FINERSR: RESTORING SMALL BUT SEMANTICALLY CRITICAL REGIONS IN GENERATIVE VIDEO SUPER-RESOLUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative video super-resolution has made remarkable progress in restoring realistic visual content from severely degraded videos. However, high global perceptual quality does not imply semantic fidelity in small but semantically critical regions, where visually plausible reconstructions may contain incorrect text or altered facial identities. To address this challenge, we propose **FinerSR**, a unified framework that strengthens small but semantically critical regions in both representation and supervision levels while dynamically coordinating local semantic fidelity with global perceptual quality. We first establish a strong four-step general restorer through supervised fine-tuning, on-policy distillation, and perceptual and reconstruction refinement. Building on this foundation, we introduce object-aware conditioning and region-normalized supervision to enhance the representation and optimization of small but semantically critical regions. We further develop discriminative objectives for two distinct semantic properties: alignment-aware identity and appearance matching for faces, and complete-string likelihood with edit-graph probability contrasts for text. Finally, we formulate semantic refinement as a quality-constrained joint optimization problem, using adaptive primal-dual updates to coordinate local semantic objectives with full-frame fidelity and perceptual constraints relative to the general restorer. Experiments on faces and text demonstrate that FinerSR enhances semantic fidelity in small but semantically critical regions while preserving overall restoration quality, bridging the gap between global visual quality and local semantic correctness.

1 INTRODUCTION

Video super-resolution (VSR) reconstructs high-resolution videos from degraded low-resolution observations. Recent generative methods have substantially improved perceptual quality, recovering realistic textures even under severe degradation (Zhou et al., 2024; Wang et al., 2026). However, perceptual realism does not guarantee semantic fidelity (Fei et al., 2026). This limitation is particularly evident in small but semantically critical regions: a restored face may appear natural while altering identity-related features, and reconstructed text may look convincing while containing incorrect characters (Figure 1). Such errors are especially challenging when degradation removes the visual cues needed to distinguish identities and characters (Feng et al., 2024; Li et al., 2026). Because these regions occupy only a small fraction of the frame, their errors may also be underrepresented in full-frame quality measures. This raises a central challenge: preserving local semantic fidelity while maintaining global visual quality.

Existing studies address different aspects of this challenge. General-purpose generative VSR methods leverage powerful priors to improve perceptual quality, while dedicated face restoration methods employ facial priors and temporal information to preserve facial structure and identity (Wang et al., 2021a; Zhou et al., 2022; Feng et al., 2024; Wang et al., 2025b). Text restoration methods similarly incorporate recognition priors, stroke structures, and glyph guidance (Ma et al., 2022; Chen et al., 2022; Min et al., 2026; Luo et al., 2026). Regional feedback and local guidance have also been explored to improve restoration quality (Liang et al., 2024; Liu et al., 2026; Kim et al., 2026b). Nevertheless, restoring heterogeneous semantic content within a general VSR model remains challenging. Global representations and aggregated objectives may insufficiently emphasize small ob-

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.

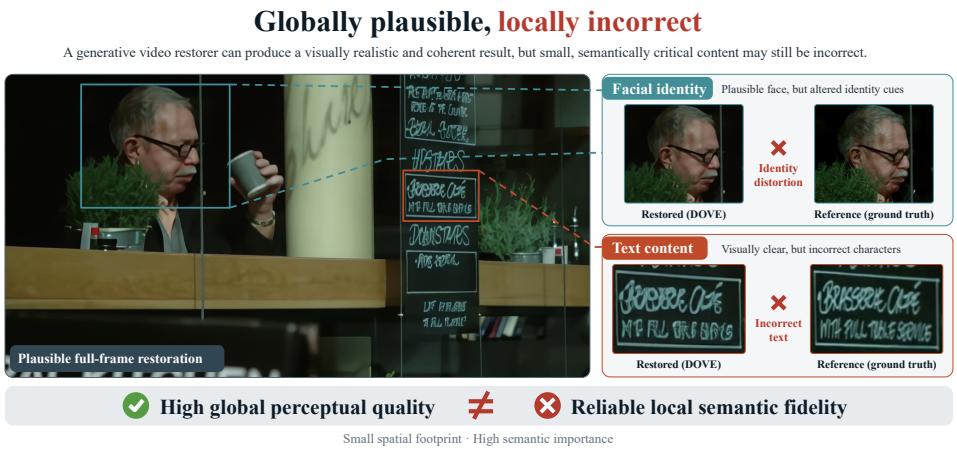

Figure 1: **Global perceptual quality does not guarantee local semantic fidelity.** In this UDM10 example, DOVE (Chen et al., 2025b) produces a visually plausible full scene, but the enlarged crops reveal altered facial features and incorrect text relative to the high-resolution reference.

jects, while perceptual feedback alone cannot reliably distinguish identity deviations or transcription errors. Moreover, specialized semantic supervision must be coordinated with full-frame restoration objectives. These challenges call for a unified approach that strengthens local representation and supervision while preserving overall restoration quality.

To address these challenges, we propose **FinerSR**, a generative VSR framework that strengthens small semantic regions in both representation and supervision while dynamically coordinating local semantic objectives with global quality constraints. We first establish a general four-step restorer through supervised fine-tuning, on-policy distillation, and subsequent reconstruction and perceptual refinement. The resulting model provides both a strong restoration foundation and a quality reference for semantic refinement.

Building on this foundation, we introduce object-aware conditioning and region-normalized supervision. Detected low-resolution face and text regions are encoded separately, and their features are spatially integrated into the shared video generator. Regional losses are aggregated across observations, tracks, and videos without area-based weighting, preventing small objects from receiving disproportionately weak supervision. We further develop discriminative objectives for two distinct semantic properties. For faces, spatial appearance matching is combined with a reference-conditioned identity metric that accounts for alignment sensitivity and emphasizes distinctions from confusable identities. For text, complete-string likelihood is complemented by an input-specific edit graph that supervises probability contrasts among competing transcriptions. Together, these designs strengthen small-object representation and provide targeted semantic supervision beyond conventional reconstruction and perceptual losses.

To coordinate local semantic refinement with global restoration, we formulate a quality-constrained joint optimization problem. Rather than enforcing pixel-level agreement with the general restorer, we use its frozen output as a quality reference and impose reconstruction and perceptual constraints across general-video, face, and text data. Stochastic primal-dual updates dynamically adjust constraint strengths according to the observed quality differences. Both local semantic objectives and full-frame quality constraints propagate through the complete four-step generation process, enabling their joint optimization within a single model. Our main contributions are summarized as follows:

- We establish a unified four-step generative VSR framework through supervised training, on-policy distillation, and reconstruction and perceptual refinement, providing a general restoration foundation and quality reference for subsequent semantic refinement.
- We introduce object-aware conditioning and region-normalized supervision to strengthen small semantic regions. We further develop discriminative objectives combining alignment-aware identity and appearance matching for faces with complete-string likelihood and edit-graph-based probability contrasts for text.

- We formulate semantic refinement as a quality-constrained joint optimization problem, using adaptive primal-dual updates to coordinate local semantic objectives with full-frame reconstruction and perceptual quality requirements.

We evaluate FinerSR on general VSR benchmarks and dedicated assessments of facial appearance, identity consistency, and text recognition. Extensive experiments on general VSR benchmarks and dedicated face and text evaluations demonstrate improved semantic fidelity in small but critical regions while maintaining overall restoration quality.

2 Related Work

Generative Image and Video Restoration. Generative restoration uses learned priors to recover plausible details from degraded observations. SR3 (Saharia et al., 2021) introduced conditional diffusion for image super-resolution; StableSR (Wang et al., 2023b) and DiffBIR (Lin et al., 2024) adapted pretrained diffusion priors to real-world restoration. ResShift (Yue et al., 2023) and OSEDiff (Wu et al., 2024) pursued efficient generation through residual shifting and one-step restoration, respectively. For video, Upscale-A-Video (Zhou et al., 2024) incorporated temporal modules and recurrent latent propagation, while SeedVR2 (Wang et al., 2026) and DOVE (Chen et al., 2025b) explored adversarial post-training and latent–pixel training. These advances support efficient generative VSR; our work adds explicit supervision for small semantic regions.

Semantic-Aware Restoration. Semantic priors can improve task-relevant content beyond full-frame perceptual quality. GFP-GAN (Wang et al., 2021a) uses facial generative priors; KEEP (Feng et al., 2024) and SVFR (Wang et al., 2025b) extend face restoration to video through temporal modeling or joint facial tasks. For text, TextZoom (Wang et al., 2020) established recognition-oriented evaluation, and Text Gestalt (Chen et al., 2022) introduced stroke-aware supervision. TIGER (Luo et al., 2026) restores glyphs before image enhancement, whereas TeReDiff (Min et al., 2026) combines text spotting with diffusion restoration. EDTR (Kim et al., 2025) and UniRestore (Chen et al., 2025a) use downstream task supervision to preserve information for visual recognition and scene understanding. Semantic Lens (Tang et al., 2024a) and SeeClear (Tang et al., 2024b) use instance semantics for cross-frame correspondence, and Tiled Prompts (Kim et al., 2026b) guides video restoration with localized descriptions. CODSR (Chen et al., 2026), FiDeSR (Kim et al., 2026a), and DynFaceRestore (Do et al., 2025) adapt guidance or objectives to regional characteristics. These approaches show the value of spatially differentiated restoration. FinerSR pairs object-level conditioning with reference-based supervision of facial identity and text content in a general video restorer.

Reward-Guided Generative Optimization. DDPO (Black et al., 2024) and Flow-GRPO (Liu et al., 2025) optimize diffusion and flow models with policy gradients, while DRaFT (Clark et al., 2024) differentiates rewards through sampling. In restoration, FinPercep-RM (Liu et al., 2026) supplies no-reference perceptual feedback, and LucidNFT (Fei et al., 2026) assesses LR-conditioned representation similarity. RefReward-SR (Song et al., 2026) combines full-image and crop assessment; SpatialFlow-GRPO (Yang et al., 2026) uses regional feedback for image-editing policy optimization. These methods illustrate ways to optimize perceptual and spatial feedback. Our formulation coordinates reference-defined regional semantic objectives with full-frame quality constraints through adaptive primal-dual updates.

3 Method

FinerSR jointly restores faces and text in a four-step video model (Fig. 2). Supervised training, on-policy distillation, and perceptual and reconstruction refinement establish general restoration (Sec. 3.1). Face supervision couples identity and spatial appearance using a metric informed by reference alignment sensitivity and competing identities; text supervision combines complete-string likelihood with edit-graph probability contrasts (Sec. 3.2). Region-level aggregation decouples loss weights from region area, while separately encoded object features provide local visual evidence. Constrained optimization couples these content targets with full-frame fidelity and perceptual supervision: both shape the reconstruction, with quality feedback relative to the general restorer adjusting full-frame supervision (Sec. 3.3).

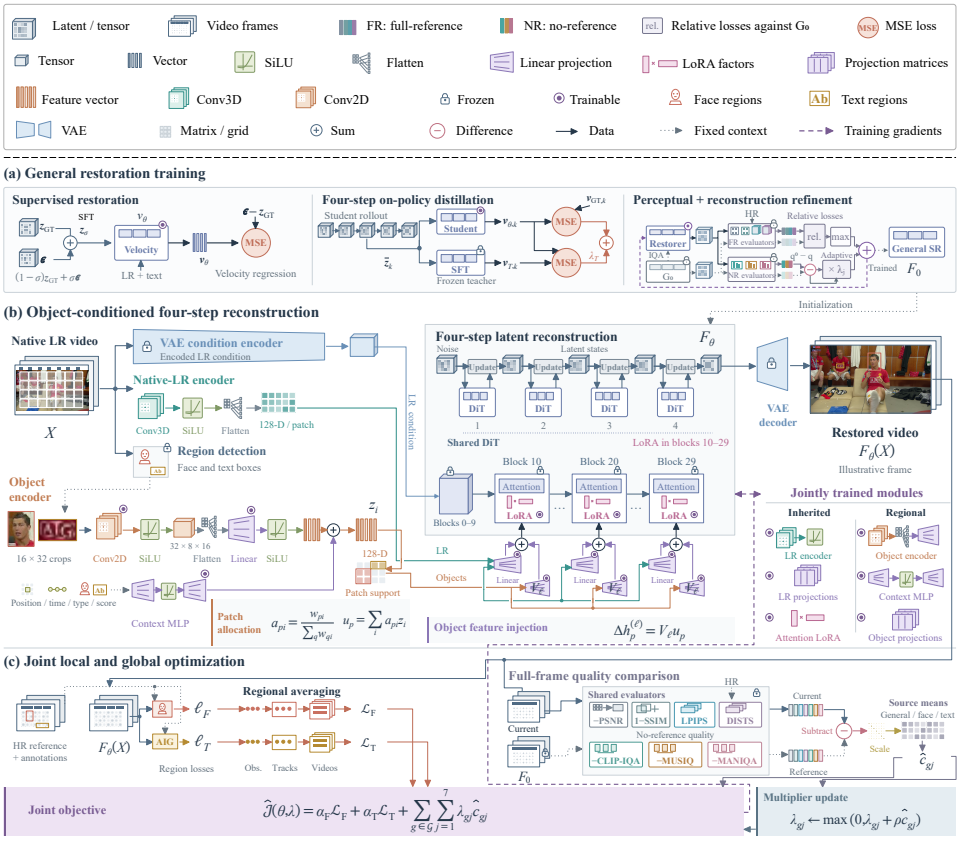

Figure 2: **Overview of FinerSR.** (a) General restoration training produces F_0; frozen G_0 provides the quality reference for reconstruction refinement. (b) Global LR and object features condition a shared four-step video generator. (c) Regional supervision and source-specific full-frame constraints jointly optimize restoration, using frozen F_0 as the quality reference.

3.1 General Restoration Training

We fine-tune Wan2.2-TI2V-5B (Wan Team, 2025) on paired videos, using mixed degradations before adapting to the target degradation and conditioning (Fig. 2(a)). These settings determine what the input preserves and what restoration must infer. On-policy distillation trains on the student's own intermediate states (Yin et al., 2024a), preparing it for the inputs encountered along its four-step trajectory. Four steps limit the cost of repeated generation for subsequent image-space objectives. Paired-reference velocity regression guides restoration, while a frozen SFT teacher constrains changes to learned behavior. Paired reconstruction losses and frozen no-reference quality predictors then produce an intermediate restorer G_0. SFT and distillation objectives, including decoded-frame supervision, are detailed in Appendix A.2; degradation and initial quality feedback are specified in Appendices A.3 and A.4.

