MATCHVR: BRIDGING THE REFERENCE UTILIZATION GAP IN GENERATIVE VIDEO RESTORATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative video restoration can produce abundant details from severely degraded video, but its strong prior may hallucinate textures inconsistent with the true scene, especially when severe degradation removes reliable high-frequency evidence. Reference-guided restoration reduces this by anchoring reconstruction to real textures from high-quality reference frames, yet reference *availability* does not imply reference *utilization*: simply appending reference frames to a video diffusion transformer leaves much of the reference unused, even when reliable correspondences still exist. To address this gap, we present ***MatchVR***, a one-step reference-guided generative video restoration framework that converts reference content into frame-specific matched guidance. Specifically, we introduce a degradation-robust feature matching strategy to estimate dense target-reference correspondences and their confidence. Additionally, we propose a spatiotemporal reference alignment module to warp and relocate the matched content to each target frame. We further introduce aligned-reference distillation to transfer full-reference teacher behavior. Experiments show that MatchVR consistently improves reference utilization and restoration quality across varying temporal distances, with particularly strong gains in fidelity.

1 INTRODUCTION

Generative video restoration has recently shown strong ability to recover realistic, sharp details from severely degraded videos. By leveraging powerful video diffusion priors, methods such as VividVR (Bai et al., 2025), SeedVR (Wang et al., 2025b), SeedVR2 (Wang et al., 2025a), and FlashVSR (Zhuang et al., 2025) can produce temporally coherent and perceptually appealing results. However, since the same degraded video may correspond to multiple plausible restoration results, the strong generative prior may synthesize visually reasonable but scene-inconsistent textures, identities, or local structures, yielding unfaithful hallucinated results to the original content.

Reference-guided restoration provides a natural way to reduce such hallucination by anchoring reconstruction to real textures. In image restoration, reference-based methods show that an external high-quality image can guide texture recovery through explicit matching and transfer. Extending this idea to video restoration is appealing because high-quality frames from the same video, the same scene, or an external restoration model can serve as real texture anchors for multiple degraded target frames. Recent reference-guided video restoration, such as SparkVSR (Yu et al., 2026), explores sparse keyframes as controllable references. In this setting, the task is no longer to guess missing details from the degraded input alone, but to selectively propagate trustworthy reference details for more faithful reconstruction.

However, our study identifies a *reference utilization gap* in existing reference-guided generative restoration methods: providing a useful reference does not ensure that the generative backbone can effectively exploit it. Existing methods typically concatenate target and reference tokens and rely on native self-attention to retrieve relevant details. As shown in Figure 1, restoration quality progressively deteriorates as the target moves farther from the reference, even when dense matching can still identify reliable correspondences. This degradation arises because target-to-reference attention weakens with temporal distance, while spatial misalignment further prevents corresponding textures from being directly accessed. Consequently, the model may fall back to its learned generative prior despite the presence of real, matchable reference content, increasing the risk of hallucinated details.

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.

In contrast, the idealized full-reference setting maintains consistently high performance, indicating that the observed degradation is not merely caused by reduced visual overlap, but by the insufficient utilization of available reference content.

To address this problem, we propose MatchVR, which actively constructs an aligned reference for each target frame instead of leaving reference use to implicit attention. The key idea is to first establish degradation-robust correspondences between the low-quality target and the reference through degradation-robust feature matching, and then warp the matched reference content in the latent space into the spatial layout of each target frame, yielding a per-frame aligned reference. The aligned reference is further relocated to the corresponding target position in time and made accessible only to that frame, thereby bypassing the temporal attention decay and spatial misalignment associated with distant references. Failed or unreliable matches are excluded from the aligned guidance, preventing incorrect reference textures from affecting the restoration. This design turns a single reference frame into frame-specific guidance that is both spatially aligned and temporally local.

Our contributions are summarized as follows:

- We reveal a *reference utilization gap* in reference-guided generative video restoration: generative backbones tend to neglect useful reference content as the temporal distance from the target frame increases.
- We propose MatchVR, a reference-guided video restoration framework that recasts reference guidance as a spatiotemporal reference alignment problem, introducing a degradation-robust matching strategy to identify reliable reference content.
- We introduce a per-frame aligned reference mechanism that warps matched reference latents into each target layout and distills full-reference teacher behavior to a single-reference student.
- Extensive experiments demonstrate that MatchVR achieves strong performance on both full-reference and no-reference metrics, with especially strong fidelity gains under high-quality references.

2 Related Work

2.1 Blind Video Restoration

Traditional video restoration methods mainly recover high-quality frames from the low-quality video itself. Representative methods improve temporal aggregation through alignment, propagation, and recurrent modeling, including EDVR (Wang et al., 2019), BasicVSR (Chan et al., 2021), and RealBasicVSR (Chan et al., 2022). Transformer-based methods further model long-range temporal dependencies, such as VRT (Liang et al., 2024) and RVRT (Liang et al., 2022). These methods are effective at suppressing degradations and exploiting neighboring frames, but they remain blind restoration systems in nature: when the degraded input lacks sufficient high-frequency evidence, especially under severe degradation, they must infer missing details from learned priors.

2.2 Diffusion-Based Video Restoration

Recent diffusion-based video restoration models significantly improve perceptual realism. SeedVR (Wang et al., 2025b) introduces a diffusion transformer for generic video restoration, and SeedVR2-3B (Wang et al., 2025a) further pursues one-step restoration through diffusion adversarial post-training. Vivid-VR (Bai et al., 2025) distills concepts from text-to-video diffusion transformers for photorealistic video restoration, while FlashVSR-Full (Zhuang et al., 2025) explores real-time diffusion-based streaming video restoration. These methods demonstrate the power of generative video priors, but strong priors may also synthesize plausible yet incorrect textures when the input is ambiguous, particularly in severely degraded regions where little faithful evidence remains.

2.3 Reference-Guided Image and Video Restoration

Reference-based image restoration transfers high-quality textures from external references. Existing methods develop diverse matching and transfer mechanisms: SRNTT (Zhang et al., 2019) performs

neural texture transfer, TTSR (Yang et al., 2020) uses transformer-based matching, MASA-SR (Lu et al., 2021) improves matching efficiency and spatial adaptation, C2-Matching (Jiang et al., 2021) models robust correspondences, TDF (Huang et al., 2022) decouples matching and reconstruction, and DATSR (Cao et al., 2022) uses deformable attention. Reference guidance has also been studied in non-diffusion video restoration: RefVSR (Lee et al., 2022) recurrently aligns and propagates reference features, ERVSR (Kim et al., 2023) uses a single reference image to guide an entire video, and RefVSR++ (Zou et al., 2023) improves reference utilization through parallel temporal aggregation. More recently, generative approaches such as SparkVSR (Yu et al., 2026) use sparse keyframes as controllable references. But generative methods may underuse reference content. MatchVR addresses this with degradation-robust matching and per-frame aligned guidance.

3 Motivation: Reference Utilization Gap

Reference-guided generative video restoration aims to import real high-frequency textures from one or more reference frames while preserving the target structure. Existing mechanisms simply concatenate target and reference tokens into one self-attention sequence and let the generative backbone borrow reference textures (Tan et al., 2025; Wan et al., 2025; Peebles & Xie, 2023); this preserves the pretrained prior but implicitly treats a reference being *available* as if it were automatically *used*.

To separate reference under-utilization from missing correspondence, we build the diagnostic curves in Figure 1 on a low-motion SPMCS subset without obvious occlusion, where distant target frames still contain visible content that can be matched to the reference. Figure 1(a) compares single-reference restoration with an idealized full-reference setting, where each target frame is paired with its own high-quality GT reference to maximize reference utilization. The three single-reference curves correspond to using the first, middle, and last frame as the only reference, respectively, and report the per-frame PSNR of the restored video. If available reference content were fully exploited, the single-reference curve would stay closer to this idealized setting. Instead, its quality drops as the target moves away from the reference, indicating that matchable reference information remains available but under-used.

