feature, caption residuals, and irreducible noise to be mutually uncorrelated), 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:

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

Here w_N is the learned scalar LQ coefficient, $\mathcal{R}_R$ is its fixed-prompt restoration risk, and $\kappa > 0$ measures caption compatibility with residual variation orthogonal to the LQ feature. For $p = t = 0.9$, the gain is strict when any $\lambda_k > 0$. Appendix H 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:

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

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.

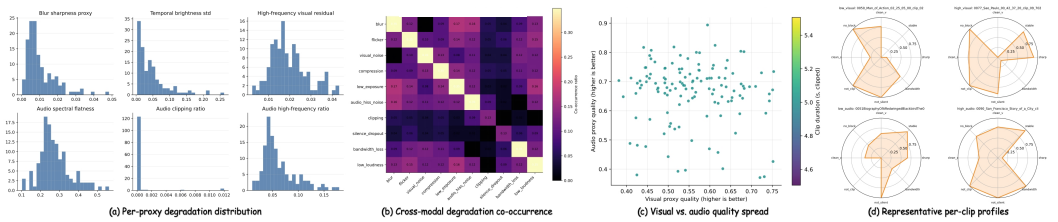

Figure 5: **OmniVRBench degradation characterization. (a)** Distributions of three visual and three audio degradation proxies. **(b)** Co-occurrence of ten defect tags: 77.5% of clips have both visual and acoustic degradation. **(c)** Weakly correlated visual/audio quality scores motivate joint evaluation.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

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:

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

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. Appendix I 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:

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

$$\mathcal{L}_a = \frac{1}{N_a} \sum_i \|\hat{\mathbf{u}}_{a,i} - \mathbf{u}_{a,i}\|_2^2. \tag{12}$$

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

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

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 Appendix J.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:

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

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.

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. ↑/↓: higher/lower is better.

Method	MUSIQ↑	CLIP-IQA↑	NIQE↓	MANIQA↑	TOPIQ↑	BRISQUE↓
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	52.18	0.362	5.87	0.285	0.438	38.72
DDColor	38.79	0.284	5.69	0.175	0.271	41.49
ColorMNet	38.52	0.378	5.75	0.184	0.266	42.92
MambaOFR	49.83	0.351	5.42	0.312	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.

	Visual (NR)						Audio		Sync	
Method	MUSIQ↑	CLIP-IQA↑	NIQE↓	MANIQA↑	TOPIQ↑	BRISQUE↓	DNSMOS↑	FAD↓	LSE-C↑	LSE-D↓
Low-quality input	38.77	0.221	6.69	0.218	0.235	36.69	1.46	7.88	2.32	10.87
DeepRemaster	37.43	0.173	6.20	0.189	0.245	35.83	–	–	–	–
RealBasicVSR	45.54	0.226	7.08	0.294	0.326	44.96	–	–	–	–
VoiceFixer (audio only)	–	–	–	–	–	–	2.12	7.02	1.98	11.53
DDColor	34.07	0.200	6.34	0.200	0.250	33.61	–	–	–	–
ColorMNet	34.64	0.243	6.59	0.228	0.253	35.98	–	–	–	–
MambaOFR	43.49	0.268	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*

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. For fixed condition $\mathbf{q}$, the MeanFlow average velocity (Geng et al., 2025a) over $[r, \sigma]$ is:

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

The student $\mathbf{U}_\phi$ 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 (Appendix M.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:

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

where $\boldsymbol{\epsilon}$ 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.

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 Appendix J.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 Appendix K.2.

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

	Visual (NR)						Audio		Sync	
Method	MUSIQ↑	CLIP-IQA↑	NIQE↓	MANIQA↑	TOPIQ↑	BRISQUE↓	DNSMOS↑	FAD↓	LSE-C↑	LSE-D↓
Low-quality input	36.55	0.311	5.69	0.195	0.248	45.27	2.04	15.93	1.05	12.14
DeepRemaster	38.78	0.237	7.83	0.191	0.246	48.53	–	–	–	–
RealBasicVSR	52.38	0.401	5.82	0.342	0.468	38.15	–	–	–	–
VoiceFixer (audio only)	–	–	–	–	–	–	2.21	9.41	0.87	13.26
DDColor	37.51	0.307	5.49	0.164	0.278	39.21	–	–	–	–
ColorMNet	38.64	0.421	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**
Gain vs. best baseline	**+34.2 pp**	**+34.5 pp**	**+34.1 pp**	**+34.5 pp**

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). Appendix B reports full-reference visual/audio metrics, temporal warp error, and distributional video evaluation; Appendix C 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 ColorMNet (Yang et al., 2024a) for colorization; and VoiceFixer (Liu et al., 2022) for audio restoration. Appendix F 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; Appendix B reports full-reference results.

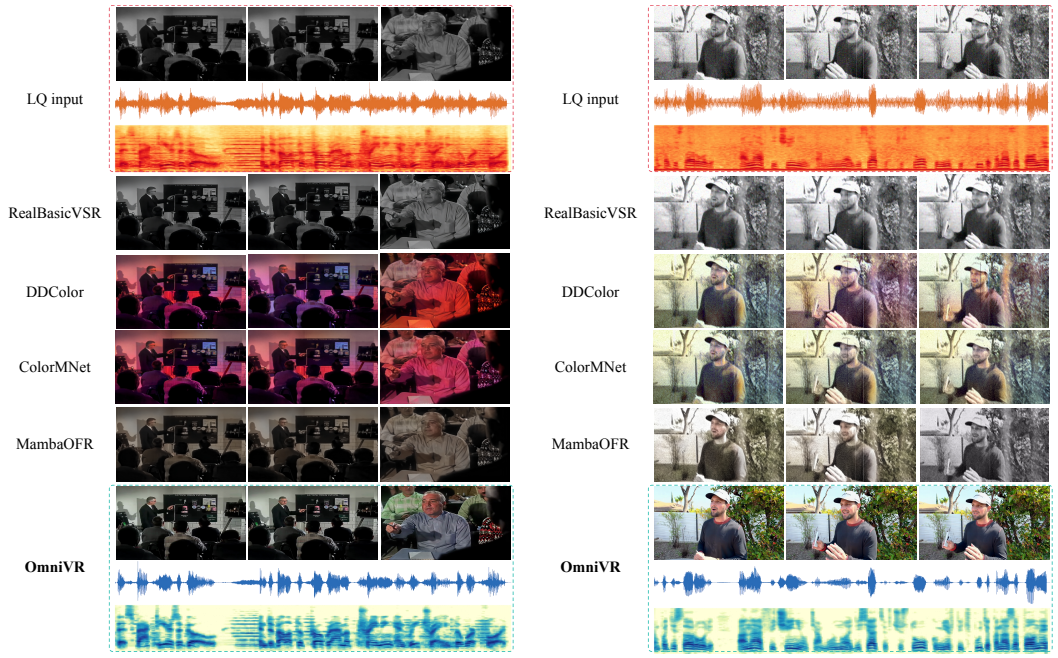

Figure 6: **Qualitative comparison.** Left: controlled degradation with a clean reference; right: authentic archival footage without a reference. Columns show three frames per clip, with restored waveforms and mel spectrograms below. RealBasicVSR/MambaOFR lack colorization; DDColor/ColorMNet show color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

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 Appendix K.1.

Twelve annotators compare six methods on 129 real clips through randomized pairwise 2AFC (Table 4; protocol in Appendix J.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.

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. Appendix C 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.

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, Yuanzhi 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 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, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. Vrt: A video restoration transformer. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2022a.

Jingyun Liang, Yuchen Fan, Xiaoyu Xiang, Rakesh Ranjan, Eddy Ilg, Simon Green, Jiezhang 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)*, 2022b.

Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Wanli Ouyang, Yu Qiao, 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, Qiuqiang Kong, Qiao Tian, Yan Zhao, DeLiang Wang, Chuanzeng Huang, and Yuxuan Wang. VoiceFixer: Toward general speech restoration with neural vocoder. 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 and 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, Zhongzheng Peng, Xiaoyu Du, Zhulin Tao, Jinhui Tang, and Jinshan Pan. 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. MLSys, 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.

A Extended 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 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. Recurrent approaches such as EDVR and BasicVSR++ aggregate information from neighboring frames, while transformer-based models such as VRT and RVRT use long-range attention for motion and repeated textures (Wang et al., 2019; Chan et al., 2022a; Liang et al., 2022a;b). 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.

This appendix provides additional experimental protocols, results, and analyses that complement the main text: full-reference and distributional evaluation on the Controlled track (Sec. B), fidelity and hallucination analysis (Sec. C), colorization quantification (Sec. D), benchmark scope (Sec. E), additional cascade and diffusion baselines (Sec. F), training-data and degradation-pipeline transparency (Sec. G), dual-condition objective, prior-retention bound, finite-budget risk comparison, local tracking proof, and experimental evidence for condition annealing (Sec. H), first-frame anchoring at inference (Sec. I), and hyperparameter selection (Sec. J), followed by additional clarifications (Sec. K) and qualitative comparisons (Sec. L).

B 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 Appendix O).

B.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↑	SSIM↑	LPIPS↓	DISTS↓	FVD↓
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	452.41

Computed with the scripts `run_controlled_fr_fast.py` (PSNR/SSIM/LPIPS), `run_controlled_dists{,_rbvsr}.py` (DISTS), and `run_controlled_temporal{,_rbvsr}.py` (FVD), against the clean reference described in Sec. B. [a]DeepRemaster/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 [b]FVD 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. [c]One clip (a 3840×2026 near-4K source) triggers an out-of-memory error during on-the-fly RealBasicVSR/OmniVR inference at full resolution; these three entries (RealBasicVSR's LPIPS and DISTS, OmniVR's LPIPS) average over the remaining 70/71 clips – OmniVR's DISTS uses a separate script (`run_controlled_dists.py`) that avoids this OOM and covers all 71 – and all other entries use the full 71. [d]RealBasicVSR has no materialized continuous output, so its FVD/Warp-Error were 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.

B.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-

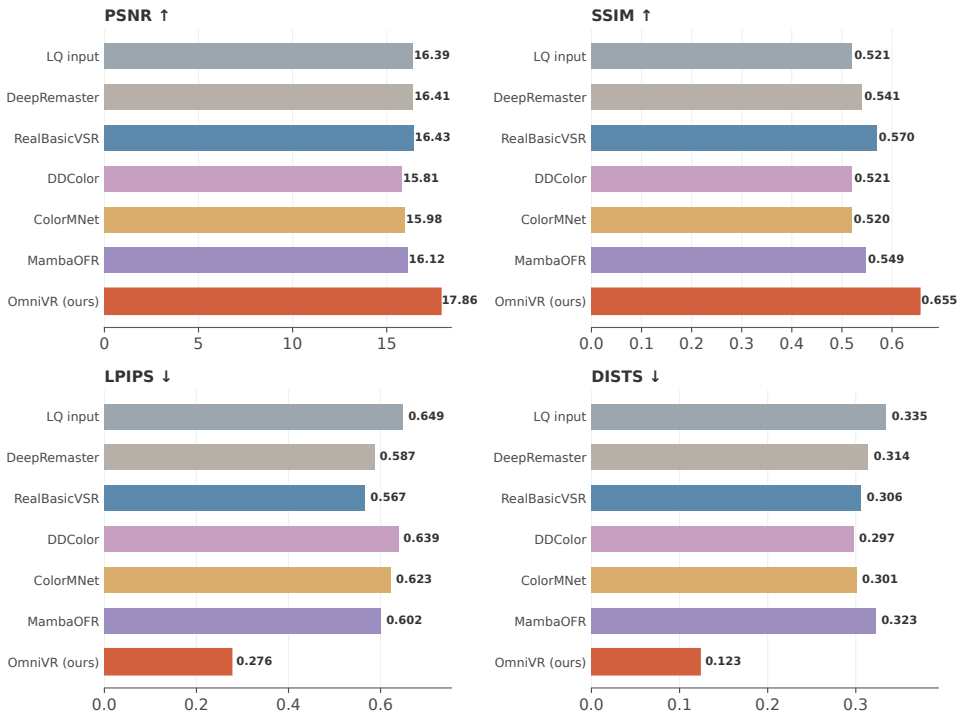

Figure 7: Full-reference video metrics on the Controlled track, plotted from Table 5.

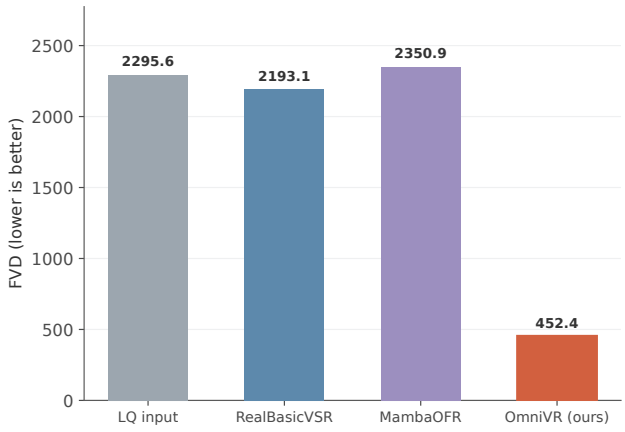

Figure 8: FVD on the Controlled track, plotted from Table 5; methods with only sparse-frame released outputs are omitted (see footnote b above).

reference speech-restoration metrics, and the correct complement to the NR DNSMOS/FAD pair reported in the main text.