Pixel, structural, and perceptual errors need not improve together; a weighted sum can hide deterioration in one criterion. We use frozen G_0 as the reference and minimize the largest relative reconstruction loss with adaptive quality feedback:

$$\mathcal{L}_{\text{ref}} = \max_{m=1,\dots,4} \mathcal{R}_m + \frac{1}{3}\sum_{j=1}^{3} \gamma_j^{\text{q}} \lambda_j \Delta q_j. \tag{1}$$

The $\mathcal{R}_m$ are relative losses from batch-mean PSNR, SSIM, LPIPS, and DISTS, zero at G_0 and negative for improvements; Δq_j is the reference-minus-current difference in batch-mean CLIP-IQA, MUSIQ, and MANIQA scores. With fixed positive scales γ_j^{q}, nonnegative multipliers λ_j increase for quality deficits and decrease toward zero for improvements. Comparing scores leaves room to correct G_0's errors without reproducing its pixels. The resulting general restorer F_0 initializes

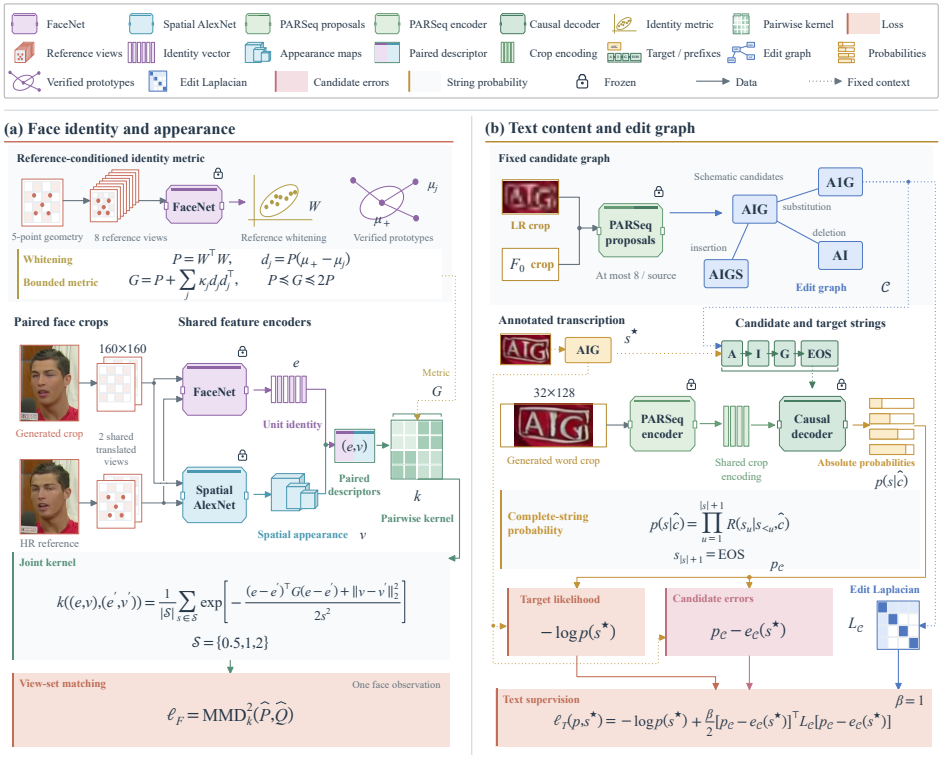

Figure 3: **Reference-defined face and text supervision.** (a) Reference-aligned face views are matched by joint identity–appearance MMD using the reference-conditioned metric G. (b) Complete-string likelihood and edit-graph probability contrasts supervise text content. Candidate strings are illustrative; the annotated target does not determine graph construction. Frozen networks pass gradients to generated crops.

joint refinement and provides its frozen quality reference. Appendix A.4 gives the relative-loss definitions, calibration, trainable modules, and sampling gradients.

3.2 Identity and Text Supervision

Visual quality alone does not specify which person or characters a region should depict. We define these targets from paired HR references: landmarks align faces, and annotated quadrilaterals define word crops. Applying the same transformations to generated videos preserves supervision even when poorly restored faces or words go undetected. Reference computations are detached; generated crops retain gradients through frozen feature and recognition networks.

Face identity and appearance. Similar identity embeddings can accompany different local appearances, while visually similar faces can depict different people. We supervise both properties with a reference-conditioned identity metric and spatial appearance features (Fig. 3(a)). Small alignment changes can shift identity embeddings without changing the person. For each reference observation, let e be the unit FaceNet embedding (Schroff et al., 2015). We estimate a normalized whitening transform W from small translations of the reference crop, adjusting distances for local alignment sensitivity. Nearby prototypes from verified different identities define directions for additional emphasis. With unperturbed reference embedding μ_+ and competitor prototypes μ_j, define

$$P = W^\top W, \qquad d_j = P(\mu_+ - \mu_j), \qquad G = P + \sum_j \kappa_j d_j d_j^\top . \tag{2}$$

The added terms increase squared metric distances for deviations along these reference–competitor directions. Normalized weights κ_j ensure $P \preceq G \preceq 2P$, retaining every base direction while bounding additional emphasis. The reference and identity gallery fix the metric independently of the reconstruction. Appendix A.5 derives the whitening, weights, and probabilistic interpretation.

For appearance, v contains spatial features from a frozen perceptual backbone (Zhang et al., 2018). Shared subpixel translations of generated and reference crops extend joint identity–appearance evaluation beyond one crop position. Their descriptors (e, v) form empirical distributions $\widehat{P}$ and $\widehat{Q}$ for the same face observation. With a fixed set $\mathcal{S}$ of positive kernel scales, we use

$$k((e,v),(e',v')) = \frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}}\exp\left[-\frac{(e-e')^\top G(e-e') + \|v-v'\|_2^2}{2s^2}\right], \qquad \ell_F = \mathrm{MMD}_k^2(\widehat{P},\widehat{Q}). \tag{3}$$

MMD compares these joint view sets within each reference frame. Appendix A.5 specifies feature normalization, kernel scales, and finite-sample MMD construction (Gretton et al., 2012).

Text content. A confidently recognized word may still be incorrect. We supervise each generated word crop with its ground-truth transcription $s^\star$ through frozen PARSeq (Bautista & Atienza, 2022). Transcription likelihood penalizes incorrect characters and stopping positions but gives identical losses for the same correct-string probability, even when different incorrect strings receive the remaining mass. We explicitly supervise these alternatives (Fig. 3(b)).

We form candidate set $\mathcal{C}$ from recognition proposals on the LR crop and F_0's corresponding restoration. With complete strings as nodes, one graph represents substitutions, insertions, and deletions: edges join strings differing by one token edit, yielding a spectrally normalized Laplacian $L_\mathcal{C}$. Fixed from this context, the graph does not follow the current model's changing predictions.

For generated crop $\hat{c}$, let $p(s) = p(s \mid \hat{c})$ be PARSeq's absolute probability of complete string s, computed with its own causal prefix and terminal EOS, without length normalization. With $\beta = 1$, we use

$$\ell_T(p, s^\star) = -\log p(s^\star) + \frac{\beta}{2}\big[p_\mathcal{C} - e_\mathcal{C}(s^\star)\big]^\top L_\mathcal{C}\big[p_\mathcal{C} - e_\mathcal{C}(s^\star)\big]. \tag{4}$$

Here $p_\mathcal{C}$ retains absolute candidate probabilities, and $e_\mathcal{C}(s^\star)$ restricts the target indicator to the candidates. Each edit edge matches predicted and target probability differences, explicitly supervising separation when it joins the true string to an alternative. With the supervision context, including crop geometry and graph, fixed, the expected combined loss is minimized by the true conditional distribution in unrestricted probability space (Gneiting & Raftery, 2007). Appendix A.6 gives the string-probability construction, graph details, and excess-risk derivation.

3.3 Joint Local and Global Optimization

Semantic fidelity and visual quality impose complementary requirements on the same reconstruction. Full-frame objectives evaluate scene structure and appearance, including faces and text; regional objectives distinguish identities and transcriptions that these measures may overlook.

Region-level supervision. An additive pixel loss weights a region's mean error by its area. We instead evaluate each annotated observation on its fixed-size reference crop and average by track and video. For content type $k \in \{F, T\}$, let $\mathcal{B}_k$ contain videos with valid type-k supervision in batch $\mathcal{B}$, $\mathcal{R}_{bk}$ their tracks, and $\mathcal{T}_{br}$ the observations of track r. We use

$$\mathcal{L}_k(\theta;\mathcal{B}) = \frac{1}{|\mathcal{B}_k|}\sum_{b\in\mathcal{B}_k}\frac{1}{|\mathcal{R}_{bk}|}\sum_{r\in\mathcal{R}_{bk}}\frac{1}{|\mathcal{T}_{br}|}\sum_{t\in\mathcal{T}_{br}}\ell_{brt,k}. \tag{5}$$

Here $\ell_{brt,k}$ is the type-k loss for observation t of track r in video b, and $\mathcal{L}_k = 0$ if $\mathcal{B}_k$ is empty. Small regions are no longer downweighted solely for occupying fewer pixels, nor long tracks upweighted for spanning more frames.

Object conditioning. The global LR encoder represents video on a shared patch grid. We encode each detected LR face or word at a fixed crop size and inject its descriptor into the corresponding DiT patches, retaining position, time, and category information (Fig. 2(b)). Per-object normalization keeps its total feature allocation independent of its fraction of the frame. LR detections provide conditioning during training and inference; HR annotations define supervision, so a missed LR detection does not remove an available annotated content loss. Appendix A.7 details feature allocation and layer injection.

Table 1: **General video super-resolution.** Scores average 80 paired clips or 50 VideoLQ clips. Per-dataset results are in Table 6. Bold/underline denote best/second best.

Set	Metric	R-ESR	RBV	RVF	UAV	VE	STAR	DOVE	SVR2	FVSR	SwVR	**Ours**
Paired (80)	PSNR ↑	23.17	23.19	<u>23.69</u>	22.68	21.75	22.92	**24.69**	22.96	22.60	23.20	23.68
	SSIM ↑	0.6279	0.6124	0.6248	0.5935	0.6068	0.6172	**0.6784**	0.6257	0.6185	0.6193	<u>0.6602</u>
	LPIPS ↓	0.3882	0.3979	0.3850	0.4060	0.5096	0.4473	0.3166	0.3461	<u>0.2902</u>	0.3087	**0.2864**
	DISTS ↓	0.2103	0.2137	0.2208	0.2194	0.2715	0.2336	0.1753	0.1844	<u>0.1529</u>	**0.1496**	0.1624
	CLIP-IQA ↑	0.4205	0.4552	0.3997	0.4191	0.3004	0.2851	0.4743	0.4307	0.5213	**0.5505**	<u>0.5337</u>
	MUSIQ ↑	57.41	63.58	61.01	56.74	37.62	38.61	62.72	55.26	67.43	**69.76**	<u>68.47</u>
	MANIQA ↑	0.3142	0.3017	0.2883	0.2852	0.2426	0.2322	0.3317	0.2965	<u>0.3663</u>	0.3637	**0.4145**
	NIQE ↓	4.266	**3.454**	3.953	4.364	7.513	6.420	5.297	4.437	3.704	<u>3.491</u>	4.556
VideoLQ	CLIP-IQA ↑	0.3619	**0.4036**	0.3466	0.2823	0.2955	0.2568	0.2911	0.2524	<u>0.4002</u>	0.3792	0.3484
	MUSIQ ↑	49.85	**56.12**	52.14	44.67	34.76	34.31	44.80	41.42	52.14	<u>55.25</u>	53.29
	MANIQA ↑	0.2986	<u>0.3084</u>	0.2844	0.2454	0.2465	0.2403	0.2628	0.2256	0.2987	0.2772	**0.3115**
	NIQE ↓	4.204	**3.585**	4.055	4.727	6.242	6.279	5.333	4.405	<u>3.936</u>	4.031	4.831

Quality-constrained joint optimization. We optimize local semantic fidelity subject to full-frame quality constraints, adapting their influence through quality feedback (Fig. 2(c)). We use F_0's quality as the reference: pixel matching would also penalize corrections to its content errors. Let r_j denote PSNR, SSIM, LPIPS, DISTS, CLIP-IQA, MUSIQ, and MANIQA, oriented so that lower is better (Appendix A.8). For paired LR and HR videos $(X, \boldsymbol{H})$ and general-video, face, and text source groups $\mathcal{D}_g$, we pose

$$\begin{aligned} &\min_{\theta} \quad \alpha_F \mathcal{L}_F^{\text{pop}}(\theta) + \alpha_T \mathcal{L}_T^{\text{pop}}(\theta) \quad \text{subject to} \quad c_{gj}(\theta) \le 0 \quad \forall g, j, \\ &c_{gj}(\theta) = s_j \, \mathbb{E}_{\mathcal{D}_g} \left[r_j(F_\theta(X), \boldsymbol{H}) - r_j(F_0(X), \boldsymbol{H}) \right]. \end{aligned} \tag{6}$$

Here $\mathcal{L}_k^{\text{pop}} = \mathbb{E}_{\mathcal{B}}[\mathcal{L}_k]$ follows the training source schedule and within-source sampling; α_F, α_T, s_j are fixed positive calibration scales. Separate constraints on source expectations prevent one source's improvement from masking another's deterioration in the constraint average.

In stochastic primal–dual optimization, primal updates combine content and quality gradients; dual updates increase a quality multiplier when its constraint is violated and decrease it toward zero when quality exceeds the reference. A positive multiplier retains the quality-gradient term even when current and reference scores coincide, so full-frame objectives participate alongside content supervision. Both objectives backpropagate through all four sampling steps. Appendix A.8 details loss-scale calibration and reuse, trainable modules, and parameter updates.

4 Experiments

4.1 Experimental setup

Training Datasets. General restoration uses a non-public video collection; SR initialization also incorporates non-public OCR-oriented videos. Joint refinement retains the general stream and adds ChokePoint videos (Wong et al., 2011) with identity labels and HierText scene-text images (Long et al., 2022) with word transcriptions. Their annotations define the content supervision in Sec. 3.2; data preparation and availability are described in Appendix A.1.

Testing Datasets. General quality is evaluated on 80 paired clips from SPMCS, UDM10, and YouHQ-40 (Zhou et al., 2024), and 50 unpaired VideoLQ clips (Chan et al., 2022). Paired inputs use the $\times 4$ degradation in Appendix A.3. From the paired benchmark, we select 24 face clips using InsightFace detections in GT middle frames. Small-face LPIPS uses the 8 clips with LR face height $[12, 32)$ pixels. The 46 text clips contain a nonempty PP-OCR prediction with confidence ≥ 0.85 in the GT middle frame, without a text-size filter. Appendix B.1 gives the filtering details.

Evaluation Metrics. General quality uses PSNR, SSIM, LPIPS, DISTS, CLIP-IQA, MUSIQ, MANIQA, and NIQE; VideoLQ uses only the last four. Face evaluation uses ArcFace cosine similarity (IDS), landmark error (AKD), FID-face, crop LPIPS, and detection rate and confidence. For text,

Table 2: **Scene-text and face restoration**: 46 text clips, 24 face clips. PP-OCR (PP) and EasyOCR (Easy) share PP-OCR reference tokens from GT frames. $\text{LPIPS}_{\text{small}}$ uses 8 clips with LR face height $[12, 32)$ pixels. IDS uses GT alignment; AKD uses matched detections and should be read with detection rate (Appendix B.1). Ranking notation follows Table 1.

