VOGVR: A NATIVE VISION-ONLY GENERATIVE VIDEO RESTORATION MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

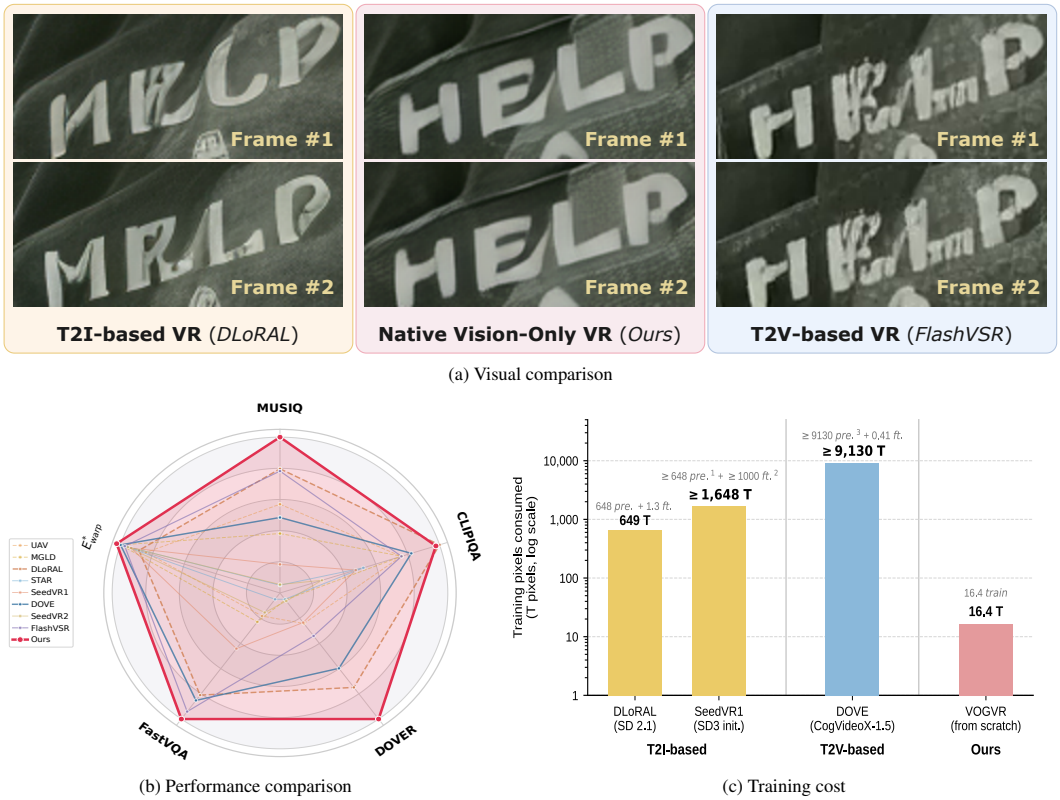
Most existing generative video super-resolution (VSR) or video restoration (VR) methods are built by adapting text-to-image (T2I) or text-to-video (T2V) diffusion models pretrained on large-scale text-visual data. However, these models often suffer from issues of temporal consistency (*e.g.*, T2I-based methods) or detail synthesis (*e.g.*, T2V-based methods). In addition, the synthesis-oriented nature of the T2I/T2V backbones, even finetuned on low-quality (LQ) and high-quality (HQ) paired data, can make it difficult to keep generated details tightly constrained by the observed input. We investigate whether a generative VR model *trained purely on visual data* can surpass these T2I and T2V-based adaptations, and propose **VOGVR**, a native **V**ision-**O**nly **G**enerative **V**ideo **R**estoration model. We first train a bidirectional VAE to capture temporal context, reducing latent ambiguity, and improving fine-detail reconstruction. Then, we extract visual features and motion cues from LQ videos using pretrained vision encoders (*e.g.*, DINOv2) and optical flow networks, and use them as visually grounded conditions to guide the training of the DiT, anchoring the generation to the LQ input for detail-rich and temporally coherent restoration. A high-quality dataset with more than 90M images and 3M video clips is built to train our VOGVR model. Extensive experiments demonstrate that VOGVR outperforms state-of-the-art generative VSR/VR models, effectively alleviating the trade-off between temporal coherence and spatial details while keeping the restoration grounded in the input video, all with one-step inference. Moreover, VOGVR requires less than 1/30 of the training cost of representative T2I-based VR methods and less than 1/400 of T2V-based ones.

1 INTRODUCTION

Classical video restoration (VR) methods (Wang et al., 2019; Chan et al., 2022a; Cao et al., 2021; Liu et al., 2022a) generally assume specific degradations (*e.g.*, bicubic downsampling, Gaussian noise, Gaussian blur) on the low-quality (LQ) video inputs, yet the trained models can hardly be applied to real-world LQ videos, which often undergo complex and heavy degradations that mix various noise, blur, scaling, compression, *etc.* (Yang et al., 2021; Wei et al., 2023). Generative VR methods aim to address this issue by reconstructing high-quality (HQ) videos from LQ inputs under real-world degradations. A representative line of research under such settings is generative video super-resolution (VSR) (Zhou et al., 2024; Yang et al., 2024a; Sun et al., 2025; Xie et al., 2025; Chen et al., 2025; Zhuang et al., 2025), where the LQ video undergoes mixed and high-order degradations (Wang et al., 2021; Chan et al., 2022b). In this paper, we treat generative VR and VSR as the same problem for convenience of discussion, since many later developed VSR methods (Xie et al., 2025; Chen et al., 2025; Sun et al., 2025; Chen et al., 2026; Zhuang et al., 2025) first interpolate the LQ video to the target resolution and then restore the details.

Early generative VR/VSR models are mostly GAN (generative adversarial network) (Ledig et al., 2017) based methods, which employ adversarial loss during training to encourage realistic detail restoration (Bulat & Tzimiropoulos, 2018; Menon et al., 2020; Bulat et al., 2018; Xu et al., 2024). However, adversarial training is unstable to complex real-world degradations, often leading to unnatural artifacts. More recently, large-scale diffusion models have significantly advanced generative restoration. Mainstream methods (Zhou et al., 2024; Yang et al., 2024a; Xie et al., 2025; Sun et al., 2025; Wang et al., 2025c; Chen et al., 2025; Zhuang et al., 2025) are typically built on pretrained text-driven diffusion backbones, including text-to-image (T2I) models (*e.g.*, Stable Diffusion (Rombach

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.



(a) Visual comparison

(b) Performance comparison

(c) Training cost

Figure 1: Overview of VOGVR. (a) Visual comparison: T2I-based and T2V-based methods often suffer from temporal inconsistency and distorted details, while our native vision-only VOGVR can achieve both visual details and temporal coherence. (b) Performance comparison on spatial quality and temporal consistency. (c) Training cost (total training pixels, log scale). The values on the right side of "≥" are conservative lower bounds (see Appendix A.1). VOGVR requires less than 1/30 of T2I-based and less than 1/400 of T2V-based VR model training cost.

et al., 2022)) and text-to-video (T2V) models (*e.g.*, CogVideoX (Yang et al., 2024b) and Wan (Wan et al., 2025)). To steer the pretrained T2I/T2V backbone toward restoration, these methods generally inject the LQ input through auxiliary conditional control (Zhou et al., 2024; Yang et al., 2024a; Xie et al., 2025) and latent modulation (Chen et al., 2025; Sun et al., 2025; Wang et al., 2025c; Zhuang et al., 2025). Based on this paradigm, T2I-based approaches (Zhou et al., 2024; Yang et al., 2024a; Sun et al., 2025) can produce sharp per-frame detail, but suffer from temporal inconsistency due to the lack of temporal modeling during backbone pretraining, even if dedicated temporal modules are introduced to compensate (see left part of Fig. 1(a)). By contrast, T2V-based approaches (Wang et al., 2025b; Chen et al., 2025; Zhuang et al., 2025) benefit from pretrained temporal priors and thus achieve smoother motion, yet sacrifice fine-grained spatial details (see right part of Fig. 1(a)). Strategies such as mixed image-video training (Chen et al., 2025; Zhuang et al., 2025) have been explored to alleviate these problems, but the recovered details still lag behind their T2I-based counterparts.

The above problems arise mainly from the mismatch between the objectives of pretrained T2I/T2V backbones and VR models. The T2I/T2V backbones are optimized for open-ended synthesis with the aim of generating visually plausible and diverse content. In contrast, VR is a highly constrained, input-conditioned restoration task. The degraded videos already provide the essential contents and motion trajectories, and the goal is to faithfully recover lost details rather than hallucinating new content. Due to this discrepancy, when adapting synthesis-oriented T2I/T2V backbones for restoration, the strong generative priors may overwhelm the input conditions, synthesizing inconsistent details across frames. To address these issues, in this work we propose **VOGVR**, a native **V**ision-**O**nly **G**enerative **V**ideo **R**estoration model, which is trained purely on visual data. VOGVR learns restoration-oriented generative priors from scratch by denoising HQ video data conditioned on their degraded LQ counterparts. Through large-scale training, VOGVR can capture HQ visual distributions to reproduce

sharp textures, faithful motion, and temporally consistent details while avoiding the cross-modal misalignment and keeping the restoration more tightly grounded in the observed visual content.