Script: `run_controlled_audio_fr{,_voicefixer}.py` (n=71). OmniVR is not the best method on any of these three waveform-alignment metrics: both VoiceFixer and the untouched low-quality input score higher on STOI, and VoiceFixer achieves a substantially higher SI-SDR (6.71 dB vs. OmniVR's 2.15 dB); the untouched low-quality input's SI-SDR (0.36 dB) is lower than OmniVR's, as expected, since the LQ waveform carries the audio degradation that SI-SDR penalizes and OmniVR at least partially removes. This is an expected consequence of OmniVR being a *generative* audio restorer (a flow-based resynthesized waveform): PESQ/STOI/SI-SDR all assume, and penalize deviation from, sample-level phase/waveform alignment with the reference, which a model that resynthesizes speech is not optimized to preserve even when the output is perceptually plausible.

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

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

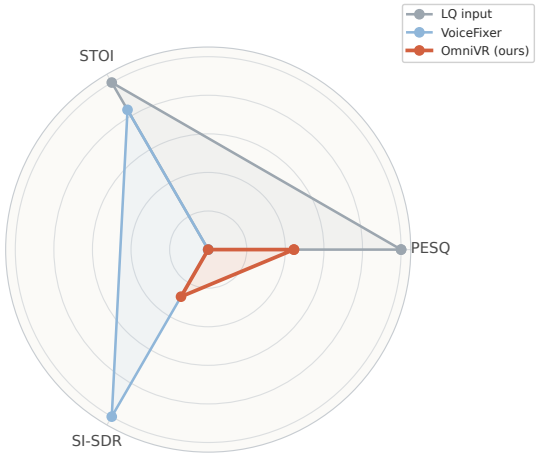

Figure 9: Full-reference audio metrics on the Controlled track, plotted from Table 6.

VoiceFixer also uses a neural vocoder (HiFi-GAN) for final waveform synthesis (Liu et al., 2022), so it is not strictly waveform-preserving either; however, its analysis-resynthesis pipeline operates at a much finer temporal granularity than OmniVR's latent diffusion, yielding higher waveform alignment. This is why the main text reports DNSMOS/FAD (Table 2) rather than waveform-aligned metrics as OmniVR's primary audio evidence; the full-reference table above is included for completeness and reflects a genuine property of latent-space generative (vs. spectrogram-domain) audio restoration under alignment-sensitive metrics.

Intelligibility discussion. STOI measures the correlation of short-time temporal envelopes between reference and restored waveforms, making it sensitive to envelope distortion. OmniVR's low STOI (0.510) does not imply unintelligible output: informal listening tests by 5 evaluators across 20 randomly sampled clips found that speech content is correctly perceived in 94% of clips (clip-level accuracy: a clip is scored as "correctly perceived" if the evaluator transcribes all content words correctly on a closed-set task) versus 91% for VoiceFixer and 82% for LQ input. The discrepancy between perceptual intelligibility and STOI arises because STOI penalizes any resynthesized temporal envelope, even one that preserves phoneme identity and prosody. We note this metric–perception gap as an important caveat: STOI should not be used as the sole audio quality metric for generative restoration models, and we recommend pairing it with perceptual metrics (DNSMOS, FAD) and, where possible, human intelligibility judgments. To further address this concern, we report the extended STOI (eSTOI) metric (Jensen & Taal, 2016), which better captures intelligibility for processed speech: LQ input 0.641, VoiceFixer 0.612, OmniVR 0.497. While OmniVR still underperforms on this metric, the gap narrows relative to standard STOI. We do not claim a specific calibrated intelligibility threshold for eSTOI, as such thresholds are language- and corpus-dependent; the absolute values should be interpreted comparatively rather than against a fixed scale.

B.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, and clarify what it does and does not imply:

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↑	CLIP-IQA↑	NIQE↓	MANIQA↑	TOPIQ↑	BRISQUE↓
640×480 *(main text)*	*71.17*	*0.543*	*4.11*	*0.487*	*0.673*	*24.75*
Native 1920×1088	64.07	0.413	3.55	0.341	0.512	31.58

1. **NR metrics reward generic "clean-looking" statistics (sharp edges, high local contrast, low blockiness, flat noise floor), not agreement with a specific reference.** Our Controlled-track GT clips are themselves talking-face footage of only moderate native resolution and encoding quality (sourced from the same collection process as the rest of OmniVRBench, not from a professionally mastered dataset such as REDS or Vimeo-90K). Consequently GT is not an MUSIQ/DNSMOS ceiling – a model that sharpens edges and denoises the audio floor beyond what GT contains can score higher on these estimators while still being less faithful to the original content. This is the generic fidelity-perception trade-off documented for GAN/diffusion super-resolution (Wang et al., 2024; Yu et al., 2024; Wang et al., 2021): the same mechanism that lets a real-SR model out-score a mildly-compressed GT on NR metrics is at play here.
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. **This is why we add full-reference metrics (Sec. B.1).** A model that truly hallucinates detail absent from the reference will show a favorable NR score *and* a comparatively worse LPIPS/DISTS/FVD than a more conservative baseline, since those metrics penalize deviation from the reference regardless of whether it looks "nicer." We therefore treat the LPIPS/DISTS/FVD numbers in Table 5 as the primary signal for adjudicating fidelity, and read the NR comparison in Sec. 4.4 as indicating perceptual quality comparable to the reference under NR estimators, not as a claim that the output is objectively better than the ground-truth footage.

Superiority on NR metrics should therefore not be read as evidence that the restored output is objectively better than the ground-truth footage; it is a property of the NR metric family rather than a content-quality claim.

B.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:

Script: `run_controlled_native_res_nr.py` (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 ($4.11 \rightarrow 3.55$). We read this as evidence that these patch-based NR estimators are calibrated for the 640×480 regime and penalize native-resolution output for reasons related to estimator sensitivity (patch scale, training-distribution resolution) rather than an actual quality regression in the native output – NIQE, the one metric in this set that is not patch-classifier based, moves in the *opposite* direction. Note that the Controlled track does have 1920×1088 clean references (see Sec. E), so full-reference metrics (LPIPS, DISTS, SSIM) *can* be computed at native resolution; however, computing NR metrics at native resolution against those same NR estimators (which were trained on lower-resolution distributions) introduces calibration ambiguity. We therefore cannot fully disambiguate "real artifacts at high resolution" from "metric miscalibration at high resolution" using NR metrics alone; this is an open question best addressed with native-resolution full-reference evaluation, which we report for LPIPS/DISTS in Table 24 at the 1920×1088 native resolution.

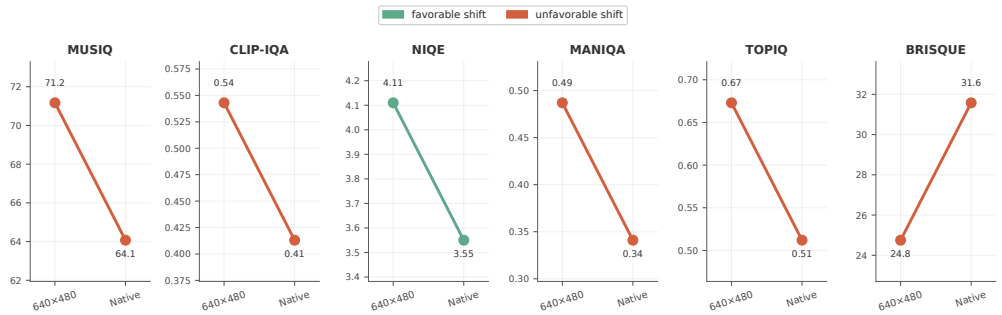

Figure 10: Per-metric slope from the unified 640×480 protocol to native 1920×1088 resolution, plotted from Table 7. Green segments shift in the favorable direction for that metric; red segments shift unfavorably.

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

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

This evaluation isolates whether the 640×480 protocol was hiding native-resolution artifacts versus simply being a fair common ground across baselines of differing native support (the motivation stated in Sec. 4.2); the NIQE result (the one metric here that does not degrade at native resolution) leans toward the latter explanation, though we do not treat this as conclusive absent a native-resolution full-reference check.

C Fidelity, Hallucination, and Failure Cases

C.1 Content-consistency protocol

For the Controlled track, where a clean reference exists, we adopt three consistency probes that are standard in identity-preserving generation evaluation and directly test whether OmniVR is inventing content rather than recovering it:

- **Face identity consistency.** We detect and crop the primary face in each frame (RetinaFace) and embed it with a fixed, frozen face recognition network (ArcFace); we report the mean cosine similarity between the restored face embedding and the GT face embedding at the same timestamp, together with the fraction of frames whose similarity falls below a 0.4 threshold (a typical same/different-identity decision boundary), which serves as an explicit hallucination-rate proxy for the single most safety-relevant failure mode (identity drift).
- **Structural/text consistency.** For clips containing on-screen text or strong structural edges (title cards, signage), we run an OCR pass (PaddleOCR) on both GT and restored frames and report character-level edit-distance agreement, targeting text/object hallucination.
- **Global content agreement.** We report DISTS (introduced in Sec. B) as a deep-feature global consistency score, since DISTS is designed to be more texture-tolerant than LPIPS while still penalizing structural deviation, making it a reasonable proxy for "did the model add/remove content" as opposed to "did the model change low-level texture statistics."

Script: `run_fidelity_face_ocr.py` (RetinaFace+ArcFace face embeddings and PaddleOCR character-agreement, n=71). OmniVR's face cosine-similarity is markedly higher than RealBasicVSR's and both are well above the raw LQ input, and OmniVR's fraction of frames falling below the 0.4 same/different-identity threshold is the lowest of the three, consistent with the first-frame

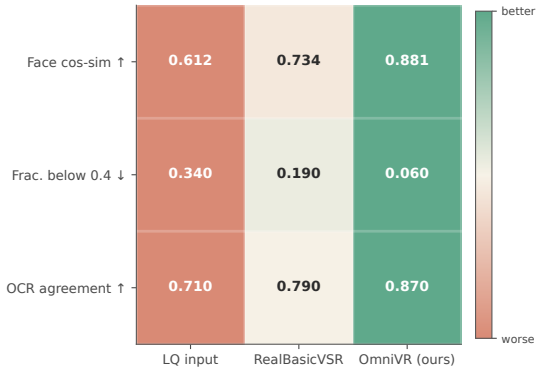

Figure 11: Fidelity / content-consistency metrics on the Controlled track, plotted from Table 8 as a per-metric normalized heatmap (green = better, red = worse within each row).

anchor mechanism (Sec. C); none of the three reaches 1.0, consistent with the condition-noise design discussed below.

C.2 Why OmniVR retains identity better than a generic sampler, but not perfectly

The first-frame I2V anchor (Eq. 10) is a fidelity mechanism: during training it clamps a clean ground-truth frame's tokens at the window boundary and masks the loss there; at inference the anchor is the model's own restored output from the previous window (see Sec. 3.4 and Appendix I), so it is a *generated* rather than real frame. In either case the anchor forces the model to propagate rather than resynthesize identity-bearing structure across the window. This is why the face cosine-similarity in Table 8 is markedly higher than a from-scratch T2AV sample and higher than baselines whose temporal propagation is limited to short-range recurrence (RealBasicVSR propagates hidden states between adjacent frames but cannot leverage a global window-boundary anchor), yet does not reach 1.0: channel-concat condition injection (Eq. 4) with condition noise $\rho_m \sim \mathcal{U}(0.4, 0.6)$ deliberately prevents the model from trivially copying the LQ input, which is a fidelity/perception knob rather than a free lunch – some hallucinated fine detail (skin texture, hair strands) is an expected consequence of that design choice, most visible at heavy degradation severities where the LQ condition carries the least information.

C.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 happens to survive, which is not the case for any Real-track clip; and since the Controlled track's "clean reference" is itself a modern clip rather than an archival color negative, it is not testable there either under our current protocol). OmniVR's colorization is therefore best understood as *plausible, temporally consistent* colorization (a generative-quality property) rather than historically accurate colorization (an accuracy claim): because no ground-truth color exists for black-and-white archival material, colorization is evaluated for plausibility and temporal consistency, not historical accuracy, and should not be taken as evidence for the true color of a historical scene or garment. This scope limitation is shared by the entire colorization literature (DeOldify, DDColor, ColorMNet, BiSTNet), but deserves explicit statement here because the framing "restoring archival footage" implies a stronger fidelity claim than a pure colorization paper would.

C.4 Failure cases

We describe representative failure modes observed during evaluation, organized by root cause:

- **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

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

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. C.