	Metric	R-ESR	RBV	RVF	UAV	VE	STAR	DOVE	SVR2	FVSR	SwVR	**Ours**
Text	$\text{TC}_{\text{PP}} \uparrow$	0.2126	0.2234	0.1884	0.1775	0.0484	0.1558	<u>0.3280</u>	0.3031	0.1938	0.1558	**0.3420**
	$\text{SS}_{\text{PP}} \uparrow$	0.3457	0.3924	0.3797	0.3786	0.2627	0.3808	0.4838	<u>0.5124</u>	0.3926	0.3888	**0.5236**
	$\text{TC}_{\text{Easy}} \uparrow$	0.0996	0.1486	0.1075	0.1159	0.0217	0.1051	0.1981	<u>0.2234</u>	0.1486	0.1268	**0.2307**
	$\text{SS}_{\text{Easy}} \uparrow$	0.2101	0.2372	0.2484	0.2783	0.1557	0.2673	0.3312	<u>0.3881</u>	0.2862	0.2719	**0.4052**
	$\text{Conf.}_{\text{PP}} \uparrow$	0.5894	0.5361	0.5532	0.5549	0.5000	0.6540	0.6576	**0.7222**	0.5738	0.5839	<u>0.7148</u>
Face	Det. conf. $\uparrow$	0.4182	0.3389	0.3251	0.3027	0.3973	0.4115	0.3932	<u>0.5125</u>	0.4237	0.4387	**0.5168**
	Det. rate $\uparrow$	0.584	0.536	0.540	0.493	0.642	0.660	0.589	<u>0.669</u>	0.660	0.629	**0.729**
	IDS $\uparrow$	0.5019	0.4999	0.4913	0.4564	0.4623	0.4814	<u>0.5465</u>	0.5293	0.5167	0.5136	**0.5619**
	AKD $\downarrow$	7.84	7.24	<u>5.33</u>	11.98	9.45	8.87	**4.16**	9.32	8.28	6.77	6.85
	$\text{FID}_{\text{face}} \downarrow$	159.6	169.5	173.4	153.8	180.9	158.4	138.3	<u>120.4</u>	127.1	134.0	**114.9**
	$\text{LPIPS}_{\text{face}} \downarrow$	0.3255	0.3068	0.3086	0.3593	0.4090	0.3457	<u>0.2709</u>	0.2920	0.2742	0.2896	**0.2458**
	$\text{LPIPS}_{\text{small}} \downarrow$	0.3548	0.3484	<u>0.3430</u>	0.4238	0.4789	0.4190	0.3485	0.3602	0.3485	0.3737	**0.2955**

PP-OCR and EasyOCR measure exact token coverage (TC) and each reference token's best string similarity to the predictions (SS), without spatial matching or penalties for extra strings. Recognition confidence is reported separately. Definitions and aggregation rules are in Appendix B.1.

Baselines. We compare full-frame restorers: Real-ESRGAN (R-ESR) (Wang et al., 2021b) for image SR; RealBasicVSR (RBV) (Chan et al., 2022) and RealViformer (RVF) (Zhang & Yao, 2024) for conventional VSR; and Upscale-A-Video (UAV) (Zhou et al., 2024), VEnhancer (VE) (He et al., 2024), STAR (Xie et al., 2025), DOVE (Chen et al., 2025b), SeedVR2-3B (SVR2) (Wang et al., 2026), FlashVSR (FVSR) (Zhuang et al., 2025), and SwiftVR (SwVR) (Yan et al., 2026) for diffusion-based video restoration.

Implementation Details. Inference uses four sampling steps. Training and inference settings are detailed in Appendices A and A.9.

4.2 Comparison with Existing Methods

General Video Quality. The face and text results in Table 2 are accompanied by overall video quality on par with the state of the art (Table 1). Ours remains competitive in LPIPS and DISTS, although DOVE attains higher PSNR and SSIM on the paired datasets. No-reference results are mixed: RealBasicVSR leads CLIP-IQA, MUSIQ, and NIQE on VideoLQ, while Ours obtains the highest MANIQA score.

Scene Text Restoration. Ours achieves the highest token coverage and string similarity with both recognizers (Table 2, text block). PP-OCR and EasyOCR token coverage reaches 0.3420 and 0.2307, exceeding DOVE (0.3280) and SeedVR2-3B (0.2234), the respective strongest baselines. SeedVR2-3B has the highest recognition confidence.

Face Restoration. Ours ranks first on 6 of 7 face measures (Table 2, face block). For LR face height $[12, 32)$ pixels, LPIPS is 0.2955, 13.8% below the strongest baseline, RealViformer (0.3430). DOVE, RealViformer, and SwiftVR obtain lower AKD.

Qualitative Comparisons. Figure 4 complements the identity and text-recognition results with visual comparisons. Ours recovers distinct eye and mouth contours and well-separated strokes in the small lettering on the shop sign. Several competing outputs exhibit unnatural facial textures or merge and deform character strokes, obscuring facial features and character shapes.

4.3 Ablation Studies

Table 3 evaluates content supervision and joint optimization against the general restorer F_0. Each ablation changes one component of the full model under matched training settings (Appendix B.4).

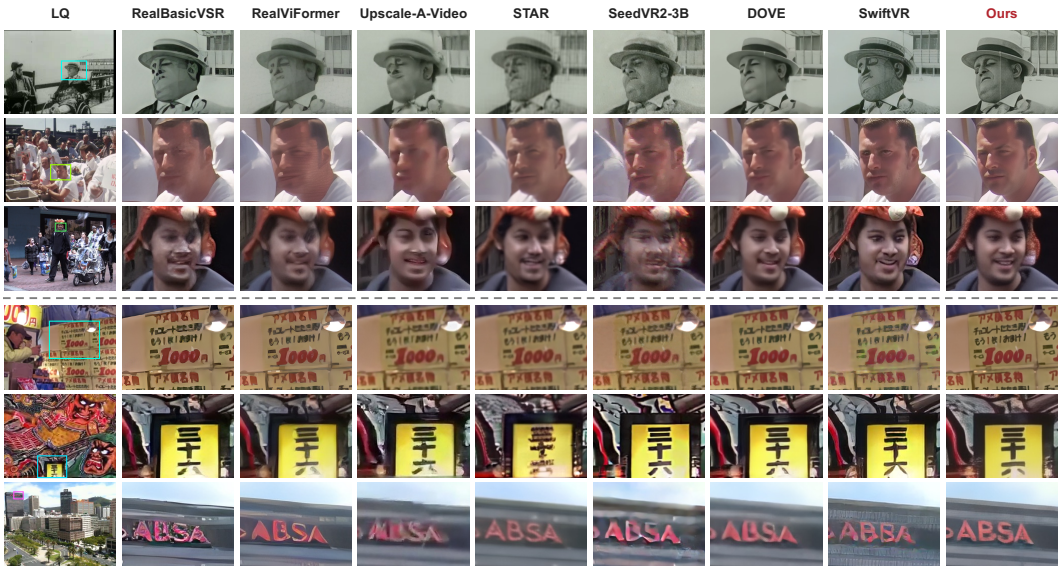

Figure 4: **Qualitative comparison of face and scene-text restoration.** Boxes in the input frames mark the enlarged regions. The top three rows show faces; the bottom three show scene text. Additional examples are in Appendix B.2.

Table 3: **Ablation of content supervision and joint optimization.** Evaluation follows Tables 1 and 2.

Setting	PSNR ↑	LPIPS ↓	IDS ↑	$\text{LPIPS}_{\text{small}}$ ↓	TC_{PP} ↑	TC_{Easy} ↑
General restorer F_0	23.72	0.2856	0.5377	0.3093	0.1938	0.1890
Cosine + feature face loss	23.68	0.2866	0.5489	0.3047	0.3414	0.2310
Text NLL only	23.69	0.2862	0.5622	0.2959	0.3193	0.2156
Fixed quality weights	23.56	0.2907	0.5624	0.2982	0.3451	0.2315
Without object conditioning	23.67	0.2875	0.5481	0.3036	0.3068	0.2119
Area-weighted region losses	23.70	0.2861	0.5530	0.3081	0.3206	0.2183
Full model	23.68	0.2864	0.5619	0.2955	0.3420	0.2307

Content supervision. With the same frozen encoders and aligned crops, replacing the face objective with cosine identity distance and spatial-feature error lowers IDS and worsens small-face LPIPS, with little change in text scores. Removing the edit-graph term while retaining complete-transcription NLL with EOS lowers token coverage with both recognizers.

Local and global optimization. Removing object conditioning while retaining region losses, or weighting these losses by their original area fractions, weakens both face and text recovery. Adaptive quality weights improve full-frame PSNR, LPIPS, and small-face LPIPS over fixed weights, with slightly lower IDS and text scores. Relative to F_0, the full model improves identity, small-face appearance, and text recovery, with slightly lower full-frame PSNR and higher LPIPS.

5 Conclusion

We presented FinerSR, a unified four-step generative video super-resolution framework for restoring small but semantically critical regions. Object conditioning and region-level aggregation strengthen their representation and supervision. Reference-conditioned identity–appearance matching and complete-string supervision with edit-graph probability contrasts address identity deviations and transcription errors. Adaptive full-frame quality constraints coordinate these content objectives with general restoration. Experiments demonstrate improved identity consistency, small-face perceptual quality, and text recovery, with overall video quality on par with the state of the art.

AI USE STATEMENT

Large language models were used to assist with translation, summarizing existing literature, searching for relevant information, improving the clarity and readability of the manuscript, and formatting references. All research methodology, experimental design, and scientific conclusions were determined by the authors. The authors reviewed all AI-assisted work and take full responsibility for the final content of this work, including its text, claims, and reported results.

ETHICS STATEMENT

Restored faces and text can appear plausible while differing from the reference. Identity similarity and text coverage assess aspects of this risk but do not establish that a restoration is factually correct. The face evaluations compare restored and reference images; they do not establish suitability for identifying unknown individuals. Restored content should not be treated as forensic evidence without independent corroboration.

Training uses non-public general and OCR-oriented video collections, alongside annotated subsets of ChokePoint and HierText (Appendix A.1). The non-public collections and their private storage indices will not be released. For the public datasets, we plan to release scripts for processing annotations, constructing training and development splits, and preprocessing samples, without distributing source images or videos. ChokePoint and HierText should be obtained from their original providers subject to the applicable access and use terms.

REPRODUCIBILITY STATEMENT

Section 3 presents the method. Appendix A provides mathematical derivations and implementation details, including data sources and training/development splits, preprocessing and degradations, optimization settings, and loss calibration. Appendix A.9 specifies inference settings. Section 4.1 and Appendix B.1 describe the evaluation datasets and baselines, face/text subset selection, metric computation, and aggregation. Appendix B.4 defines the component changes and shared training and evaluation settings used in the ablations.

We plan to release training and inference code, configurations, index-construction recipes, and scripts for processing annotations, constructing data splits, and preprocessing the public datasets. The non-public video collections and their private indices are excluded from the release; full retraining requires access to these collections. Data availability is detailed in Appendix A.1.

REFERENCES

Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. Character region awareness for text detection. In *CVPR*, 2019.

Darwin Bautista and Rowel Atienza. Scene text recognition with permuted autoregressive sequence models. In *ECCV*, 2022.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training Diffusion Models with Reinforcement Learning. In *ICLR*, 2024. URL https://arxiv.org/abs/2305.13301.

Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, 2018.

Kelvin C.K. Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Investigating tradeoffs in real-world video super-resolution. In *CVPR*, 2022.

Hao Chen, Junyang Chen, Jinshan Pan, and Jiangxin Dong. Bridging Fidelity-Reality with Controllable One-Step Diffusion for Image Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30584–30594, June 2026.

I-Hsiang Chen, Wei-Ting Chen, Yu-Wei Liu, Yuan-Chun Chiang, Sy-Yen Kuo, and Ming-Hsuan Yang. UniRestore: Unified perceptual and task-oriented image restoration model using diffusion prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17969–17979, 2025a. URL https://arxiv.org/abs/2501.13134.

Jingye Chen, Haiyang Yu, Jianqi Ma, Bin Li, and Xiangyang Xue. Text gestalt: Stroke-aware scene text image super-resolution. In *AAAI*, 2022.

Zheng Chen, Zichen Zou, Kewei Zhang, Xiongfei Su, Xin Yuan, Yong Guo, and Yulun Zhang. DOVE: Efficient one-step diffusion model for real-world video super-resolution. In *NeurIPS*, 2025b.

Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly Fine-Tuning Diffusion Models on Differentiable Rewards. In *ICLR*, 2024. URL https://arxiv.org/abs/2309.17400.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *CVPR*, 2019.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

Huu-Phu Do, Yu-Wei Chen, Yi-Cheng Liao, Chi-Wei Hsiao, Han-Yang Wang, Wei-Chen Chiu, and Ching-Chun Huang. DynFaceRestore: Balancing Fidelity and Quality in Diffusion-Guided Blind Face Restoration with Dynamic Blur-Level Mapping and Guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10432–10441, October 2025.

Song Fei, Tian Ye, Sixiang Chen, Zhaohu Xing, Jianyu Lai, and Lei Zhu. LucidNFT: LR-anchored multi-reward preference optimization for flow-based real-world super-resolution. *arXiv preprint arXiv:2603.05947*, 2026.

Ruicheng Feng, Chongyi Li, and Chen Change Loy. Kalman-inspired feature propagation for video face super-resolution. In *ECCV*, 2024.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL https://jmlr.org/papers/v13/gretton12a.html.

Jingwen He, Tianfan Xue, Dongyang Liu, Xinqi Lin, Peng Gao, Dahua Lin, Yu Qiao, Wanli Ouyang, and Ziwei Liu. VEnhancer: Generative space-time enhancement for video generation. *arXiv preprint arXiv:2407.07667*, 2024.

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 *ICLR*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *ICCV*, 2021.

Aro Kim, Myeongjin Jang, Chaewon Moon, Youngjin Shin, Jinwoo Jeong, and Sang-hyo Park. FiDeSR: High-Fidelity and Detail-Preserving One-Step Diffusion Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 38270–38280, June 2026a.

Bryan Sangwoo Kim, Jonghyun Park, and Jong Chul Ye. Tiled prompts: Overcoming prompt misguidance in image and video super-resolution. *arXiv preprint arXiv:2602.03342*, 2026b. URL https://arxiv.org/abs/2602.03342.

Jaeha Kim, Junghun Oh, and Kyoung Mu Lee. Exploiting Diffusion Prior for Task-driven Image Restoration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10151–10161, October 2025.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004. URL https://www.ledoit.net/ole1_abstract.htm.

Jinlong Li, Jiaming Ding, Dingfu Lu, Malcolm Hsiu, Chuang Ke, Kangning Yang, Bochen Guan, Lan Fu, Jie Cai, Huiming Sun, and Zibo Meng. ClearText-Video: A large-scale text-centric video dataset bridging video restoration and scene-text enhancement. *arXiv preprint arXiv:2608.28784*, 2026. URL https://arxiv.org/abs/2608.28784.

Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, Junjie Ke, Krishnamurthy Dj Dvijotham, Katherine M. Collins, Yiwen Luo, Yang Li, Kai J Kohlhoff, Deepak Ramachandran, and Vidhya Navalpakkam. Rich human feedback for text-to-image generation. In *CVPR*, pp. 19401–19411, 2024.

Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. DiffBIR: Toward Blind Image Restoration with Generative Diffusion Prior. In *European Conference on Computer Vision*, 2024. doi: 10.1007/978-3-031-73202-7_25. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/7690_ECCV_2024_paper.php.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-GRPO: Training Flow Matching Models via Online RL. In *NeurIPS*, 2025. URL https://arxiv.org/abs/2505.05470.

Yidi Liu, Zihao Fan, Jie Huang, Jie Xiao, Dong Li, Wenlong Zhang, Lei Bai, Xueyang Fu, and Zheng-Jun Zha. FinPercep-RM: A fine-grained reward model and co-evolutionary curriculum for RL-based real-world super-resolution. In *CVPR*, pp. 4839–4849, 2026.

Shangbang Long, Siyang Qin, Dmitry Panteleev, Alessandro Bissacco, Yasuhisa Fujii, and Michalis Raptis. Towards end-to-end unified scene text detection and layout analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Minxing Luo, Linlong Fan, Qiushi Wang, Ge Wu, Yiyan Luo, Yuhang Yu, Jinwei Chen, Yaxing Wang, Qingnan Fan, and Jian Yang. Restore text first, enhance image later: Two-stage scene text image super-resolution with glyph structure guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30553–30563, June 2026.

Jianqi Ma, Zhetong Liang, and Lei Zhang. A text attention network for spatial deformation robust scene text image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5911–5920, 2022.

Jaewon Min, Jin Hyeon Kim, Paul Hyunbin Cho, Jaeeun Lee, Jihye Park, Minkyu Park, Sangpil Kim, Hyunhee Park, and Seungryong Kim. Text-Aware Image Restoration with Diffusion Models. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/6e5112eaa45f8c30b242c5f576213a92-Abstract-Conference.html.

Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 2013.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image Super-Resolution via Iterative Refinement. *arXiv preprint arXiv:2104.07636*, 2021. URL https://arxiv.org/abs/2104.07636.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *CVPR*, 2015.

Yichun Shi and Anil K. Jain. Probabilistic face embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6902–6911, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Shi_Probabilistic_Face_Embeddings_ICCV_2019_paper.html.

Yushuai Song, Weize Quan, Weining Wang, Jiahui Sun, Jing Liu, Meng Li, Pengbin Yu, Zhentao Chen, Wei Shen, Lunxi Yuan, and Dong-ming Yan. RefReward-SR: LR-conditioned reward modeling for preference-aligned super-resolution. *arXiv preprint arXiv:2603.24198*, 2026. URL https://arxiv.org/abs/2603.24198.

Qi Tang, Yao Zhao, Meiqin Liu, Jian Jin, and Chao Yao. Semantic lens: Instance-centric semantic alignment for video super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5154–5161, 2024a. doi: 10.1609/aaai.v38i6.28321. URL https://ojs.aaai.org/index.php/AAAI/article/view/28321.

Qi Tang, Yao Zhao, Meiqin Liu, and Chao Yao. SeeClear: Semantic distillation enhances pixel condensation for video super-resolution. In *Advances in Neural Information Processing Systems*, volume 37, 2024b. URL https://papers.nips.cc/paper_files/paper/2024/hash/f358b2a880adf34939d2d6f926e54d2a-Abstract-Conference.html.

Wan Team. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Jianyi Wang, Kelvin C.K. Chan, and Chen Change Loy. Exploring CLIP for assessing the look and feel of images. In *AAAI*, 2023a.

Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin C. K. Chan, and Chen Change Loy. Exploiting Diffusion Prior for Real-World Image Super-Resolution. *arXiv preprint arXiv:2305.07015*, 2023b. URL https://arxiv.org/abs/2305.07015.

Jianyi Wang, Zhijie Lin, Meng Wei, Yang Zhao, Ceyuan Yang, Chen Change Loy, and Lu Jiang. SeedVR: Seeding infinity in diffusion transformer towards generic video restoration. In *CVPR*, 2025a.

Jianyi Wang, Shanchuan Lin, Zhijie Lin, Yuxi Ren, Meng Wei, Zongsheng Yue, Shangchen Zhou, Hao Chen, Yang Zhao, Ceyuan Yang, Xuefeng Xiao, Chen Change Loy, and Lu Jiang. SeedVR2: One-step video restoration via diffusion adversarial post-training. In *ICLR*, 2026.

Wenjia Wang, Enze Xie, Xuebo Liu, Wenhai Wang, Ding Liang, Chunhua Shen, and Xiang Bai. Scene text image super-resolution in the wild. In *ECCV*, 2020.

Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *CVPR*, 2021a.

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In *ICCV Workshops*, 2021b.

Zhiyao Wang, Xu Chen, Chengming Xu, Junwei Zhu, Xiaobin Hu, Jiangning Zhang, Chengjie Wang, Yuqi Liu, Yiyi Zhou, and Rongrong Ji. SVFR: A unified framework for generalized video face restoration. In *CVPR*, 2025b.

Yongkang Wong, Shaokang Chen, Sandra Mau, Conrad Sanderson, and Brian C. Lovell. Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition. In *2011 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 81–88, 2011.

Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-Step Effective Diffusion Network for Real-World Image Super-Resolution. In *Advances in Neural Information Processing Systems*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/a8223b0ad64007423ffb308b0dd92298-Abstract-Conference.html.

Rui Xie, Yinhong Liu, Penghao Zhou, Chen Zhao, Jun Zhou, Kai Zhang, Zhenyu Zhang, Jian Yang, Zhenheng Yang, and Ying Tai. STAR: Spatial-temporal augmentation with text-to-video models for real-world video super-resolution. In *ICCV*, 2025.

Jiaqi Yan, Xiangyu Chen, Xinlin Zhong, Haibin Huang, Chi Zhang, Jie Liu, Jiantao Zhou, and Xuelong Li. SwiftVR: Real-time one-step generative video restoration. *arXiv preprint arXiv:2606.09516*, 2026.

Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. MANIQA: Multi-dimension attention network for no-reference image quality assessment. In *CVPR Workshops*, 2022.

Yankai Yang, Yancheng Long, Wei Chen, Xingyu Lu, Hongyang Wei, Bin Wen, Fan Yang, Tingting Gao, Han Li, and Shuo Yang. SpatialFlow-GRPO: Where Spatial Credit Drives Image Editing. *arXiv preprint arXiv:2606.26872*, 2026. URL https://arxiv.org/abs/2606.26872.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Frédo Durand, and William T. Freeman. Improved distribution matching distillation for fast image synthesis. In *NeurIPS*, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *CVPR*, 2024b.

Zongsheng Yue, Jianyi Wang, and Chen Change Loy. ResShift: Efficient Diffusion Model for Image Super-resolution by Residual Shifting. In *Advances in Neural Information Processing Systems*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/2ac2eac5098dba08208807b65c5851cc-Abstract-Conference.html.

Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 2016.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Yuehan Zhang and Angela Yao. RealViformer: Investigating attention for real-world video super-resolution. In *ECCV*, 2024.

Shangchen Zhou, Kelvin C. K. Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. In *Advances in Neural Information Processing Systems*, 2022.

Shangchen Zhou, Peiqing Yang, Jianyi Wang, Yihang Luo, and Chen Change Loy. Upscale-a-video: Temporal-consistent diffusion model for real-world video super-resolution. In *CVPR*, 2024.

Junhao Zhuang, Shi Guo, Xin Cai, Xiaohui Li, Yihao Liu, Chun Yuan, and Tianfan Xue. FlashVSR: Towards real-time diffusion-based streaming video super-resolution. *arXiv preprint arXiv:2510.12747*, 2025.

APPENDIX

A IMPLEMENTATION DETAILS

A.1 TRAINING DATA AND AVAILABILITY

The general training corpus is a non-public video collection assembled from multiple sources, including web videos. Supervised adaptation, distillation, and perceptual and reconstruction refinement use this corpus. SR initialization additionally uses a non-public OCR-oriented video collection. Joint refinement preserves the general stream and introduces separately annotated face and scene-text sources.

Face supervision uses NICTA's ChokePoint dataset (Wong et al., 2011), with identity labels checked against the original frame annotations. The identity gallery is built from training identities; its construction and exclusion of development sources are described in Appendix A.5. Text supervision uses a recognizer-compatible subset of the official HierText training images and annotations (Long et al., 2022). Accepted words have nonempty transcriptions, explicit legibility labels, quadrilaterals, and track identifiers. OCR predictions supply candidate strings, not target transcriptions.

Table 4 summarizes the annotated subsets before spatial visibility filtering. ChokePoint is split by identity and HierText by original image, with source identifiers and frame hashes checked for overlap. Development samples are used for validation without model updates; semantic-loss calibration and identity-gallery construction use training samples only.

Table 4: **Annotated data for joint refinement.** Counts describe the admitted subsets before spatial crop filtering. Video windows may come from the same sequence; they are not independent source videos.

Source	Unit	Train	Development	Frames/sample
ChokePoint	Video windows	813	180	33
	Identities	23	6	—
HierText	Images	4,065	237	1

Annotated samples retain their actual frame count, so a still image remains a one-frame sample. We choose the nearest-aspect spatial bucket that retains a complete annotated region of the requested type, and transform frames and annotations together. No implicit frame repetition is used for these samples. The non-public video corpus and its private indices are not part of the planned release.

A.2 TRAINING OBJECTIVES AND SETTINGS

Supervised restoration objective. Initial SR training and its continuation optimize the same velocity MSE (Sec. 3.1). With reference latent $\boldsymbol{z}_{\mathrm{GT}}$, $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$, and $\sigma \in [0, 1]$, the input is $\boldsymbol{z}_\sigma^{\mathrm{SFT}} = (1-\sigma)\boldsymbol{z}_{\mathrm{GT}} + \sigma\boldsymbol{\epsilon}$; the velocity target is $\boldsymbol{\epsilon} - \boldsymbol{z}_{\mathrm{GT}}$. Let q_{SFT} denote the joint training distribution induced by the clean videos, degradation process, conditioning policy, and sampled time and noise. For a fixed q_{SFT}, minimizing the population MSE over square-integrable predictors gives the conditional mean:

$$v^\star_{q_{\mathrm{SFT}}}(\boldsymbol{z}_\sigma^{\mathrm{SFT}}, \sigma; \boldsymbol{c}) = \mathbb{E}_{q_{\mathrm{SFT}}}\left[\boldsymbol{\epsilon} - \boldsymbol{z}_{\mathrm{GT}} \,\middle|\, \boldsymbol{z}_\sigma^{\mathrm{SFT}}, \sigma, \boldsymbol{c}\right].$$

This standard squared-loss identity explains why an unchanged loss can learn different restoration fields: changing the degradation or conditioning can change the clean-content distribution consistent with the observed inputs. We therefore continue supervised training under the intended input configuration before freezing the model for OPD, consistent with the sensitivity to degradations and LQ conditioning reported in prior restoration work (Chan et al., 2022; Wang et al., 2025a). OPD subsequently matches this teacher on states visited by the four-step student, alongside paired GT supervision.

Initialization and optimization. We initialize the video backbone from Wan2.2-TI2V-5B, retain its input projection, and zero-initialize the added weights for the 48 video conditioning channels.

SFT, OPD, and the initial IQA refinement update all DiT parameters, while the VAE and text encoder remain frozen. Supervised training uses velocity MSE, first with mixed degradations and then with the target $\times 4$ configuration (Appendix A.3). The learning rate is 10^{-4} initially and 10^{-5} during this adaptation, followed by 2×10^{-6} for OPD and 10^{-6} for the initial IQA refinement. These updates use AdamW with $(\beta_1, \beta_2) = (0.9, 0.95)$, $\epsilon = 10^{-10}$, zero weight decay, and bfloat16 computation. Reconstruction refinement freezes this backbone and uses the separate settings in Appendix A.4.

The video condition is formed as $(1 - \alpha_{\text{cond}})\boldsymbol{z}_{\text{LQ}} + \alpha_{\text{cond}}\boldsymbol{\epsilon}_c$, where $\boldsymbol{\epsilon}_c \sim \mathcal{N}(0, I)$ has the same shape as $\boldsymbol{z}_{\text{LQ}}$. The degraded video is resized to reference resolution and VAE-encoded to obtain $\boldsymbol{z}_{\text{LQ}}$; the video condition is concatenated with the noisy output latent. Initially, $\alpha_{\text{cond}} \sim U[0.1, 0.5]$; the condition is dropped with probability $p_{\text{drop}} \sim U[0.39, 0.41]$, and captions are independently replaced by a fixed negative prompt with probability 0.1. Supervised adaptation, distillation, and refinement use $\alpha_{\text{cond}} \sim U[0, 0.1]$, with condition dropout and caption replacement disabled, providing lower-noise LQ conditioning for the teacher used in distillation.

Video preprocessing. Initial SR training uses temporal buckets of 33–81 frames in increments of eight and spatial buckets with side lengths from 480 to $1{,}920$ pixels. From supervised adaptation onward, the general-video stream uses 33-frame clips, with spatial buckets ranging from $512 \times 2{,}048$ through $1{,}024 \times 1{,}024$ to $2{,}048 \times 512$.

The loader uses a nominal 16 fps setting with integer-stride indexing, followed by uniformly spaced sampling over the available sequence (probability 0.4) or a random contiguous window (0.6). Short sequences are extended by repeating frames. Frames are bilinearly resized with antialiasing to cover the selected bucket and then center-cropped.

SFT, OPD, and initial IQA refinement separately VAE-encode reference and full-resolution degraded videos, and use dataset captions. Reconstruction and joint regional refinement use RGB references and empty prompts. Annotated face and text samples follow the frame-count and geometry rules in Appendix A.1; inference also uses empty prompts.

Velocity distillation and endpoint gradients. For distillation, the SFT-initialized student takes $K = 4$ sampling steps from Gaussian noise $\boldsymbol{z}_0$, with $\sigma_0 = 1 > \cdots > \sigma_K = 0$:

$$\boldsymbol{z}_{k+1} = \boldsymbol{z}_k + (\sigma_{k+1} - \sigma_k) v_\theta(\boldsymbol{z}_k, \sigma_k; \boldsymbol{c}), \qquad k = 0, \ldots, K-1. \tag{7}$$

The training noise levels are $[1, 15/16, 5/6, 5/8, 0]$. We retain the four states before the updates without gradients and sample k uniformly from $\{0, \ldots, K-1\}$. Write $\bar{\boldsymbol{z}}_k = \text{sg}(\boldsymbol{z}_k)$, where sg denotes stop-gradient. The student and frozen SFT teacher, which share the same architecture, predict velocities at that state: $\boldsymbol{v}_{\theta,k} = v_\theta(\bar{\boldsymbol{z}}_k, \sigma_k; \boldsymbol{c})$ and $\boldsymbol{v}_{T,k} = v_T(\bar{\boldsymbol{z}}_k, \sigma_k; \boldsymbol{c})$. The paired-reference target and velocity objective are

$$\boldsymbol{v}_{\text{GT},k} = \frac{\bar{\boldsymbol{z}}_k - \boldsymbol{z}_{\text{GT}}}{\sigma_k}, \qquad \mathcal{L}_v = \text{MSE}(\boldsymbol{v}_{\theta,k}, \boldsymbol{v}_{\text{GT},k}), \tag{8}$$

$$\mathcal{L}_{\text{vel},k} = \mathcal{L}_v + \lambda_T \,\text{MSE}(\boldsymbol{v}_{\theta,k}, \boldsymbol{v}_{T,k}). \tag{9}$$

Both MSE terms average over the full latent video, with GT and teacher weights 1 and 0.5. The target satisfies $\bar{\boldsymbol{z}}_k - \sigma_k \boldsymbol{v}_{\text{GT},k} = \boldsymbol{z}_{\text{GT}}$.