A plausible mechanism behind this under-utilization is limited direct access to distant reference tokens in the transformer's self-attention. For a target token i with query q_i and a reference token j with key k_j, the attention weight is

$$A_{ij} = \frac{\exp(q_i^\top k_j/\sqrt{d})}{\sum_m \exp(q_i^\top k_m/\sqrt{d})}, \quad (1)$$

where d is the feature dimension and A_{ij} is the fraction of attention that token i pays to token j. Figure 1(b) visualizes the resulting target-to-reference attention distribution over video time, averaged over queries, heads, layers, and videos. Under the controlled low-motion setting above, the attention to the chosen reference peaks at zero temporal distance and is weaker for distant frames. When attention values become similar beyond a moderate distance, restoration can still keep degrading because later frames increasingly rely on intermediate frames that are already reference-limited, causing errors to accumulate instead of directly using the original high-quality reference.

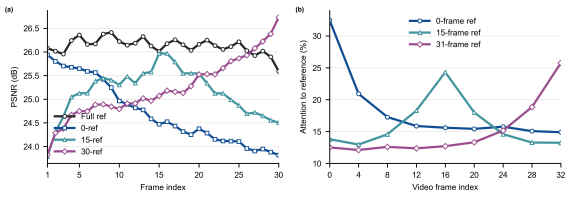

Figure 1: Evidence for the reference utilization gap. (a) An idealized full-reference setting stays high and flat, while single-reference curves degrade away from their reference frame. (b) Target-to-reference attention is the attention mass assigned to reference tokens, averaged over queries, heads, layers, and videos; it peaks near the reference frame.

The weakened access to distant reference tokens can be understood through two factors. Object and camera motion steadily reduce the genuine overlap between a distant frame and the reference, imposing an *availability ceiling* that is hard for any reference mechanism to surpass. Yet even under strong motion a distant frame still shares some *still-matchable* regions with the reference; spatial misalignment together with the temporal bias steers attention away from them, leaving a *utilization gap* that is, in principle, addressable.

Our goal is not to eliminate the performance decline with temporal distance, but to raise the entire curve by narrowing the utilization gap at every distance. Given a single reference frame, we achieve this by redirecting each distant frame's attention to the regions that remain matchable with the reference.

4 METHOD

In this section, we introduce our video restoration method, MatchVR, which uses high-quality reference to guide one-step generative video restoration. MatchVR uses *dense matching* results to construct latent-space *aligned reference guidance* for restoring each target frame.

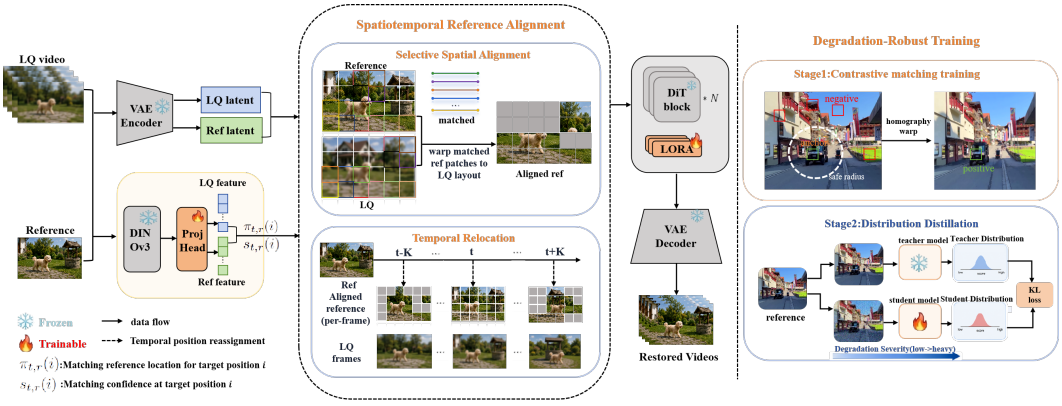

Figure 2: Overview of MatchVR. The left-to-right pipeline encodes the LQ video and reference into latents while performing pixel-domain dense matching in parallel; the two branches meet at reference alignment to build aligned reference guidance for the DiT restoration backbone. The two right panels illustrate the two-stage matcher training: contrastive matching and distribution distillation.

4.1 FRAMEWORK OVERVIEW

As shown in Figure 2, MatchVR takes a low-quality video $V^{\mathrm{lq}} = \{I_t^{\mathrm{lq}}\}_{t=1}^{T}$ and high-quality reference frame $\{I_r^{\mathrm{ref}}\}_{r=1}^{R}$ as input, and outputs a restored high-quality video $\hat{V}$. During inference, the ***degradation-robust feature matcher*** first estimates content correspondences and confidence maps between reference frames and degraded target frames, identifying which reference regions remain reliable. Guided by these correspondences, the ***spatiotemporal reference alignment*** mechanism resamples reliable reference content into each target frame's latent layout and feeds the resulting aligned reference together with the low-quality latent into the DiT denoising backbone, which predicts the restored latent in a single step and decodes it into $\hat{V}$. For training, the restoration backbone is learned with a two-phase strategy, comprising reference-guided baseline adaptation and aligned-reference distillation, while the matcher is learned with a two-stage strategy, comprising contrastive matching training and degradation-robust distillation.

4.2 DEGRADATION-ROBUST FEATURE MATCHING

Dense Feature Matching. The matcher estimates dense pixel-domain correspondences from degraded target frames to reference frames. Given a low-quality target I_t^{lq} and a reference I_r^{ref}, indexed by t and r. A frozen DINOv3 backbone ϕ extracts visual features, which are mapped by a trainable projection head ψ into dense descriptor maps: $F_t^{\mathrm{lq}} = \psi(\phi(I_t^{\mathrm{lq}}))$ and $F_r^{\mathrm{ref}} = \psi(\phi(I_r^{\mathrm{ref}}))$. For target location i and reference location j, descriptors $f_i^{\mathrm{lq}} \in F_t^{\mathrm{lq}}$ and $f_j^{\mathrm{ref}} \in F_r^{\mathrm{ref}}$ are matched by cosine similarity,

$$S_{ij}^{t,r} = \frac{(f_i^{\mathrm{lq}})^{\top} f_j^{\mathrm{ref}}}{\|f_i^{\mathrm{lq}}\|_2 \, \|f_j^{\mathrm{ref}}\|_2}, \tag{2}$$

where $S_{ij}^{t,r}$ is the similarity for frame pair (t, r). We take the best reference location and its score as the correspondence and confidence,

$$\pi_{t,r}(i) = \arg\max_j S_{ij}^{t,r}, \qquad s_{t,r}(i) = \max_j S_{ij}^{t,r}, \tag{3}$$

where $\pi_{t,r}(i)$ matches target location i to the reference and $s_{t,r}(i)$ is its confidence. Over all target locations, they form the correspondence matrix $\Pi_{t,r}$ and confidence score matrix $C_{t,r}$; when unambiguous, we write S_{ij}, $\pi_t(i)$, and $s_t(i)$.

Degradation-Robust Training. The matcher is trained to be accurate on clean geometry and robust to degradation. (1) In contrastive matching training, a homography-warped high-quality pair provides positives (i, j^+), where j^+ is the true match of anchor i, and hard negatives j^- are sampled outside a safe radius. For unit-normalized descriptors u and v, $d(u, v) = 2 - 2\,u^\top v$, and the projection head is optimized by

$$\mathcal{L}_{\text{tri}} = \frac{1}{|\Omega|} \sum_{(i,j^+)\in\Omega} \max\big(0,\ m + d_i^+ - d_i^-\big)\,, \tag{4}$$

where Ω is the set of valid positives, m is the margin, and d_i^+ and d_i^- are the positive and hardest-negative distances (Jiang et al., 2021). (2) In degradation-robust distillation, the frozen teacher matches a high-quality target-reference pair, while the student matches the degraded target with the same reference. For target location i, let $S_i^{\text{hq,ref}}$ and $S_i^{\text{lq,ref}}$ be the teacher and student similarity rows, and τ the softmax temperature; the distributions $P_i^T = \text{softmax}(S_i^{\text{hq,ref}}/\tau)$ and $P_i^S = \text{softmax}(S_i^{\text{lq,ref}}/\tau)$ are aligned by

$$\mathcal{L}_{\text{match}} = \sum_i \text{KL}\big(P_i^T \,\|\, P_i^S\big)\,, \tag{5}$$

where KL denotes the Kullback–Leibler divergence. A curriculum with increasing degradation, temporal offset, and distractor hardness further improves robustness.