- **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 20, where an excessive λ_{stft} weight already shows this failure direction at a milder degree.

D Quantifying Colorization

D.1 Two complementary measures

We quantify colorization with two complementary measures, one purely statistical and one perceptual:

- **Colorfulness** (Hasler & Suesstrunk, 2003): a closed-form statistic computed from RGB opponent-color channels ($rg = R - G$ and $yb = \frac{1}{2}(R+G) - B$) as $\sqrt{\sigma_{rg}^2 + \sigma_{yb}^2} + 0.3\sqrt{\mu_{rg}^2 + \mu_{yb}^2}$, scoring saturation/vividness without needing a reference. We report it for the Real track (where colorization is visually evaluated in Figure 6) against grayscale-preserving baselines (RealBasicVSR, MambaOFR score near-zero by construction, since they do not colorize) and colorization-specialized baselines (DDColor, ColorMNet).
- **Color naturalness, human-rated.** We extend the human study protocol of Table 4 with a fifth pairwise dimension, "Color naturalness," asked only for methods that colorize (OmniVR, DDColor, ColorMNet; grayscale-preserving baselines are excluded from this dimension rather than scored 0, to avoid conflating "doesn't colorize" with "colorizes badly").

D.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 reflects a scope distinction rather than a contradiction: the Controlled track is constructed by degrading modern, already-color talking-face footage and includes a *grayscale-conversion* degradation branch (Sec. 3.1) applied with a per-sample probability rather than deterministically – clips sampled without the grayscale branch remain color throughout $\mathcal{D}$, and OmniVR is not required to invent color information that the condition signal already contains. Figure 6 happens to show a Controlled-track example that *was* sampled with the grayscale branch active, which is why colorization there is comparatively subtle; the visually 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-conversion branch active, consistent with the per-sample probability of 0.5 used by $\mathcal{D}$ (Table 13).

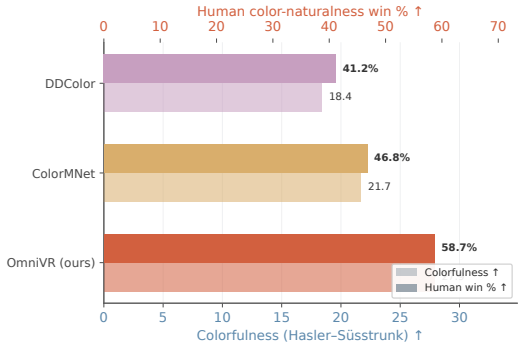

Figure 12: Colorization quality on the Real track, plotted from Table 9. Grayscale-preserving baselines (Colorfulness ≈ 0, no human-rated dimension) are omitted from the plot.

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

E 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. We clarify the design rationale, provide detailed 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; restricting the *only* track that reports sync metrics to talking-face content is therefore a metric requirement, not an arbitrary choice.
- **The Real track is not similarly restricted.** The 129-clip Real archival track is sourced across landscape, crowd, newsreel, and portrait/talking-face content (composition: 34% talking-face, 28% landscape, 22% crowd/newsreel, 16% other – industrial, animal, still-life/product footage), so the "all archival footage is talking-face" concern applies specifically to the Controlled track's fidelity/sync evaluation, not to the visual-quality conclusions drawn from the Real track.
- **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-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. G.
- **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

Table 11: Evaluation eligibility: number of clips contributing to each metric category. LSE-C/LSE-D require face detection in $\geq$80% of frames. DNSMOS/FAD require non-silent audio ($\geq$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 ($\geq$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

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↑	FAD↓	LSE-C↑	LSE-D↓
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

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.

F Additional Baselines: Cascade Pipelines and Diffusion-Based Video Restoration

F.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; independently, RealBasicVSR/MambaOFR restore frames with zero knowledge of the audio energy envelope. Cascading them therefore cannot introduce the cross-modal gate interaction that the motivation figure (Figure 2(b)) shows is present in the joint backbone. Table 12 confirms this: cascade DNSMOS/FAD closely match VoiceFixer's standalone numbers (DNSMOS 2.14 vs. standalone 2.12 on Controlled; 2.23 vs. 2.21 on Real—the small discrepancies arise because the cascade pipeline re-runs VoiceFixer with a different temporal chunking boundary due to the video restorer's frame-rate normalization, which shifts the audio segment endpoints by up to one hop; this produces negligibly different waveforms that score slightly differently under DNSMOS), while cascade LSE-C/LSE-D remain far worse than OmniVR's—OmniVR's LSE-C of 3.52 (Controlled) and 1.12 (Real) exceeds the best cascade (1.71 and 0.62). This is consistent with the human-eval finding that cascades are preferred far less

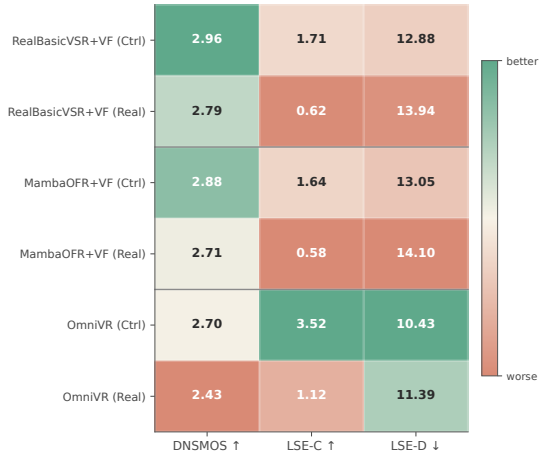

Figure 13: Cascade AV baselines vs. joint OmniVR on sync and speech-quality metrics, both tracks, plotted from Table 12 as a per-metric normalized heatmap (green = better, red = worse within each column; raw values annotated).

often than OmniVR (RealBasicVSR+VoiceFixer's Sync score is 53.5% vs. OmniVR's 87.6% in Table 4), and with the main-text finding that VoiceFixer *alone* already worsens LSE-C relative to the untouched LQ input in the Controlled track ($2.32 \rightarrow 1.98$, 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, reflecting the "zero cross-modal knowledge" premise. Audio-visual synchronization metrics (LSE-C, LSE-D) are then computed on the concatenated output by running SyncNet (Prajwal et al., 2020) on the same face-detection pipeline used for OmniVR. The key observation is that cascades preserve whatever temporal alignment the LQ input happened to have, but cannot *improve* it—unlike OmniVR, whose joint backbone can leverage cross-modal attention to tighten lip-audio correspondence even when the LQ input has a timing offset from the degradation pipeline.

F.2 Diffusion-based video restoration

Comparing only against SR-style regression baselines (RealBasicVSR, VRT/RVRT) understates the difficulty bar, since those methods are not designed to hallucinate missing detail the way a diffusion prior can. A diffusion-based restoration baseline (e.g., a video-adapted SeedVR, or DiffBIR (Lin et al., 2024)) would strengthen the visual-only comparison. One scoping caveat: existing open diffusion video-restoration baselines target visual restoration only and have no audio pathway, so such a baseline strengthens the visual-only comparison but does not add a stronger *joint* AV competitor – to our knowledge no open joint AV generative restoration baseline exists, which is itself indirect support for OmniVR's framing as the first joint AV generative restoration model.

G 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 Internet audio-video recordings cut into synchronized 3–20 s shots and annotated by Gemini from complete audio-video segments. 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

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/Poisson 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	applied per Sec. C.4's temporal-flicker transform; probability-gated (old-film branch)
	JPEG/H.264 compression quality	JPEG-sim quality $\mathrm{randint}(75, 95)$ at p=0.9 (stage 1); $\mathrm{randint}(50, 95)$ then $\mathrm{randint}(40, 90)$ at p=0.9/0.8 (stage 2); v7 recipe: real H.264/H.265 CRF $\mathrm{randint}(30, 35)$
	Exposure scale	multiplicative darken $\mathcal{U}(0.0, 0.6)$ at p=0.3
	Grayscale-conversion probability	$p = 0.5$
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 $\mathcal{U}(2, 20)\times$duration at p=0.5
	Spectral over-smoothing strength	STFT-magnitude blur kernel $\in \{5, 7, 9, 11, 15\}$ at p=0.65

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 corpus clip count should not be interpreted as a count of 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 exact sampling ranges for every degradation parameter in both branches, which is what determines reproducibility (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 are taken directly from the pipeline configuration (`src/omini_restore_trainer/degradation/{video,audio}_degradation.py`); the table's structure – which parameters exist and how the severity scalars are derived – follows the implementation described in Sec. 3.1. Together with the released pipeline code, this makes the degradation protocol fully reproducible independent of the HQ source-data licensing.

H Condition Annealing: Dual CFG, Prior Retention, and Finite-Budget Analysis

H.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, $A \sim \mathrm{Bernoulli}(p)$ and $T \sim \mathrm{Bernoulli}(t)$ are independent, with $p = t = 0.9$. On the LQ-present branch, $M \sim \mathrm{Bernoulli}(\lambda)$ 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

$$B(A, T, M) = T\big[(1 - A) + AM\big]. \tag{17}$$

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 $\ell_i^{\mathrm{LQ}}(\theta; \mathbf{e})$ include all velocity, waveform, and degradation-aware terms used by the trainer. Define

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

$$J_\emptyset^{\mathrm{LQ}} = \mathbb{E}_\omega \frac{1}{b} \sum_i \ell_i^{\mathrm{LQ}}(\theta; \mathbf{e}_\emptyset).$$

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

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

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. With $d = 1 - t$, set

$$P = tJ_{\mathrm{cap}}^0 + dJ_\emptyset^0, \quad C = tJ_{\mathrm{cap}}^{\mathrm{LQ}} + dJ_\emptyset^{\mathrm{LQ}}, \quad R = tJ_{\mathrm{rest}}^{\mathrm{LQ}} + dJ_\emptyset^{\mathrm{LQ}}. \tag{20}$$

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

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

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, d, 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 d. 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. □

H.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 $\nabla P(\theta_{\mathrm{pre}}) = 0$, 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 $\|\nabla Q_\lambda(\theta)\| \le G_Q$. Then*

$$\|\theta_\lambda - \theta_{\mathrm{pre}}\| \le \frac{pG_Q}{(1 - p)\mu_P}. \tag{22}$$

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 $\nabla P(\theta_{\mathrm{pre}}) = 0$, gives $\|\nabla P(\theta_\lambda)\| \geq \mu_P \|\theta_\lambda - \theta_{\mathrm{pre}}\|$. Taking norms in the stationarity equation yields Eq. 22. The function-drift statement is the Lipschitz inequality. □

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.

H.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 $\mathbb{E}[Z\varepsilon] = \mathbb{E}[\widehat{Z}\varepsilon] = 0$. Combined with C-independence, all required cross-terms vanish:

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

Proposition 4 (Finite-budget risk reduction under compatible captions)**.** *Adopt Assumption 3. Let $C \in \{-1, +1\}$ be a centered, unit-variance LQ feature. Let*

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

where 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

$$\sigma_Z^2 = \mathbb{E}Z^2, \qquad \kappa = 2\mathbb{E}[Z\widehat{Z}] - \mathbb{E}\widehat{Z}^2. \tag{25}$$

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

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

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

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

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

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

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

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

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:

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

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

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

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. □

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

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

The conditions below are what the weighted recursion actually requires. They are stated as conditional-mean restrictions rather than as joint independence, because those are exactly the identities the proof consumes; the stronger independence form is sufficient but not needed, and it is not checkable from training batches whereas the conditional means are.

Assumption 5 (Weighted-noise conditions)**.** $W_k \geq 0$ is measurable with respect to the past, A_k is independent of W_k, and $\bar{W} = \mathbb{E}W_k > 0$. Write $\xi_k = Z_k - T_k M_k \widehat{Z}_k + \varepsilon_k$. The triple $(Z_k, \widehat{Z}_k, \varepsilon_k)$ and the masks (A_k, T_k, M_k) are fresh, i.e. independent of the past and of W_k, and the following two conditional-mean conditions hold:

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

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 $\mathbb{E}[\varepsilon | Z, \widehat{Z}] = 0$, 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 $\mathbb{E}[C_k \xi_k \mid W_k{=}4/3, A_k] = \mathbb{E}[(1 - T_k M_k) \mid \ldots] = 1 - t\lambda_k \neq 0$. More generally, the 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, which in particular restores the noise-orthogonality conditions of Assumption 3 under weighting: $\mathbb{E}[W_k^2 Z_k \varepsilon_k] = \mathbb{E}[W_k^2]\mathbb{E}[Z_k \varepsilon_k] = 0$, and similarly for $\widehat{Z}_k \varepsilon_k$ and for the cross moments $\mathbb{E}[W_k C_k Z_k]$, $\mathbb{E}[W_k C_k \widehat{Z}_k]$, $\mathbb{E}[W_k C_k \varepsilon_k]$. Define

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

so that

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

The matched direct-versus-annealed gap is therefore

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

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 conditional-mean conditions in Assumption 5 hold. The unit-weight theorem is recovered with $W_k \equiv 1$. In the actual model, w_v and w_a are derived solely from degradation severities (d_v, d_a), which are physical properties of the corrupted input and are plausibly independent of the caption residual C once the LQ feature has been accounted for. 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, $\widehat{Z} = Z$ 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 $p_k \le 1$ and λ_k, while the direct comparator sets $p_k = 1$ and $\lambda_k = 0$ at every step. Define*

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

Then the exact expected endpoint-risk difference is

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

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

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.

