SCOPE-VSR: A BROAD-COVERAGE TRAINING DATASET FOR VIDEO SUPER-RESOLUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion models have shown superior performance in video super-resolution (VSR). However, restoring diverse real-world content remains challenging due to the lack of balanced semantic coverage in existing training data. In this paper, we introduce a structured curation pipeline for open-domain VSR, SCOPE. In this pipeline, we combine task-oriented video selection with hierarchical semantic rebalancing. Applying this pipeline to a large-scale collection of data sources, we construct the SCOPE-VSR dataset for training and SCOPE-test for testing. SCOPE-VSR has approximately 27k videos with a balanced and diverse coverage of various real-world scenarios and subjects. In addition, we propose GraCo-VSR, a novel framework for one-step video restoration. GraCo-VSR adopts a reliability-aware design with explicit supervision to reduce degradation-induced distortions in motion and texture guidance. This design enables the model to better leverage the diverse training content in SCOPE-VSR, thereby supporting generalization across varied scenarios. Experiments demonstrate that our model exhibits strong performance on both perceptual fidelity and visual quality.

1 INTRODUCTION

The growing demand for high-quality visual content in digital media production has made effective video restoration increasingly important. Video super-resolution (VSR) aims to reconstruct high-quality, ground-truth (GT) videos from degraded, low-resolution (LR) inputs. Conventional VSR methods mainly model spatial-temporal dependencies with CNN- or Transformer-based architectures (Jo et al., 2018; Fuoli et al., 2019; Wang et al., 2019; Lucas et al., 2019). However, they are commonly developed under limited and predefined degradation settings, restricting their effectiveness on real-world videos with complex and unknown degradations.

Recently, diffusion models have shown strong potential in image and video restoration due to their powerful generative priors (Wu et al., 2024b;a; Blattmann et al., 2023). While diffusion priors enhance the synthesis of fine-grained details, their generative nature can also introduce frame-to-frame variations that undermine temporal consistency. A common strategy is therefore to incorporate spatial and temporal cues as complementary guidance and constraints, balancing perceptual fidelity with temporal coherence (Zhou et al., 2024; Wang et al., 2025; Xie et al., 2025).

However, progress in open-domain VSR is impeded by a lack of task-oriented datasets. Broadening the semantic scope of training data provides richer textural priors, enabling better generalization to complex real-world conditions. Different scenes and subjects expose models to distinct textures, structures, and motion patterns. A useful training set should therefore combine broad semantic coverage with informative restoration content, rather than simply increase the number of visually similar clips. Existing VSR datasets often lack fine-grained category organization and tend to exhibit long-tailed distributions across scenarios and subjects (Zhao et al., 2025; Yang et al., 2021; An et al., 2026). Even large-scale datasets can still suffer from semantic imbalance. For instance, YouHQ (Zhou et al., 2024) remains dominated by a few frequent categories and provides limited coverage of content such as sports, as shown in Fig. 4. Such imbalance can bias models toward frequently observed content. Consequently, their generalization to open-domain scenarios is limited.

To address this limitation, we propose SCOPE, a structured curation pipeline for open-domain and perceptually enriched videos. We firstly filter candidate videos using image quality assessment metrics and wavelet-frequency responses to retain high-quality content with rich details. A foreground-