Specifically, VOGVR consists of a dedicated VAE and a DiT tailored to VR. Prior methods (Zhou et al., 2024; Chen et al., 2025; Wang et al., 2025c; Zhuang et al., 2025) typically adopt image-based VAE or causal video VAE, both lacking full temporal context in latent space. However, in VR we can leverage both past and future information to enhance each frame. We therefore introduce a bidirectional VAE to capture richer temporal dependencies, which is unavailable in existing pretrained backbones commonly adopted by VR and thus needs to be trained from scratch. On the other hand, existing VR methods (Chen et al., 2025; Sun et al., 2025; Zhuang et al., 2025) often retain text-aligned high-level semantic representations inherited from pretrained T2I/T2V backbones, whereas restoration is naturally grounded in the visual content and motion of the observed LQ video. We therefore condition the generative process on robust visual features and explicit motion trajectories extracted from the LQ input using pretrained vision encoders (*i.e.,* DINOv2 (Oquab et al., 2023)) and optical flow networks (*i.e.,* Unimatch (Xu et al., 2023)), promoting detail recovery under strong spatio-temporal constraints (see middle part of Fig. 1(a)). To support the training of VOGVR, we curate a large-scale, high-quality dataset comprising more than 90M images and 3M video clips across various scenes and motion patterns. Following prior works (Wan et al., 2025; Wang et al., 2025c), we adopt a progressive training strategy that transitions from images to videos, from low to high resolutions, and from short to long temporal clips. As shown in Figs. 1(b) and (c), VOGVR achieves state-of-the-art performance while consuming less than 1/30 of the training cost of T2I-based VR methods and 1/400 of T2V-based ones.

Our main contributions are summarized as follows. First, we propose a new generative VR training framework, which learns from purely visual data to resolve the inherent issues of T2I/T2V adapted VR models. Second, we introduce a tailored architecture comprising a bidirectional VAE and a restoration-oriented DiT, anchoring the generation to the degraded input through explicit structural, semantic, and motion cues. Finally, VOGVR achieves state-of-the-art VR performance with rich per-frame spatial details and smooth temporal coherence, while remaining closely grounded in the degraded input and requiring much less training cost than representative T2I-based and T2V-based VR models.

2 Related Work

Real-World VSR/VR. Classical VSR/VR methods (Wang et al., 2019; Fuoli et al., 2019; Li et al., 2020) typically assume simple and specific degradations (*e.g.*, bicubic downsampling), leading to a significant gap on real-world videos. To bridge this gap, prior works either collect real-world LQ-HQ pairs (Yang et al., 2021; Wei et al., 2023) to train the model, or develop realistic degradation pipelines (Chan et al., 2022b) to synthesize training data, or improve robustness through architectural designs (Chan et al., 2021; 2022a). Nevertheless, these approaches still produce over-smoothed results under complex real-world degradations. On the other hand, adversarial training can improve video sharpness but may introduce visible artifacts (Chan et al., 2022b; Xu et al., 2024).

Diffusion-based VSR/VR. Recent generative VSR/VR methods mainly adapt pretrained text-driven diffusion backbones. T2I (*e.g.*, Stable Diffusion (Rombach et al., 2022))-based methods, such as Upscale-A-Video (Zhou et al., 2024), MGLD (Yang et al., 2024a) and DLoRAL (Sun et al., 2025) leverage strong image priors but require additional temporal modeling. SeedVR (Wang et al., 2025c) addresses this by training a causal video VAE, which is however still initialized from a T2I backbone. Another line of research (Xie et al., 2025; Chen et al., 2025; Zhuang et al., 2025) leverages pretrained T2V models (*e.g.*, CogVideoX (Yang et al., 2024b), Wan (Wan et al., 2025)), inheriting stronger temporal priors but often sacrificing fine-grained spatial details. Various detail-enhancement modules, decoders, and mixed image-video training strategies have been explored to alleviate this trade-off (Chen et al., 2025; Zhuang et al., 2025). However, they remain largely influenced by the synthesis-oriented priors inherited from pretrained backbones.

Vision-Only Generative Restoration. Diffusion-based image super-resolution (SR) methods (Saharia et al., 2022; Li et al., 2022; Yue et al., 2023; Wang et al., 2023b) show that strong restoration can be achieved without text supervision. More recently, VOSR (Wu et al., 2026) further incorporates visual semantic features from pretrained vision encoders to improve image-level restoration quality. However, all of these methods are designed for single images, encoding each frame independently without considering inter-frame motion and temporal consistency. Video restoration poses funda-

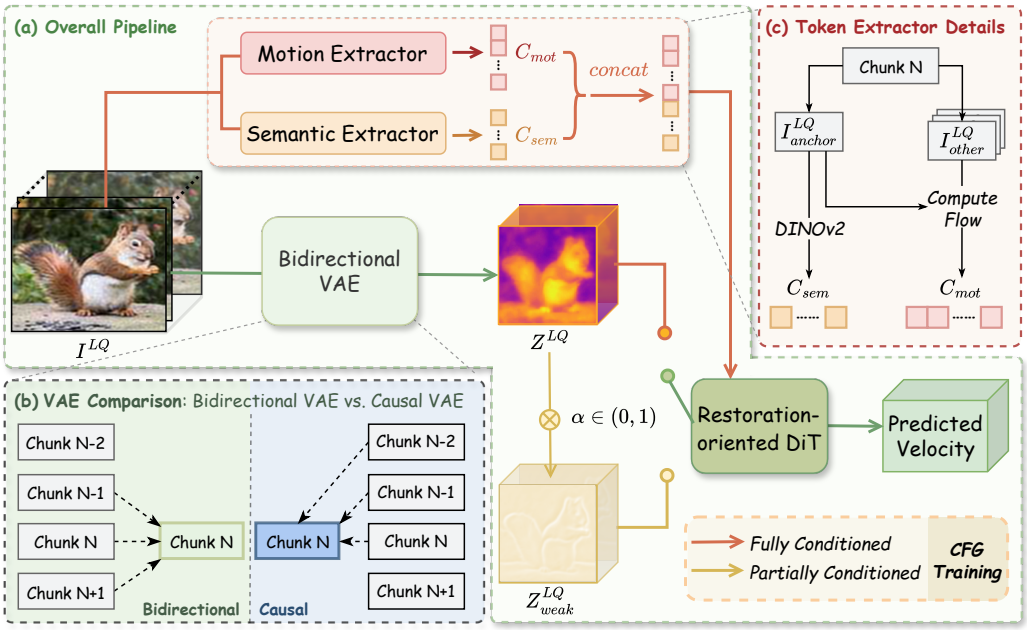


Figure 2: Overview of VOGVR. (a) Overall pipeline. Given an LQ video, VOGVR extracts structural latent z^{LQ}, semantic tokens c_{sem} and motion tokens c_{mot}, and injects them into a restoration-oriented DiT for velocity prediction. A partially conditioned guidance with weakened $z^{\mathrm{LQ}}_{\mathrm{weak}}$ replaces the unconditional branch to keep outputs input-anchored. (b) Bidirectional vs. causal VAE: unlike causal VAE that only sees preceding chunks, our bidirectional VAE leverages both past and future contexts. (c) Token extractors: c_{sem} is from the anchor frame of each chunk via DINOv2; c_{mot} is from the optical flow between the anchor and remaining frames.

mentally different challenges: the model must simultaneously handle temporal dynamics, maintain cross-frame coherence, and recover spatially fine details, which demands rethinking the latent representation, the conditioning mechanism, and the generative backbone altogether. In this work, we propose VOGVR, a native vision-only generative model for video restoration, which introduces a bidirectional VAE for temporally enriched latent representations, and a restoration-oriented DiT conditioned on explicit motion and visual cues.

3 Methodology

3.1 Preliminary

Latent Diffusion Formulation. Latent diffusion models (Rombach et al., 2022) perform generative modeling in a compressed latent space spanned by a variational autoencoder (VAE). Given an HQ video $\mathbf{I}^{\mathrm{HQ}} = \{I^{\mathrm{HQ}}_n\}_{n=1}^{N}$, the VAE encoder $\mathcal{E}$ maps it to a latent code $z_0 = \mathcal{E}(\mathbf{I}^{\mathrm{HQ}})$. Rectified flow (Liu et al., 2022b) defines a straight-line interpolation between z_0 and Gaussian noise $z_1 \sim \mathcal{N}(0, \mathbf{I})$: $z_t = (1-t)\, z_0 + t\, z_1, t \in [0,1]$. A denoising network v_θ is trained using the following loss

$$\mathcal{L} = \mathbb{E}_{z_0, z_1, t}\left[\left\| v_\theta(z_t, t, c) - (z_1 - z_0) \right\|_2^2\right] \tag{1}$$

to predict the velocity field $v_t = z_1 - z_0$ from the noisy latent z_t, the timestep t, and a task-specific condition c. At inference, starting from $z_1 \sim \mathcal{N}(0, \mathbf{I})$, the learned velocity field is numerically integrated to recover z_0, which is then decoded by the VAE decoder $\mathcal{D}$ to produce the output. In our VOGVR, z_0 corresponds to the latent of the HQ video, and c is derived from the LQ input.

Classifier-Free Guidance. Classifier-free guidance (CFG) (Ho & Salimans, 2022) enables guidance without an external classifier by randomly dropping the condition c during training, allowing the same network to learn both a conditional prediction $v_\theta(z_t, t, c)$ and an unconditional prediction $v_\theta(z_t, t, \varnothing)$. In standard generation, the unconditional branch learns a condition-free data prior from the noisy latent alone. At inference, CFG extrapolates from this unconditional prediction toward the conditional

one:

$$v_{\text{cfg}} = v_\theta(z_t, t, \varnothing) + s\left(v_\theta(z_t, t, c) - v_\theta(z_t, t, \varnothing)\right), \tag{2}$$

where s is the guidance scale. However, this condition-free branch is not well suited for the restoration task, where the degraded input anchors the reconstruction. A natural remedy is to replace the unconditional branch with a partially conditioned branch that retains weakened structural cues of the LQ input (Wu et al., 2026), making both branches input-anchored. We adopt this restoration-oriented CFG strategy and extend it to the video setting in Sec. 3.2, incorporating temporal and motion-aware conditions.