First-frame GT LPIPS has weight 0.5 and is evaluated every fourth update. To limit activation memory for image losses, we recompute only the final sampling update:

$$\hat{\boldsymbol{z}} = \bar{\boldsymbol{z}}_{K-1} - \sigma_{K-1} v_\theta(\bar{\boldsymbol{z}}_{K-1}, \sigma_{K-1}; \boldsymbol{c}). \tag{10}$$

The first latent frame decodes to $\hat{\boldsymbol{x}}$, with reference $\boldsymbol{h}$, both in $[0, 1]$. Image gradients pass through the frozen decoder and final student update via truncated backpropagation (Clark et al., 2024); earlier trajectory states are treated as constants. Initial IQA uses this path.

Distillation objective. Equation (9) uses $\lambda_T = 0.5$. The teacher is evaluated without gradients at the same stored state as the student, supplying a local velocity prediction as a soft constraint on changes to the SFT field. Paired-reference and endpoint losses allow the student to adapt its short trajectory without being required to reproduce the teacher's outputs exactly.

Choice of supervision. On-policy specifies that training states come from the current student, independently of the teacher count or supervision objective. DMD (Yin et al., 2024b) matches teacher and student output distributions using a fixed teacher score and an additional score model fitted to current student outputs. DMD2 (Yin et al., 2024a) also uses student-generated intermediate states; on-policy sampling and distribution matching are compatible choices. Distribution objectives can be combined with paired video restoration, as in FlashVSR (Zhuang et al., 2025). Here, the paired clean video already provides a reconstruction target at each sampled state. We use this target with fixed teacher guidance, without fitting an additional model of the evolving student distribution.

The SFT teacher provides general restoration guidance during step reduction. Identity and text criteria are subsequently applied to the student's own reconstructions using corresponding reference crops. This directly optimizes the desired content constraints without first training separate face and text restoration teachers and transferring their generated predictions.

A.3 Degradation recipes

Initial SR degradation. Supervised training initially uses a mixture of degradation branches. Each clip is first encoded as an MPEG-4 intermediate and transcoded with H.264 or H.265 at CRF 25–30. It then receives either two rounds of image degradation (probability 0.70), one round (0.05), or a spatial CutMix of both (0.25), with a shared patch across frames. The image operations include unsharp masking, isotropic Gaussian blur, random resizing, and Gaussian noise. Image compression is approximated by bicubic downsampling and upsampling, without JPEG encoding. Final downsampling uses a factor of two or four; in the CutMix branch, the two regions can use different factors. All branches return to the reference pixel dimensions before VAE encoding.

Target $\times 4$ degradation. Supervised adaptation uses the benchmark's second-order $\times 4$ degradation protocol, also retained in subsequent distillation and refinement. Each of the two rounds applies blur, resize, noise, JPEG, and video compression, in that order. Final bicubic resizing reduces height and width to one quarter of the reference dimensions. Table 5 lists the round-specific settings. Some parameters vary across frames through clipped random walks; the table gives maximum absolute increments. Degraded pixels are quantized to eight bits. SFT, OPD, and initial IQA enlarge them bicubically before VAE encoding. Reconstruction and joint refinement retain the native LR video for their learned encoders and form the enlarged VAE condition as described in Appendix A.9.

Table 5: **Target $\times 4$ degradation parameters.** Gaussian noise standard deviations use $[0, 255]$ intensity units; Poisson scales are dimensionless. Each interpolation method is selected with probability $1/3$.

Setting	First round	Second round
Blur standard deviations, σ_x, σ_y	$[0.2, 3]$	$[0.2, 1.5]$
Resize up/down/keep probabilities	$0.2/0.7/0.1$	$0.3/0.4/0.3$
Resize scale bounds	$[0.15, 1.5]$	$[0.3, 1.2]$
Resize scale walk	0.015	0.03
Gaussian noise standard deviation	$[1, 30]$	$[1, 25]$
Poisson noise scale	$[0.05, 3]$	$[0.05, 2.5]$
JPEG quality	$[30, 95]$	$[30, 95]$
Interpolation choices	bilinear, area, bicubic	

Blur support is sampled uniformly from odd sizes between 7 and 21. The kernel-family probabilities for isotropic Gaussian, anisotropic Gaussian, generalized isotropic/anisotropic, plateau isotropic/anisotropic, and circular low-pass are respectively 0.405, 0.225, 0.108, 0.027, 0.108, 0.027, and 0.1. The circular cutoff range is $[\pi/3, \pi]$ for support below 13 and $[\pi/5, \pi]$ otherwise.

Gaussian and Poisson noise are chosen with equal probability, each with gray-noise probability 0.4. JPEG uses actual encoding and decoding. Both rounds apply video compression with H.264 (probability $2/3$) or MPEG-4 ($1/3$), at 25 fps with GOP 10 and up to two B-frames. The base bitrate is sampled from 10–100 kbit/s and scaled with resolution when the encoder area exceeds 256^2 pixels.

A.4 PERCEPTUAL AND RECONSTRUCTION REFINEMENT

Initial quality feedback. We start from the distilled model and retain its degradation, conditioning, and four-step sampling schedule. For the first decoded frame $\hat{\boldsymbol{x}}$ and reference $\boldsymbol{h}$ in $[0,1]$, the objective is

$$\begin{aligned}\mathcal{L}_{\mathrm{IQA}} &= \mathcal{L}_v + \lambda_{\mathrm{LF}}\,\mathrm{MSE}(B_{\mathrm{LF}}(\hat{\boldsymbol{x}}), B_{\mathrm{LF}}(\boldsymbol{h})) \\ &\quad + \lambda_{\mathrm{LPIPS}}\,\mathrm{LPIPS}(\hat{\boldsymbol{x}}, \boldsymbol{h}) + \lambda_{\mathrm{DISTS}}\,\mathrm{DISTS}(\hat{\boldsymbol{x}}, \boldsymbol{h}) + \mathcal{L}_{\mathrm{NR}}, \\ \mathcal{L}_{\mathrm{NR}} &= \sum_j w_j\big[1 - \tilde{q}_j(\hat{\boldsymbol{x}})\big].\end{aligned} \tag{11}$$

Here $\lambda_{\mathrm{LF}} = 5$ and $\lambda_{\mathrm{LPIPS}} = \lambda_{\mathrm{DISTS}} = 0.5$. The $\tilde{q}_j$ are CLIP-IQA, MUSIQ$/100$, and MANIQA, with respective weights $w_j = 0.2, 0.4$, and 0.4.

Here $\mathcal{L}_v$ is the paired-reference velocity MSE in Eq. (8), evaluated on one uniformly sampled student state over the full latent video, with coefficient 1. B_{LF} is a separable 7×7 Gaussian filter with standard deviation 1.5 and zero padding. It constrains low-frequency reconstruction over the entire frame; the quality predictors receive the decoded image without this filtering. This objective does not include OPD's velocity matching to the SFT teacher.

At every update, the student generates the trajectory without gradients and recomputes the final sampling update with gradients. Only the first latent frame is decoded for all image losses. The decoder output is clamped to $[-1,1]$ and rescaled to $[0,1]$. Gradients pass through the quality models and frozen decoder to the final student update; the earlier trajectory is treated as constant. The pretrained quality models remain fixed, and only the video model is optimized. LPIPS and DISTS use the paired reference; MUSIQ, MANIQA, and CLIP-IQA use only the generated image.

Relative reconstruction loss. Let $\hat{\boldsymbol{Y}}_i$ and $\boldsymbol{Y}_i^0$ be outputs of the current model and G_0, and let $\boldsymbol{H}_i$ be the clean video. Write $\bar{P}, \bar{S}, \bar{L}, \bar{D}$ for batch means of PSNR, SSIM, LPIPS, and DISTS against the reference, and $\bar{q}_j$ for CLIP-IQA, MUSIQ, and MANIQA when $j = 1, 2, 3$. Superscript 0 denotes scores of G_0. The reconstruction objective is

$$\begin{aligned}\mathcal{R}_{\mathcal{B}} &= \max\left\{\frac{\ln 10}{10}(\bar{P}^0 - \bar{P}),\, \ln\frac{1-\bar{S}}{1-\bar{S}^0},\, \ln\frac{\bar{L}}{\bar{L}^0},\, \ln\frac{\bar{D}}{\bar{D}^0}\right\}, \\ \mathcal{L}_{\mathrm{ref}} &= \mathcal{R}_{\mathcal{B}} + \frac{1}{3}\sum_{j=1}^{3}\gamma_j^{\mathrm{q}}\lambda_j(\bar{q}_j^0 - \bar{q}_j).\end{aligned} \tag{12}$$

The four entries in the maximum define $\mathcal{R}_m$ in Eq. (1), with $\Delta q_j = \bar{q}_j^0 - \bar{q}_j$. All means are computed over the full synchronized training batch with equal weight per video. For $1 - \mathrm{SSIM}$, LPIPS, and DISTS, averaging precedes the logarithm and maximum; each relative term is zero at the reference. Fixed positive scales γ_j^{q} and adaptive nonnegative multipliers λ_j weight the no-reference terms.

PSNR is averaged over frames before averaging over videos. For videos of T frames, let e_{it} and e_{it}^0 be the current and G_0 RGB mean squared errors of frame t in video i, after removing the four-pixel border. The PSNR contribution satisfies

$$\frac{\ln 10}{10}(\bar{P}^0 - \bar{P}) = \frac{1}{|\mathcal{B}|T}\sum_{i\in\mathcal{B}}\sum_{t=1}^{T}\ln\frac{e_{it} + 10^{-8}}{e_{it}^0 + 10^{-8}}. \tag{13}$$

Thus the PSNR term compares geometric means of stabilized per-frame MSE, whereas the other three terms compare arithmetic mean losses. The maximum retains negative values, so it continues to emphasize the least relative improvement after all four losses improve. Exact ties share the maximum's gradient equally. The maximum is over metric criteria, not individual videos; a parameter update can still affect the other criteria through the shared generator.

Quality constraints and multiplier updates. The quality predictors use raw scores, including unnormalized MUSIQ. Define the batch quality deficit as $\Delta q_j = \bar{q}_j^0 - \bar{q}_j$, with zero tolerance relative to G_0. At optimization iteration n, AdamW minimizes Eq. (12) over the trainable reconstruction

modules with the multipliers fixed. After each successful model update, the deficits evaluated before that update determine

$$\lambda_j^{(0)} = 1, \qquad \lambda_j^{(n+1)} = \max\Big\{0, \lambda_j^{(n)} + \rho_j \, \Delta q_j^{(n)}\Big\}. \tag{14}$$

For CLIP-IQA, MUSIQ, and MANIQA, respectively, the multiplier step sizes are approximately $(4.899,\ 0.04378,\ 7.261)$. The update uses the raw deficit Δq_j; the factors $\gamma_j^{\mathrm{q}}/3$ scale its contribution to the model objective. Signed deficits are retained: scores below the reference increase the corresponding multiplier, while scores above it reduce the multiplier toward zero. These are constraints on batch-average scores, not per-video guarantees.

Metric calibration. Before updating the reconstruction modules, we calibrate metric scales on the first eight synchronized training batches. Each frozen output $\boldsymbol{Y}_i^0$ is treated as an independent RGB variable. Let r_j, $j = 1, \ldots, 7$, be the raw full-frame measures in Eq. (40), ordered as $-\mathrm{PSNR}$, $1 - \mathrm{SSIM}$, LPIPS, DISTS, $-\mathrm{CLIP\text{-}IQA}$, $-\mathrm{MUSIQ}$, and $-\mathrm{MANIQA}$. Calibration uses these underlying measures before relative-loss or maximum aggregation; negating a score does not change its gradient norm. For the calibration videos $\mathcal{V}_{\mathrm{cal}}$, define

$$\gamma_j = \underset{i \in \mathcal{V}_{\mathrm{cal}}}{\mathrm{median}} \frac{\|\nabla_{\boldsymbol{Y}_i^0}[-\mathrm{PSNR}(\boldsymbol{Y}_i^0, \boldsymbol{H}_i)]\|_2}{\|\nabla_{\boldsymbol{Y}_i^0} r_j(\boldsymbol{Y}_i^0, \boldsymbol{H}_i)\|_2}, \tag{15}$$

The quality coefficients in Eq. (12) are $\gamma_j^{\mathrm{q}} = \gamma_{j+4}$ for $j = 1, 2, 3$, corresponding to CLIP-IQA, MUSIQ, and MANIQA. Its four reconstruction terms use their relative losses directly, without calibrated weights. We retain all seven scales for the full-frame quality constraints in joint optimization. Calibration does not update model parameters, and the scales remain fixed. These RGB-gradient scales do not imply balanced parameter gradients.

Trainable reconstruction modules. The IQA backbone and VAE are frozen. Self-attention query, key, value, and output projections in zero-indexed DiT blocks 10–29 receive LoRA updates (Hu et al., 2022) $\Delta W = (\alpha_L / r_L) B_L A_L$, with $r_L = \alpha_L = 16$, randomly initialized A_L, and zero-initialized B_L. A native-LR encoder uses a $3 \to 128$ channel Conv3d with kernel and stride $(4, 8, 8)$ followed by SiLU. Prepending three copies of the first frame maps 33 input frames to 9 temporal positions. Its spatial grid matches the DiT tokens. Separate zero-initialized $128 \to 3072$ projections add these features at the inputs of blocks 10, 20, and 29. The original VAE conditioning path remains active. These modules contain $9{,}151{,}616$ trainable parameters.

Sampling and image gradients. Frozen and current outputs share the same realized initial noise and conditioning, then follow separate four-step sampling trajectories. The learned modules are active at every current-model call. We backpropagate through all four updates, including their dependence on earlier predictions, and decode the full latent video with a frozen VAE. Block and decoder activation recomputation limits memory use without truncating this trajectory. Losses use the continuous output after range clamping and the per-frame AdaIN described in Appendix A.9; PNG quantization is excluded from the training graph.