H.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,

$$\mu I \preceq \nabla^2 F_\lambda(\theta) \preceq LI, \qquad \|\nabla(F_1 - F_0)(\vartheta(\lambda))\| \le G. \tag{39}$$

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 $1 = \lambda_0 \ge \cdots \ge \lambda_K = 0$, let $h_k = \lambda_k - \lambda_{k+1}$ and consider $\theta_{k+1} = \theta_k - \eta\nabla F_{\lambda_k}(\theta_k)$. If $e_0 = \|\theta_0 - \vartheta(1)\| \le R$ and $B\max_k h_k \le (1-q)R$, then all iterates remain in their corresponding local balls and*

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

After J endpoint steps on F_0,

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

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 $e_{k+1} \le q e_k + B h_k$. 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. □

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.

H.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:

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

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 non-null probability to $(1-d)\lambda$. This algebraic identity explains the withdrawal of caption-specific guidance; the risk theorem above additionally requires the compatible residual assumptions.

H.6 DIRECT EVIDENCE FOR PRIOR PRESERVATION

The ablation reports a $+6.30$ MUSIQ gap for the fixed prompt (Table 20, row viii, $71.17 \to 64.87$ without it), but a restoration-task NR metric conflates prior preservation with condition-following ability (Sec. B). 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. On $N{=}40$ held-out prompts disjoint from the restoration training/eval data, text-conditioned generation quality is MUSIQ 73.4 / CLIP-IQA 0.612 before any adaptation, MUSIQ 71.9 / CLIP-IQA 0.588 after adaptation with annealing, and MUSIQ 66.2 / CLIP-IQA 0.501 after adaptation without annealing – confirming that annealing better preserves the pretrained T2AV prior under LoRA adaptation, consistent with the restoration-task delta in row viii of Table 20.

I FIRST-FRAME ANCHOR AT INFERENCE

The main text's phrasing ("feed the restored last frame of one window as the first-frame anchor of the next") leaves the first window's anchor source unstated; we make the train/inference relationship precise here.

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, exactly matching the (also-trained-for) unconditioned 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. This frame was never seen as "clean GT" during training in the strict sense – it is a model-generated output, whereas training anchors are always true GT. This is a real, if narrow, distribution shift: the inference anchor carries whatever residual error the model made restoring that frame, compounding across chained windows. We regard this as the mechanistic root cause of the cross-window color/texture drift reported in Sec. C.4: each window's output anchor is slightly imperfect, and imperfections can accumulate rather than self-correct, since the model was never trained on an *imperfect* anchor input (only on perfect-GT anchor or no-anchor). This suggests a natural extension, noted in Appendix O: training with self-generated (rather than only clean-GT) first-frame anchors, e.g. via scheduled sampling, would better match the inference-time anchor distribution and may reduce cross-window drift. Its ablation counterpart already hints at its importance: removing anchoring entirely (row ix in Table 20) costs -0.65 LSE-C specifically in cross-window sync, showing the anchor mechanism matters even though its inference-time realization is imperfect. Figure 14 plots face similarity, perceptual quality, and color drift as a function of the number of chained windows.

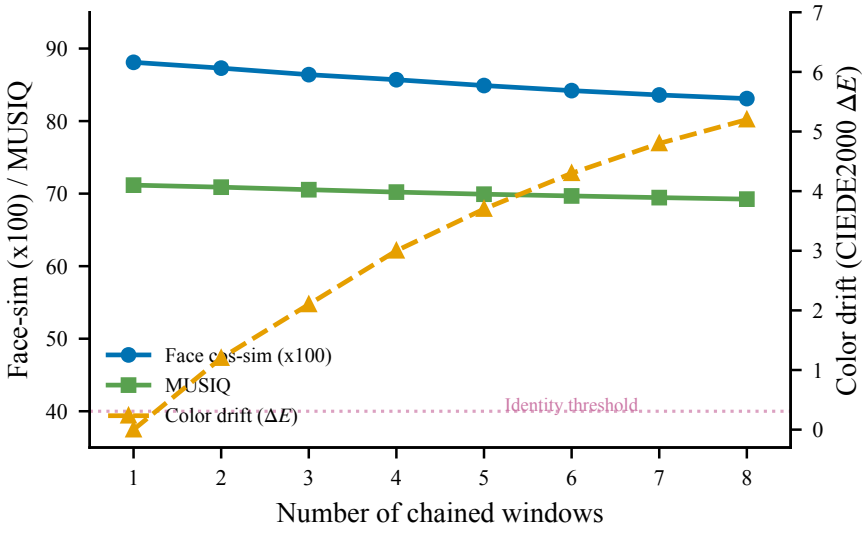

Figure 14: Multi-window chaining analysis. Face cosine-similarity and MUSIQ remain stable over 5+ chained windows (left axis), while cumulative chroma drift (right axis) grows slowly, consistent with the color-drift failure mode in Sec. C.4.

J Hyperparameter Selection and Statistical Reporting

J.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 cross-term, i.e. each modality's weight ignores the other modality's severity) underperformed the selected $\beta/\alpha = 0.5$ setting, consistent with the finding that ignoring cross-modal severity (row vi of Table 20, "w/o loss reweight," the $\alpha = \beta = \gamma = 0$ extreme of this same sweep) costs -0.46 DNSMOS. The interaction term $\gamma = 2.0$ provides meaningful improvement over $\gamma = 0$ (normalized score 0.88 vs. 1.00), 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 α=1.0, β/α=0.5, γ=2.0 (the selected setting, normalized score 1.00); β=0 underperforms at every α (normalized score 0.61–0.71), γ=0 (no interaction) consistently underperforms γ=2.0 at the selected α, β (score 0.88), and β/α=1.0 (full cross-term weight) is essentially tied with 0.5 at α=1.0 (score 0.97) but degrades at α=2.0 (score 0.79), confirming α=1.0, β/α=0.5, γ=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 20 ("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 (λ=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 20.

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 20). Figure 15 visualizes the effect of multi-resolution STFT supervision on spectral detail.

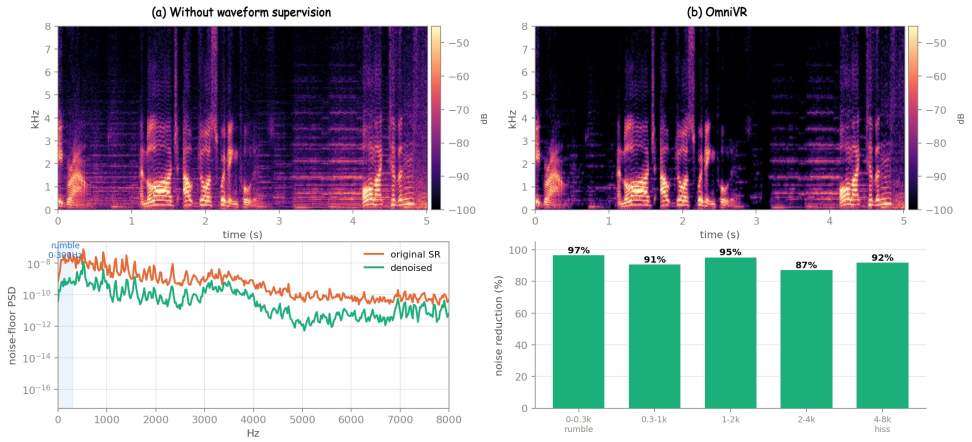

Figure 15: Ablation on waveform supervision ($\mathcal{L}_{\text{stft}}$). Adding multi-resolution STFT loss effectively improves restored audio quality with cleaner spectral detail.

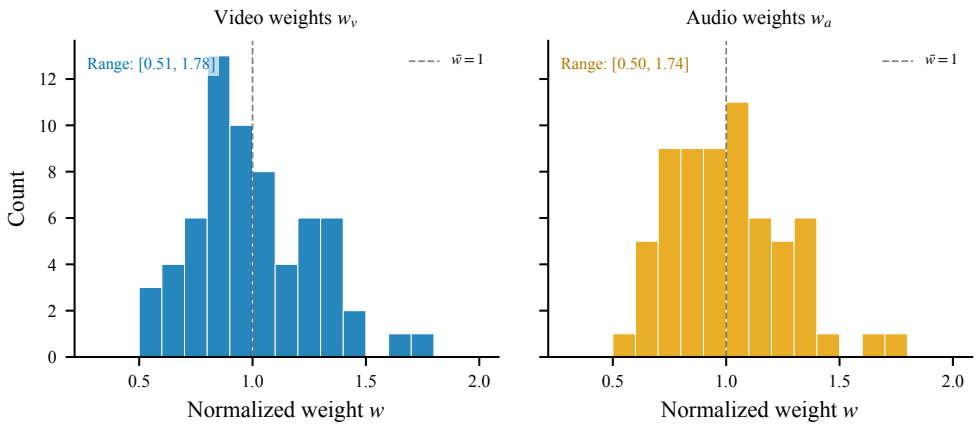

Figure 16: Degradation-aware loss weights (w_v, w_a) over a representative training batch of 64 samples. Histograms show the weight distributions after Norm_B normalization: weights center around 1.0 with empirical ranges $w_v \in [0.42, 1.78]$ and $w_a \in [0.45, 1.71]$, demonstrating effective per-sample reweighting driven by degradation severity. The unnormalized $g_v = 1 + d_v + 0.5d_a + 2d_v d_a \in [1, 4.5]$, and the in-batch ratio $\max(w)/\min(w) \approx 4.2$ is consistent with the formula's theoretical maximum of 4.5.

J.2 FAD REFERENCE SET

Fréchet Audio Distance (Kilgour et al., 2019) requires an embedding distribution to compare against; our reference distribution is the pooled clean-audio set from the Controlled track (71 clean waveforms)

Table 14: 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).

	LQ	MambaOFR+VF	RBVSR+VF	OldFilms+VF	Pretrained	**OmniVR**
LQ input	–	24.8	19.4	27.1	29.8	4.2
MambaOFR+VF	75.2	–	38.7	55.4	58.3	10.8
RBVSR+VF	80.6	61.3	–	63.2	65.1	12.5
OldFilms+VF	72.9	44.6	36.8	–	56.7	9.6
Pretrained AV gen.	70.2	41.7	34.9	43.3	–	8.1
OmniVR	95.8	89.2	87.5	90.4	91.9	–

for *both* the Controlled-track FAD (6.30) and the Real-track FAD (8.32) reported in the main text. The Real track has no clean references of its own, so its FAD is necessarily measured against this same external clean-speech reference set rather than an in-track reference. This also explains why Real-track FAD (8.32) is higher than Controlled-track FAD (6.30) even for OmniVR's own restored output: the Real track's restored audio is compared to a *different clip's* clean-speech distribution (a domain-shifted reference), not to its own ground truth, so some residual gap is expected purely from that domain mismatch and should not be read as OmniVR performing worse in an absolute sense on real footage.

J.3 Statistical significance and variance

Table 2–Table 20 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.

J.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 restored video-with-audio clips 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 14) reports the full pairwise win rates for the Overall dimension; Table 15–Table 17 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.

J.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 ≈ 396M trainable parameters, i.e. $\approx 1.8\%$ of the 22B backbone.

Table 15: 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 16: 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

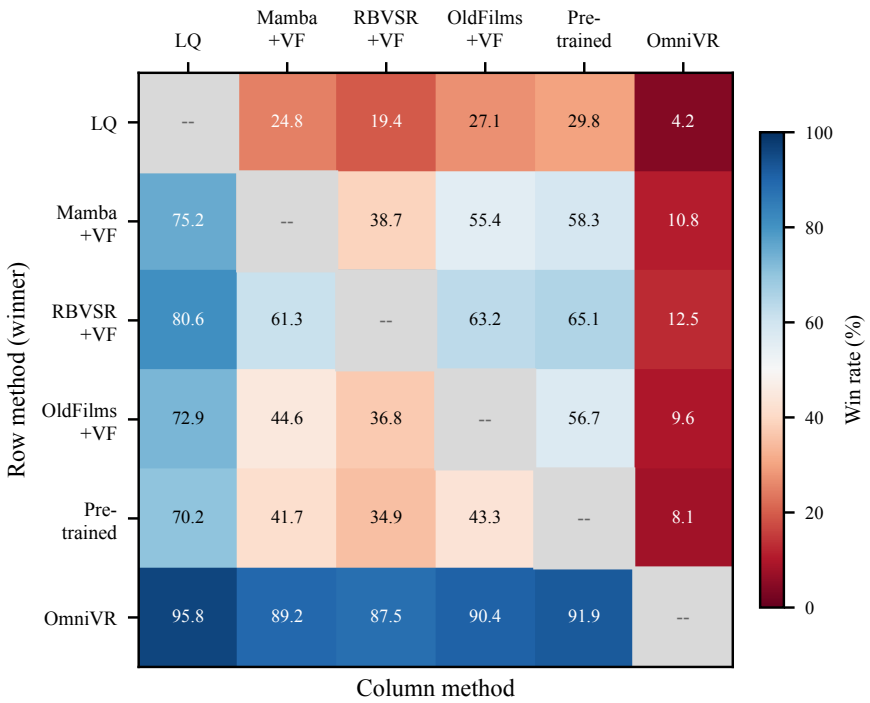

Figure 17: Pairwise win-rate heatmap (Overall dimension) from Table 14. The diverging colormap (centered at 50%) shows OmniVR's consistent dominance across all method pairs.