3.2 Model Architecture

Overall Framework. As illustrated in Fig. 2, given an LQ video $\mathbf{I}^{\text{LQ}} = \{I_n^{\text{LQ}}\}_{n=1}^{N}$, our goal is to recover its HQ counterpart $\mathbf{I}^{\text{HQ}}$, which is expected to be perceptually realistic, temporally coherent, and content faithful to the input. We first encode the LQ video with a 3D bidirectional VAE encoder $\mathcal{E}$ to obtain the latent representation $z^{\text{LQ}} = \mathcal{E}(\mathbf{I}^{\text{LQ}})$, which provides a temporally informed structural condition spatially aligned with the noisy HQ latent z_t. In parallel, we extract semantic tokens c_{sem} with a pretrained vision encoder (*i.e.*, DINOv2 (Oquab et al., 2023)) and motion tokens c_{mot} with a pretrained optical flow network (*i.e.*, Unimatch (Xu et al., 2023)). The former provides degradation-robust visual context, while the latter supplies explicit inter-frame correspondences. These structural, semantic and motion cues constitute the full visual condition to guide the DiT to denoise z_t toward the clean latent z_0. Specifically, z^{LQ} is concatenated with z_t as the spatial input, while c_{sem} and c_{mot} are injected via cross-attention, leading to the fully conditioned velocity prediction:

$$\hat{v}_t = v_\theta\big(z_t,\ t,\ z^{\text{LQ}},\ c_{\text{sem}},\ c_{\text{mot}}\big). \tag{3}$$

To equip the model with restoration-oriented generative priors while keeping the output tightly anchored to the input, we tailor CFG (Eq. 2) by replacing the standard unconditional branch with a *partially conditioned* branch, while referring to the standard conditional prediction (Eq. 3) as the *fully conditioned* branch. Specifically, the partially conditioned branch keeps only a weakened structural input $z^{\text{LQ}}_{\text{weak}} = \alpha \cdot z^{\text{LQ}}$ ($\alpha \in (0, 1)$, we empirically set it to 0.05 following (Wu et al., 2026)) concatenated with z_t, dropping the semantic and motion tokens. Compared to the standard unconditional branch in CFG, this weakened conditioning preserves coarse structural cues from the LQ input, while allowing the model to refine missing details from HQ video priors. Meanwhile, the fully conditioned branch constrains the prediction with semantic and motion tokens, making the denoising trajectory adhere closely to the observed structure and dynamics. During training, we randomly sample either branch; during multi-step sampling, the guidance is computed as:

$$v_{\text{cfg}} = v_\theta(z_t,\ t,\ z^{\text{LQ}}_{\text{weak}}) \ +\ s\left(v_\theta(z_t,\ t,\ z^{\text{LQ}},\ c_{\text{sem}},\ c_{\text{mot}}) - v_\theta(z_t,\ t,\ z^{\text{LQ}}_{\text{weak}})\right). \tag{4}$$

Under this formulation, VOGVR is first trained as a multi-step rectified flow model with a progressive curriculum, and then distilled into a one-step generator for efficient inference. Both the VAE and the DiT follow the architectural design of Wan2.1-1.3B (Wan et al., 2025) but are trained from scratch, with restoration-specific modifications and training strategies detailed in the following subsections.

Bidirectional Video VAE. The VAE defines the latent space in which restoration is performed. Existing generative VR methods usually inherit the VAE from pretrained T2I or T2V backbones. The VAE of T2I-based methods (Zhou et al., 2024; Yang et al., 2024a; Sun et al., 2025) encodes frames independently, leaving temporal modeling entirely to the diffusion backbone. The VAE of T2V-based methods (Xie et al., 2025; Chen et al., 2025; Wang et al., 2025b; Zhuang et al., 2025) considers temporal modeling but follows a strict causal (*i.e.*, future frames do not influence past frames) protocol. As illustrated in Fig. 2, where we denote the VAE temporal processing unit as Chunk (Wan et al., 2025), the latent of Chunk N is constructed from Chunks $\leq N$, while Chunk $N + 1$ is excluded. For the offline VR task, however, the whole degraded video can be accessed. When local structures or textures are highly corrupted in a frame, useful cues may appear in either previous or subsequent frames. We therefore relax these generation-oriented restrictions and allow the VAE to exploit both preceding and subsequent contexts for latent construction. Since such a bidirectional video VAE is unavailable in the pretrained backbones commonly adopted by existing VR methods, we train the VAE from scratch in this work.

Specifically, we train the bidirectional VAE following the overall architecture of Wan2.1 (Wan et al., 2025), with two key modifications to remove the causal constraint. First, we replace causal temporal

convolutions with non-causal temporal convolutions, so that each temporal kernel aggregates features from both past and future positions. Second, we extend the chunk-wise processing to incorporate both preceding and subsequent chunks, as illustrated in Fig. 2. The same bidirectional encoder is used to obtain z^{HQ} from the HQ video and z^{LQ} from the LQ video, providing full-context latent representations for both the HQ learning target and the LQ structural condition of the DiT (see ablation in Appendix A.2 for more details).

Restoration-oriented DiT. Existing generative VR methods (Xie et al., 2025; Wang et al., 2025c; Chen et al., 2025; Zhuang et al., 2025) typically inherit the DiTs from pretrained T2I/T2V backbones, which are designed for open-ended synthesis. In VR, however, the observed LQ video serves as the primary condition, providing restoration contents, structural layout and motion trajectories that the output is expected to follow. Therefore, we explicitly ground the restoration-oriented DiT in visual observations from the degraded video. To this end, we remove the text-conditioning branch from the Wan2.1-1.3B DiT architectural template and construct a native vision-only DiT with three complementary conditions, as illustrated in Fig. 2. (1) **Structural condition** z^{LQ}. The LQ latent z^{LQ} is concatenated with the noisy latent z_t, providing a spatially aligned anchor that preserves the coarse layout of the degraded input throughout denoising. (2) **Semantic tokens** c_{sem}. We extract intermediate-layer patch features from the anchor frame of each temporal chunk $\mathbf{I}^{\mathrm{LQ}}_{\mathrm{anchor}}$ (*i.e.*, the first frame of each chunk) using a pretrained DINOv2 (Oquab et al., 2023) encoder. Compared with final-layer features, these intermediate representations preserve more spatial detail while remaining robust to image degradation. Since the entire video is available offline, we temporally average the extracted features across all chunks and project them with an MLP to obtain a global c_{sem}. We adopt averaging rather than concatenating per-chunk features because fine-grained spatial details are already captured by the spatially aligned z^{LQ}, while c_{sem} serves as a global semantic context; concatenation would linearly increase cross-attention token length with video duration yet yields no noticeable quality gain (see ablation in Appendix A.2). (3) **Motion tokens** c_{mot}. We estimate dense optical flow from the LQ video using Unimatch (Xu et al., 2023), computing displacements of each frame relative to the anchor frame within each temporal chunk. The flow maps are spatially downsampled, flattened into token sequences, and linearly projected to obtain c_{mot}, which provides explicit inter-frame correspondence for temporally consistent restoration.

Within each DiT block, the concatenated spatial input (the noisy latent z_t and the structural condition z^{LQ}) is processed via self-attention, while the semantic and motion tokens are concatenated and injected via cross-attention. This tailored architecture ensures that the rich generative priors are anchored to the visual and temporal observations of the degraded input.

3.3 Training Data and Strategy

Training Data. To learn the data-hungry restoration priors, we build an image-video data pool to train both the VAE and the DiT. The pool contains images and videos from the web [1] and public datasets (Li et al., 2024; Kirillov et al., 2023; Ravi et al., 2024; Qiao et al., 2025; Xue et al., 2025; Wang et al., 2025a). Following (Wan et al., 2025; Wu et al., 2026), we filter the data by quality and diversity, using image-entropy filtering, motion-richness filtering, and category-level de-duplication and balancing. The resulting dataset contains over 90M images and 3M video clips (each clip contains more than 100 frames). We synthesize LQ-HQ training pairs with RealESRGAN degradation (Wang et al., 2021) for images and RealBasicVSR-like degradation (Sun et al., 2025) for videos.

Training Strategy. Since the bidirectional VAE and restoration-oriented DiT depart from the corresponding pretrained T2I/T2V components, we train the entire VR model from scratch. Training such a model directly on high-resolution videos is resource-consuming and unstable. We therefore adopt a progressive curriculum following (Wan et al., 2025; Wang et al., 2025c). Specifically, for the VAE, we first train a 2D image VAE that shares the spatial architecture with the target bidirectional 3D VAE on image data, allowing the model to establish a stable spatial representation. We then inflate the 2D VAE into a bidirectional 3D VAE and train it on 9-frame video clips to introduce temporal modeling. Finally, we fine-tune the VAE on high-quality videos with varying numbers of frames, enabling it to process videos of arbitrary length.

With the bidirectional VAE fixed, we train the DiT in the learned space. We begin with images repeated into short low-resolution static clips, *i.e.*, 5 frames at 256×256, to learn the basic restoration mapping. We then introduce video data and raise the resolution to 512×512 with 5 frames to

[1] `https://www.unsplash.com/`; `https://www.pexels.com/`

strengthen high-resolution detail synthesis. Finally, the temporal extent is progressively increased to 9 and then 21 frames at 512 × 512 to extend temporal context and long-range consistency. This progressive curriculum leads to rapid convergence and yields a multi-step rectified flow VOGVR model that performs restoration via iterative denoising. More details are in Appendix A.3.

Distillation to One-Step Inference. While the multi-step model can achieve strong restoration quality, its iterative sampling incurs non-trivial inference latency. To enable efficient deployment, we distill it into a one-step generator using a GAN-based one-step distillation strategy inspired by (Wu et al., 2025). Specifically, we initialize the one-step generator from the multi-step model and fine-tune the generator with LoRA (Hu et al., 2022). Given Gaussian noise z_1 and the visual conditions (*i.e.*, z^{LQ}, c_{sem} and c_{mot}) extracted from the LQ video, the generator predicts the clean latent in a single denoising step from $t = 1$ to $t = 0$. We optimize the one-step generator with reconstruction, perceptual, and adversarial objectives. An ℓ_1 loss and a perceptual loss computed on frozen DINOv3-ConvNeXt (Siméoni et al., 2025) features preserve content fidelity and perceptual structure, while two adversarial objectives improve spatial and temporal realism, respectively: a frame-level 2D discriminator leveraging frozen DINOv3-ConvNeXt (Siméoni et al., 2025) features, and a video-level 3D discriminator leveraging frozen V-JEPA2.1 (Mur-Labadia et al., 2026) features, each paired with a lightweight trainable head. The distillation is conducted at the same resolution and frame count as the final stage (21 frames at 512×512). The distilled one-step model is used as our default model. More details are provided in Appendix A.3.