4.3 Spatiotemporal Reference Alignment

Selective Spatial Alignment. The aligned reference is constructed by placing reliable reference content onto the target latent grid. For target frame t and target latent location i, the Degradation-Robust Feature Matching module provides a correspondence $\pi_t(i)$ and confidence $s_t(i)$. We resample the reference latent z^{ref} at the matched location and keep only confident matches,

$$A_t(i) = M_t(i)\, z^{\text{ref}}(\pi_t(i))\,, \qquad M_t(i) = \mathbf{1}[s_t(i) \geq \eta]\,, \tag{6}$$

where $A_t(i)$ is the aligned reference token for target location i, $M_t(i)$ is the binary validity mask, $\mathbf{1}[\cdot]$ is the indicator function, and η is the confidence threshold. Low-confidence locations with $M_t(i) = 0$ are not exposed to the target tokens: their attention-mask entries are set to invisible, so unreliable reference content is not injected into unmatched target regions.

Temporal Relocation. The aligned reference is injected as target-time guidance rather than as a distant reference frame. Let p_i denote the spatial coordinate of target location i and t its temporal index; after spatial alignment, the token $A_t(i)$ is assigned the relocated position

$$\text{pos}(A_t(i)) = (t, p_i), \tag{7}$$

where $\text{pos}(\cdot)$ denotes the spatiotemporal position used by the DiT backbone. Thus the content of $A_t(i)$ comes from the matched reference location $\pi_t(i)$, but its position encoding is moved to the target location (t, p_i). Equivalently, the aligned reference token $A_t(i)$ and its paired LQ latent token $z_t^{\text{lq}}(i)$ share the same target spatial-temporal position, while the original reference time is not kept for $A_t(i)$. The final denoising sequence is arranged as

$$Z = \Big[\, z_1^{\text{lq}}, \ldots, z_T^{\text{lq}},\ A_1, \ldots, A_T \,\Big]\,, \tag{8}$$

where z_t^{lq} denotes the low-quality latent of frame t and A_t collects all aligned reference tokens for that frame. In the attention mask, A_t is visible only to its paired target frame t, ensuring that each aligned reference provides local guidance exclusively to its corresponding target frame.

4.4 Reference Distillation Training Strategy

Phase 1: Reference-Guided Baseline. Throughout training, we directly optimize the inserted LoRA parameters for one-step video restoration rather than relying on multi-step diffusion sampling. In the first phase, we train the LoRA with ordinary reference tokens, without using the aligned reference. Given the restored video $\hat{x}$ and the ground truth video x, the model is supervised by a pixel-perceptual objective,

$$\mathcal{L}_{\text{rec}} = \lambda_1 \|\hat{x} - x\|_1 + \lambda_p \,\text{LPIPS}(\hat{x}, x), \tag{9}$$

where λ_1 and λ_p weight the ℓ_1 and LPIPS terms. This phase gives the LoRA-adapted backbone a stable reference-guided restoration ability, but its native reference utilization can still decay with temporal distance.

Phase 2: Aligned-Reference Distillation. We then freeze the Phase-1 baseline as a full-reference teacher and continue training the inserted LoRA as a single-reference student equipped with the aligned reference. The teacher receives one high-quality GT reference for each target frame, while the student receives a single reference and the aligned tokens A_t constructed above. The one-step student output is optimized to match the teacher output,

$$\mathcal{L}_{\text{distill}} = \lambda_1 \|\hat{x}^S - \hat{x}^T\|_1 + \lambda_p \,\text{LPIPS}(\hat{x}^S, \hat{x}^T), \tag{10}$$

where $\hat{x}^S$ and $\hat{x}^T$ are the student and teacher outputs, and λ_1, λ_p use the same weights as in Phase 1. This distillation transfers the full-reference teacher's utilization behavior to the single-reference student while preserving the pretrained generative prior.

Table 1: Main quantitative comparison under different reference sources. Reference-guided methods use only the first frame as reference. GT Ref. denotes the ground-truth first frame as a reliable high-quality reference, while ODTSR Ref. and PiSA-SR Ref. denote restored first-frame references. Full-reference metrics are omitted for VideoLQ, where ground truth is unavailable.

Dataset	Metric	STAR	FlashVSR-Full	Vivid-VR	SeedVR2-3B	ODTSR Ref.		PiSA-SR Ref.		GT Ref.	
						SparkVSR	Ours	SparkVSR	Ours	SparkVSR	Ours
SPMCS	PSNR ↑	20.98	21.67	20.12	22.19	20.64	21.74	20.58	21.93	25.27	**25.51**
	SSIM ↑	0.5386	0.5450	0.4799	0.6367	0.5196	0.5746	0.5054	0.5670	0.7857	**0.7919**
	LPIPS ↓	0.4966	0.3004	0.4086	0.3194	0.3254	0.2933	0.3384	0.2985	0.1252	**0.1053**
	ARNIQA ↑	0.2965	0.4441	0.4477	0.4128	0.5155	**0.5179**	0.4862	0.4696	0.4912	0.4853
	ILNIQE ↓	44.1295	27.6612	28.6963	33.8686	27.2499	25.2626	28.2297	25.4799	25.4616	**24.6547**
	DOVER ↑	0.2245	0.4179	0.4740	0.3584	**0.5320**	0.4729	0.4651	0.3925	0.4932	0.4433
YouHQ40	PSNR ↑	21.57	21.94	20.89	22.07	22.28	22.96	21.69	22.79	24.64	**24.94**
	SSIM ↑	0.6009	0.5750	0.5292	0.6656	0.5840	0.6290	0.5632	0.6199	0.7344	**0.7458**
	LPIPS ↓	0.3812	0.2508	0.3431	0.2680	0.3001	0.2657	0.2985	0.2572	0.1652	**0.1360**
	ARNIQA ↑	0.3496	0.4565	0.4779	0.3973	0.5009	**0.5153**	0.5125	0.4930	0.4413	0.4455
	ILNIQE ↓	28.9348	23.3989	23.2158	28.8563	25.0435	23.6336	24.9497	**22.3906**	22.7622	23.6230
	DOVER ↑	0.5331	0.6501	**0.6809**	0.6038	0.6447	0.6286	0.6599	0.6345	0.6736	0.6478
UDM10	PSNR ↑	23.51	24.37	23.28	25.39	24.53	25.08	24.39	25.17	27.83	**28.10**
	SSIM ↑	0.7376	0.7124	0.6384	0.7818	0.7013	0.7481	0.7017	0.7475	0.8270	**0.8386**
	LPIPS ↓	0.3542	0.2532	0.3351	0.2410	0.3104	0.2616	0.3213	0.2681	0.1607	**0.1215**
	ARNIQA ↑	0.3407	0.4888	0.4536	0.3844	0.5345	**0.5393**	0.5276	0.5139	0.4764	0.4624
	ILNIQE ↓	47.9492	**25.6138**	29.2697	35.0530	28.5242	26.9967	28.9865	27.0885	27.3864	28.7470
	DOVER ↑	0.3335	0.5468	0.5473	0.3639	**0.5779**	0.5228	0.5749	0.5174	0.5691	0.5109
VFHQ	PSNR ↑	24.76	26.46	25.83	26.34	26.98	27.34	26.84	27.12	29.49	**29.68**
	SSIM ↑	0.7765	0.7767	0.7748	0.8219	0.7822	0.8134	0.7825	0.8062	0.8643	**0.8769**
	LPIPS ↓	0.2844	0.1701	0.1955	0.1519	0.1761	0.1566	0.1797	0.1677	0.0960	**0.0803**
	ARNIQA ↑	0.4393	0.5356	0.5450	0.4985	0.5815	**0.5819**	0.5721	0.5607	0.5279	0.5097
	ILNIQE ↓	46.3693	**26.3833**	28.8351	33.3233	29.8118	30.9558	30.4892	29.5168	29.7401	34.6208
	DOVER ↑	0.5388	0.7274	0.7514	0.5898	**0.7950**	0.7194	0.7835	0.7081	0.7524	0.6561
	IDS ↑	0.7897	0.8339	0.7596	0.8732	0.7884	0.8338	0.7782	0.8247	0.8979	**0.9282**
VideoLQ	ARNIQA ↑	0.3806	0.4041	0.4586	0.4008	0.4280	**0.4754**	0.4224	0.4543	-	-
	ILNIQE ↓	34.1585	**24.9256**	28.6396	27.7080	28.2896	27.6766	28.1239	26.5370	-	-
	DOVER ↑	0.3456	0.3975	0.4563	0.4286	**0.5204**	0.4855	0.5134	0.4681	-	-