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

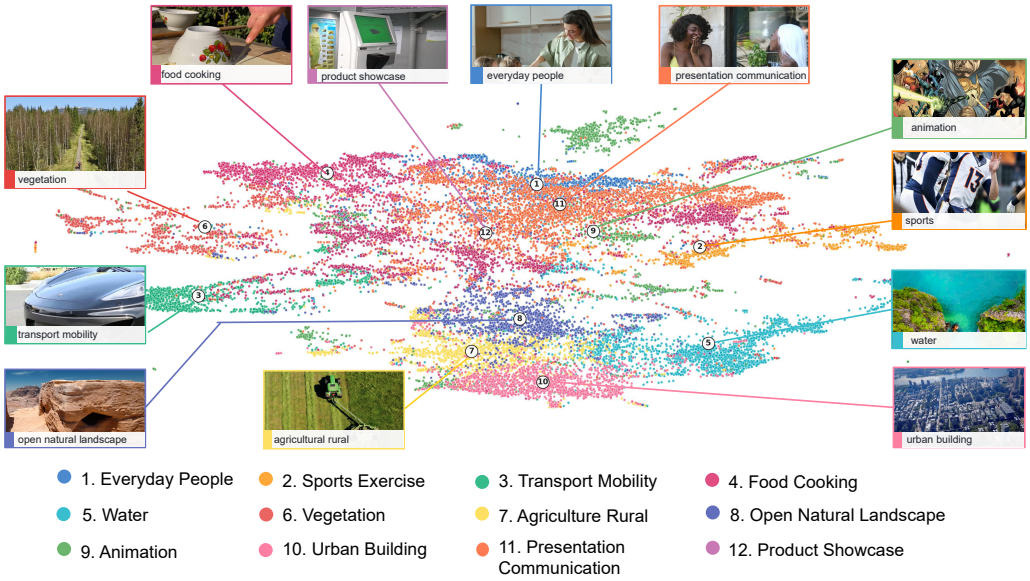

Figure 1: The semantic "atlas" of SCOPE-VSR. We use InternVideo2 (Wang et al., 2024c) to extract video embeddings and assign topic labels, then project the embeddings into 2D. Colors denote topics, with representative frames illustrating the diverse semantic content of our dataset.

sensitive mining mechanism then prioritizes clips with informative spatial details and meaningful motion, while semantic rebalancing improves the distribution of scenarios and subjects. Applying SCOPE to five existing video collections yields SCOPE-VSR, a high-quality dataset of 27,351 videos. As shown in Fig. 1, SCOPE-VSR exhibits broad and balanced semantic coverage, exposing models to a range of real-world content while limiting the dominance of frequent categories.This design couples semantic rebalancing with task-oriented selection, so that broader category coverage is pursued while retaining videos with useful spatial and temporal information.

Despite the importance of high-quality data, real-world VSR still requires a well-designed architecture to effectively exploit cues from degraded inputs. Recent diffusion-based VSR methods commonly introduce auxiliary conditioning modules for temporal coherence and detail recovery (He et al., 2024; Wu et al., 2026). However, these modules often rely on priors estimated from LR input, whose reliability can be affected by complex degradation patterns, particularly when useful signals are weak or ambiguous. For example, degradation artifacts may resemble fine textures, while blurred boundaries can reduce the accuracy of motion estimation. Consequently, such priors may introduce uncertainty into the conditioning process and limit restoration quality.

To this end, we propose GraCo-VSR, a one-step restoration model that aligns reliable dynamic and texture guidance with their natural granularities. For dynamic guidance, the Clip-Level Dynamics Conditioning (CLDC) distills a lightweight multi-scale flow network from a RAFT model to obtain robust motion cues from adjacent LR frames. It then aggregates flow representation into a clip-level global dynamics embedding, guiding the DiT to better capture overall video motion during restoration. For texture guidance, the Reliability-Aware Texture Encoder (RATE) extracts frequency responses from low-resolution (LR) videos with a fixed frequency bank to characterize local textures. It further learns pixel-wise confidence under GT-frequency supervision, encouraging the texture prior to emphasize authentic visual details rather than degradation-induced noise.

Our contributions can be summarized as follows:

- We propose SCOPE, a structured curation pipeline for open-domain VSR that selects high-quality content with rich visual details and diverse semantics. Using SCOPE, we construct SCOPE-VSR, a large-scale VSR dataset broadly covering real-world scenarios.
- We introduce GraCo-VSR, a granularity-aligned conditioning model that improves the reliability of priors extracted from degraded inputs through GT-supervised texture confidence estimation and RAFT-distilled clip-level dynamics modeling.
- Extensive experiments demonstrate that our method achieves superior performance and robustness across diverse video content under challenging real-world degradations.

2 Related Work

Video Super-Resolution. Early VSR methods, including sliding-window-based (Wang et al., 2019) and recurrent-based (Fuoli et al., 2019) frameworks, commonly exploit temporal redundancy across multiple frames to provide complementary information for restoration (Li et al., 2020; Chan et al., 2021; 2022a). These methods are usually optimized on synthetic degradations with pixel-wise objectives, which often leads to over-smoothed textures in real-world VSR tasks. To address limitation in these scenarios, GAN-based restoration methods introduce adversarial learning to synthesize sharper details (Wang et al., 2021; Xu et al., 2025). Other real-world VSR methods, such as RealBasicVSR (Chan et al., 2022b), focus on complex degradation modeling with temporal propagation to handle noise and compression artifacts. Despite the improvements, these methods still struggle in open-domain video restoration, where recovering faithful details is challenging.

Diffusion Based VSR. The impressive performance of diffusion models in image super-resolution (Wang et al., 2024b; Wu et al., 2024b;a) has motivated their adaptation in recent VSR methods. To adapt diffusion priors to video restoration, existing methods introduce explicit temporal guidance to preserve the perceptual advantages of generative restoration while suppressing inter-frame inconsistencies (Li et al., 2025; Zhou et al., 2024; Yang et al., 2024; Chen et al., 2026b). More recent methods exploit the spatial and temporal priors encoded in pretrained text-to-video (T2V) models (Xie et al., 2025; Zhuang et al., 2026; Chen et al., 2026a). VEnhancer (He et al., 2024) adopts a ControlNet-based design to inject features from low-resolution key frames, while PS-SR (Wu et al., 2026) further introduces speculative refinements with frequency-domain constraints to inject high-frequency details and preserve content consistency. Nevertheless, the extracted guidance may still entangle useful restoration evidence with degradation-induced responses, limiting the reliability of conditioning under complex real-world degradations.