PSNR averages per-frame RGB scores over all 33 frames; PSNR and SSIM exclude a four-pixel image border. The other six metrics use frames $\{(11i + n) \bmod 33 : i = 0, 1, 2\}$ at zero-indexed update n, and LPIPS uses AlexNet. These are full-frame spatial objectives, without an explicit temporal consistency term. The frozen reference is scored on the same frames and with the same metric random seed as the current output.

Reconstruction optimization. Training uses the general-video stream, the $\times 4$ degradation protocol, 33-frame spatial buckets, and empty prompts. Sampling uses the configuration in Appendix A.9. AdamW updates only the new FP32 parameters, with learning rate 10^{-4}, $(\beta_1, \beta_2) = (0.9, 0.999)$, weight decay 0.01, and gradient clipping at 1. The frozen backbone and VAE use bfloat16 computation. The exported generator retains the backbone and learned modules; inference executes one four-step trajectory without generating the frozen quality reference.

A.5 REFERENCE-CONDITIONED FACE SUPERVISION

This section specifies the face metric and its finite-view objective. All reference quantities are constructed separately for each GT face observation $o = (r, t)$, where r is a track and t is a frame; we suppress o below. Let c and $\widehat{c}$ denote the aligned GT and restored crops, $e(c) \in \mathbb{R}^d$ the unit embedding of the frozen FaceNet InceptionResnetV1 identity network (Schroff et al., 2015), pretrained on VGGFace2 (Cao et al., 2018), and $v(c)$ the frozen spatial appearance descriptor. Reference descriptors, covariance estimates, and gallery selections are held fixed during differentiation.

Shared geometry and perturbations. For each observation, a least-squares similarity transform maps the five GT landmarks to the ordered template

$$\frac{160}{112}\begin{bmatrix} 38.2946 & 51.6963 \\ 73.5318 & 51.5014 \\ 56.0252 & 71.7366 \\ 41.5493 & 92.3655 \\ 70.7299 & 92.2041 \end{bmatrix}. \tag{16}$$

The inverse transform samples both images on the same 160×160 crop grid, with pixel centers $(i + \frac{1}{2}, j + \frac{1}{2})$, bilinear interpolation, border padding, and `align_corners=False`. The grid depends only on the GT landmarks. We subsequently sample translated crops $T_\xi c$ by evaluating the crop at coordinates displaced by $\xi \in [-0.25, 0.25]^2$ pixels, using the same interpolation convention. Each translation component is drawn uniformly with a fixed CPU pseudorandom generator. The covariance uses $M_c = 8$ GT-only views with seed 1; the objective uses $M_{\mathrm{F}} = 2$ views with seed 0 and shares each translation between GT and restoration. The generator is reset for each call, so the seed and observation order determine the translations. The identity network receives RGB values mapped from $[0, 1]$ to $[-1, 1]$, and its output is normalized to unit length.

Shrinkage covariance and normalized whitening. Let $e_m^{\mathrm{cov}} = e(T_{\xi_m} c)$, $\bar{e} = M_c^{-1} \sum_m e_m^{\mathrm{cov}}$, and $u_m = e_m^{\mathrm{cov}} - \bar{e}$. We use the scaled-identity Ledoit–Wolf estimator (Ledoit & Wolf, 2004):

$$S = \frac{1}{M_c} \sum_{m=1}^{M_c} u_m u_m^\top, \qquad \nu = \frac{\mathrm{tr}(S)}{d},$$

$$\Delta_{\mathrm{LW}} = \left[\|S\|_F^2 - d\nu^2\right]_+, \qquad \beta_{\mathrm{LW}} = \left[\frac{M_c^{-1} \sum_m \|u_m\|_2^4 - \|S\|_F^2}{M_c}\right]_+, \tag{17}$$

$$\alpha_{\mathrm{sh}} = \begin{cases} 1, & \Delta_{\mathrm{LW}} \le \epsilon_{64} \|S\|_F^2, \\ \min\{1, \beta_{\mathrm{LW}} / \Delta_{\mathrm{LW}}\}, & \text{otherwise}, \end{cases}$$

$$C^{\mathrm{LW}} = (1 - \alpha_{\mathrm{sh}}) S + \alpha_{\mathrm{sh}} \nu I. \tag{18}$$

Here $[x]_+ = \max\{x, 0\}$ and ϵ_{64} is float64 machine precision. Centering and the singular value decomposition are computed in float64. The rank budget is 8, retaining the full empirical covariance rank, which is at most $M_c - 1 = 7$.

For $\nu > 0$, write $C^{\mathrm{LW}} = U \,\mathrm{diag}(\lambda_1, \ldots, \lambda_d) U^\top$ and set

$$\lambda_{\mathrm{floor}} = \max\left\{\alpha_{\mathrm{sh}} \nu,\ \nu \epsilon_{\mathrm{num}},\ \frac{\max_i \lambda_i}{100}\right\},$$

$$C = U \,\mathrm{diag}\big(\max\{\lambda_i, \lambda_{\mathrm{floor}}\}\big)_{i=1}^{d} U^\top, \tag{19}$$

$$s_W = \sqrt{\frac{\mathrm{tr}(C^{-1})}{d}}, \qquad W = \frac{C^{-1/2}}{s_W}, \qquad P = W^\top W. \tag{20}$$

Here ϵ_{num} is machine precision in the feature computation dtype (float32 unless the input embeddings are float64). If $\nu = 0$, we set $C = I$. Consequently, C is positive definite with condition number at most 100, and $\mathrm{tr}(P)/d = 1$. The working covariance associated with the actual whitening is $P^{-1} = s_W^2 C$. The eigenvalue floor and RMS normalization are additional numerical and scale choices. This covariance describes sensitivity to the prescribed translations of one GT observation. Probabilistic face embeddings (Shi & Jain, 2019) instead learn input-dependent diagonal uncertainty and compare embeddings by mutual likelihood; no uncertainty predictor is trained here.

Verified identity directions and the LLR interpretation. The positive reference is the unperturbed current-frame embedding $\mu_+ = e(c)$. For each gallery identity, we average its unit embeddings and normalize the mean to obtain a prototype. The training gallery uses ChokePoint faces with identity labels verified against the original dataset annotations, with at most four examples per identity selected at evenly spaced indices after sorting by sample and frame identifiers. Sources, identity labels, and frame hashes are checked against the designated exclusion sets, and repeated gallery frames are removed. Among gallery identities whose verified identity label differs from the current GT label, we retain the $J \le 8$ prototypes μ_j with largest $\mu_+^\top \mu_j$. Both construction and selection use fixed reference data; inference requires no gallery.

For $J > 0$, define

$$\begin{aligned} &\delta_j = \mu_+ - \mu_j, \qquad b_j = W\delta_j, \qquad d_j = P\delta_j, \\ &\kappa_j = \frac{\eta}{J(\|b_j\|_2^2 + \tau)}, \quad \eta = 1, \quad \tau = 10^{-6}. \end{aligned} \tag{21}$$

Under an equal-prior Gaussian working model with means μ_+ and μ_j and shared covariance P^{-1}, the positive-to-negative log-likelihood ratio is

$$\begin{aligned} \Lambda_j(e) &= \log \frac{\mathcal{N}(e; \mu_+, P^{-1})}{\mathcal{N}(e; \mu_j, P^{-1})} \\ &= d_j^\top e - \frac{1}{2}\left(\mu_+^\top P \mu_+ - \mu_j^\top P \mu_j\right). \end{aligned} \tag{22}$$

Its gradient with respect to e is d_j; increasing Λ_j supports the positive identity relative to identity j in this working model. The implemented identity descriptor and its pairwise distance are

$$\begin{aligned} h(e) &= \left[We; \sqrt{\kappa_1}\Lambda_1(e); \ldots; \sqrt{\kappa_J}\Lambda_J(e)\right], \\ \|h(e) - h(e')\|_2^2 &= (e - e')^\top G (e - e'), \qquad G = P + \sum_{j=1}^{J} \kappa_j d_j d_j^\top. \end{aligned} \tag{23}$$

The affine constants cancel in pairwise differences, establishing exact equivalence to the metric form in the main text. The objective matches these coordinates to GT descriptors; it does not maximize a likelihood ratio. If no verified negative identity is available, the sum is empty and $h(e) = We$, $G = P$.

The prescribed normalization bounds the additional metric. For $J > 0$ and every $q \in \mathbb{R}^d$, Cauchy–Schwarz gives

$$0 \le \sum_{j=1}^{J} \kappa_j (q^\top b_j)^2 \le \frac{\eta}{J} \sum_{j=1}^{J} \frac{\|b_j\|_2^2}{\|b_j\|_2^2 + \tau} \|q\|_2^2 \le \eta \|q\|_2^2. \tag{24}$$

Since $G = W^\top (I + \sum_j \kappa_j b_j b_j^\top) W$, congruence by W yields

$$P \preceq G \preceq (1 + \eta) P, \qquad G \succ 0. \tag{25}$$

This is a bound on squared feature distances for a fixed reference context. It does not bound the image-gradient norm or establish separation of identity changes from alignment perturbations.

Spatial appearance and finite-view MMD. Let $F_l(c) \in \mathbb{R}^{C_l \times H_l \times W_l}$ be layer l of the frozen AlexNet backbone used by LPIPS, after its prescribed input scaling. For L returned layers, define

$$\begin{aligned} \widetilde{F}_l(c)_{:,p} &= \frac{F_l(c)_{:,p}}{\max\{\|F_l(c)_{:,p}\|_2, 10^{-8}\}}, \\ v(c) &= \frac{1}{\sqrt{L}} \operatorname{concat}_{l=1}^{L} \frac{\operatorname{vec}(\widetilde{F}_l(c))}{\sqrt{H_l W_l}}. \end{aligned} \tag{26}$$

Spatial positions are retained in the flattened descriptor. The learned LPIPS output heads are not used. For $z = (e, v)$ and $z' = (e', v')$, the joint kernel is

$$k(z, z') = \frac{1}{3} \sum_{s \in \{0.5, 1, 2\}} \exp\left[-\frac{(e - e')^\top G (e - e') + \|v - v'\|_2^2}{2s^2}\right]. \tag{27}$$

The identity and appearance bandwidths are paired at the same scale, giving three terms. For the $M_{\mathrm{F}} = 2$ shared translations, let $\widehat{z}_m = (e(T_{\xi_m}\widehat{c}), v(T_{\xi_m}\widehat{c}))$ and $z_m = (e(T_{\xi_m}c), v(T_{\xi_m}c))$. With empirical measures $\widehat{P} = M_{\mathrm{F}}^{-1}\sum_m \delta_{\widehat{z}_m}$ and $\widehat{Q} = M_{\mathrm{F}}^{-1}\sum_m \delta_{z_m}$, the per-observation loss is

$$\begin{aligned}\ell_F &= \mathrm{MMD}_k^2(\widehat{P}, \widehat{Q}) \\ &= \frac{1}{M_{\mathrm{F}}^2}\sum_{m,n=1}^{M_{\mathrm{F}}} k(\widehat{z}_m, \widehat{z}_n) + \frac{1}{M_{\mathrm{F}}^2}\sum_{m,n=1}^{M_{\mathrm{F}}} k(z_m, z_n) \\ &\quad - \frac{2}{M_{\mathrm{F}}^2}\sum_{m,n=1}^{M_{\mathrm{F}}} k(\widehat{z}_m, z_n). \end{aligned} \tag{28}$$

This is the squared-MMD V-statistic, including its diagonal terms (Gretton et al., 2012). It is the exact discrepancy of the two empirical measures, rather than an unbiased estimator of a population discrepancy. Shared translations also couple the two view sets. Gradients pass through both restored arguments of the first term and the restored argument of the cross term; all GT arguments remain detached. The loss is computed separately for each observation. We average first over observations within each track, then over included tracks within a video, and finally over videos with face supervision in the synchronous batch. A missing track identifier makes that observation a singleton track. These averages use equal weights at each level and introduce no cross-frame kernel comparisons.

Properties for a fixed reference context. For fixed W, gallery prototypes, and weights, $G \succ 0$ makes Equation (27) a mixture of Gaussian kernels after the invertible coordinate map $(e, v) \mapsto (G^{1/2}e, v)$. Each component is characteristic, and the MMD squared of the mixture is the positive weighted sum of the component MMDs squared (Gretton et al., 2012). Thus, for probability measures ρ_1, ρ_2 on this joint descriptor space,

$$\mathrm{MMD}_k^2(\rho_1, \rho_2) \geq 0, \qquad \mathrm{MMD}_k^2(\rho_1, \rho_2) = 0 \iff \rho_1 = \rho_2. \tag{29}$$

For the empirical loss, $0 \leq \ell_F \leq 2$ in exact arithmetic because $0 < k \leq 1$; the implementation applies no nonnegative clamp, so roundoff may produce small negative values. Zero loss equates descriptor empirical measures, not necessarily pixels or true identities. The reference determines the covariance and kernel, so these are conditional statements for a fixed context. They do not establish an unconditional proper score for a latent identity posterior or guarantee reconstruction of information absent from the input.

A.6 TEXT SUPERVISION

Annotated crops. We supervise text with human or dataset transcriptions, retaining words explicitly annotated as legible whose complete quadrilaterals remain inside the transformed reference frame. Let $s^\star$ denote the transcription and a_0, a_1, a_2, a_3 its top-left, top-right, bottom-right, and bottom-left corners. We map a 32×128 sampling grid to the annotated quadrilateral by bilinear interpolation,

$$g(u, v) = (1-u)(1-v)a_0 + u(1-v)a_1 + uva_2 + (1-u)va_3, \tag{30}$$

where $u = (j + \frac{1}{2})/128$ and $v = (i + \frac{1}{2})/32$ index crop-pixel centers. Bilinear image sampling with this common grid extracts the predicted crop $\hat{c}$, an upsampled LR crop, and a crop from the output of the frozen general restorer F_0. The geometry is fixed during differentiation. Losses are averaged over observations within each annotated track and then over tracks.

Complete-string probabilities. We use frozen PARSeq (Bautista & Atienza, 2022) with causal decoding to evaluate strings. Let R denote its next-token conditional distribution. For $s = (s_1, \ldots, s_m)$, each token is conditioned on that string's own prefix, with BOS supplied initially and EOS appended as s_{m+1}:

$$p(s \mid \hat{c}) = \prod_{j=1}^{m+1} R(s_j \mid s_{<j}, \hat{c}), \qquad -\log p(s \mid \hat{c}) = -\sum_{j=1}^{m+1} \log R(s_j \mid s_{<j}, \hat{c}). \tag{31}$$

Candidate and target strings share the crop encoding and are evaluated with causal teacher forcing and their respective prefixes. PAD positions are excluded, whereas EOS contributes to the score.

We apply no length normalization. Training labels use the recognizer's 94 non-whitespace printable ASCII characters, with maximum length $M_{\mathrm{T}} = 25$. Unsupported or overlong labels are rejected rather than silently modified.