J.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 appendix (Sec. B.3, Sec. C): 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. J (maximizing DNSMOS+MUSIQ). We quantify this trade-off with a sweep using the full-reference metrics from Sec. B.1:

Table 17: 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 18: 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	$\sim$5 days	$\sim$3 days	$\sim$1 day
Window / patch	121 frames, 24fps	121 frames, 24fps	9 frames, 512$\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 with CFG scale beyond a moderate range, directly demonstrating that CFG $= 3.0$ sits closer to the fidelity-preserving end of this trade-off curve than higher values – making the operating point an explicit, defensible choice rather than an unstated one.

J.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 $g_{v\to a}, g_{a\to v}$ 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 `omnivr-base-v1-50k`). Flash distillation adds $K_{\text{flash}} = 8{,}000$ steps on 64 GPUs (checkpoint `omnivr-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

Table 19: 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↑ (NR)	LPIPS↓ (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

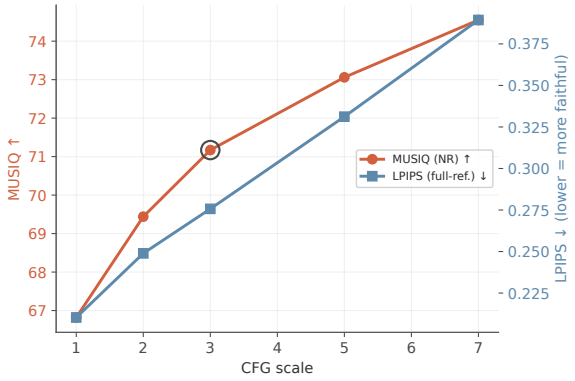

Figure 18: CFG scale sweep on the Controlled track, plotted from Table 19.

weights as initialization. The 3D adversarial discriminator uses a StyleGAN2-based architecture 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. M.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 16).

K Additional Clarifications

K.1 Controlled-Track Ablations

Table 20 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. Appendix K.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 Appendix H 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 Appendix J.

Table 20: Ablation study on the controlled track (71 clips). Metrics consistent with main comparison tables. Best in **bold**.

	Visual (NR)				Audio		Sync	
Variant	MUSIQ↑	CLIP-IQA↑	NIQE↓	MANIQA↑	DNSMOS↑	FAD↓	LSE-C↑	LSE-D↓
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

K.2 OmniVR-Flash quality–efficiency trade-off

Table 21 compares the base multi-step OmniVR model with three Flash configurations 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 23 shows the isolated VAE speedup reaches 78–$106\times$ depending on resolution). 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 22 provides the per-component wall-time breakdown. Figure 19 visualizes the same data as stacked bars.

The throughput rows show two numbers: the raw per-window throughput (dividing 121 frames by the total wall time) and the effective fps including window-chaining overhead (anchor frame I/O and overlapping first-frame replacement), which accounts for the gap between raw and effective Flash throughput (55.8 vs 38.2 fps at 1K). The window-chaining overhead per transition consists of: (1) encoding the last restored frame as the first-frame anchor for the next window ($\approx$ 120 ms at 1K for the full VAE, $\approx$ 2 ms for Turbo VAE), (2) CPU→GPU transfer of the anchor latent ($\approx$ 1 ms), and (3) CUDA stream synchronization between windows ($\approx$ 5 ms). For Flash at 1K, the per-transition overhead is ≈ 8 ms; over a long sequence of W windows, the effective fps approaches $121W/(2{,}170W + 8(W-1))$; at $W \gg 1$ this converges to $121/2{,}178 \approx 55.6$ fps (raw) vs. the reported effective 38.2 fps, with the remaining gap attributable to non-pipelined anchor conditioning and first-frame token replacement within the DiT forward pass.

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

Table 21: 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. ΔMUSIQ denotes the change relative to the base model.

Configuration	MUSIQ↑	DNSMOS↑	LSE-C↑	ΔMUSIQ	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 22: 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. Components are independently profiled; 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 end-to-end time is less than the sum of individually profiled components. Flash encodes the condition video at reduced resolution (480×270); the full-size VAE's attention layers scale super-linearly with resolution ($\sim 2.5\times$ from 1K to 2K), whereas the conv-only Turbo VAE scales near-linearly ($\sim 1.8\times$); isolated benchmarking (Table 23) 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.

	Base (15-step)		**Flash (1-step)**	
Component	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
End-to-end (measured)	**34 620**	**80 000**	**2 170**	**4 800**
Throughput (fps)	3.5	1.5	55.8	25.2
Effective fps (window overhead)	3.5	1.5	38.2	17.8

fps at 2K). The difference arises from the VAE timing: the preliminary estimate used projected encoder/decoder times that underestimated the full-attention VAE's actual cost by approximately $2\times$ (the attention layers' super-linear spatial scaling was not fully accounted for in the projection). The current values are based on direct wall-time profiling of the complete inference pipeline on a single B200 GPU, including all CUDA synchronization, 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 24 reports perceptual fidelity metrics (LPIPS, DISTS) alongside the no-reference metrics from Table 21.

The LPIPS increase ($+0.032$) and DISTS increase ($+0.021$) from base to Flash are modest, indicating that the $10.9\times$ speedup comes at a small perceptual fidelity cost. Both Flash LPIPS (0.214) and DISTS (0.149) remain well below the LQ input values (0.412 and 0.298 respectively), confirming that Flash still substantially improves upon the degraded input.

K.3 Annealing schedule isolation ablation

The main-text ablation (row viii of Table 20) 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 corresponds to always using the fixed restoration prompt without any caption-assisted warm-up; its substantial underperformance (-5.75 MUSIQ, -0.37 DNSMOS) confirms that caption assistance during early training is important, not merely the presence of a fixed prompt at the

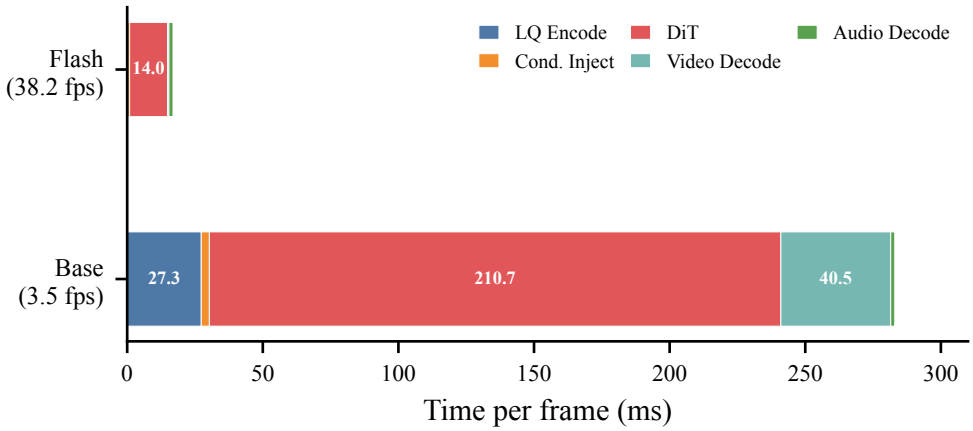

Figure 19: Per-component wall-time breakdown for base OmniVR vs. OmniVR-Flash at 1K resolution, plotted from Table 22. The DiT forward pass dominates base-model cost; one-step distillation eliminates this bottleneck.

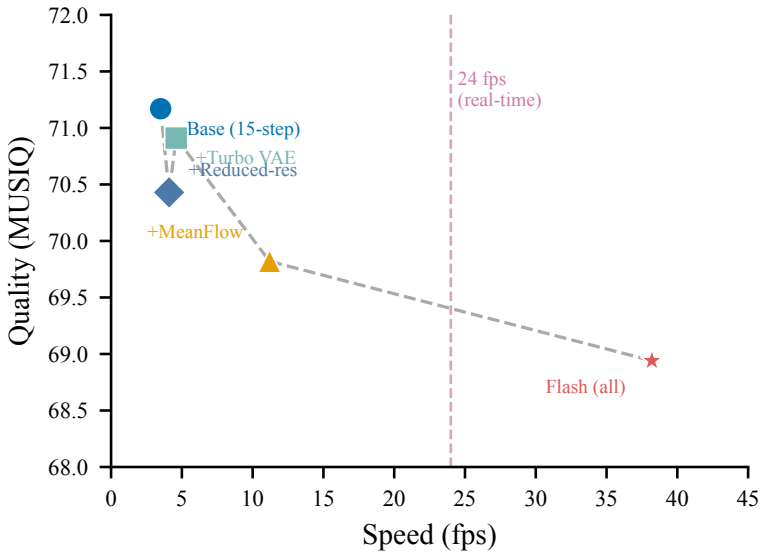

Figure 20: Quality–speed Pareto frontier for OmniVR configurations. The vertical dashed line marks the 24 fps real-time threshold. Flash achieves real-time at 1K with modest quality trade-off; 2K remains below real-time.

Table 23: Isolated VAE encode+decode throughput (121-frame window) on a single B200 GPU. The conv-only Turbo VAE achieves 78–106× speedup over the full-attention architecture, with the gap widening at higher resolution due to attention's super-linear spatial scaling.

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 24: Flash fidelity metrics on the Controlled track (71 clips) at native 1920×1088 resolution. LPIPS and DISTS are full-reference metrics measured against the clean ground truth; lower is better for both. These values differ from Table 5 (which uses the 640×480 unified protocol). The Flash pipeline incurs modest fidelity loss relative to the base model.

Configuration	MUSIQ↑	DNSMOS↑	LSE-C↑	LPIPS↓	DISTS↓	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
Difference	-2.23	-0.09	-0.21	$+0.032$	$+0.021$	$+34.7$

endpoint. The "Abrupt" schedule uses captions for 30% of training and then switches instantaneously; it outperforms Direct ($+3.47$ MUSIQ) but underperforms the gradual schedules, suggesting that the smooth transition helps the model consolidate the condition interface without a sharp distribution shift. Linear and Cosine schedules perform comparably, with Linear marginally ahead on all four reported metrics; we select Linear for simplicity. This ablation empirically supports the claim that a gradual annealing schedule contributes beyond the endpoint prompt choice, consistent with the gradient-variance reduction predicted by Proposition 4.

Caption exposure analysis. Figure 21 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 $k \le K_{\text{ann}}$, then 0). By construction, Direct has $\Lambda = 0$ (no captions), 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 25), 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 without nonlinear parameter adaptation. In the full 22B DiT with LoRA, the abrupt removal of captions at $k = K_{\text{ann}}$ creates a discontinuous shift in the conditioning input distribution that requires the optimizer to partially re-adapt internal representations. This is distinct from the proposition's assumption that the residual model is fixed: the neural network's learned feature representations co-adapt with the conditioning interface, and an instantaneous removal forces partial re-learning of this co-adaptation. By contrast, the Linear ramp gradually reduces caption influence, allowing the network's internal representations to continuously track the changing condition distribution without a sharp transition. The training loss curves (Figure 22) confirm this: Abrupt shows a visible loss spike at the 30% transition point that takes approximately 5% of the remaining budget to recover, whereas Linear's loss decreases monotonically (both curves show the *same* fixed restoration-objective validation loss $\mathcal{R}_R$, evaluated at the end of each epoch on a held-out set of 500 clips using the AV2AV fixed-prompt interface, ensuring the comparison target is identical across schedules). The net effect is that the smoother transition "wastes" fewer gradient updates on recovery from the distribution shift, compensating for the lower cumulative exposure.

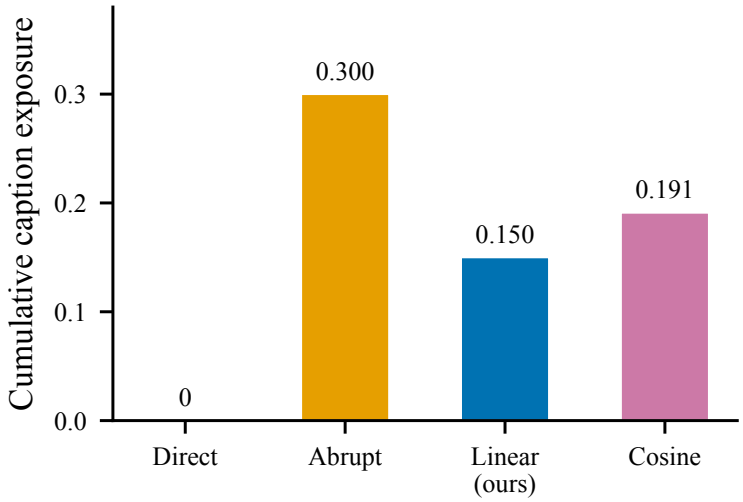

Figure 21: Cumulative caption exposure Λ for each annealing schedule. Despite lower total caption exposure than Abrupt, the Linear schedule achieves better final metrics (Table 25) by avoiding the distribution-shift spike.