5 Experiments

5.1 Implementation Details

MatchVR is implemented on the pretrained Wan2.2-TI2V-5B video diffusion transformer (Wan et al., 2025). We freeze the backbone and optimize rank-256 LoRA adapters (Hu et al., 2022) for

one-step restoration. The dense matcher uses frozen DINOv3 features (Siméoni et al., 2025) with a trainable projection head, following the two-stage training in Degradation-Robust Feature Matching. The restoration model is trained on 53-frame clips at 320×320 using RealBasicVSR-style degradations (Chan et al., 2022). We use AdamW (Loshchilov & Hutter, 2017) with a learning rate of 3×10^{-5}, bf16 precision, a per-GPU batch size of 1, gradient accumulation of 2, and 4 GPUs. The restoration objective is ℓ_1+LPIPS with λ_1=1 and λ_p=5 in both training phases.

5.2 Datasets and Metrics

Datasets. We evaluate on four benchmarks with ground-truth references: SPMCS (Tao et al., 2017), YouHQ40 (Zhou et al., 2024), UDM10 (Yi et al., 2019), and VFHQ (Xie et al., 2022). VFHQ is included to test whether reference guidance preserves facial identity. We also report results on the real-world VideoLQ set (Chan et al., 2022), where ground truth is unavailable.

Metrics. For fidelity, we use PSNR, SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018). For perceptual quality, we use ARNIQA (Agnolucci et al., 2023) with the quality regressor trained on TID2013 (Ponomarenko et al., 2015), ILNIQE (Zhang et al., 2015), and DOVER (Wu et al., 2023). On VFHQ, we additionally report face identity similarity (IDS). Higher PSNR, SSIM, ARNIQA, DOVER, and IDS are better, while lower LPIPS and ILNIQE are better. In all tables, the best result in each row is marked in bold and the second-best is underlined.

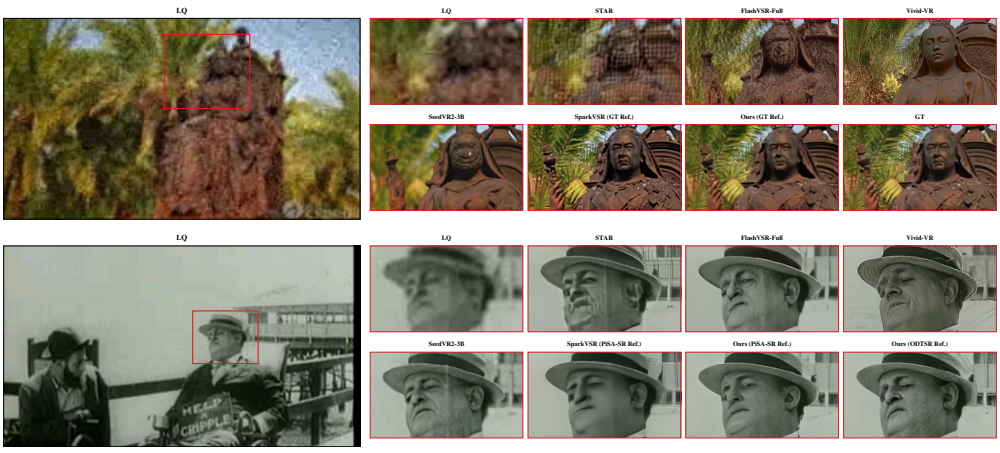

Figure 3: The first row is from synthetically degraded SPMCS, while the second row is from VideoLQ, which contains real-world degraded videos without ground-truth (GT) counterparts.

5.3 Comparison with State-of-the-Art Methods

Quantitative Comparisons. We compare MatchVR with recent blind generative video restoration methods, including STAR (Xie et al., 2025), FlashVSR-Full (Zhuang et al., 2025), Vivid-VR (Bai et al., 2025), and SeedVR2-3B (Wang et al., 2025a). Following their official settings, STAR and Vivid-VR use their default 15-step and 50-step sampling, respectively. For reference-guided restoration, we compare with SparkVSR (Yu et al., 2026). Since the reference-guided setting assumes an available high-quality reference, our main setting uses the ground-truth first frame as the single reference for every test video. We also report practical variants where the first frame is restored by ODTSR (Fang et al., 2025) or PiSA-SR (Sun et al., 2024) and then used as the reference. Table 1 shows the quantitative results. With GT references, MatchVR achieves the best PSNR, SSIM, and LPIPS on all four full-reference benchmarks, showing that one high-quality reference frame is enough to guide faithful video restoration and reduce hallucination. On VFHQ, MatchVR also gives the best IDS, further confirming its identity preservation in face restoration. For no-reference metrics, the scores are strongly tied to reference quality: GT references may yield relatively limited perceptual scores, while restored references often lead to more favorable results. Notably, MatchVR with ODTSR Ref. consistently achieves the best ARNIQA scores across all five benchmarks, demonstrating strong perceptual quality across diverse datasets. MatchVR also

performs competitively on ILNIQE, achieving multiple first- or second-best results under different reference settings. These results suggest that GT references do not necessarily improve no-reference perceptual scores, while restored references can be more favorable. Further analysis of different reference sources is provided in the ablation study.

Qualitative Comparisons. Figure 3 presents qualitative results on synthetically degraded SPMCS and real-world degraded VideoLQ videos. We visualize the last frame of each video, which has the largest temporal distance from the first-frame reference, and zoom in on local regions. Under relatively strong degradation, blind restoration methods tend to produce plausible but less faithful structures, while reference-guided methods recover more reliable details with the help of the reference frame. This is visible in the first row, where the leaf textures behind it are more faithfully restored. Compared with SparkVSR, MatchVR further produces results closer to the GT, especially in the statue material.On the real-world VideoLQ example, where GT is unavailable, SparkVSR exhibits noticeable grid-pattern artifacts on the face when using the same reference frame, whereas MatchVR reconstructs a cleaner facial appearance with better-preserved details.

Table 2: Ablation on the aligned reference.

Dataset	Variant	PSNR ↑	SSIM ↑	LPIPS ↓	ARNIQA ↑	ILNIQE ↓
SPMCS	w/o aligned ref	24.71	0.7739	0.1159	**0.4886**	24.7714
	w/ aligned ref	**25.51**	**0.7919**	**0.1053**	0.4853	**24.6547**
YouHQ40	w/o aligned ref	24.36	0.7352	0.1426	0.4439	23.6404
	w/ aligned ref	**24.94**	**0.7458**	**0.1360**	**0.4455**	**23.6230**

5.4 Ablation Study

Aligned Reference. We evaluate the aligned reference on SPMCS and YouHQ40 using the ground-truth first frame as the high-quality reference. As shown in Table 2, adding the aligned reference improves all full-reference metrics on both datasets. In particular, PSNR increases from 24.71 dB to 25.51 dB on SPMCS and from 24.36 dB to 24.94 dB on YouHQ40, showing that the aligned reference improves fidelity by letting each frame better exploit the given reference. Figure 4 provides complementary evidence from attention and restoration quality. The left panel shows that adding the aligned reference substantially raises target-to-reference attention for distant frames, indicating that MatchVR indeed encourages the model to attend to the reference. The right panel further shows that the resulting per-frame PSNR curve is consistently shifted toward the behavior of the idealized full-reference setting.This comparison is intended to measure how much of the reference-utilization gap is recovered by aligned guidance.

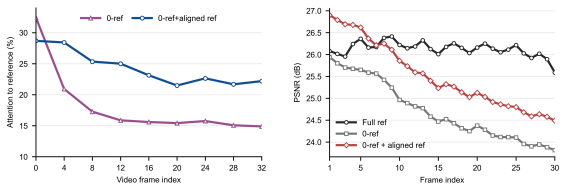

Figure 4: Effect of the aligned reference on reference utilization and restoration quality. Left: the aligned reference increases the attention assigned to reference tokens, especially for frames far from the original reference. Right: the same design lifts the per-frame PSNR curve toward the behavior observed under the idealized full-reference setting.