4 Experiment

4.1 Experimental Settings

Implementation Details. We train VOGVR purely from visual data, without relying on pretrained T2I and T2V diffusion backbones. Our VOGVR model has approximately 1.6B parameters, including a 0.13B bidirectional VAE and a 1.4B DiT, and frozen visual encoders (DINOv2-B and Unimatch). The DiT consists of 30 transformer blocks, each comprising a self-attention layer, a cross-attention layer, and a gated feed-forward network, with a hidden dimension of 1536, 12 attention heads, and a feed-forward dimension of 8960. Following the progressive curriculum in Sec. 3.3, we first train the VAE to establish a stable video latent space, and then freeze it to train the DiT in the learned latent space. All models are trained using the PyTorch framework on 32 NVIDIA A800 GPUs. During inference, we use the distilled one-step model and process long videos with overlapping temporal chunks to fit GPU memory limits. More details of the network architecture, training configurations, and the training cost analysis are provided in the Appendix A.1 and A.3.

Testing Datasets. We evaluate our VOGVR model on two synthetic benchmarks, UDM10 (Yi et al., 2019) (10 sequences) and SPMCS (Tao et al., 2017) (30 sequences), and two real-world benchmarks, RealVSR (Yang et al., 2021) (50 sequences with paired HQ references) and VideoLQ (Chan et al., 2022b) (50 sequences without ground truth). For UDM10 and SPMCS, we synthesize LQ-HQ pairs following the RealBasicVSR-like degradation pipeline adopted in (Sun et al., 2025), consistent with our training setup.

Evaluation Metrics. We report full-reference metrics (PSNR, SSIM, LPIPS (Zhang et al., 2018)), no-reference image quality metrics (MUSIQ (Ke et al., 2021), CLIPIQA (Wang et al., 2023a), MANIQA (Yang et al., 2022)), no-reference video quality metrics (DOVER (Wu et al., 2023), FastVQA (Wu et al., 2022)), and the warping error E^*_{warp} ($\times 10^{-3}$) (Sun et al., 2025) for temporal consistency.

4.2 Experimental Results

To demonstrate the effectiveness of our VOGVR algorithm, we compare it with eight representative and state-of-the-art methods, including three T2I-based VR models (Upscale-A-Video (UAV) (Zhou et al., 2024), MGLD (Yang et al., 2024a), and DLoRAL (Sun et al., 2025)) and five T2V-based VR models (STAR (Xie et al., 2025), SeedVR1-7B (Wang et al., 2025c) (categorized here as T2V-based due to its video diffusion formulation), DOVE (Chen et al., 2025), SeedVR2-7B (Wang et al., 2025b), and FlashVSR (Zhuang et al., 2025)).

Quantitative Comparison. We show the quantitative comparison on both synthetic and real-world video benchmarks in Tab. 1, from which several key observations can be made. First, VOGVR achieves a strong balance between perceptual quality and temporal consistency across all benchmarks.

Table 1: Quantitative comparison of competing VR models across datasets. **Red** and **blue** highlight the best and second-best results, respectively.

Datasets	Metrics	Text-Driven Methods								Vision-Only
		T2I-based Methods			T2V-based Methods					
		UAV	MGLD	DLoRAL	STAR	SeedVR1	DOVE	SeedVR2	FlashVSR	Ours
UDM10	PSNR ↑	20.357	25.878	25.775	**26.087**	22.559	24.113	23.518	22.972	**26.346**
	SSIM ↑	0.626	0.748	0.768	**0.773**	0.722	0.759	0.755	0.697	**0.783**
	LPIPS ↓	0.295	0.323	0.327	0.405	0.297	**0.265**	0.287	0.299	**0.266**
	MUSIQ ↑	64.496	58.881	65.618	38.490	56.008	61.626	50.179	**67.219**	**67.528**
	CLIPIQA ↑	0.442	0.497	**0.651**	0.259	0.346	0.478	0.307	0.498	**0.505**
	MANIQA ↑	0.283	0.272	0.372	0.221	0.273	0.326	0.236	**0.379**	**0.393**
	DOVER ↑	45.221	40.104	43.448	30.746	46.561	53.812	42.211	**54.709**	**57.420**
	FastVQA ↑	0.604	0.668	0.722	0.455	0.709	0.722	0.661	**0.796**	**0.794**
	E^*_{warp} ↓	2.793	**1.317**	1.940	1.933	2.004	1.954	2.035	2.185	**1.893**
SPMCS	PSNR ↑	22.578	22.918	22.130	22.331	20.171	**23.237**	21.264	22.678	**23.049**
	SSIM ↑	0.486	0.556	0.569	0.534	0.500	**0.599**	0.550	0.549	**0.586**
	LPIPS ↓	0.473	0.391	0.366	0.595	0.372	**0.321**	0.359	0.324	**0.312**
	MUSIQ ↑	66.312	62.149	67.375	32.062	63.814	65.922	55.173	**69.915**	**70.795**
	CLIPIQA ↑	0.494	0.448	**0.580**	0.234	0.417	0.528	0.336	0.475	**0.529**
	MANIQA ↑	0.412	0.309	0.340	0.225	0.349	0.392	0.290	**0.412**	**0.418**
	DOVER ↑	40.283	31.247	35.006	17.303	**46.530**	43.064	35.353	**45.645**	42.712
	FastVQA ↑	0.569	0.578	**0.702**	0.283	0.693	0.672	0.576	**0.697**	0.691
	E^*_{warp} ↓	3.560	**0.478**	1.092	1.106	1.238	1.086	1.268	1.173	**1.068**
RealVSR	PSNR ↑	20.060	21.120	20.360	15.080	22.832	**23.393**	**22.955**	20.465	21.304
	SSIM ↑	0.591	0.646	0.606	0.433	0.706	**0.740**	**0.726**	0.558	0.633
	LPIPS ↓	0.263	0.219	0.242	0.409	0.196	**0.194**	**0.193**	0.246	0.229
	MUSIQ ↑	67.714	70.734	**70.908**	67.947	64.101	70.650	63.852	70.879	**73.332**
	CLIPIQA ↑	0.436	0.530	**0.617**	**0.532**	0.309	0.526	0.288	0.417	0.483
	MANIQA ↑	0.414	**0.484**	0.386	0.446	0.354	0.376	0.359	0.473	**0.487**
	DOVER ↑	36.136	42.044	**49.646**	30.214	41.430	**49.200**	43.081	40.454	45.607
	FastVQA ↑	0.558	0.612	0.580	0.475	**0.654**	0.648	0.633	0.632	**0.728**
	E^*_{warp} ↓	3.152	**2.421**	2.678	2.608	2.784	2.668	2.890	2.445	**2.307**
VideoLQ	MUSIQ ↑	57.489	53.092	**63.846**	49.305	42.839	48.583	38.592	59.720	**67.133**
	CLIPIQA ↑	0.377	0.315	**0.567**	0.333	0.235	0.305	0.218	0.364	**0.481**
	MANIQA ↑	0.327	0.247	**0.343**	0.186	0.203	0.247	0.193	0.323	**0.431**
	DOVER ↑	36.860	31.899	**38.505**	29.400	27.008	34.262	23.897	35.982	**42.838**
	FastVQA ↑	0.482	0.465	**0.615**	0.392	0.433	0.557	0.374	0.573	**0.631**
	E^*_{warp} ↓	19.438	17.989	19.062	18.565	17.974	**17.677**	18.322	18.296	**17.595**

It ranks first on MUSIQ and MANIQA on all four datasets, while achieving the best or second-best E^*_{warp} throughout. Consistently, on the two synthetic paired benchmarks, UDM10 and SPMCS, VOGVR ranks within the top two across PSNR, SSIM, and LPIPS. Second, VOGVR performs strongly on video-level quality assessment, particularly on FastVQA. It ranks first on RealVSR and VideoLQ, second on UDM10, and remains close to the top-performing methods on SPMCS. On DOVER, VOGVR achieves the best results on UDM10 and VideoLQ and remains competitive on the other benchmarks. Third, on RealVSR, DOVE, SeedVR1, and SeedVR2 achieve stronger full-reference fidelity, whereas VOGVR obtains substantially better perceptual/video quality on several metrics while maintaining the best temporal consistency.

Qualitative Comparison. We visualize the VR results and temporal profiles of T2V-based methods in Fig. 3. We see that across different regions, existing methods tend to over-smooth fine textures, distort local structures, or introduce visible artifacts. In contrast, VOGVR recovers sharper fur strands, cleaner artwork patterns, and better-preserved structural details. The temporal profiles further show that VOGVR maintains stable transitions across frames while preserving fine spatial details. This qualitative evidence aligns with our quantitative results, demonstrating the ability of VOGVR to preserve rich visual details while ensuring natural temporal consistency. More visual comparisons and video demos are provided in Appendix A.5 and the Supplementary File.