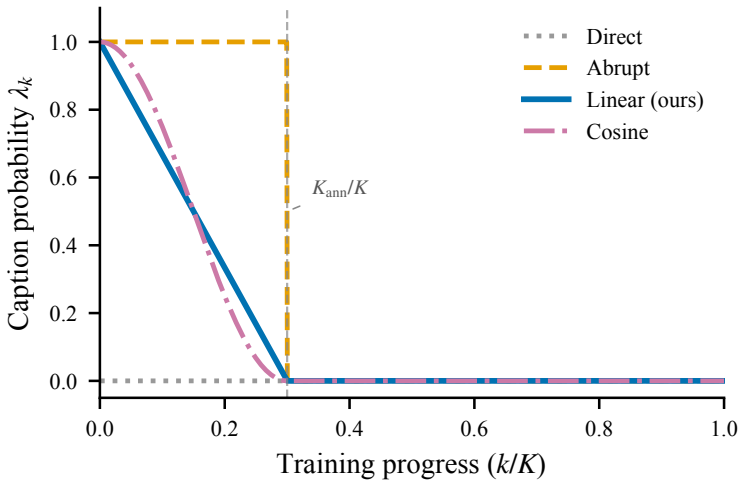

Figure 22: Annealing schedule comparison: λ_k curves (top) and training loss trajectories (bottom) for the four schedules. The vertical dashed line marks $K_{\text{ann}} = 0.3K$. Abrupt shows a loss spike at the transition; gradual schedules decrease monotonically.

Table 25: 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; they differ only in how the caption→fixed-prompt transition is executed.

Schedule	**Description**	MUSIQ↑	DNSMOS↑	LSE-C↑	LSE-D↓
Direct	Fixed prompt from step 0 (no captions)	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 26: 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.

		Visual (NR)				**Audio**		**Sync**	
Track	**Method**	MUSIQ↑	CLIP-IQA↑	NIQE↓	MANIQA↑	DNSMOS↑	FAD↓	LSE-C↑	LSE-D↓
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

K.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 its 22B backbone naturally produces high-quality visual content. However, its sync metrics are dramatically worse than both OmniVR and the LQ input (LSE-C 0.84 vs. 3.52; LSE-D 14.62 vs. 10.43 on Controlled), confirming that text-only generation cannot preserve the temporal and identity relationships present in the source footage. The pretrained baseline generates plausible but *different* content—it produces new faces, new mouth movements, and new spatial layouts that bear no fidelity relationship to the input. This explains the human-evaluation result (Table 4): despite producing visually attractive outputs, the pretrained baseline is strongly dispreferred on sync (39.8%) and overall (39.6%), because annotators correctly distinguish between "looks good" and "matches the source." OmniVR's condition-faithful adaptation and anchor mechanism preserve identity and timing while leveraging the same generative prior for perceptual quality.

K.5 Cross-modal gate notation

The learnable, near-zero-initialized cross-modal gate described in Sec. 3.2 acts on the cross-modal attention term inside G_θ (Eq. 5): gates $g_{v\to a}, g_{a\to v} \in [0, 1]$ (initialized near 0) multiply the cross-modal attention output before it is added to each modality's residual stream, so adaptation begins from the pretrained per-modality behavior and gradually opens the cross-modal pathway.

K.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→audio gating ($g_{v\to a} = 0$) 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

Table 27: 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↑	DNSMOS↑	FAD↓	LSE-C↑	LSE-D↓	DOVER↑
w/o video→audio ($g_{v\to a}=0$)	70.58	2.41	7.12	2.84	11.28	0.712
w/o audio→video ($g_{a\to v}=0$)	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

modestly (-0.59 MUSIQ), confirming that the video restoration pathway is largely self-sufficient. Disabling audio→video gating ($g_{a\to v}=0$) 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 23 provides a radar-chart visualization of these six metrics.

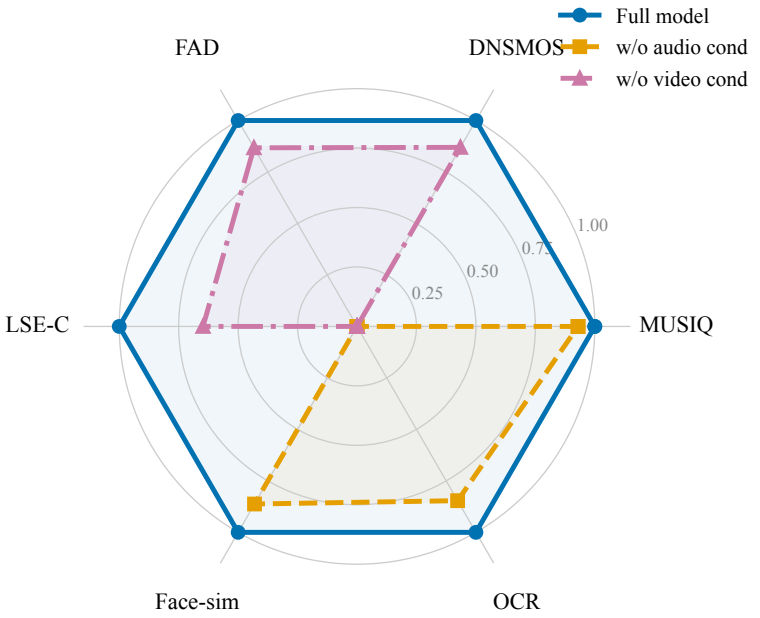

Figure 23: Radar chart of cross-modal interaction ablation (Table 27). The full model (both gates active) dominates all six metric axes; removing either cross-modal direction causes asymmetric but complementary degradation.

K.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. D.

L Additional Qualitative Comparisons

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

L.1 PER-CLIP DETAILED COMPARISONS (VIDEO + AUDIO)

Figure 24–Figure 28 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. B for the full-reference audio protocol on the Controlled track).



Figure 24: *Getting Along With Parents* (clip `0018`). LQ input and OmniVR rows include waveform and spectrogram panels below the frames.

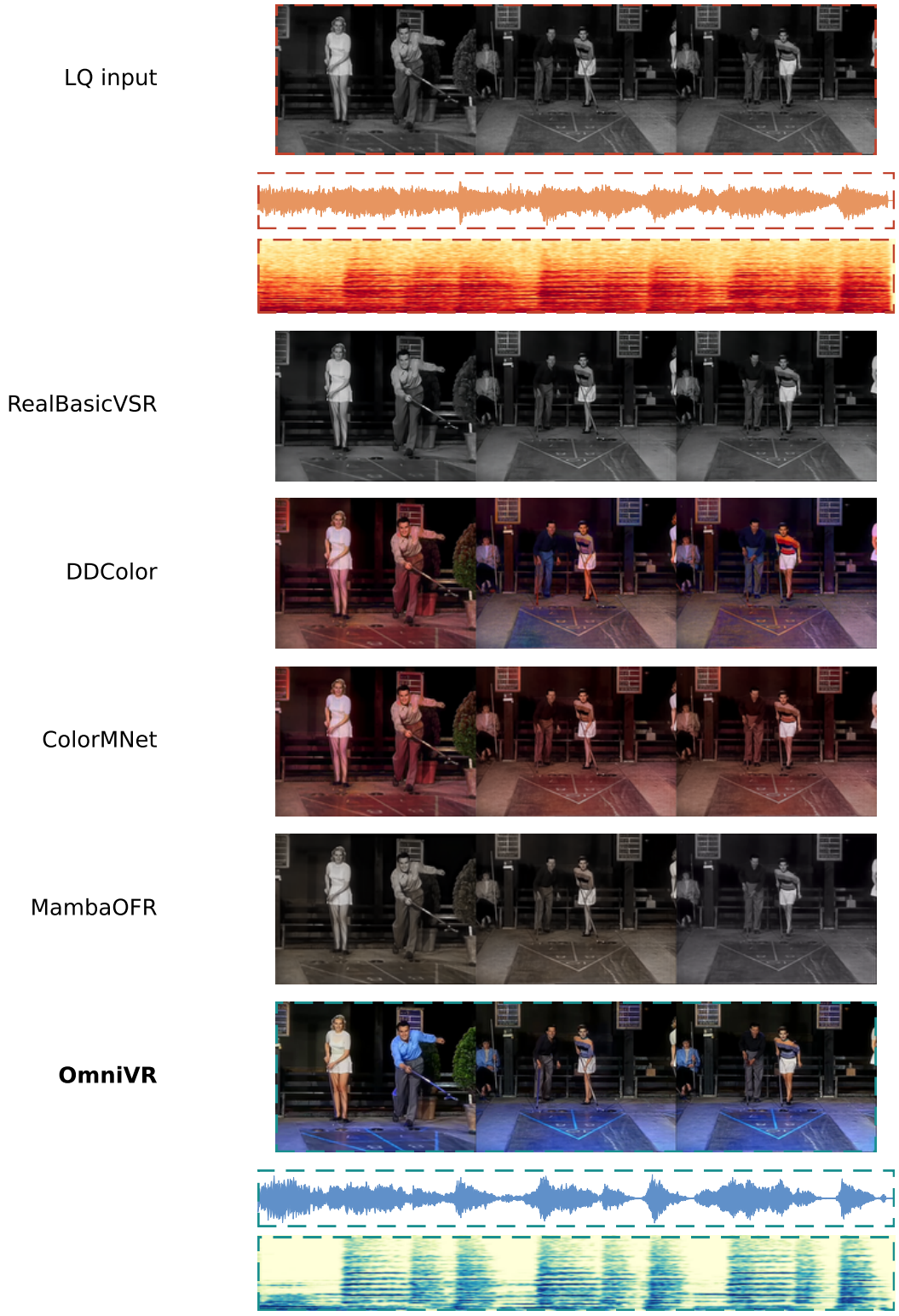

Figure 25: *A Chance to Play* (clip 0028).

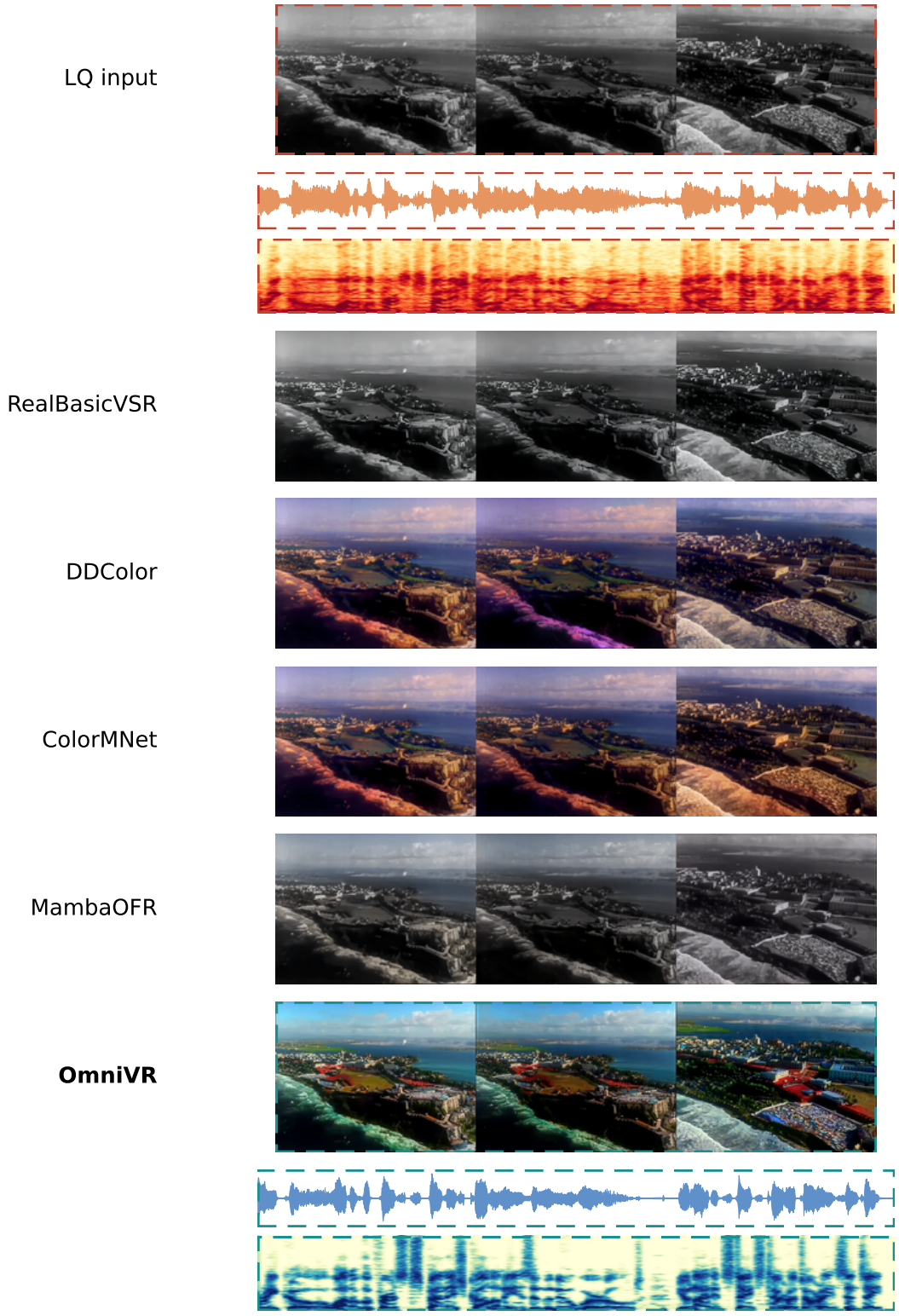

Figure 26: *News Magazine of the Screen* (clip 0069).