Reference Quality. We further analyze how different references affect the final restoration quality. Figure 5 provides a visual example on YouHQ40, where higher-quality references lead to clearer and more faithful panda fur details in the restored frame. Table 3 further reports the five metrics of the ODTSR/PiSA-SR restored reference frames and the corresponding MatchVR outputs on SPMCS and YouHQ40.The output metrics are highly consistent with the reference metrics: references with better PSNR, SSIM, and LPIPS usually lead to better full-reference video results, while the no-reference metrics also vary with the perceptual quality of the reference. Compared with the GT-reference setting in Table 1, the drop under restored references mainly comes from the lower fidelity of the reference itself, not from a failure to use it. This confirms that MatchVR follows the reference

Table 3: Reference-quality ablation on SPMCS and YouHQ40. "Ref." reports the metrics of the restored first-frame reference itself, and "Ours" reports the video restored using that reference.

Dataset	Source	Type	PSNR ↑	SSIM ↑	LPIPS ↓	ARNIQA ↑	ILNIQE ↓
SPMCS	ODTSR	Ref.	21.48	0.5618	0.2988	0.5432	26.2806
		Ours	21.74	0.5746	0.2933	0.5179	25.2626
	PiSA-SR	Ref.	21.52	0.5464	0.3244	0.4687	26.3711
		Ours	21.93	0.5670	0.2985	0.4696	25.4799
YouHQ40	ODTSR	Ref.	23.06	0.6336	0.2638	0.5389	24.1566
		Ours	22.96	0.6290	0.2657	0.5153	23.6336
	PiSA-SR	Ref.	22.21	0.5933	0.2675	0.5256	22.8713
		Ours	22.79	0.6199	0.2572	0.4930	22.3906

guidance instead of blindly generating textures, and explains why the final video quality in Table 1 is strongly reference-dependent.



Figure 5: Visual effect of reference quality on YouHQ40. We visualize the last video frame. Different-quality references lead to different panda fur details in the restored frame.

Efficiency Analysis. We further examine whether the aligned reference introduces a large inference overhead. Table 4 compares the inference speed and peak GPU memory with and without the aligned reference. Adding the aligned reference has almost no impact on overall inference speed or memory usage, because standard self-attention already has $\mathcal{O}(n^2)$ complexity, while the local visibility of aligned-reference tokens introduces only $\mathcal{O}(n)$ additional attention interactions. Therefore, the overall complexity remains dominated by $\mathcal{O}(n^2)$, resulting in only minor additional inference overhead.

Table 4: Efficiency analysis on 720p videos with a single reference frame, measured on one A100 GPU.

Variant	FPS ↑	Peak Mem. (GB) ↓
w/o aligned ref	2.77	19.9
w/ aligned ref	2.70	20.1

6 Conclusion

We present MatchVR, a reference-guided generative video restoration framework that converts one high-quality reference into per-frame aligned guidance. It estimates degradation-robust correspondences, warps matched reference content to target layouts in latent space, and distills reference-utilization behavior from an idealized full-reference setting to a single-reference student. Spatially aligned, temporally local references narrow the utilization gap. Experiments show strong fidelity, better face identity preservation, and consistent gains from higher-quality references, reducing hallucination.

AI USE STATEMENT

In this work, we did not use generative AI tools for any tasks with required disclosure. We have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, interpret results. Additionally, we used generative AI tools for editing the manuscript to improve readability by correcting grammatical errors, refining wording, and improving overall language fluency. We have reviewed all AI-assisted work. All AI-assisted edits were carefully reviewed by the authors to ensure that they did not alter the intended scientific meaning, technical content, reported results, or conclusions. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Implementation and evaluation details, including the model backbone, LoRA configuration, training data construction, and degradation settings, are provided in the Experiments section and appendix. We plan to release the source code and trained models to facilitate reproduction of all reported results.

REFERENCES

Lorenzo Agnolucci, Leonardo Galteri, Marco Bertini, and A. Bimbo. Arniqa: Learning distortion manifold for image quality assessment. *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 188–197, 2023. URL https://api.semanticscholar.org/CorpusID:264436513.

Haoran Bai, Xiaoxu Chen, Canqian Yang, Zongyao He, Sibin Deng, and Ying Chen. Vivid-vr: Distilling concepts from text-to-video diffusion transformer for photorealistic video restoration. *arXiv preprint arXiv:2508.14483*, 2025.

Jiezhang Cao, Jingyun Liang, Kai Zhang, Yawei Li, Yulun Zhang, Wenguan Wang, and Luc Van Gool. Reference-based image super-resolution with deformable attention transformer. *Lecture Notes in Computer Science*, 13678:325–342, 2022.

Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Basicvsr: The search for essential components in video super-resolution and beyond. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4947–4956, 2021.

Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Investigating tradeoffs in real-world video super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5962–5971, 2022.

Yushun Fang, Yuxiang Chen, Shibo Yin, Qiang Hu, Jiangchao Yao, Ya Zhang, Xiaoyun Zhang, and Yanfeng Wang. One-step diffusion transformer for controllable real-world image super-resolution. *arXiv preprint arXiv:2511.17138*, 2025.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Yixuan Huang, Xiaoyun Zhang, Yu Fu, Siheng Chen, Ya Zhang, Yan-Feng Wang, and Dazhi He. Task decoupled framework for reference-based super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5931–5940, 2022.

Yuming Jiang, Kelvin CK Chan, Xintao Wang, Chen Change Loy, and Ziwei Liu. Robust reference-based super-resolution via c2-matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2103–2112, 2021.

Youngrae Kim, Jinsu Lim, Hoonhee Cho, Minji Lee, Dongman Lee, Kuk-Jin Yoon, and Ho-Jin Choi. Efficient reference-based video super-resolution (ervsr): Single reference image is all you need. *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1828–1837, 2023. URL https://api.semanticscholar.org/CorpusID:256653410.

Junyong Lee, Myeonghee Lee, Sunghyun Cho, and Seungyong Lee. Reference-based video super-resolution using multi-camera video triplets. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17803–17812, 2022. URL https://api.semanticscholar.org/CorpusID:247762826.

Jingyun Liang, Yuchen Fan, Xiaoyu Xiang, Rakesh Ranjan, Eddy Ilg, Simon Green, Jiezhang Cao, Kai Zhang, Radu Timofte, and Luc V Gool. Recurrent video restoration transformer with guided deformable attention. *Advances in Neural Information Processing Systems*, 35:378–393, 2022.

Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. Vrt: A video restoration transformer. *IEEE Transactions on Image Processing*, 33:2171–2182, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Liying Lu, Wenbo Li, Xin Tao, Jiangbo Lu, and Jiaya Jia. Masa-sr: Matching acceleration and spatial adaptation for reference-based image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6368–6377, 2021.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Nikolay N. Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Vasilyevich Lukin, Karen O. Egiazarian, Jaakko Astola, Benoît Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, and C.-C. Jay Kuo. Image database tid2013: Peculiarities, results and perspectives. *Signal Process. Image Commun.*, 30:57–77, 2015. URL https://api.semanticscholar.org/CorpusID:1768724.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Lingchen Sun, Rongyuan Wu, Zhiyuan Ma, Shuaizheng Liu, Qiaosi Yi, and Lei Zhang. Pixel-level and semantic-level adjustable super-resolution: A dual-lora approach. *arXiv preprint arXiv:2412.03017*, 2024.

Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14940–14950, 2025.

Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia. Detail-revealing deep video super-resolution. In *Proceedings of the IEEE international conference on computer vision*, pp. 4472–4480, 2017.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Jianyi Wang, Shanchuan Lin, Zhijie Lin, Yuxi Ren, Meng Wei, Zongsheng Yue, Shangchen Zhou, Hao Chen, Yang Zhao, Ceyuan Yang, et al. Seedvr2: One-step video restoration via diffusion adversarial post-training. *arXiv preprint arXiv:2506.05301*, 2025a.