Fixed candidate graph. We construct candidates separately from the LR crop and the frozen F_0 crop. For each source, causal greedy decoding yields an initial string. We rank its token positions by the gap between the largest and second-largest next-token probabilities. In ascending gap order, we propose a substitution using the highest-probability different non-EOS token, followed by a deletion. If greedy decoding terminates before the length limit, we also propose appending its highest-probability non-EOS token. After deduplication, we retain the first eight proposals per source, including the greedy string, and take their union, deduplicated by token sequence, to form $\mathcal{C}$. This bounded proposal procedure produces at most sixteen candidates. The true transcription is evaluated for the log score even when absent from $\mathcal{C}$; it is not added to the graph.

Two candidates are connected when they differ by exactly one token substitution, insertion, or deletion. Let L_0 be the unit-edge graph Laplacian. We set

$$L_{\mathcal{C}} = \begin{cases} L_0/\lambda_{\max}(L_0), & \text{if the graph has an edge,} \\ 0, & \text{otherwise.} \end{cases} \tag{32}$$

This bounds the graph's spectral norm by one as its size and connectivity vary. Candidate proposals and graph construction use neither the transcription nor the current model's prediction. They do use the annotated crop geometry; their fixed context therefore includes this geometry and the sample-selection rule, as well as LR and F_0.

Absolute probabilities and residual mass. Let $p_{\mathcal{C}}$ collect the complete probabilities in Eq. (31), and let $e_{\mathcal{C}}(s^\star)$ have entry one at $s^\star$ if it belongs to $\mathcal{C}$ and zero elsewhere. Candidate probabilities are not renormalized within $\mathcal{C}$. For the recognizer's non-special token alphabet Σ, we account for all probability mass on the finite outcome space

$$\Omega = \bigcup_{m=0}^{M_{\mathrm{T}}} \Sigma^m \cup \{\perp\}, \qquad p(\perp \mid \hat{c}) = 1 - \sum_{s \in \Omega \setminus \{\perp\}} p(s \mid \hat{c}). \tag{33}$$

Each string outcome includes its terminal EOS; $\perp$ denotes failure to emit EOS by the last permitted position. The residual mass $1 - \sum_{s \in \mathcal{C}} p(s \mid \hat{c})$ contains all noncandidate strings and this overflow event. It is not an additional penalty: these outcomes have no graph edges. The implemented loss, with $\beta = 1$, is

$$\ell_T(p, s^\star) = -\log p(s^\star \mid \hat{c}) + \frac{\beta}{2}\big(p_{\mathcal{C}} - e_{\mathcal{C}}(s^\star)\big)^\top L_{\mathcal{C}}\big(p_{\mathcal{C}} - e_{\mathcal{C}}(s^\star)\big). \tag{34}$$

Conditional scoring property. Fix the supervision context Z, comprising LR, the frozen restoration, annotated crop geometry, and selection rule, so that $\mathcal{C}$ and $L_{\mathcal{C}}$ are fixed. Let q be the conditional distribution of accepted transcriptions on Ω, with $q(\perp) = 0$, and write $q_{\mathcal{C}}$ for its candidate coordinates. In the following, p denotes an arbitrary distribution on Ω, and the context is suppressed. Since $\mathbb{E}_{s\sim q} e_{\mathcal{C}}(s) = q_{\mathcal{C}}$, expansion of the quadratic gives

$$\begin{aligned} \mathbb{E}_{s\sim q}\big[\ell_T(p, s) - \ell_T(q, s)\big] &= \mathrm{KL}(q\|p) + \frac{\beta}{2}(p_{\mathcal{C}} - q_{\mathcal{C}})^\top L_{\mathcal{C}}(p_{\mathcal{C}} - q_{\mathcal{C}}) \\ &\geq 0. \end{aligned} \tag{35}$$

For $\beta \geq 0$, both terms are nonnegative because $L_{\mathcal{C}}$ is positive semidefinite. Thus the additional graph term preserves q as a population minimizer in distribution space. Equivalently, the excess risk is the Bregman divergence $D_\Phi(q, p)$ generated by the convex potential $\Phi(p) = \sum_{s\in\Omega} p_s \log p_s + \frac{\beta}{2} p_{\mathcal{C}}^\top L_{\mathcal{C}} p_{\mathcal{C}}$, following the proper-scoring construction of Gneiting & Raftery (2007).

The log score specifies the correct transcription, while the graph term additionally supervises probability differences among the proposed one-edit alternatives. The graph covers only these proposals, not every possible editing error. The conditional scoring property concerns distributions: a frozen recognizer composed with an SR model may not realize q, and its best attainable solution can depend on the added graph term. It therefore does not by itself guarantee human readability or the removal of all recognition errors.

A.7 OBJECT CONDITIONING

LR inputs and object descriptors. The regional branch reads only the native LR video, during both training and inference. Fixed MTCNN (Zhang et al., 2016) and CRAFT (Baek et al., 2019) detectors operate on every supplied frame. We retain face detections with confidence at least 0.85; MTCNN uses a minimum face size of 20 LR pixels and cascade thresholds $(0.6, 0.7, 0.85)$. CRAFT uses text and low-text thresholds 0.6 and 0.3, respectively, through the EasyOCR wrapper with its recognition network disabled. Boxes are clipped to the LR canvas, and free-form text polygons are represented by their axis-aligned bounds. A detected text box is assigned score 1; this is not recognition confidence.

For each object i, bilinear sampling maps its LR box to a 16×32 RGB crop. A shared encoder uses a 3×3 convolution with stride 2 and 32 channels, SiLU, flattening, a linear projection to 128 dimensions, and SiLU. A two-layer MLP maps 13 metadata values to the same dimension: the four normalized box coordinates, two sub-patch center coordinates, log aspect ratio, normalized frame index, position within its temporal group, a first-frame flag, two face/text indicators, and detection score. Their sum is the object descriptor z_i. The same encoders serve faces and text.

Temporal and spatial support. The LR evidence grid groups $4 \times 8 \times 8$ pixels in time and space. To match the causal VAE, frame 0 forms the first temporal group; frame $t > 0$ belongs to group $1 + \lfloor (t-1)/4 \rfloor$. Let b_i be an LR box and P_p a spatial patch in its temporal group. Writing d_i for the detection score and V_ℓ for a block-specific object projection, we use

$$\begin{aligned} w_{pi} &= d_i \, |b_i \cap P_p|, \qquad a_{pi} = \frac{w_{pi}}{\sum_q w_{qi}}, \\ u_p &= \sum_i a_{pi} z_i, \qquad \Delta h_p^{(\ell)} = V_\ell u_p. \end{aligned} \tag{36}$$

Only positive-overlap patches receive the descriptor. An object with zero total support contributes nothing. Otherwise $\sum_p a_{pi} = 1$: the object's total allocation is independent of its fraction of the full frame. The confidence factor cancels in this normalization when it is positive, although it remains an input to the metadata encoder. Contributions of different objects at the same patch are added. This normalization controls feature allocation, rather than ensuring equal parameter gradients or preserving all information in an unresolved LR object.

Injection into the shared generator. We concatenate u_p with the global LR evidence along the feature dimension, retaining the original DiT token grid. At blocks 10, 20, and 29 (zero-based indices), the bias-free linear map V_ℓ projects u_p to the hidden dimension. Its output $\Delta h_p^{(\ell)}$ is added to the hidden state at patch p before applying the global reconstruction adapter. The object projections are initialized to zero, so the initial regional model reproduces F_0. The inherited global adapter and the new object and metadata encoders and projections are jointly trainable; the original backbone remains frozen.

The direct object residual is zero outside its box support, but shared attention and adapter updates can affect the rest of the frame. Accordingly, object support does not replace the global constraints below. The generator uses LR detections even during training; annotated GT geometry is used only to define the loss crops. An LR detection miss does not remove an available annotated loss on the shared generator.

A.8 JOINT OPTIMIZATION

Reference and loss reduction. Let F_0 denote the completed general restorer from Section 3.1, including its reconstruction adapter. It is distinct from the intermediate quality-refined backbone G_0. A separate frozen copy of the F_0 adapter supplies the reference output; the current model F_θ starts from the same adapter and adds the object branch. Both outputs use the same LR input and initial noise, with matched generation and postprocessing settings.

For content type $k \in \{F, T\}$, let $\mathcal{B}_k$ contain videos in batch $\mathcal{B}$ with valid type-k supervision, $\mathcal{R}_{bk}$ their annotated tracks, and $\mathcal{T}_{br}$ the observations in a track. The content mean in Equation (5) is

$$\mathcal{L}_k(\theta; \mathcal{B}) = \frac{1}{|\mathcal{B}_k|} \sum_{b \in \mathcal{B}_k} \frac{1}{|\mathcal{R}_{bk}|} \sum_{r \in \mathcal{R}_{bk}} \frac{1}{|\mathcal{T}_{br}|} \sum_{t \in \mathcal{T}_{br}} \ell_{brt,k}. \tag{37}$$

Untracked observations each form a separate track. Videos without valid type-k supervision are excluded from $|\mathcal{B}_k|$; if none remain, we set $\mathcal{L}_k = 0$. We omit $(\theta; \mathcal{B})$ from batch content losses below for brevity. Region area does not weight these averages, and a longer track does not receive more total weight merely because it has more observations. Across workers, sums and valid-video counts are reduced separately. The local loss-gradient factor is the number of workers divided by the global valid count; subsequent averaging of parameter gradients gives the gradient of the stated video mean.

Region area and loss weighting. For an additive mean over N image pixels, a region R with $0 < A = |R| < N$ contributes

$$\frac{1}{N} \sum_{p=1}^{N} e_p = \frac{A}{N} \overline{e}_R + \frac{N - A}{N} \overline{e}_{\overline{R}}, \tag{38}$$

where e_p is a pixel error and the two regional means are taken inside and outside R. This illustrates area weighting for additive spatial averages; it is not a decomposition of every nonlinear quality measure. In Equation (37), an annotated observation instead has coefficient $1/(|\mathcal{B}_k|\,|\mathcal{R}_{bk}|\,|\mathcal{T}_{br}|)$, with no factor A/N. Its aggregation coefficient depends on supervision counts, not its fraction of the image. This removes explicit area weighting from the content objective; the actual gradient also depends on crop resampling and the feature and generator Jacobians.

Loss scales. The seven positive global scales satisfy $s_j = \gamma_j$ in Eq. (15), with the metric order fixed by Eq. (40). The two content scales are calibrated before regional updates on frozen F_0 outputs from training data. For each content type k, let $\mathcal{V}_{\mathrm{cal},k}$ contain one calibration video from each of at least four distinct original training sources. We set

$$\alpha_k = \mathrm{median}_{b \in \mathcal{V}_{\mathrm{cal},k}} \left. \frac{\|\nabla_{\boldsymbol{Y}} \left[-\mathrm{PSNR}(\boldsymbol{Y}, \boldsymbol{H}_b)\right]\|_2}{\|\nabla_{\boldsymbol{Y}}\, \ell_{b,k}(\boldsymbol{Y})\|_2} \right|_{\boldsymbol{Y} = F_0(X_b)}. \tag{39}$$

Here $\ell_{b,k}$ already includes the within-video reduction above, and PSNR uses all frames. Calibration differentiates with respect to the RGB output, without updating model parameters, and the resulting α_k remain fixed. The training stream then continues after these calibration samples. Matching these scales does not equate parameter gradient directions or guarantee equal influence for every example.

Reference-relative quality constraints. We write all seven full-frame measures in a lower-is-better convention:

$$\boldsymbol{r}(\boldsymbol{Y}, \boldsymbol{H}) = \big[-\mathrm{PSNR},\ 1 - \mathrm{SSIM},\ \mathrm{LPIPS},\ \mathrm{DISTS},\ -\mathrm{CLIP\text{-}IQA},\ -\mathrm{MUSIQ},\ -\mathrm{MANIQA}\big]^\top, \tag{40}$$

where the last three measures depend only on $\boldsymbol{Y}$. Let $\mathcal{G} = \{\mathrm{general}, \mathrm{face}, \mathrm{text}\}$ denote training-source groups. They are data-source categories, not spatial masks. For videos $\mathcal{B}_g$ from group g, the batch estimate is

$$\widehat{c}_{gj} = \frac{s_j}{|\mathcal{B}_g|} \sum_{b \in \mathcal{B}_g} \left[r_j(F_\theta(X_b), \boldsymbol{H}_b) - r_j(F_0(X_b), \boldsymbol{H}_b) \right]. \tag{41}$$

Here $c_{gj} = s_j \mathbb{E}_g[r_j(F_\theta(X), \boldsymbol{H}) - r_j(F_0(X), \boldsymbol{H})]$ is the corresponding source expectation. The target is $c_{gj} \le 0$ for all 3×7 source–metric pairs, with zero tolerance. Separate source constraints prevent improvements on one source from cancelling degradations on another within the same quality average. Full-frame PSNR uses all frames; the other six measures use the same set of up to three sampled frames for the current and reference outputs. The training objective uses unquantized RGB outputs.

Stochastic primal–dual updates. For groups present in a batch, we optimize

$$\widehat{\mathcal{J}}(\theta, \lambda) = \alpha_F \mathcal{L}_F + \alpha_T \mathcal{L}_T + \sum_{g \in \mathcal{G}} \sum_{j=1}^{7} \lambda_{gj} \widehat{c}_{gj}. \tag{42}$$

The frozen reference sets the constraint value but has zero derivative with respect to θ. Each positive multiplier thus weights the gradient of the current output's quality measure. Equality of current and reference scores does not switch off this gradient term: the signed residual enters Eq. (42) directly, without a hinge. Quality feedback therefore participates in learning the reconstruction even before a constraint is violated.

Absent groups have zero batch contribution. The multipliers start at $\lambda_{gj} = 1$ and are held fixed during an AdamW update of the shared and regional adapter parameters. After that update, the pre-update residuals produce

$$\lambda_{gj} \leftarrow \max(0, \lambda_{gj} + \rho\, \widehat{c}_{gj}), \qquad \rho = 0.01, \tag{43}$$

only for groups present in the batch. The repeating source schedule is general, face, general, text, with the same source type on all workers at a given update. The expectation defining $\mathcal{L}_k^{\mathrm{pop}}$ averages uniformly over the four cycle positions and over within-source sampling, including zero contributions from batches without valid type-k supervision. A violated constraint increases its multiplier; improvement decreases it toward zero. This procedure approximates the constrained training objective; it does not establish feasibility for every generated frame or for unseen data.

The optimizer is AdamW with learning rate 10^{-5}, $(\beta_1, \beta_2) = (0.9, 0.999)$ and weight decay 0.01; the synchronized parameter gradient is clipped to norm 1. Both content and global losses differentiate through all four sampling updates and the frozen VAE decoder. The fixed initial state, LR detections, reference outputs and supervision contexts are outside this gradient path. During inference, only the current full-video generator and its LR MTCNN/CRAFT detectors are needed; the frozen F_0 reference, recognition networks and identity gallery are training components.

A.9 Inference configuration

The final model uses four sampling steps, flow shift 5, guidance scale 1, empty text prompts, and conditioning noise in $[0, 0.1]$. Native LR pixels are bilinearly enlarged by four and replicate-padded to satisfy VAE and latent-patch divisibility. Any latent resizing uses nearest-neighbor interpolation. Per-frame AdaIN matches output color statistics to the enlarged LR condition before spatial cropping and eight-bit PNG export.