VSR Datasets. To address the complex degradations encountered in real-world videos, several datasets have been developed specifically for VSR. RealVSR (Yang et al., 2021) captures paired LR-HR videos with smartphones, thereby reflecting degradations arising from practical imaging systems. YouHQ (Zhou et al., 2024) collects a large-scale collection of real-world videos from YouTube and applies a mixed degradation pipeline, whereas RealisVSR (Zhao et al., 2025) focuses on fine-grained details by constructing a high-resolution 4K video dataset. Recent studies have also curated data for specialized restoration tasks (Zou et al., 2026; An et al., 2026; Wang et al., 2024a). Nevertheless, existing datasets rarely consider diversity and balance across semantic categories, limiting their ability to support stable VSR performance over broad open-domain scenarios.

3 Method

Our approach comprises two components:SCOPE, a pipeline for constructing the large-scale, high-quality SCOPE-VSR dataset, and GraCo-VSR, which incorporates GT-supervised motion and texture guidance at complementary granularities for open-domain VSR.

3.1 Dataset Curation Pipeline

High-quality datasets are essential for fully exploiting the generative priors of pretrained models in VSR. However, existing VSR training datasets (Nah et al., 2019; Yang et al., 2021; Zhou et al., 2024) are either limited in scale or curated without sufficient consideration of task-specific characteristics, such as spatial detail and temporal variation. Broad and balanced semantic coverage is also necessary for models to generalize to complex real-world scenarios. Nevertheless, large-scale video curation pipelines (Chen et al., 2026c) often prioritize visual quality alone, which may weaken the basis for models to learn restoration priors across diverse scenarios.

To address these limitations, we propose **SCOPE**, a systematic pipeline that combines multi-stage curation with semantic rebalancing to construct training data for open-domain VSR. As shown in Fig. 2, SCOPE comprises two stages of filtering, followed by a foreground-sensitive mining stage and a hierarchical semantic rebalancing stage. Details are as follows.

Meta Filtering. We first perform basic data cleaning by removing videos below 720p, thereby ensuring sufficient spatial information for subsequent processing. To ensure temporal consistency, we detect shot boundaries and segment the retained videos into coherent clips.

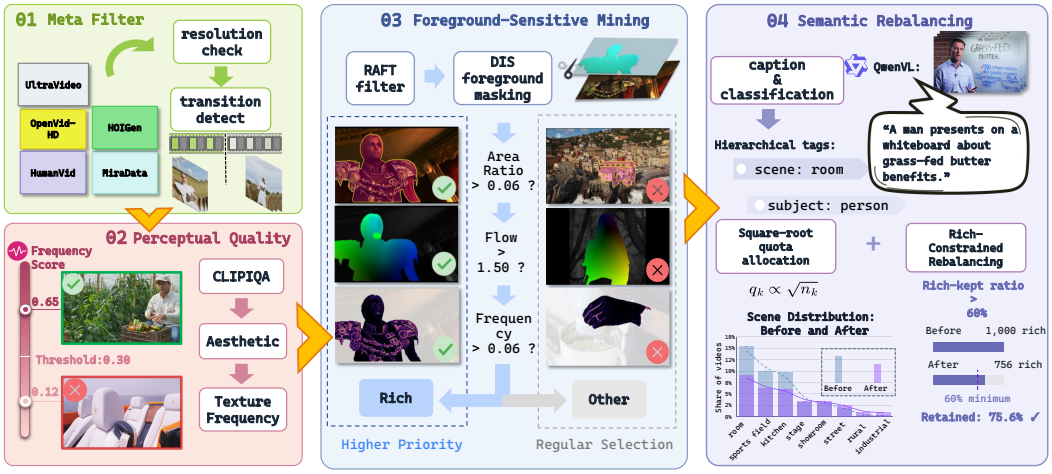

Figure 2: Overview of our dataset curation pipeline, SCOPE. The four stages produce SCOPE-VSR, a dataset of 27k high-quality videos with rich visual details and balanced semantic coverage.

Perceptual Quality Screening. We evaluate the perceptual quality of each video by applying CLIP-IQA (Wang et al., 2023) and LAION (Schuhmann et al., 2022). Frequency-domain responses reveal the richness of spatial details. We therefore apply a wavelet transform (Farge et al., 1992) to quantify the texture richness of each candidate clip and retain videos with a score higher than 0.03.