Jianyi Wang, Zhijie Lin, Meng Wei, Yang Zhao, Ceyuan Yang, Fei Xiao, Chen Change Loy, and Lu Jiang. Seedvr: Seeding infinity in diffusion transformer towards generic video restoration. *arXiv preprint arXiv:2501.01320*, 2025b.

Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1954–1963. IEEE, 2019.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 20144–20154, 2023.

Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 657–666, 2022.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17108–17118, 2025.

Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5791–5800, 2020.

Peng Yi, Zhongyuan Wang, Kui Jiang, Junjun Jiang, and Jiayi Ma. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3106–3115, 2019.

Jiongze Yu, Xiangbo Gao, Pooja Verlani, Akshay Gadde, Yilin Wang, Balu Adsumilli, and Zhengzhong Tu. Sparkvsr: Interactive video super-resolution via sparse keyframe propagation. *arXiv preprint arXiv:2603.16864*, 2026.

Lin Zhang, Lei Zhang, and Alan Conrad Bovik. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing*, 24:2579–2591, 2015. URL `https://api.semanticscholar.org/CorpusID:262546756`.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Zhifei Zhang, Zhaowen Wang, Zhe Lin, and Hairong Qi. Image super-resolution by neural texture transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7982–7991, 2019.

Weisong Zhao, Jingkai Zhou, Xiangyu Zhu, Weihua Chen, Xiao-Yu Zhang, Zhen Lei, and Fan Wang. Realisvsr: Detail-enhanced diffusion for real-world 4k video super-resolution. *arXiv preprint arXiv:2507.19138*, 2025.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2535–2545, 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.

Hang Zou, Masanori Suganuma, and Takayuki Okatani. Refvsr++: Exploiting reference inputs for reference-based video super-resolution. *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2756–2765, 2023. URL https://api.semanticscholar.org/CorpusID:259361051.

We provide further implementation and evaluation details for MatchVR. We first document the degradation-robust feature matcher, restoration training, LoRA adaptation, spatiotemporal reference alignment, and inference settings. We then evaluate degradation-robust matching, disentangle Selective Spatial Alignment from Temporal Relocation, analyze the influence of ground-truth quality on no-reference metrics, present additional real-world results, and discuss limitations.

A Reproducibility Details

A.1 Degradation-Robust Feature Matching

Our degradation-robust feature matcher uses a frozen DINOv3 ViT-B/16 backbone (Siméoni et al., 2025) and two stream-specific instances of the trainable projection head ψ. Each projection head follows a $768 \to 512 \to 256$ architecture with a GELU activation. The backbone remains frozen throughout the two-stage matcher training; only the projection heads are optimized.

In *contrastive matching training*, we generate image pairs with known correspondences by applying random homographies to 320×320 crops. We train the projection heads with the triplet margin objective $\mathcal{L}_{\text{tri}}$ (Jiang et al., 2021), using a margin of 1.0 and a safe negative radius of three feature cells. The maximum perspective perturbation is 0.06 of the image extent. Training uses Adam with a learning rate of 10^{-3}, batch size 4, gradient-norm clipping at 1.0, and 8,000 iterations.

In *degradation-robust distillation* (Hinton et al., 2015), the frozen teacher matches a high-quality target-reference pair, while the student matches the degraded target with the same reference. The student matches the teacher's pairwise similarity distribution through $\mathcal{L}_{\text{match}}$ with temperature 0.15. We apply a three-stage degradation curriculum: light degradation for iterations 0–15k, medium degradation for 15k–30k, and the full heavy degradation pipeline thereafter. The student is trained for 50k iterations with Adam, learning rate 10^{-3}, batch size 4, and distillation weight 15. The learning rate is reduced by a factor of 0.1 at 97% of training.

A.2 Restoration Training and LoRA Configuration

We build MatchVR on Wan2.2-TI2V-5B (Wan et al., 2025) and keep the pretrained backbone frozen. LoRA adapters (Hu et al., 2022) are inserted into linear layers whose module names contain `self_attn`, `cross_attn`, `ffn.0`, or `ffn.2`; bias and normalization parameters are excluded. We use rank-256 LoRA adapters with Gaussian initialization. The degradation-robust feature matcher is also frozen during restoration training, so only the inserted LoRA parameters are updated. The two stage training is optimized with the same ℓ_1+LPIPS objective (Zhang et al., 2018) introduced in the Method section, using λ_1=1 and λ_p=5.

Training samples contain 53 frames and are randomly cropped to 320×320. We use online blur, resizing, Gaussian/Poisson noise, JPEG compression, and video compression following RealBasicVSR-style degradations (Chan et al., 2022). Optimization uses AdamW (Loshchilov & Hutter, 2017) with learning rate 3×10^{-5}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 10^{-2}, and $\epsilon = 10^{-8}$. The per-GPU batch size is 1 with gradient accumulation of 2.

A.3 Spatiotemporal Reference Alignment and Inference Settings

For each target latent cell, the degradation-robust feature matcher retrieves the top three reference candidates. Candidate coordinates are fused with a softmax temperature of 0.1, and candidates farther than 1.5 latent-grid cells from the local peak are rejected. A 2×2 latent group is retained only when all four constituent cells are valid. We use a matching-confidence threshold of $\eta = 0.70$ in all restoration experiments. As analyzed in next Section, this operating point balances correspondence accuracy and spatial coverage across degradation strengths. During inference, large frames are processed with 512×512 spatial tiling.

Table 5: Matching accuracy and coverage on SPMCS (Tao et al., 2017) under three degradation strengths. Each entry is accuracy / coverage (%). Raw uses frozen DINOv3 features (Siméoni et al., 2025) without the learned projection heads; Ours is the degradation-robust feature matcher.

Feature	Degradation	τ=0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85
Raw	Light	88.00/100.00	88.00/100.00	88.00/100.00	88.02/99.97	88.09/99.82	88.16/99.46	88.34/98.52	88.78/95.55
Ours	Light	88.71/99.80	88.82/99.49	89.02/98.59	89.53/95.62	90.23/88.87	91.22/73.38	92.80/47.52	93.52/16.58
Raw	Medium	77.02/100.00	77.02/100.00	77.06/99.93	77.76/98.10	78.77/95.97	79.10/94.70	80.24/89.80	82.82/73.89
Ours	Medium	87.03/99.14	87.36/98.19	87.79/96.25	88.32/92.33	89.29/83.52	90.77/67.18	92.03/41.66	93.89/14.19
Raw	Heavy	60.70/99.21	62.10/96.29	63.75/91.97	66.24/84.53	68.55/76.27	70.34/66.92	72.60/55.31	78.64/32.96
Ours	Heavy	83.54/97.87	83.98/95.22	84.75/90.29	85.58/82.66	87.02/69.84	88.75/50.55	90.64/27.20	89.90/7.95

Table 6: Matching accuracy and coverage on YouHQ40 (Zhou et al., 2024). The evaluation protocol and notation are identical to Table 5.

Feature	Degradation	τ=0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85
Raw	Light	89.00/100.00	89.05/99.91	89.31/99.27	89.60/98.28	89.78/97.63	90.03/96.88	90.30/95.70	90.64/92.85
Ours	Light	89.29/99.80	89.33/99.38	89.50/98.32	89.81/95.60	90.64/88.49	91.56/73.91	92.40/47.40	91.64/15.94
Raw	Medium	81.78/98.88	81.86/98.38	82.16/97.48	82.63/96.52	83.05/95.42	83.34/93.58	83.82/90.78	85.75/82.56
Ours	Medium	87.86/99.57	88.01/98.82	88.30/97.21	88.81/93.54	89.84/85.79	90.98/70.10	92.11/44.12	90.92/14.97
Raw	Heavy	68.07/91.89	69.18/88.21	70.55/84.16	72.16/79.71	73.80/74.16	75.68/67.91	79.64/55.93	84.26/37.64
Ours	Heavy	84.68/98.07	85.21/96.14	86.01/91.89	86.78/83.93	88.25/70.76	89.71/51.32	90.48/28.55	88.48/9.22

B Degradation-Robust Matching Evaluation

B.1 Evaluation Protocol

We evaluate correspondence robustness on SPMCS (Tao et al., 2017) and YouHQ40 (Zhou et al., 2024) using image pairs with known ground-truth geometry. We randomly sample one PNG image from each sequence folder, resize and crop it to the training-aligned evaluation size, and use it as the high-quality reference. The target is generated by an affine transformation with rotation in $[-10°, 10°]$, scale in $[0.9, 1.1]$, and horizontal and vertical translations within 5% of the image size. Because the affine matrix is known, its inverse maps every valid target patch center to its ground-truth reference location. We then independently apply light, medium, or heavy degradation to the transformed target and compare matching with frozen raw DINOv3 features (Siméoni et al., 2025) against our degradation-robust feature matcher. Both representations use cosine-similarity Top-1 retrieval, and a prediction is correct when it lies within one DINOv3 feature cell of the ground-truth location.