All frames are processed in nonoverlapping windows of up to 81 frames. The final window is padded by repeating its last frame to a length of $4m + 1$ for an integer $m \geq 0$, and excess outputs are removed. VAE encoding and decoding use spatial tiling for inputs larger than 1400^2 pixels; the DiT processes the full latent field.

B Evaluation details and additional results

B.1 Metric definitions and aggregation

General benchmark. Full-reference measures are PSNR, SSIM, LPIPS (Zhang et al., 2018), and DISTS (Ding et al., 2020); no-reference measures are CLIP-IQA (Wang et al., 2023a), MUSIQ (Ke et al., 2021), MANIQA (Yang et al., 2022), and NIQE (Mittal et al., 2013). Full-reference scores use RGB float images in $[0, 1]$. PSNR and SSIM crop four pixels from each border and do not convert to luminance; LPIPS uses AlexNet. Scores are averaged across valid frames within a clip and then across clips. Table 1 pools the 30 SPMCS, 10 UDM10, and 40 YouHQ-40 clips, assigning each clip the same weight; Table 6 reports these *clip-macro* scores separately by dataset. VideoLQ lacks paired reference frames: its 50 clips are averaged separately and contribute only no-reference metrics.

Reference-side subset selection. For each paired clip, the reference middle frame defines eligibility; face and text membership are determined independently and can overlap. The face-crop channel

uses InsightFace `buffalo_l`, detection threshold 0.3, and detector size 640×640. It chooses the face with the widest bounding box, expands width and height by a factor 1.3, and requires each clipped crop side to be at least 16 reference pixels. The tiny-face subchannel additionally requires the original box height divided by four to lie in $[12, 32)$. Thus the full face panel is not a 12–32 LR-pixel subset. Its 24 clips include 8 tiny-face clips. The 46 text-bearing clips have at least one nonempty reference PP-OCR token with confidence at least 0.85; this text subset has no size filter. These reference tokens are OCR-derived pseudo labels without manual filtering. These face and text subsets also informed model selection during development.

Crop metrics and detection confidence. The reference face ROI is applied at the same coordinates in each method output, with both crops resized to 256×256 before LPIPS evaluation. Detection confidence is instead the maximum score of any face detected in the full method frame, with zero for no detections; it is not constrained to match the reference ROI. These metrics are averaged over eligible clips. Outputs must match the reference dimensions; mismatched outputs are rejected before scoring.

Identity, landmarks, and face distribution. For IDS, AKD, detection rate, and FID-face, we select the widest detected face in each middle reference frame and retain the clip only if both original box dimensions are at least 12 reference pixels. This rule and the center-frame crop rule select the same 24 clips in this benchmark. For a clip with n frames, we take up to 24 frames using the first 24 entries of `frames[::max(1,n//24)]`. In each sampled frame, we again choose the widest detected face and retain the frame only if both original box dimensions pass the same size threshold. Both reference and restored frames undergo the same five-point similarity transform defined by reference landmarks, producing 112×112 crops for the InsightFace ArcFace recognition model (Deng et al., 2019). IDS is their embedding cosine similarity, averaged directly over the 450 available reference-selected frames without first computing per-clip means. It does not require method-side face detection, but still depends on reference detection.

Method-side detections are evaluated separately. The largest-IoU detection with IoU at least 0.3 is matched to the reference box. Det-rate is the fraction of readable reference-selected frames with such a match; AKD is the mean Euclidean error of the five corresponding landmarks only on matched frames, in reference-grid pixels without face-size normalization. They must be read together. FID-face uses the same aligned face images but InceptionV3 features, not identity embeddings: clean bicubic resizing to 299×299, $2{,}048$-dimensional features, and Fréchet distance between empirical means and covariances. It is a descriptive distribution score on correlated video frames.

Token coverage and string similarity. For ASCII strings, normalization uppercases and removes non-alphanumeric characters; for non-ASCII strings it removes whitespace. Reference PP-OCR tokens are deduplicated within each frame of the text subset. For an engine's predicted string set P and reference tokens G, the reported measures are

$$\begin{aligned} \mathrm{TC} &= \frac{1}{|G|} \sum_{g \in G} \mathbf{1}[g \in P], \\ \mathrm{SS} &= \frac{1}{|G|} \sum_{g \in G} \max_{p \in P} s(g, p). \end{aligned} \tag{44}$$

Here s is Python `SequenceMatcher.ratio()`, additionally maximized with the shorter/longer length ratio when one string contains the other; exact equality gives 1 and an empty prediction set gives 0. TC is exact token coverage, and SS is string similarity. Neither performs spatial matching, one-to-one assignment, or false-positive penalization. SS uses sequence similarity rather than Levenshtein edit distance.

On the text subset, PP-OCR runs through PaddleOCR's Chinese-language configuration and EasyOCR uses its English configuration. Both are evaluated against the same PP-OCR-derived reference tokens. Predicted nonempty strings require confidence at least 0.3. Each frame's token average is computed first, then these scores are averaged across clips. OCR confidence is the mean confidence of all predictions above that threshold, with zero for no predictions, followed by a clip mean; it is not conditioned on matching any reference token.

B.2 Additional qualitative comparisons

Figures 5 and 6 extend Figure 4 with additional face and scene-text examples, respectively. The enlarged crops show differences in facial contours and texture, and in character shapes and stroke separation.

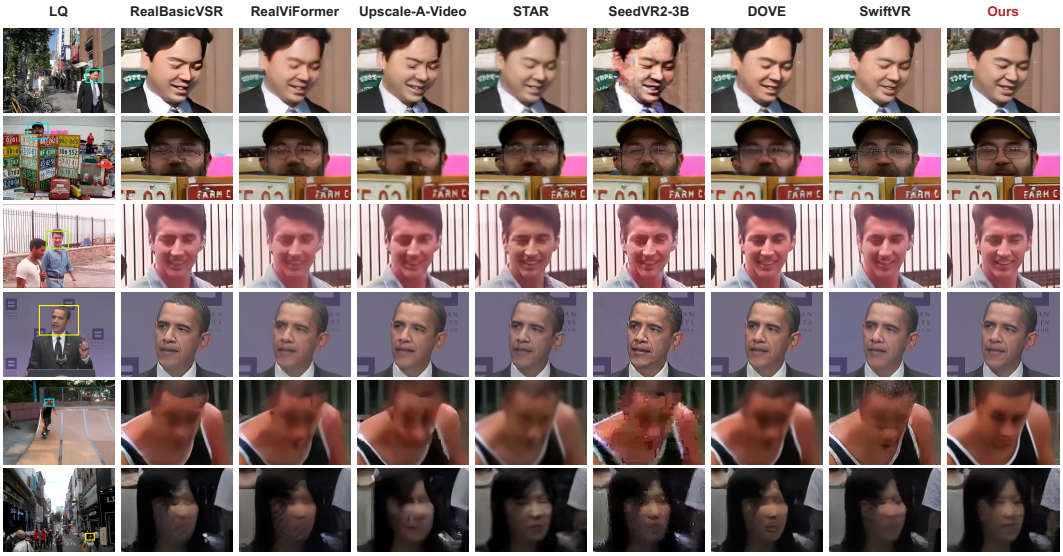

Figure 5: **Additional qualitative comparisons of face restoration.** The first column shows LQ input frames, with colored boxes marking the face regions enlarged in the remaining columns. Each row compares the same region across methods.



Figure 6: **Additional qualitative comparisons of scene-text restoration.** The first column shows LQ input frames, with colored boxes marking the text regions enlarged in the remaining columns. Each row compares the same region across methods.

B.3 Per-dataset general quality

Table 6 provides the per-dataset results underlying the summary in Table 1, with the same methods and metrics.

Table 6: **General video super-resolution by dataset.** Scores average clips within each dataset. Ranking notation follows Table 1.

Set	Metric	R-ESR	RBV	RVF	UAV	VE	STAR	DOVE	SVR2	FVSR	SwVR	**Ours**
SPMCS	PSNR ↑	22.29	22.82	22.87	21.42	21.41	21.94	**23.76**	23.02	22.11	22.71	23.06
	SSIM ↑	0.5829	0.5890	0.5850	0.5370	0.5781	0.5614	**0.6450**	0.6294	0.5916	0.5908	0.6302
	LPIPS ↓	0.3887	0.3805	0.3790	0.4243	0.5752	0.4775	0.3147	0.3042	**0.2825**	0.3052	0.2857
	DISTS ↓	0.2278	0.2250	0.2320	0.2489	0.3133	0.2611	0.1822	0.1733	0.1682	**0.1657**	0.1738
	CLIP-IQA ↑	0.4581	0.4315	0.3781	0.4380	0.2597	0.2848	**0.5385**	0.4756	0.5189	0.5110	0.5054
	MUSIQ ↑	63.61	64.70	62.18	60.62	33.02	41.19	68.41	62.12	70.11	**71.33**	68.40
	MANIQA ↑	0.3537	0.3125	0.2964	0.3241	0.2184	0.2364	0.3767	0.3550	0.3862	0.3792	**0.4241**
	NIQE ↓	3.848	3.621	4.072	4.119	8.097	5.808	5.182	4.474	3.736	**3.494**	4.665
UDM10	PSNR ↑	23.37	23.71	24.26	23.42	22.65	23.56	**25.35**	23.31	23.03	23.86	23.87
	SSIM ↑	0.6782	0.6728	0.6889	0.6593	0.6846	0.6830	**0.7154**	0.6550	0.6765	0.6834	0.7015
	LPIPS ↓	0.4410	0.4376	0.4132	0.4250	0.5261	0.4674	0.3575	0.4350	0.3655	0.3652	**0.3382**
	DISTS ↓	0.2573	0.2446	0.2442	0.2426	0.3364	0.2651	0.2366	0.2757	0.2329	**0.2029**	0.2036
	CLIP-IQA ↑	0.3881	0.4177	0.3597	0.3637	0.2683	0.2500	0.4302	0.3062	0.4611	**0.5464**	0.5401
	MUSIQ ↑	51.42	57.09	53.39	53.25	31.59	37.67	52.39	38.62	58.52	65.62	**68.22**
	MANIQA ↑	0.3061	0.2931	0.2665	0.2610	0.2473	0.2156	0.3064	0.2011	0.3380	0.3511	**0.4229**
	NIQE ↓	4.801	**3.936**	4.654	5.079	9.585	7.392	6.338	5.277	4.735	4.222	5.118
YouHQ-40	PSNR ↑	23.77	23.34	24.16	23.44	21.77	23.49	**25.22**	22.84	22.87	23.40	24.09
	SSIM ↑	0.6492	0.6149	0.6386	0.6195	0.6088	0.6426	**0.6941**	0.6156	0.6243	0.6247	0.6723
	LPIPS ↓	0.3745	0.4010	0.3824	0.3875	0.4563	0.4197	0.3079	0.3553	0.2772	0.2972	**0.2740**
	DISTS ↓	0.1853	0.1974	0.2065	0.1914	0.2239	0.2052	0.1549	0.1699	**0.1215**	0.1241	0.1436
	CLIP-IQA ↑	0.4005	0.4824	0.4259	0.4189	0.3389	0.2941	0.4372	0.4282	0.5380	**0.5812**	0.5533
	MUSIQ ↑	54.25	64.37	62.04	54.71	42.57	36.92	61.03	54.27	67.65	**69.61**	68.58
	MANIQA ↑	0.2866	0.2958	0.2876	0.2620	0.2596	0.2333	0.3043	0.2764	0.3584	0.3551	**0.4052**
	NIQE ↓	4.446	**3.207**	3.688	4.369	6.557	6.637	5.122	4.199	3.423	3.306	4.333
VideoLQ	CLIP-IQA ↑	0.3619	**0.4036**	0.3466	0.2823	0.2955	0.2568	0.2911	0.2524	0.4002	0.3792	0.3484
	MUSIQ ↑	49.85	**56.12**	52.14	44.67	34.76	34.31	44.80	41.42	52.14	55.25	53.29
	MANIQA ↑	0.2986	0.3084	0.2844	0.2454	0.2465	0.2403	0.2628	0.2256	0.2987	0.2772	**0.3115**
	NIQE ↓	4.204	**3.585**	4.055	4.727	6.242	6.279	5.333	4.405	3.936	4.031	4.831

B.4 Ablation protocol

The comparisons in Sec. 4.3 use the completed general restorer F_0 as a common initialization and keep the face, text, and general training streams fixed. Loss replacements retain the same annotations and source exposure. Calibration uses the same training examples and rule for each variant; its resulting coefficients are then fixed, with the subsequent training sequence held constant. The update budget, endpoint rule, degradation, initial noise, and inference configuration are shared across all models refined from F_0.

Direct face supervision. Using the $M_{\mathrm{F}} = 2$ paired translated views in Appendix A.5, let $(\hat{e}_m, \hat{v}_m)$ and (e_m, v_m) denote the restored and GT identity and appearance descriptors for view m. The direct alternative is

$$\ell_F^{\mathrm{direct}} = \frac{1}{M_{\mathrm{F}}} \sum_{m=1}^{M_{\mathrm{F}}} \left[2\big(1 - \hat{e}_m^{\top} e_m\big) + \|\hat{v}_m - v_m\|_2^2 \right]. \tag{45}$$

The identity term equals $\|\hat{e}_m - e_m\|_2^2$ because both embeddings are unit normalized. The crops, frozen networks, descriptor normalization, and paired translations are unchanged. This replaces the reference-conditioned joint kernel objective as a whole; it does not remove appearance supervision. The spatial term uses backbone descriptors, not the learned scalar LPIPS heads.

Text and quality controls. The text control sets $\beta = 0$ in Eq. (4), retaining the same transcriptions, recognizer, causal prefixes, and EOS term. The quality control fixes $\lambda_{gj} = 1$ for every

source–metric pair and omits the dual update, while retaining all seven full-frame objectives and their calibrated scales. It tests adaptive feedback against fixed weighting, rather than removing full-frame supervision.

Object representation and area weighting. The object control disables only the additional object descriptors and their residual projections from the beginning of training. The inherited global LR encoder, restoration adapter, and reference-defined content losses remain active. The area-weighted control instead keeps the full architecture and replaces each observation loss $\ell_{brt,k}$ in Eq. (5) by $\rho_{brt,k}\ell_{brt,k}$, where $\rho_{brt,k} = A_{brt,k}/(H_b W_b)$. Here A is the annotated face-box or text-quadrilateral area on the training reference canvas, before canonical crop resizing, and $H_b W_b$ is that canvas's area. The track and video denominators are unchanged. Task-level calibration controls the overall loss scale while retaining the relative area weights.

Evaluation. The general PSNR and LPIPS columns in Table 3 first average frames within each paired clip and then average all 80 clips across SPMCS, UDM10, and YouHQ-40. This is a pooled clip mean, not an equal average of the three dataset means. IDS, small-face LPIPS, and PP-OCR/EasyOCR token coverage retain the exact evaluation populations and aggregation rules of Table 2. Each variant is assessed on all three aspects.