User Study. To complement the automatic metrics, we conduct a user study on all nine methods in the main comparison. Ten participants evaluate 12 test cases with method outputs presented in

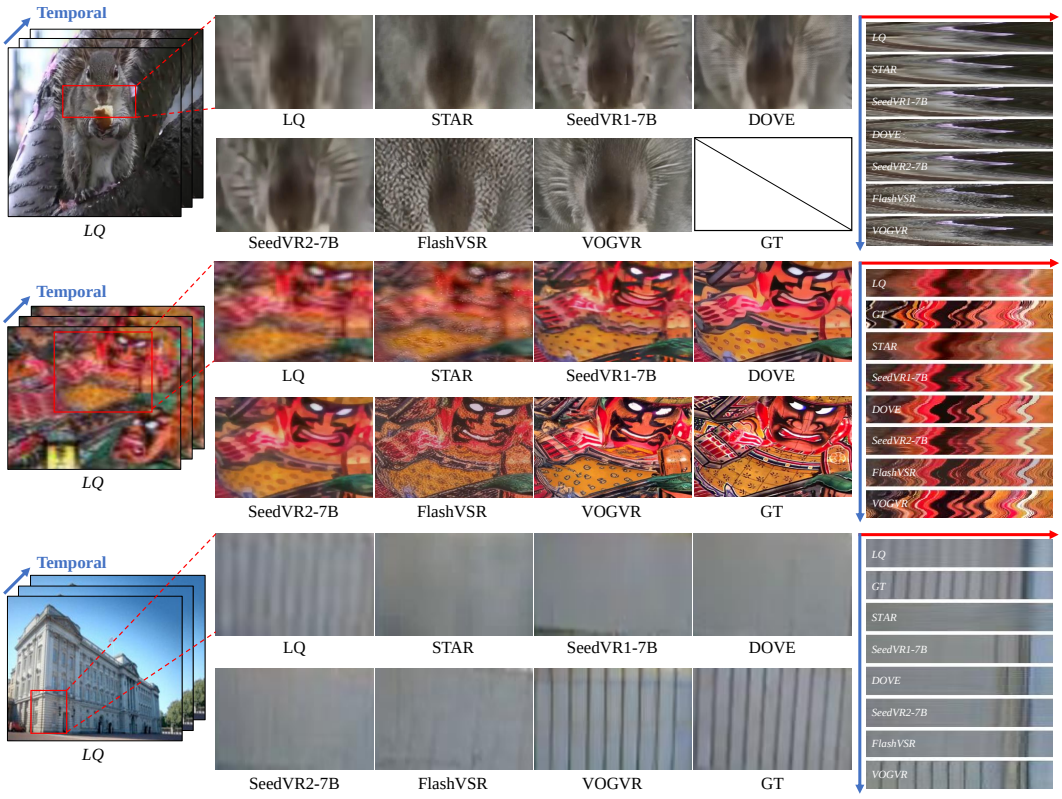


Figure 3: Qualitative comparison on real-world and synthetic datasets. Left: spatial details on cropped regions. Right: temporal profiles along the red scanline; note that smooth stripes indicate stable temporal transitions.

Table 2: User study results. Ten participants evaluate 12 test cases, yielding 120 selections in total.

Method	UAV	MGLD	DLoRAL	STAR	SeedVR1	DOVE	SeedVR2	FlashVSR	Ours
Votes	2	2	4	3	12	13	8	8	**68**

randomized order, selecting the best result for each case based on perceptual quality, detail fidelity, and temporal consistency. As shown in Tab. 2, VOGVR receives 68 of 120 selections (56.7%), compared with 13 (10.8%) for the strongest competitor. This study provides complementary evidence that human observers prefer the overall perceptual–fidelity–temporal balance of VOGVR.

Ablation Study. We further conduct ablation studies to validate the contribution of each key component in VOGVR, including the bidirectional VAE and restoration-oriented DiT. Detailed results and analysis can be found in Appendix A.2.

5 Conclusion

We proposed VOGVR to achieve temporally consistent and detail-rich generative video restoration without relying on pretrained T2I or T2V diffusion backbones. To resolve the trade-off between spatial detail synthesis and temporal coherence that plagues existing T2I/T2V-based methods, we first developed a bidirectional VAE to capture richer temporal context from both past and future frames, reducing latent ambiguity and improving fine-detail reconstruction. We then developed a restoration-oriented DiT conditioned on visual semantics and explicit motion cues extracted from the LQ input, anchoring the generation to the degraded video. The resulting VOGVR model demonstrated superior performance to previous generative VR methods, achieving rich spatial details and strong temporal coherence while remaining closely grounded in the degraded input, all with one-step inference and at a fraction of the training cost of existing approaches. There remain some limitations of VOGVR, which are discussed in Appendix A.6.

REFERENCES

Adrian Bulat and Georgios Tzimiropoulos. Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 109–117, 2018.

Adrian Bulat, Jing Yang, and Georgios Tzimiropoulos. To learn image super-resolution, use a gan to learn how to do image degradation first. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 185–200, 2018.

Jiezhang Cao, Yawei Li, Kai Zhang, and Luc Van Gool. Video super-resolution transformer. *arXiv preprint arXiv:2106.06847*, 2021.

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. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5972–5981, 2022a.

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, 2022b.

Junyang Chen, Jiangxin Dong, Long Sun, Yixin Yang, and Jinshan Pan. Stcdit: Spatio-temporally consistent diffusion transformer for high-quality video super-resolution. In *CVPR*, 2026.

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. *arXiv preprint arXiv:2505.16239*, 2025.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorber, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. *arXiv preprint arXiv:2403.03206*, 2024.

Dario Fuoli, Shuhang Gu, and Radu Timofte. Efficient video super-resolution through recurrent latent space propagation. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 3476–3485. IEEE, 2019.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *ArXiv*, abs/2207.12598, 2022. URL `https://api.semanticscholar.org/CorpusID:249145348`.

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

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5148–5157, 2021.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4681–4690, 2017.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022.

Wenbo Li, Xin Tao, Taian Guo, Lu Qi, Jiangbo Lu, and Jiaya Jia. Mucan: Multi-correspondence aggregation network for video super-resolution. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pp. 335–351. Springer, 2020.

Yawei Li, Kai Zhang, Jingyun Liang, Jiezhang Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demandolx, et al. Lsdir: A large scale dataset for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1775–1787, 2023.

Chengxu Liu, Huan Yang, Jianlong Fu, and Xueming Qian. Learning trajectory-aware transformer for video super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5687–5696, 2022a.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022b.

Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 2437–2445, 2020.

Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mahmoud Assran, Koustuv Sinha, Michael Rabbat, Yann LeCun, Nicolas Ballas, and Adrien Bardes. V-jepa 2.1: Unlocking dense features in video self-supervised learning. *arXiv preprint arXiv:2603.14482*, 2026.

Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 0–0, 2019.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

Qianqian Qiao, DanDan Zheng, Yihang Bo, Bao Peng, Heng Huang, Longteng Jiang, Huaye Wang, Jingdong Chen, Jun Zhou, and Xin Jin. Vadb: A large-scale video aesthetic database with professional and multi-dimensional annotations. *arXiv preprint arXiv:2510.25238*, 2025.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. URL https://arxiv.org/abs/2408.00714.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2022.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. URL https://arxiv.org/abs/2508.10104.

Yujing Sun, Lingchen Sun, Shuaizheng Liu, Rongyuan Wu, Zhengqiang Zhang, and Lei Zhang. One-step diffusion for detail-rich and temporally consistent video super-resolution. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 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.

Jiahao Wang, Yufeng Yuan, Rujie Zheng, Youtian Lin, Jian Gao, Lin-Zhuo Chen, Yajie Bao, Yi Zhang, Chang Zeng, Yanxi Zhou, et al. Spatialvid: A large-scale video dataset with spatial annotations. *arXiv preprint arXiv:2509.09676*, 2025a.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 2555–2563, 2023a.

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*, 2025b.

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*, 2025c.

Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 0–0, 2019.

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1905–1914, 2021.

Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C Kot, and Bihan Wen. Sinsr: Diffusion-based image super-resolution in a single step. *arXiv preprint arXiv:2311.14760*, 2023b.

Pengxu Wei, Yujing Sun, Xingbei Guo, Chang Liu, Guanbin Li, Jie Chen, Xiangyang Ji, and Liang Lin. Towards real-world burst image super-resolution: Benchmark and method. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13233–13242, 2023.

Haoning Wu, Chaofeng Chen, Jingwen Hou, Liang Liao, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Fast-vqa: Efficient end-to-end video quality assessment with fragment sampling. *Proceedings of European Conference of Computer Vision (ECCV)*, 2022.

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.

Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 25456–25467, 2024.

Rongyuan Wu, Lingchen Sun, Zhengqiang Zhang, Xiangtao Kong, Jixin Zhao, Shihao Wang, and Lei Zhang. Vosr: A vision-only generative model for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Zhiqiang Wu, Zhaomang Sun, Tong Zhou, Bingtao Fu, Ji Cong, Yitong Dong, Huaqi Zhang, Xuan Tang, Mingsong Chen, and Xian Wei. Omgsr: You only need one mid-timestep guidance for real-world image super-resolution, 2025. URL https://arxiv.org/abs/2508.08227.

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. *arXiv preprint arXiv:2501.02976*, 2025.

Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Yiran Xu, Taesung Park, Richard Zhang, Yang Zhou, Eli Shechtman, Feng Liu, Jia-Bin Huang, and Difan Liu. Videogigagan: Towards detail-rich video super-resolution. *arXiv preprint arXiv:2404.12388*, 2024.

Zhucun Xue, Jiangning Zhang, Teng Hu, Haoyang He, Yinan Chen, Yuxuan Cai, Yabiao Wang, Chengjie Wang, Yong Liu, Xiangtai Li, et al. Ultravideo: High-quality uhd video dataset with comprehensive captions. *arXiv preprint arXiv:2506.13691*, 2025.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1191–1200, 2022.

Xi Yang, Wangmeng Xiang, Hui Zeng, and Lei Zhang. Real-world video super-resolution: A benchmark dataset and a decomposition based learning scheme. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4781–4790, 2021.

Xi Yang, Chenhang He, Jianqi Ma, and Lei Zhang. Motion-guided latent diffusion for temporally consistent real-world video super-resolution. In *European Conference on Computer Vision*, pp. 224–242. Springer, 2024a.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024b.

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.

Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *arXiv preprint arXiv:2307.12348*, 2023.

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.

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.