The three degradation levels follow the curriculum used in degradation-robust distillation. Light degradation uses a single weak stage with mild blur, limited resizing, weak Gaussian/Poisson noise, and high-quality JPEG compression, while disabling video compression. Medium degradation uses a moderate two-stage pipeline with broader blur, resizing, noise, and JPEG ranges, but without video compression or the shuffled final degradation. Heavy degradation uses the complete two-stage pipeline with the broadest parameter ranges, including blur, resizing, Gaussian/Poisson noise, JPEG compression, video compression, and shuffled second-stage operations.

For a confidence threshold τ, accuracy is computed among retained Top-1 matches, whereas coverage is the fraction of geometrically valid target locations whose confidence exceeds τ. Reporting both metrics prevents a strict threshold from appearing artificially strong by retaining only a few easy correspondences. SPMCS contains 30 evaluated image pairs and 11,077 valid target locations; YouHQ40 contains 40 pairs and 14,788 valid locations. The same images, affine transforms, degradation realizations, and valid locations are used for raw DINOv3 and our matcher.

B.2 Results and Analysis

The learned projection heads consistently improve matching under degradation, with gains that grow as the target quality deteriorates. At $\tau = 0.60$, where coverage remains comparable, our matcher improves accuracy over raw DINOv3 by 10.73 and 21.00 percentage points under medium and heavy degradation on SPMCS, and by 6.14 and 15.46 points on YouHQ40. The heavy-degradation im-

Table 7: Ablation of Selective Spatial Alignment and Temporal Relocation on SPMCS (Tao et al., 2017) and YouHQ40 (Zhou et al., 2024). All variants use the first GT frame as the single reference.

			SPMCS			YouHQ40		
Variant	Spatial	Temporal	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
w/o aligned ref	–	–	24.71	0.7739	0.1159	24.36	0.7352	0.1426
Spatial alignment only	✓	–	<u>25.03</u>	<u>0.7866</u>	<u>0.1106</u>	<u>24.57</u>	<u>0.7393</u>	0.1389
Temporal relocation only	–	✓	24.37	0.7649	0.1111	23.91	0.7224	<u>0.1369</u>
w/ aligned ref	✓	✓	**25.51**	**0.7919**	**0.1053**	**24.94**	**0.7458**	**0.1360**

provement therefore cannot be explained by retaining substantially fewer locations: coverage differs by only 1.68 points on SPMCS, while our matcher has 7.73 points higher coverage on YouHQ40. These results directly verify that degradation-robust distillation preserves correspondence information that is lost by the raw backbone under strong corruption.

We use $\tau = 0.70$ in all restoration experiments because it provides a favorable accuracy–coverage balance for our matcher. At this threshold, coverage remains between 69.84% and 88.87% across the two datasets and three degradation levels, while accuracy remains between 87.02% and 90.64%. Relative to raw DINOv3 at the same threshold, the accuracy gains are 2.14, 10.52, and 18.47 points for light, medium, and heavy degradation on SPMCS, and 0.86, 6.79, and 14.44 points on YouHQ40. More importantly, from light to heavy degradation, raw DINOv3 loses 19.54 points on SPMCS and 15.98 points on YouHQ40, whereas our matcher loses only 3.21 and 2.39 points, respectively. This substantially smaller degradation-induced drop demonstrates the intended robustness of the learned matching space.

The learned matcher has lower coverage than raw DINOv3 at some identical numerical thresholds because the projection heads reshape the cosine-similarity distribution and generally produce lower absolute similarity scores. Thus, a fixed τ is stricter in the projected feature space, and coverage values across the two representations should not be interpreted as calibrated probabilities. This is why we report the complete accuracy–coverage trade-off rather than accuracy alone. Thresholds above 0.85 are omitted: at $\tau = 0.90$, our matcher retains only 0.81–1.38% of valid locations across both datasets and all degradation levels, and at $\tau = 0.95$ it retains virtually none. Accuracy in this regime is dominated by a negligible number of matches and has little practical meaning for aligned-reference construction.

C Ablation of Spatiotemporal Reference Alignment

We separately ablate Selective Spatial Alignment and Temporal Relocation while keeping the restoration backbone, LoRA checkpoint, input reference, visibility policy, and all inference settings fixed. We report PSNR, SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018) to measure full-reference fidelity and perceptual similarity. For *spatial alignment only*, we retain degradation-robust feature matching and latent resampling, but assign every aligned reference the original reference-frame temporal position instead of the target-frame position. For *temporal relocation only*, we disable feature matching and spatial resampling, replicate the unwarped reference latent for every target latent frame, and assign each copy the temporal position of its target.In both variants, each aligned reference remains locally visible only to its designated target. The full model combines matched spatial resampling with target-time relocation.

Selective Spatial Alignment improves reference adherence by ensuring that injected details originate from geometrically corresponding regions. Compared with the model without aligned references, spatial alignment alone improves PSNR by 0.32 dB on SPMCS and 0.21 dB on YouHQ40, increases SSIM by 0.0127 and 0.0041, and reduces LPIPS by 0.0053 and 0.0037, respectively. These consistent gains confirm that degradation-robust matching and latent resampling allow the restoration model to follow spatially relevant reference content. Nevertheless, this variant retains the original reference-frame temporal position for every aligned reference. The remaining gap to the full model is therefore consistent with temporally distant targets assigning insufficient attention to otherwise correctly aligned reference evidence.

Temporal Relocation alone increases the temporal accessibility of the reference but cannot guarantee content correspondence. Although it reduces LPIPS from 0.1159 to 0.1111 on SPMCS and from 0.1426 to 0.1369 on YouHQ40, it decreases PSNR by 0.34 and 0.45 dB and SSIM by 0.0090 and 0.0128, respectively. This variant copies the unwarped reference latent to every target time; consequently, stronger temporal access may introduce reference textures at incorrect spatial locations. The resulting fidelity degradation shows that improving attention to the reference without first establishing spatial correspondence can weaken reference adherence rather than improve it.

Combining spatial and temporal alignment is therefore necessary for effective reference utilization. Relative to spatial alignment alone, Temporal Relocation further improves PSNR by 0.48 dB on SPMCS and 0.37 dB on YouHQ40, increases SSIM by 0.0053 and 0.0065, and reduces LPIPS by 0.0053 and 0.0029. The complete model achieves the best PSNR, SSIM, and LPIPS on both datasets; compared with the model without aligned references, it yields gains of 0.80 and 0.58 dB in PSNR, 0.0180 and 0.0106 in SSIM, and reductions of 0.0106 and 0.0066 in LPIPS. Spatial alignment determines which reference content should be used, while temporal alignment ensures that the target frame allocates sufficient attention to this reference evidence. Their combination consequently increases both the utilization of the reference and the model's ability to faithfully follow reference details.

Table 8: No-reference quality scores of the ground-truth videos and MatchVR outputs using the first GT frame as the single reference (Ours-GT-Ref) on SPMCS (Tao et al., 2017), UDM10 (Yi et al., 2019), and YouHQ40 (Zhou et al., 2024). DOVER (Wu et al., 2023) is reported as $\mathrm{sigmoid}(\mathrm{overall})$.