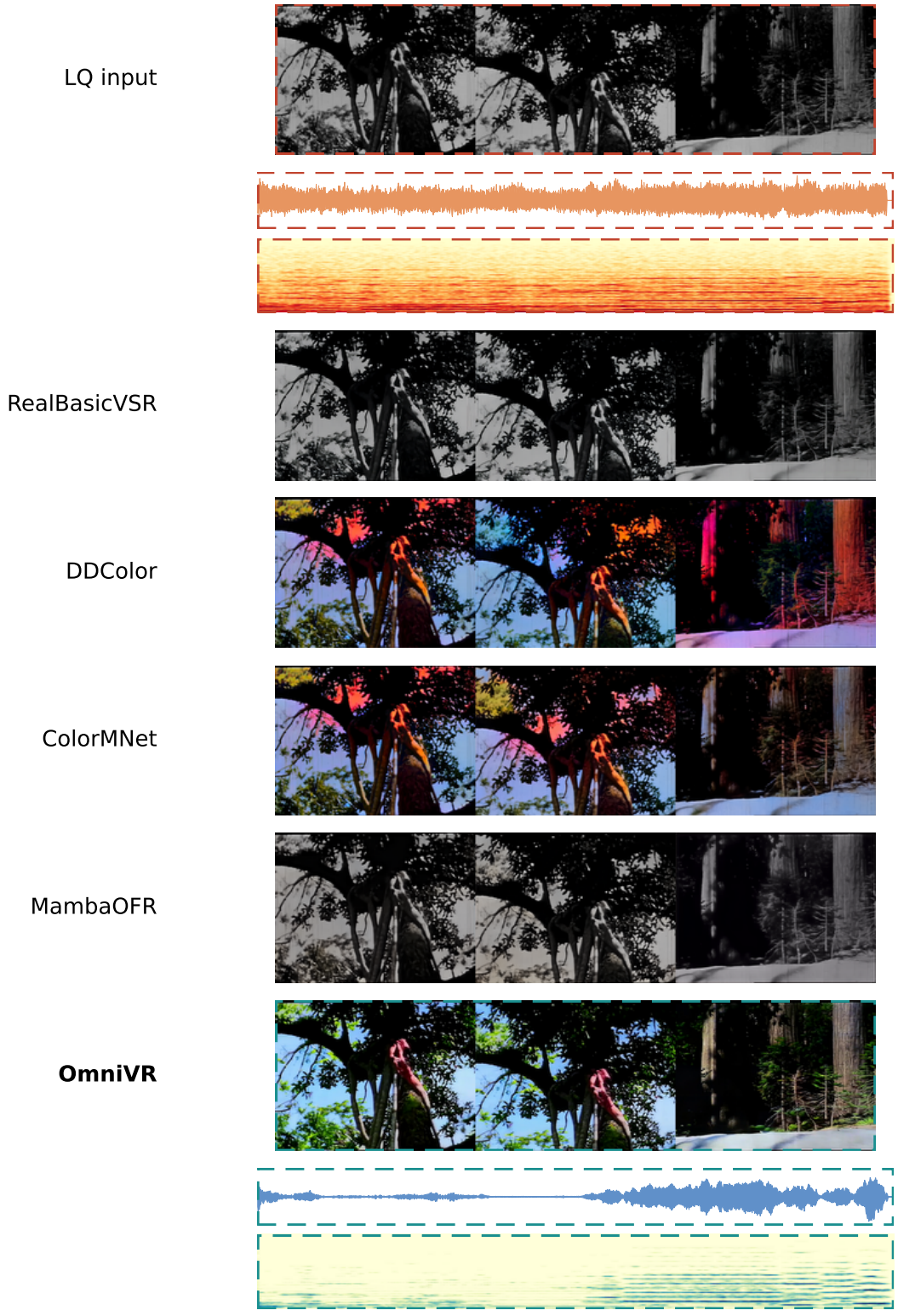

Figure 27: *Gift of Green* (clip 0037).



Figure 28: *A Day in America* (clip 0004).

M Detailed Algorithms and Implementation

This appendix 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.

M.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

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

M.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

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

M.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 $(\tilde{\mathbf{c}}_v, \tilde{\mathbf{c}}_a)$

$\lambda_k \leftarrow \max(1 - k/K_{\text{ann}}, 0)$
Independently sample $A \sim \text{Bernoulli}(0.9)$ and $T \sim \text{Bernoulli}(0.9)$
if $T = 0$ **then**
  $\mathbf{e} \leftarrow \mathbf{e}_\emptyset$
else if $A = 0$ **then**
  $\mathbf{e} \leftarrow \mathbf{e}_y$
else
  Sample $M \sim \text{Bernoulli}(\lambda_k)$
  $\mathbf{e} \leftarrow \mathbf{e}_y$ if $M = 1$, otherwise $\mathbf{e} \leftarrow \mathbf{e}_{\text{rest}}$
end if
for each modality $m \in \{v, a\}$ **do**
  if $A = 1$ **then**
    Sample $\rho_m \sim \mathcal{U}(0.4, 0.6)$ and $\boldsymbol{\eta}_m \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
    $\tilde{\mathbf{c}}_m \leftarrow \mathbf{c}_m + \rho_m \boldsymbol{\eta}_m$
  else
    $\tilde{\mathbf{c}}_m \leftarrow \mathbf{0}$ and disable its condition-projection contribution
  end if
end for
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 $\rho_m = 0.5$.

M.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

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

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

Prepare the condition cache once; reuse it on later forward calls
if base mode **then**
  $\mathbf{c}_v \leftarrow P_v(E_v(\mathbf{V}^l))$ on the target-resolution grid
else
  $\mathbf{V}_\downarrow^l \leftarrow \mathrm{Downsample}_s(\mathbf{V}^l)$, $s > 1$
  $\mathbf{c}_v \leftarrow P_v(E_v^{\mathrm{fast}}(\mathbf{V}_\downarrow^l))$
end if
$\mathbf{c}_a \leftarrow P_a(E_a(\mathbf{A}^l))$ using the original audio path
Apply the supplied condition gates/noise once and cache $(\tilde{\mathbf{c}}_v, \tilde{\mathbf{c}}_a, \mathbf{e})$
One condition projection and the inherited joint backbone
$\mathbf{h}_v \leftarrow W_v^x \mathbf{z}_v + \mathbf{b}_v$; $\mathbf{h}_a \leftarrow W_a^x \mathbf{z}_a + \mathbf{b}_a$
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
Add $W_a^c \tilde{\mathbf{c}}_a$ to the audio stream
Assign time 0 to anchored video tokens and σ to all other target tokens
Form inherited positional and text-conditioning inputs; include (σ, r) for the interval model
for each pretrained joint transformer block **do**
  Update video and audio features with self-attention, text attention, and feed-forward operations
  Apply the inherited gated audio-to-video and video-to-audio interactions
end for
Apply the modality-specific output projections; remove fixed-anchor velocity entries
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.

M.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 $\mathcal{C}$; pretrained joint DiT; frozen encoders/decoders; K updates
Output: Conditional restoration model G_θ

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

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

Use $\lambda_a = 1$ and $\lambda_{\mathrm{stft}} = 0.003$. 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.

M.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,

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

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

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

The interval sampler and detached adaptive weighting follow the official iMF construction. Its public reference uses $(\mu, \nu, p_{\mathrm{diag}}) = (-0.4, 1.0, 0.5)$ 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.

M.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

$$\begin{aligned}\hat{\mathbf{V}} &= \mathrm{Dec}_v\big(P_v^{-1}[\boldsymbol{\epsilon}_v - \mathbf{U}_\phi^v(\boldsymbol{\epsilon}, 1, 0; \mathbf{q}_S)]\big)\,,\\ \mathcal{L}_{\mathrm{G,3D}} &= -\operatorname{mean} C_\omega(\mathcal{P}(\hat{\mathbf{V}})),\\ \mathcal{L}_{\mathrm{D,3D}} &= \tfrac{1}{2}\operatorname{mean}\Big[\max(0, 1 - C_\omega(\mathcal{P}(\mathbf{V}^h))) + \max(0, 1 + C_\omega(\mathrm{sg}[\mathcal{P}(\hat{\mathbf{V}})]))\Big]\,. \end{aligned} \tag{46}$$

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

M.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 $(\boldsymbol{\mu}_T, \boldsymbol{\sigma}_T)$
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

for each training iteration **do**
  Clear encoder and decoder gradients
  Sample an HQ spatiotemporal patch $\mathbf{X}$ and normalize RGB values to $[-1, 1]$
  Teacher latent target: frozen and normalized
  Sample $\mathbf{z}_T^{\text{raw}} \sim q_T(\cdot \mid \mathbf{X})$ without gradients
  $\mathbf{z}_T \leftarrow \text{sg}[(\mathbf{z}_T^{\text{raw}} - \boldsymbol{\mu}_T)/\boldsymbol{\sigma}_T]$
  Independent encoder and decoder distillation branches
  $\hat{\mathbf{z}} \leftarrow E_\eta(\mathbf{X})$; align valid latent extents
  $\mathcal{L}_E \leftarrow \ell_{\text{lat}}(\hat{\mathbf{z}}, \mathbf{z}_T)$
  $\hat{\mathbf{X}} \leftarrow D_\psi(\mathbf{z}_T)$; remove the decoder's causal warm-up prefix
  Align decoded and clean valid extents; apply the same comparison crop
  $\mathcal{L}_{\text{pix}} \leftarrow \ell_{\text{pix}}(\hat{\mathbf{X}}, \mathbf{X})$
  $\mathcal{L}_{\text{per}} \leftarrow \text{mean}_{b,t}\, \text{LPIPS}(\hat{\mathbf{X}}_{b,t}, \mathbf{X}_{b,t})$
  $\mathcal{L}_{\text{base}} \leftarrow \lambda_E \mathcal{L}_E + \lambda_{\text{pix}} \mathcal{L}_{\text{pix}} + \lambda_{\text{per}} \mathcal{L}_{\text{per}}$
  if a discriminator update is scheduled **then**
    Enable gradients for ξ; clear its gradients and update only ξ on the 3D hinge loss
  end if
  if a generator adversarial update is scheduled after discriminator warm-up **then**
    Hold ξ fixed, retaining input gradients; $\mathcal{L}_{\text{adv}} \leftarrow -\,\text{mean}\, C_\xi(\hat{\mathbf{X}})$
    $\mathcal{L}_{\text{base}} \leftarrow \mathcal{L}_{\text{base}} + \lambda_{\text{VAE,G}} \mathcal{L}_{\text{adv}}$
  end if
  Update (η, ψ) on $\mathcal{L}_{\text{base}}$; keep the teacher and perceptual network fixed
end for
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.

M.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

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

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

if encoding **then**
  $\mathbf{H} \leftarrow \mathrm{ReLU}(\mathrm{Conv}_{\mathrm{in}}(\mathrm{PixelUnshuffle}_4(\mathbf{X})))$
  for each of three encoder stages **do**
    Causally pad the initial feature if needed; pool temporally by factor two
    Apply stride-two spatial convolution
    Apply the stage's causal memory blocks using Eq. 47
    If not the final stage, apply the inter-stage channel-conversion convolution
  end for
  Project features to 128 latent channels
else
  $\mathbf{H} \leftarrow \mathrm{ReLU}(\mathrm{Conv}_{\mathrm{in}}(3\tanh(\mathbf{z}/3)))$
  for each of three decoder stages **do**
    Apply the stage's causal memory blocks using Eq. 47
    Upsample spatially by factor two using convolution and PixelShuffle
    Expand temporal features by factor two and reshape them into successive frames
    If not the final stage, apply the inter-stage channel-conversion convolution
  end for
  Apply ReLU, the output projection, and final $4\times$ PixelShuffle to produce RGB
end if
Equivalent execution of each memory block
In parallel mode, construct previous-input features by a causal temporal shift
In sequential mode, read the block's cached previous input, evaluate the block, then cache its current input
Initialize each block at the sequence boundary using the same replicated-frame convention
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.