A APPENDIX

In this appendix, we provide the following materials:

- Training cost comparison (referring to Introduction in the main paper).
- Ablation studies of key components in VOGVR (referring to Section 3.2 and Section 4.2 in the main paper).
- More training details (referring to Section 3.3 and Section 4.1 in the main paper).
- Complexity comparison of different methods (referring to Section 4.1 in the main paper).
- More real-world visual comparisons (referring to Section 4.2 in the main paper).
- Limitations.
- AI use statement.

In addition, the **Supplementary Materials** provide video demos on four scenes with different motion levels, comparing the LQ input, DOVE (Chen et al., 2025), DLoRAL (Sun et al., 2025), SeedVR2 (Wang et al., 2025b), FlashVSR (Zhuang et al., 2025), and VOGVR. The four sequences contain 29, 100, 100, and 100 frames, respectively.

A.1 TRAINING COST COMPARISON

Fig. 1(c) of the main paper compares the total training cost of representative video restoration methods, measured by the total number of training pixels consumed. We adopt a unified metric to measure the total training pixels consumed:

$$C = \sum_{s=1}^{S} N_s \times B_s \times H_s \times W_s \times F_s, \tag{5}$$

where S is the number of training stages, and N_s, B_s, $H_s \times W_s$, F_s denote the number of iterations, global batch size, spatial resolution, and number of frames per sample in stage s, respectively. For fairness, we count only the training cost of the core generative backbone (*i.e.*, UNet or DiT); all auxiliary modules, including VAEs and semantic/text encoders, are excluded for all methods regardless of whether they are pretrained, reused, or trained from scratch. All values marked with "$\geq$" are *conservative lower bounds* derived from publicly available information.

DLoRAL (SD 2.1). ***Base model.*** The Stable Diffusion 2.1 pre-training cost is fully documented on the HuggingFace model card[2]: Phase-1 trains for 550k iterations with batch size 2048 at 256^2 resolution (73.9 T pixels); Phase-2 for 850k iterations with batch size 2048 at 512^2 (456.3 T); the SD 2.1 continuation for 220k iterations with batch size 2048 at 512^2 (118.2 T). The total pre-training cost is 648 T pixels. ***Fine-tuning.*** The DLoRAL (Sun et al., 2025) LoRA fine-tuning uses 4 A100 GPUs, batch size 16, 3 frames at 512^2 (Sun et al., 2025); the number of iterations is $\sim$100k based on typical LoRA training schedules, yielding $\sim$1.3 T pixels. The total is $648 + 1.3 \approx 649$ T pixels.

SeedVR1 (SD3 init.). ***Base model.*** SeedVR1 (Wang et al., 2025c) initializes from SD3-Medium (Esser et al., 2024). SD3 only discloses the low-resolution pre-training phase (500k iterations $\times$ batch size 4096 $\times$ 256^2 = 134 T pixels); the high-resolution fine-tuning stages do not have publicly available iteration counts or batch sizes. We therefore use the fully documented SD 2.1 cost (648 T) as a conservative lower bound; the actual SD3 cost is strictly higher due to its additional high-resolution stages. ***Fine-tuning.*** The SeedVR1 paper reports training on 256 H100 GPUs for 30K H100 GPU-hours with $\sim$150 720p frames per batch per GPU, but does not disclose the number of iterations or the global batch size. We derive a lower bound as follows: the wall-clock time is $30{,}000/256 \approx 117$ hours; assuming a conservatively slow 15 s/iteration (an upper bound for this hardware and batch configuration), this yields $\sim$28k iterations; the per-iteration pixel count is $256 \times 150 \times 1280 \times 720 \approx 35.4$ G. The resulting lower bound is $28{,}000 \times 35.4\,\text{G} \approx 1{,}000$ T pixels. The total is $\geq 648 + 1{,}000 = 1{,}648$ T pixels.

DOVE (CogVideoX-1.5). ***Base model.*** DOVE (Chen et al., 2025) builds upon CogVideoX-1.5 (Yang et al., 2024b), which is further trained from CogVideoX-1.0. Since CogVideoX-1.5 does not have a

[2] https://huggingface.co/stabilityai/stable-diffusion-2-1

Table A.1: Reconstruction comparison between the causal VAE and BiVAE on clean HQ videos. The two VAEs share the same training data and optimization setting. The evaluation covers diverse motion regimes, including 9 static scenes from LSDIR, 40 mild-motion clips from RealBSR and YouHQ, and 12 substantial-motion clips from REDS and UDM10.

Clip Length	VAE	PSNR↑	SSIM↑	LPIPS↓	FID↓	E^*_{warp}↓
5	Causal	32.16	0.924	0.042	9.01	3.4
	BiVAE	**32.62**	**0.928**	**0.036**	**8.33**	**3.0**
9	Causal	32.32	0.926	0.042	7.97	3.2
	BiVAE	**32.81**	**0.931**	**0.037**	**7.74**	**3.0**
21	Causal	33.09	0.933	0.038	5.56	2.9
	BiVAE	**33.72**	**0.939**	**0.034**	**5.30**	**2.8**

Table A.2: Ablation study on key components of VOGVR. All variants are trained under a reduced schedule for controlled comparison. "Full Configuration" denotes the complete VOGVR architecture under the reduced 5-frame training setting with bidirectional VAE, semantic tokens c_{sem}, and motion tokens c_{mot}. The last row replaces temporal averaging with concatenation for c_{sem} aggregation. The best result of each metric is highlighted in **bold**. The results are evaluated on VideoLQ.

Configuration	MUSIQ ↑	CLIPIQA ↑	DOVER ↑	E^*_{warp} ↓
Full Configuration	60.53	0.339	**31.35**	14.4
w/ Causal VAE	55.51	0.296	27.16	14.3
w/o Semantic Tokens c_{sem}	60.43	0.328	27.67	14.4
w/o Motion Tokens c_{mot}	60.23	0.337	31.14	14.8
w/ Concatenated c_{sem}	**60.78**	**0.342**	31.34	**14.2**

separate technical report, we use the CogVideoX-1.0 training details from Table 5 of (Yang et al., 2024b) as a lower bound (CogVideoX-1.5 undergoes additional training and thus costs strictly more). CogVideoX employs a mixed image-video curriculum across four stages with Frame Packing. Assuming a 1:1 image:video ratio within each batch and 16 fps for video samples, the total cost across all four stages is ≥9,130 T pixels. ***Fine-tuning.*** DOVE fine-tuning parameters are fully specified (Chen et al., 2025): Stage-1 runs for 10k iterations with batch size 8, 25 frames at 320×640 (0.41 T); Stage-2 for 500 iterations with batch size 8, ∼5.8 frames at 320×640 (∼0.005 T). The total cost is $\geq 9{,}130 + 0.41 \approx 9{,}130$ T pixels.

VOGVR (from scratch). All training parameters are transparent: Stage-1 trains for 300k iterations with batch size 32, 5 frames at 256^2 (3.15 T); Stage-2 for 200k iterations with batch size 16, 5 frames at 512^2 (4.19 T); Stage-3 for 100k iterations with batch size 16, 9 frames at 512^2 (3.77 T); Stage-4 for 50k iterations with batch size 16, 21 frames at 512^2 (4.40 T). The one-step distillation runs for another 10k iterations with batch size 16, 21 frames at 512^2 (0.88 T). The total is 16.4 T pixels, with no base-model pre-training cost.

A.2 Ablation Study

To validate the key architectural and conditioning design choices of VOGVR, we conduct a series of controlled ablation studies. We first evaluate the reconstruction capability of the bidirectional VAE independently of the DiT. For the subsequent DiT-based studies, we adopt a reduced 5-frame setting, in which both the VAE and DiT are trained from scratch up to the 5-frame stage. Specifically, we use 13M images and 400K video clips, with 100K image-training iterations followed by 50K video-training iterations, and evaluate all variants on 5-frame inputs to maintain consistent temporal lengths between training and evaluation. Unless otherwise specified, all variants use the same training data, optimization settings, and evaluation protocol. The full-scale model in Tab. 1 instead follows the complete progressive training schedule described in Sec. 3.3.

Effectiveness of Bidirectional VAE. We first isolate the reconstruction capability of the VAE by directly comparing BiVAE with its causal counterpart on clean HQ videos. The two models share the same architecture and optimization settings, differing only in whether the temporal convolutions and chunk-wise processing are causal or bidirectional. The evaluation set covers diverse motion regimes.

Table A.3: Controlled ablation of restoration-oriented conditioning with a fixed pretrained Wan2.1 VAE. All variants use the same randomly initialized Wan2.1 DiT, training data, and optimization schedule; only the conditioning signals differ. Results are evaluated on VideoLQ.

Condition	MUSIQ ↑	DOVER ↑
z^{LQ} + Text	55.54	26.59
z^{LQ} + Text + c_{mot}	57.94	28.01
z^{LQ} + c_{mot}	57.95	26.02
z^{LQ} + c_{sem} + c_{mot}	**63.44**	**36.64**

Table A.4: Controlled comparison between T2V-pretrained and our native vision-only (VO) training settings. Both settings use the same Wan2.1 DiT architecture, pretrained Wan2.1 VAE, training data, and optimization schedule. The results are evaluated on VideoLQ and RealVSR.

Configuration	VideoLQ		RealVSR				
	MUSIQ↑	DOVER↑	PSNR↑	SSIM↑	LPIPS↓	MUSIQ↑	DOVER↑
T2V-pretrained DiT + Text	**64.80**	36.16	19.06	0.548	0.290	73.80	37.37
Randomly Init. DiT + VO	63.44	**36.64**	**20.29**	**0.587**	**0.254**	**74.01**	**39.90**

The static subset consists of 9 randomly sampled single-frame scenes from LSDIR (Li et al., 2023), each repeated temporally to form a video sequence. The mild-motion subset contains 40 clips from RealBSR (Wei et al., 2023) and YouHQ (Zhou et al., 2024) with relatively mild global and local motion, while the substantial-motion subset contains 12 clips selected from REDS (Nah et al., 2019) and UDM10 (Yi et al., 2019) with the largest motion magnitude. As shown in Tab. A.1, BiVAE consistently improves reconstruction fidelity, perceptual quality, and temporal consistency across all evaluated clip lengths (5, 9, and 21 frames). For example, at 21 frames, BiVAE improves PSNR from 33.09 to 33.72 dB and reduces E^*_{warp} from 2.9 to 2.8. These results show that exploiting both preceding and subsequent observations enables the VAE to construct more faithful and temporally coherent latent representations.