Dataset	Source	ARNIQA ↑	ILNIQE ↓	DOVER ↑
SPMCS	GT	0.4981	24.9020	0.4461
	Ours-GT-Ref	0.4853	24.6547	0.4433
UDM10	GT	0.4560	27.8613	0.5127
	Ours-GT-Ref	0.4624	28.7470	0.5109
YouHQ40	GT	0.4435	22.8819	0.6329
	Ours-GT-Ref	0.4455	23.6230	0.6478

D Influence of Ground-Truth Quality on No-Reference Metrics

The quantitative results in the Experiments section reveal a trade-off between full-reference fidelity and no-reference perceptual scores when the first ground-truth frame is used as the reference. MatchVR is explicitly designed to follow reliable reference evidence, so its restored videos are encouraged to approach the ground-truth content rather than maximize the preferences encoded by a no-reference quality predictor. Importantly, no-reference metrics do not treat the ground-truth video as an upper bound: authentic GT videos can themselves receive modest ARNIQA (Agnolucci et al., 2023), ILNIQE (Zhang et al., 2015), or DOVER (Wu et al., 2023) scores, as these metrics reflect perceptual or natural-image priors rather than fidelity to the target sequence.

Table 8 shows that the no-reference scores of Ours-GT-Ref closely follow the quality characteristics of the GT videos. Across the three datasets, the ARNIQA gap between Ours-GT-Ref and GT remains within 0.0020–0.0128, while the ILNIQE gap ranges from 0.2473 to 0.8857. DOVER is nearly unchanged on SPMCS and UDM10, with differences of only 0.0028 and 0.0018, and Ours-GT-Ref is 0.0149 higher on YouHQ40. More importantly, the dataset-level ranking is consistently preserved across ARNIQA, ILNIQE, and DOVER. This suggests that the no-reference scores of the restored outputs largely inherit the quality characteristics of the GT content being followed.

Full-reference and no-reference rankings therefore need not agree. Full-reference metrics directly measure whether the restored result approaches the target video, whereas no-reference metrics may favor perceptually enhanced textures even when they deviate from the GT. The consistent score tendencies between GT and Ours-GT-Ref indicate that GT quality also influences the no-reference evaluation of our outputs. Together with the stronger PSNR, SSIM, and LPIPS results, this evidence indicates that MatchVR prioritizes faithful use of reference details rather than altering them solely to obtain higher no-reference scores. We therefore use full-reference metrics to assess reference ad-

herence on paired datasets and treat no-reference metrics as complementary indicators of perceptual quality.

E Additional Qualitative Results

Figure 6 provides more qualitative examples on both paired degraded benchmarks and the real-world VideoLQ dataset (Chan et al., 2022). The compared restoration baselines include STAR (Xie et al., 2025), FlashVSR (Zhuang et al., 2025), Vivid-VR (Bai et al., 2025), SeedVR2 (Wang et al., 2025a), and SparkVSR (Yu et al., 2026). For the paired examples, MatchVR produces restorations that are visibly closer to the GT than competing methods, especially in the enlarged regions where reference-consistent structures and textures are easier to inspect. For the real-world VideoLQ examples, ground truth is unavailable; nevertheless, using restored first-frame references from PiSA-SR (Sun et al., 2024) or ODTSR (Fang et al., 2025) still provides useful guidance, and MatchVR obtains strong visual restoration quality under this practical reference setting.

F Limitations

As a reference-guided restoration method, MatchVR naturally benefits most from high-quality references that cover the content appearing in the target frames. When the supplied reference has limited visual quality, the reliable details available for transfer are correspondingly reduced, which may constrain the improvement achievable through reference guidance. Similarly, some target regions may contain newly observed content that does not appear in the reference and therefore cannot be directly supported by spatial correspondence. In such regions, the model continues to rely on the degraded target and its generative prior, while the strongest reference-driven gains remain concentrated on content covered by the available reference. In future work, we will continue to improve this aspect by incorporating multiple complementary references and constructing a dynamic reference library that progressively stores previously restored high-quality frames. These additional references can expand content coverage over time and provide direct guidance when later frames reveal regions that were absent from the initial reference.

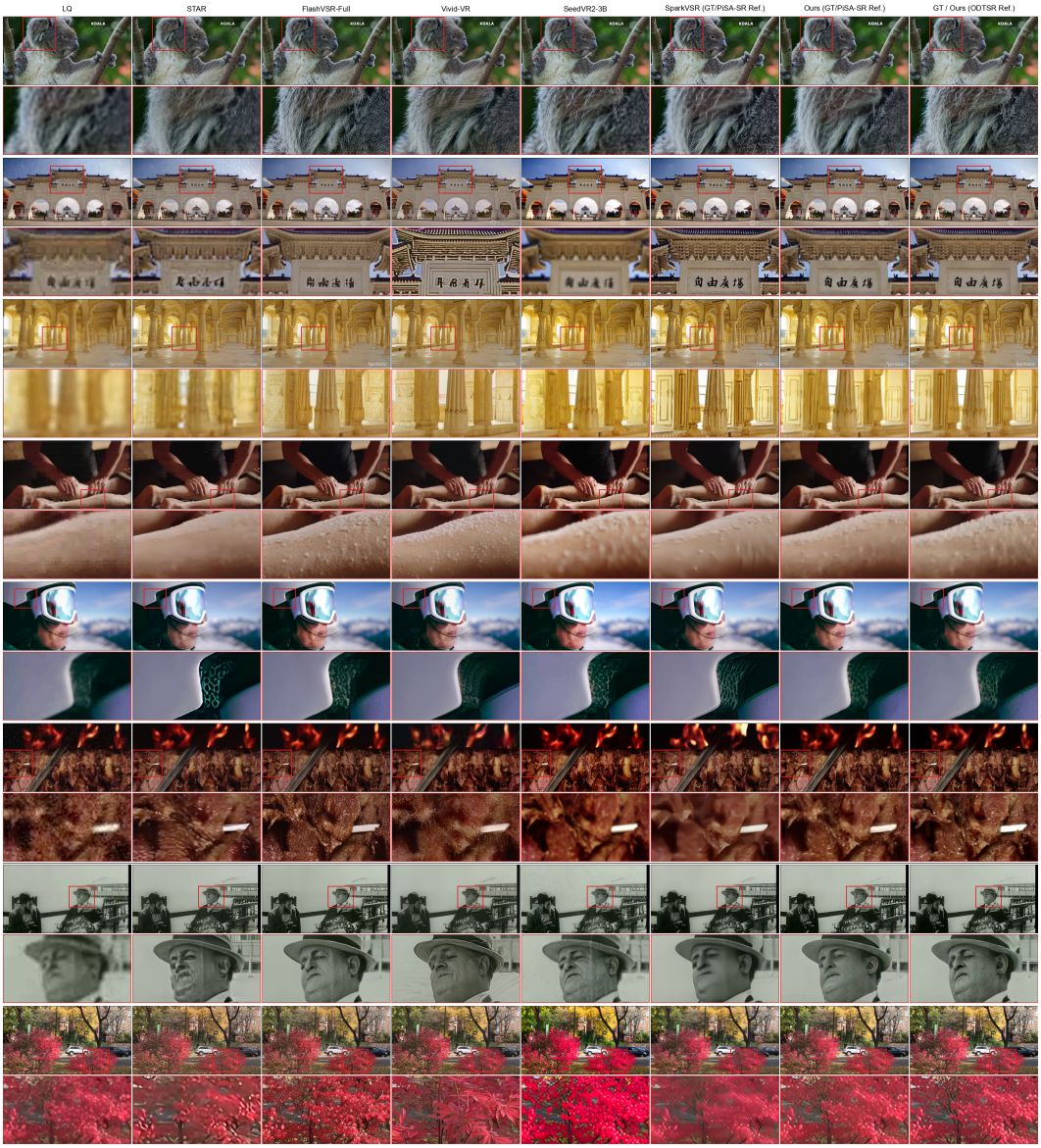

Figure 6: Additional qualitative comparisons with enlarged crops. Each case contains a full-frame row and a zoomed crop row. Row 1 is from YouHQ40 (Zhou et al., 2024); rows 2–3 are from SPMCS (Tao et al., 2017); rows 4–6 are from RealisVideo-720p (Zhao et al., 2025); rows 7–8 are from VideoLQ (Chan et al., 2022). The first six rows use the GT frame in the last column. For the VideoLQ rows, GT is unavailable: the last three columns are SparkVSR (Yu et al., 2026) with a PiSA-SR (Sun et al., 2024) restored reference, MatchVR with a PiSA-SR restored reference, and MatchVR with an ODTSR (Fang et al., 2025) restored reference.