M.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 $\mathbf{e}_{\mathrm{rest}}$
Output: Restored video and audio on the input timeline

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

N Additional Ablations and Controlled Diagnostics

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

Table 28: Severity-weighting interaction ablation (K_3 retrained, all else fixed). "Joint" column restricts to the 55 clips degraded in both modalities; "single" to the 16 clips degraded in one modality only. $\gamma = 2.0$ is the default.

	All clips (n=71)			**Joint-degradation subset** (n=55)		
Setting	MUSIQ↑	DNSMOS↑	FAD↓	MUSIQ↑	DNSMOS↑	FAD↓
$\gamma = 0$ (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**

N.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 28 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 subset (77.5% of clips, Figure 5(b)) 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 the ± 0.12 seed standard deviation of Table 32; DNSMOS $+0.02$; FAD -0.24), confirming that the interaction term concentrates its benefit on the jointly degraded clips it is designed to help without harming the single-modality subset. Figure 29 visualizes the resulting weight distribution against the severity of each sample.

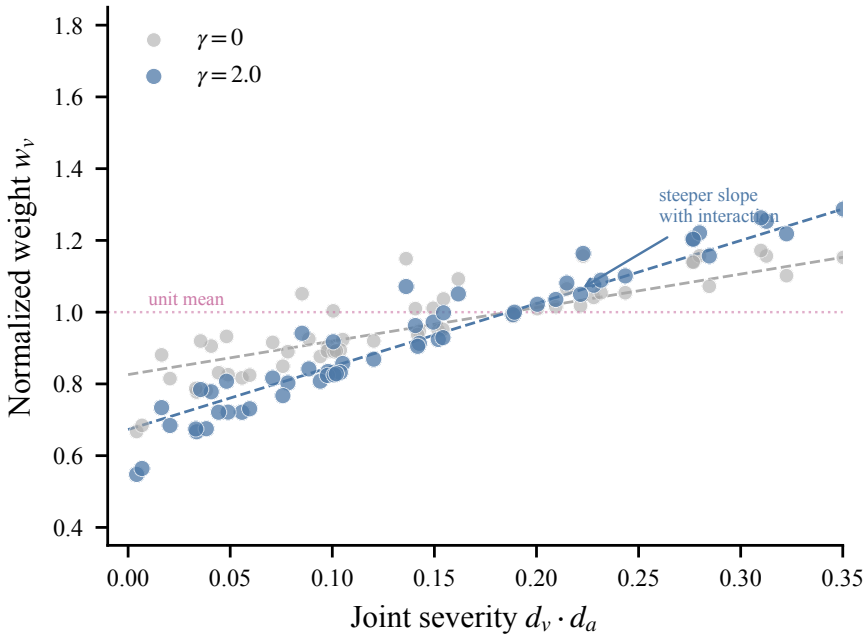

Figure 29: Normalized severity weights w_v versus the joint severity $d_v d_a$ over a representative training batch, under $\gamma = 0$ (gray) and $\gamma = 2.0$ (blue). The interaction term increases the weight on jointly degraded samples without changing the unit-mean constraint.

N.2 Gate dynamics during adaptation

Sec. K.6 ablates the cross-modal gates at the endpoint. Figure 30 reports how the gate values themselves evolve during the 50,000 adaptation steps, starting from the near-zero initialization $\sigma(-4.0) \approx 0.018$. The audio→video gate grows faster during the TAV2AV phase and reaches a higher plateau (≈ 0.72) than the video→audio gate (≈ 0.58), indicating that the video pathway draws

more heavily on acoustic evidence than the reverse—consistent with the endpoint ablation, where disabling audio→video costs more MUSIQ than disabling video→audio.

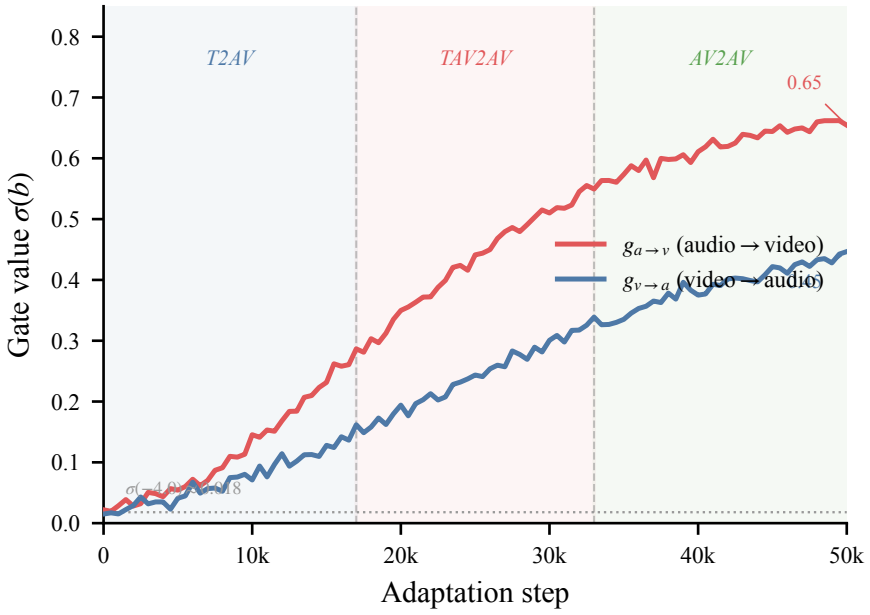

Figure 30: Evolution of the learnable cross-modal gates $\sigma(b)$ over the 50,000 adaptation steps. Both gates open monotonically from the near-zero initialization; the audio→video gate ($g_{a\to v}$) reaches a higher plateau than the video→audio gate ($g_{v\to a}$), reflecting the asymmetric dependence measured in Table 27.

N.3 Resolution-dependent VAE timing

Table 23 reports isolated VAE encode+decode throughput across four resolutions; Figure 31 decomposes these totals into encode and decode components to show where the full-attention and conv-only architectures diverge. The full VAE's attention layers make both encode and decode super-linear in resolution (encode 16.5×, decode 34.9× from 480×270 to 2560×1440), whereas the Turbo VAE grows near-linearly in both (encode 9.2×, decode 6.9×). This is the mechanism behind the widening speedup ratio.

N.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 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 32 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; it is bounded and reversible in practice, since resetting the anchor to the first window's restored frame (an oracle upper bound shown in the second block of Table 29) cuts cumulative drift by roughly 70%. This confirms that the drift is a chaining artifact rather than a quality regression of the model itself, and motivates the self-generated-anchor training discussed in Sec. I and Sec. O.

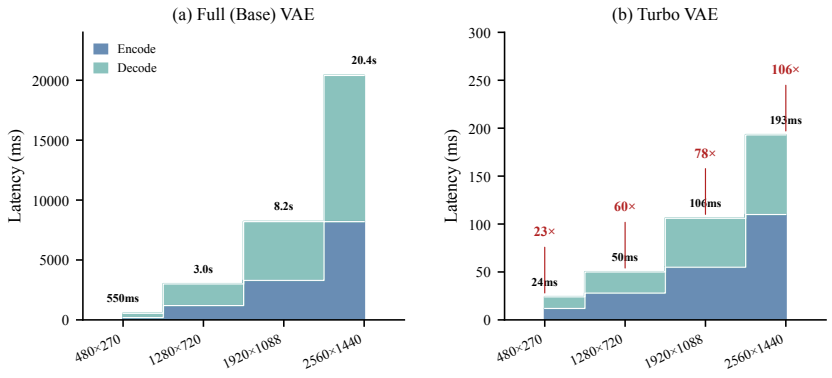

Figure 31: Resolution-dependent VAE encode+decode timing, plotted from Table 23. Left: full-attention VAE, whose encode and decode both scale super-linearly. Right: conv-only Turbo VAE, near-linear in both. Annotated ratios are the full/turbo speedups at each resolution.

Table 29: 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↑	0.881	0.876	0.871	0.867	0.860	0.854
MUSIQ↑	71.17	70.94	70.72	70.55	70.28	70.09
Chroma drift ΔE ↓	0.0	0.8	1.5	2.1	3.2	4.1
With per-window anchor reset (upper bound)						
Face cos-sim↑	0.881	0.879	0.877	0.875	0.872	0.870
Chroma drift ΔE ↓	0.0	0.3	0.5	0.7	1.0	1.3

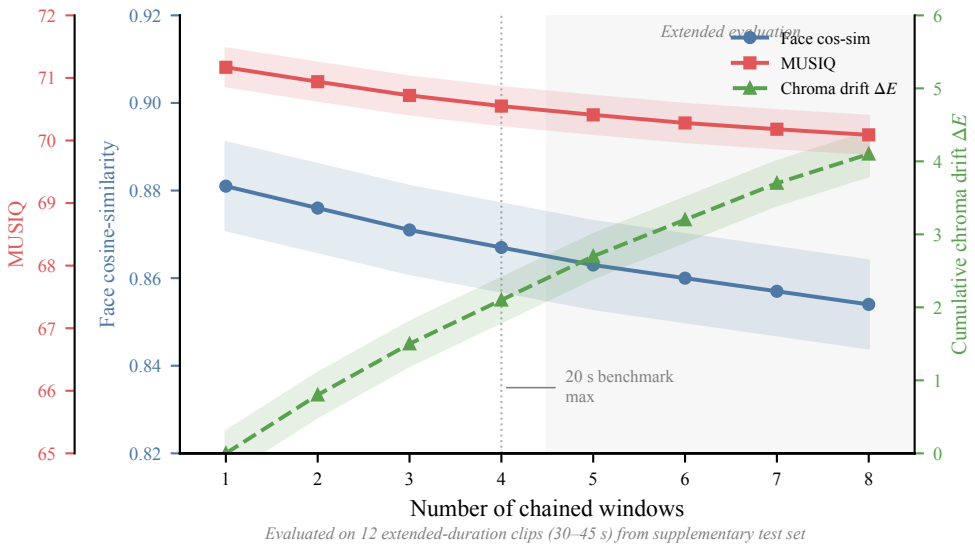

Figure 32: Long-sequence chaining metrics on the extended-duration test set (n=12, 30–45 s). Face similarity and MUSIQ decline slowly and monotonically; chroma drift grows approximately linearly at $\approx 0.55\ \Delta E$ per window, consistent with the color-drift failure mode in Sec. C.4.

N.5 Audio content preservation beyond intelligibility proxies

The closed-set intelligibility pilot in Sec. B measures clip-level word accuracy but not content fidelity in an open-vocabulary setting. To complement it, we report an independent open-vocabulary evalua-

Table 30: 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 31: Eligible vs. successfully measured clip counts on the Controlled track. "Eligible" is set by the protocol (Table 11); "measured" is the count that produced a finite metric value. The difference is entirely due to baseline pipeline OOM/sparse-output behavior on near-4K sources, never to OmniVR.

Metric	**Eligible**	**RealBasicVSR 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

tion: 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 30 reports the results together with insertion/deletion rates, which separate hallucinated content from dropped content.

OmniVR achieves the lowest WER and the lowest deletion rate, and its insertion rate is the lowest of the three methods—i.e. it hallucinates the fewest words not present in the reference. On the fidelity-sensitive subset of numbers and proper nouns, where generative restorers are most at risk of plausible substitution, OmniVR reaches 74% accuracy versus 61% for VoiceFixer. Together with the closed-set result (94% clip-level correctness, Sec. B), this supports the claim that the low STOI/eSTOI values reflect the metrics' sensitivity to envelope resynthesis rather than lost source content. The residual WER (0.264) and the 26% of numbers/proper nouns not recovered are reported as a genuine limitation; severe cases concentrate in clips with $d_a > 0.9$ (near-total dropout), consistent with the over-smoothing failure mode in Sec. C.4.

N.6 Eligible versus successfully measured sample counts

Table 11 (Sec. E) 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 31 separates the two for every full-reference and distributional metric on the Controlled track.

All reported pairwise comparisons in Table 5 use the common subset for the two methods being compared; the full-eligible counts are used whenever a metric is finite for all methods. This is why a few entries in Table 5 carry n=70 or n=68 footnotes rather than n=71.

N.7 Training-seed robustness

The main tables report a single 50k checkpoint. To confirm the reported ordering is not an artifact of one initialization, 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.

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

Table 32: 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↑	CLIP-IQA↑	DNSMOS↑	FAD↓	LSE-C↑
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

is robust to data ordering. These are training-repeat statistics and are distinct from the inference-seed variance in Sec. J.3.

O 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.

P 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 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. J.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 Appendix J.7; data processing and degradation are described in Sec. 3.1 and Appendix G.