We further examine whether this advantage translates to downstream restoration by replacing BiVAE with its causal counterpart and training the corresponding DiT from scratch under the same reduced 5-frame schedule. As shown in Tab. A.2, the causal variant decreases MUSIQ from 60.53 to 55.51, CLIPIQA from 0.339 to 0.296, and DOVER from 31.35 to 27.16, while E^*_{warp} remains comparable (14.4 vs. 14.3). Together, the reconstruction and downstream results show that bidirectional temporal modeling improves VAE reconstruction and substantially benefits downstream restoration quality.

Effectiveness of Restoration-Oriented Conditioning. We then evaluate the two categories of restoration-specific conditioning: semantic tokens c_{sem} from DINOv2 and motion tokens c_{mot} from optical flow. Removing c_{sem} slightly decreases MUSIQ from 60.53 to 60.43 and CLIPIQA from 0.339 to 0.328 while causing a larger drop in DOVER from 31.35 to 27.67; E^*_{warp} remains unchanged at 14.4. Removing c_{mot} leads to smaller changes in perceptual metrics but increases E^*_{warp} from 14.4 to 14.8, indicating the benefit of explicit motion guidance for temporal consistency.

We also examine the aggregation strategy for c_{sem}. Concatenating per-chunk features yields only marginal differences compared with temporal averaging (60.78 vs. 60.53 MUSIQ, 0.342 vs. 0.339 CLIPIQA, 31.34 vs. 31.35 DOVER, and 14.2 vs. 14.4 E^*_{warp}), while linearly increasing the cross-attention token length with video duration. We therefore adopt temporal averaging as a more compact aggregation strategy with comparable restoration quality.

To further isolate the effect of the conditioning design from the proposed BiVAE, we conduct an additional controlled study using a fixed pretrained Wan2.1 VAE. All variants use the same randomly initialized Wan2.1 DiT, training data, optimization schedule, and pretrained VAE; only the conditioning signals differ. The text condition is extracted from the first LQ frame using a fixed lightweight degradation-aware tagger (Wu et al., 2024). As shown in Tab. A.3, adding the motion condition c_{mot} to the text-conditioned baseline improves MUSIQ from 55.54 to 57.94 and DOVER from 26.59 to 28.01. Starting from $z^{\mathrm{LQ}} + c_{\mathrm{mot}}$, further introducing the visual semantic condition c_{sem} substantially improves MUSIQ from 57.95 to 63.44 and DOVER from 26.02 to 36.64. Moreover, under the same motion condition, replacing text conditioning with c_{sem} improves MUSIQ/DOVER from 57.94/28.01 to 63.44/36.64. These results provide additional evidence for the effectiveness of the proposed semantic and motion conditioning under a fixed pretrained VAE.

Table A.5: Ablation study of the one-step distillation strategy. All one-step variants are initialized from the same full-scale multi-step model and distilled for the same number of iterations.

Configuration	VideoLQ				RealVSR					
	MUSIQ↑	DOVER↑	FastVQA↑	$E^*_{\text{warp}}\downarrow$	PSNR↑	LPIPS↓	MUSIQ↑	DOVER↑	FastVQA↑	$E^*_{\text{warp}}\downarrow$
Multi-step Model	59.75	35.21	0.552	18.200	22.44	**0.197**	72.13	41.20	0.661	2.700
$\mathcal{L}_1 + \mathcal{L}_{\text{DISTS}}$	44.54	26.97	0.428	18.100	**23.64**	0.274	60.55	35.59	0.598	2.710
$+\mathcal{L}^{3D}_{\text{GAN}}$	65.68	42.68	0.624	18.000	20.72	0.283	70.78	43.40	0.700	2.650
$+\mathcal{L}^{2D}_{\text{GAN}} + \mathcal{L}^{3D}_{\text{GAN}}$	**67.37**	**43.95**	**0.637**	**17.400**	20.79	0.248	**74.34**	**47.53**	0.723	2.530
$+\mathcal{L}^{2D}_{\text{GAN}} + \mathcal{L}^{3D}_{\text{GAN}}$ (pure video)	67.13	42.84	0.631	17.595	21.30	0.229	73.33	45.61	**0.728**	**2.307**

Effectiveness of One-Step Distillation. We further analyze the one-step distillation strategy using the full-scale multi-step model as the reference. Inspired by (Wu et al., 2025), we adopt reconstruction and perceptual objectives together with adversarial supervision, and extend this framework to video restoration with video-level adversarial learning. As shown in Tab. A.5, when the multi-step model is distilled using only the reconstruction and perceptual objectives, i.e., $\mathcal{L}_1 + \mathcal{L}_{\text{DISTS}}$, distortion-oriented fidelity is largely preserved, but perceptual and video quality degrade substantially. Introducing the video-level adversarial objective $\mathcal{L}^{3D}_{\text{GAN}}$ substantially improves the one-step model. On VideoLQ, MUSIQ and DOVER increase from 44.54 and 26.97 to 65.68 and 42.68, respectively, while FastVQA improves from 0.428 to 0.624. Consistent gains are observed on RealVSR, where MUSIQ, DOVER, and FastVQA improve from 60.55, 35.59, and 0.598 to 70.78, 43.40, and 0.700. These results demonstrate the importance of video-level adversarial supervision for preserving perceptual and spatio-temporal quality during one-step distillation.

We further complement the video-level discriminator with the frame-level adversarial objective $\mathcal{L}^{2D}_{\text{GAN}}$. Combining the 2D and 3D adversarial objectives consistently improves perceptual and video quality over using the 3D objective alone. On RealVSR, MUSIQ, DOVER, and FastVQA increase from 70.78, 43.40, and 0.700 to 74.34, 47.53, and 0.723, respectively. Similar improvements are observed on VideoLQ, where MUSIQ, DOVER, and FastVQA increase from 65.68, 42.68, and 0.624 to 67.37, 43.95, and 0.637. This suggests that frame-level spatial realism and video-level spatio-temporal realism provide complementary supervision.

Finally, replacing the mixed image-video distillation data inherited from SFT stage with pure video data yields a better fidelity–temporal balance on RealVSR. Specifically, PSNR improves from 20.79 to 21.30, LPIPS decreases from 0.248 to 0.229, FastVQA increases from 0.723 to 0.728, and E^*_{warp} decreases from 2.530 to 2.307, while the perceptual metrics remain strong. We therefore adopt the pure video data setting with both 2D and 3D adversarial objectives for the final one-step model.

Effectiveness of Native Vision-Only Paradigm. We further compare the native vision-only configuration with a conventional T2V-pretrained counterpart under the same restoration-training protocol. The native vision-only configuration corresponds to the full visual-conditioning setting in Tab. A.3, using a randomly initialized DiT with z^{LQ}, c_{sem}, and c_{mot} as conditions. Both configurations use the same Wan2.1 DiT architecture, pretrained Wan2.1 VAE, training data, and optimization schedule, while the T2V-pretrained counterpart additionally benefits from large-scale T2V pretraining and retains text conditioning.

As shown in Tab. A.4, the T2V-pretrained configuration achieves slightly higher MUSIQ on VideoLQ (64.80 vs. 63.44), while our native vision-only variant attains slightly better DOVER (36.64 vs. 36.16). On RealVSR, our variant maintains comparable perceptual quality in terms of MUSIQ (74.01 vs. 73.80), while achieving better video quality and substantially stronger input fidelity: DOVER improves from 37.37 to 39.90, PSNR from 19.06 to 20.29, SSIM from 0.548 to 0.587, and LPIPS from 0.290 to 0.254. Since both configurations share exactly the same restoration data and training settings, while the T2V-pretrained counterpart additionally benefits from large-scale T2V pretraining, these results support the effectiveness of the proposed native vision-only paradigm rather than attributing its performance to a larger restoration training set.

A.3 Training Details

Both the VAE and the DiT are trained from scratch in a progressive manner, transitioning from images to videos and from short to long temporal clips. For the DiT, we additionally increase the spatial

Table A.6: Training recipe for VOGVR.

	Multi-step DiT				Distill to
	PT $5\times256_{px}$	PT $5\times512_{px}$	SFT $9\times512_{px}$	SFT $21\times512_{px}$	One-step
Learning rate	1.0×10^{-4}	1.0×10^{-4}	1.0×10^{-4}	1.0×10^{-4}	1.0×10^{-4}
LR scheduler	Constant	Constant	Constant	Constant	Constant
Warm-up steps	0	0	0	0	0
Training steps	300K	200K	100K	50K	10K
Global batch size	32	16	16	16	16
Weight decay	0.01	0.01	0.01	0.01	0.01
Gradient clip	1.0	1.0	1.0	1.0	1.0
EMA	0.9999	0.9999	0.9999	0.9999	0.999
Optimizer	AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$)				

resolution from 256×256 to 512×512. Starting from the video-training stages of both components, we adopt a *mixed image-video training strategy*: each batch consists of N-frame video clips and single images repeated to N frames. This paradigm has been widely adopted in prior works (Sun et al., 2025; Wan et al., 2025) to strengthen spatial detail synthesis while enabling temporal modeling. We empirically set the image-video mixing ratio to 50%, which provides a good balance between spatial quality and temporal modeling capability.

VAE Training. We train our bidirectional VAE from scratch in a progressive manner. The training objective is $\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_{\text{LPIPS}}$ throughout all stages, and the optimizer is Adam with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. In the first stage, we train the model as a 2D image autoencoder on images at 256×256 resolution for 320K steps with a batch size of 480, which allows the encoder and decoder to establish a stable spatial representation. In the second stage, we inflate the model into a 3D bidirectional video VAE and train it on 9-frame video clips and repeated images at 256×256 resolution for 50K steps, introducing temporal modeling. In the third stage, we fine-tune the VAE on video clips and repeated images with variable frame counts uniformly sampled from $\{5, 9, 13, 17, 21\}$ at 256×256 resolution for 90K steps with a batch size of 16, enabling it to generalize to variable video lengths. Across all stages, we use a constant learning rate of 1.0×10^{-4}, no weight decay, a gradient clipping threshold of 1.0, and bf16 mixed precision. The latent dimension is 16 with a spatial downsampling factor of $8\times$ and a temporal downsampling factor of $4\times$.

DiT Training. With the bidirectional VAE frozen, we train the DiT backbone from scratch in the learned latent space. As shown in Table A.6, we train VOGVR in a progressive manner. Specifically, the multi-step model is first pretrained at 256×256 resolution for 300K steps with a global batch size of 32 and a constant learning rate of 1.0×10^{-4}. In this stage, we use only image data, where each image is repeated into a 5-frame static clip so that the model can learn the basic restoration mapping in a video-shaped format while the optical flow tokens are naturally zero. From the second stage onward, the mixed image-video training strategy is applied. The model is further pretrained at 512×512 resolution on 5-frame dynamic video clips and repeated images for 200K steps with a global batch size of 16 and the same learning rate, strengthening high-resolution detail synthesis and introducing real video dynamics. Next, we fine-tune the model on high-quality filtered data with 9-frame video clips and repeated images at 512×512 resolution for 100K steps, extending temporal context. Finally, we extend to 21-frame clips and repeated images at 512×512 resolution for 50K steps to improve long-range temporal consistency. The optical flow condition is enabled across all DiT training stages. For the DiT backbone, we use a hidden dimension of 1536, depth of 30, 12 attention heads and a patch size of (1,2,2). During both multi-step training and distillation, the partially conditioned branch is sampled with a probability of 0.1 (*i.e.*, 10% of training iterations use the weakened condition $z^{\text{LQ}}_{\text{weak}}$ while the remaining 90% use the fully conditioned mode).

One-Step Distillation. After obtaining the converged multi-step model, we adapt it into a one-step generator by using the GAN-based distillation strategy. For the distillation stage, we use pure video training data rather than the mixed image-video data adopted during multi-step training, as validated in Tab. A.5. The one-step generator is initialized from the EMA checkpoint of the multi-step model, which serves only as initialization during distillation. Given Gaussian noise $z_1 \sim \mathcal{N}(0, I)$ and the visual conditions $(z^{\text{LQ}}, c_{\text{sem}}, c_{\text{mot}})$ extracted from the LQ video, the generator directly recovers the

Table A.7: Complexity comparison among different methods. All methods are evaluated using 50 2048×2048 frames for the VR task. Inference time is measured on an A100 GPU and includes the entire pipeline: data loading, processing, and result storage.

	UAV	MGLD	DLoRAL	STAR	DOVE	SeedVR1	SeedVR2	FlashVSR	**Ours**
Inference Step	30	50	1	15	1	50	1	1	1
Inference Time (s/50 frames)	3640	4146	346	2830	338	37206	304	**222**	294
# Total Param (M)	14442	1430	**1300**	2492	10549	8240	8240	1780	1628

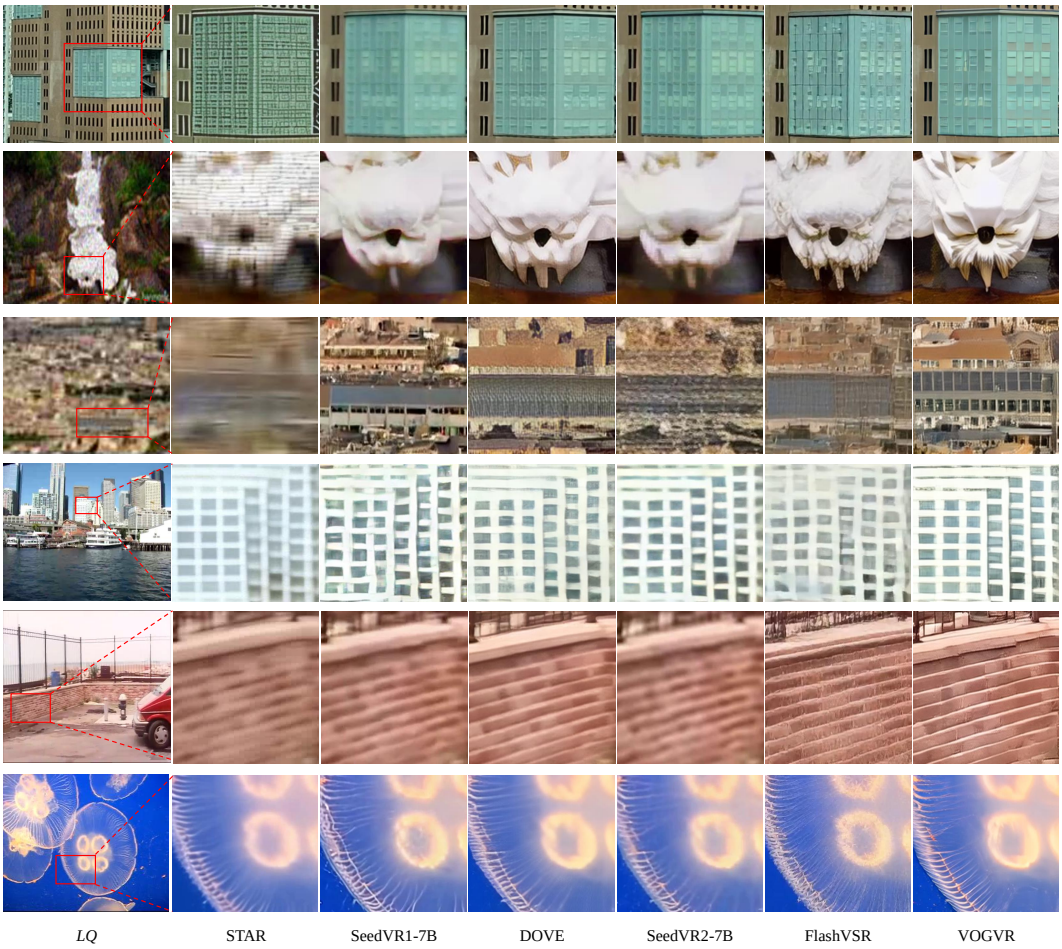


Figure A.1: Additional qualitative comparisons with T2V-based methods on videos from UDM10, SPMCS, RealVSR, and VideoLQ. VOGVR produces richer spatial details and more realistic textures while preserving temporal coherence.

clean latent with:

$$\hat{z}_0 = z_1 - v_\theta(z_1, t = 1, z^{\mathrm{LQ}}, c_{\mathrm{sem}}, c_{\mathrm{mot}}). \tag{6}$$

We optimize the one-step generator with LoRA using four complementary objectives:

$$\mathcal{L}_G = \mathcal{L}_1 + \mathcal{L}_{\mathrm{DISTS}} + \mathcal{L}^{2D}_{\mathrm{GAN}} + \mathcal{L}^{3D}_{\mathrm{GAN}}, \tag{7}$$

with all loss weights set to 1.0. The ℓ_1 reconstruction loss preserves pixel-level fidelity, while the perceptual loss $\mathcal{L}_{\mathrm{DISTS}}$ preserves perceptual structures by computing DISTS over the features from the first three stages of a frozen DINOv3-ConvNeXt (Siméoni et al., 2025) backbone. For adversarial supervision, we employ both image- and video-level discriminators. The 2D discriminator evaluates individual frames using frozen multi-level DINOv3-ConvNeXt (Siméoni et al., 2025) features followed by trainable discriminator heads, encouraging realistic spatial details. The 3D discriminator instead operates on video clips, using a frozen V-JEPA 2.1 (Mur-Labadia et al., 2026)

backbone followed by trainable multi-level heads to capture temporal dynamics and encourage video-level realism. Both discriminators are optimized with multi-level binary cross-entropy losses, using 0.8 and 0 as the real and fake labels, respectively.

A.4 Complexity Comparison

As shown in Table A.7, we compare the inference efficiency and model size of all competing methods. Multi-step methods such as Upscale-A-Video (*i.e.,* UAV) (30 steps), MGLD (50 steps), STAR (15 steps), and SeedVR1 (50 steps) incur substantially higher inference latency, ranging from 2,830s to over 37,000s for 50 frames, which entails significant computational overhead. VOGVR achieves comparable model size and inference speed to DLoRAL (1,300M, 346s) and FlashVSR (1,780M, 222s), with 1,628M parameters and 294s per 50 frames, while delivering superior restoration quality as shown in Table 1 of the main paper.

A.5 More Visual Comparisons

We provide additional qualitative comparisons using videos from our evaluation benchmarks (UDM10, SPMCS, RealVSR, and VideoLQ) in Fig. A.1. Video demos are provided in Supplementary Materials. Across diverse scenes and degradation types, VOGVR consistently recovers sharper textures and better-preserved structural details than competing T2V-based methods, while maintaining smooth temporal transitions.

A.6 Limitation

Despite its strong performance, VOGVR has some limitations. First, optical flow estimated from heavily degraded LQ videos may be inaccurate, propagating errors into the motion condition. Second, the spatial and temporal compression of the VAE may sacrifice some fine-scale structures (*e.g.*, small text). Pixel-space diffusion, which learns task-adaptive compression implicitly, offers a promising direction to address this issue. We leave these challenges for future work.

A.7 AI use statement

In this work, generative AI tools were used for manuscript editing, language polishing, and code assistance. All AI-assisted outputs were reviewed and verified by the authors, including language and code. The authors take full responsibility for the final content of this work.