Foreground-Sensitive Mining. To exclude nearly static videos, we estimate optical flow with a RAFT model (Wang et al., 2024d) and measure the overall motion intensity. Videos dominated by camera motion, however, may bias the model toward simple global transformations and limit its ability to capture interactions and non-rigid deformations within a scene. We therefore employ a segmentation model (Qin et al., 2022) to identify foreground subjects and separate them. Using the resulting foreground masks, we characterize foreground area ratio, locally independent motion and subject-level texture richness. The retained clips are divided into two subsets (*rich* and *other*) according to these three criteria, and the *rich* subset is given higher priority in the following rebalancing procedure. By favoring clips with clearly identifiable subjects, this foreground-aware selection also provides structured semantic information and substantial motion characteristics at multiple levels.

Semantic Rebalancing. Reliable rebalancing requires semantic partitions that capture diverse content while remaining mutually exclusive within each balancing axis. However, classification schemes in previous datasets are often organized around a single criterion (Zhou et al., 2024; Zhao et al., 2025) and overlook the hierarchical structure of video semantics. The Holistic Video Understanding (HVU) taxonomy (Diba et al., 2020) uses a multi-level category system and has proven effective in large-scale downstream classification tasks. Building on HVU, we characterize video semantics along five complementary dimensions: scene, object, action, event and domain.

Directly balancing the joint label space, however, would produce highly fragmented and sparsely populated groups, complicating quota allocation across categories. We therefore select two axes (*scene* and *subject*) for hierarchical rebalancing to reduce fragmentation and preserve meaningful group sizes, while retaining the full taxonomy for semantic annotation. We further consolidate the categories within each dimension into a fixed label set and enforce single-label assignment, yielding a compact and well-defined semantic distribution for stable data rebalancing.

Based on this taxonomy, we use Qwen3-VL (Bai et al., 2025) to caption each video and generate multi-level classification tags. The annotated videos are then organized hierarchically, with scene and subject serving as the first- and second-level indices, respectively. At each level, square-root rebalanced allocation based on observed category sample counts reduces the dominance of frequent categories while retaining sufficient samples from rare categories. This procedure yields a more diverse semantic distribution in the open-domain VSR training dataset.

SCOPE-VSR. We first assemble publicly available datasets from multiple domains, including OpenVid-HD (Nan et al., 2025), UltraVideo (Xue et al., 2025), MiraData (Ju et al., 2024), HumanVid (Wang et al., 2024e) and HOIGen (Liu et al., 2025). The original collection contains ap-



Figure 3: Representative samples of SCOPE-VSR assembled from the individual category images.

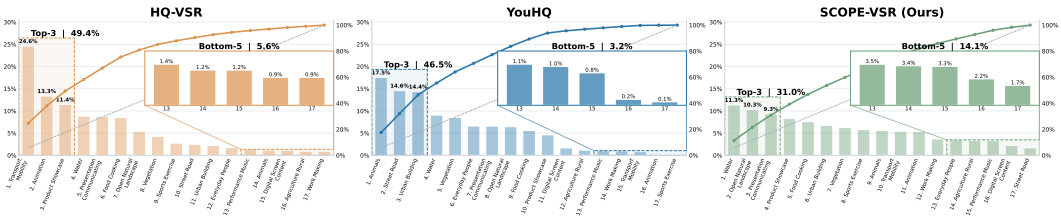

Figure 4: Ranked topic distributions of HQ-VSR, YouHQ40, and SCOPE-VSR, as predicted by InternVideo2 (Wang et al., 2024c). Bars denote per-topic shares; solid curves indicate cumulative shares, and dashed lines mark the uniform-distribution reference.

proximately 0.6 million videos. By applying the proposed SCOPE pipeline, we extract 27,351 clips to form the final SCOPE-VSR dataset. Details of pipeline settings are presented in Appendix C

Dataset Analysis. We analyze the category coverage and balance of SCOPE-VSR. As illustrated in Fig. 3, the dataset contains diverse subjects and scenes. We further employ InternVideo2 (Wang et al., 2024c) to generate video-level representations and classify each video into different topics. The representations are visualized in Fig. 1 in two dimensions, with points colored by predicted topic. The distribution spans multiple semantic regions and exhibits substantial intra-topic variation.

To compare with existing high-quality video datasets, we apply the same classification procedure to HQ-VSR (Chen et al., 2026c) and YouHQ40 (Zhou et al., 2024). We present the classification statistics and category concentrations in Fig. 4. In SCOPE-VSR, the top three topics account for 31.0%, which is remarkably less than the other datasets. Similarly, its five least frequent topics account for a larger share. Its cumulative category distribution also rises more gradually and remains closer to the uniform reference, demonstrating a more balanced category distribution.

3.2 GraCo-VSR: Granularity-Aligned Conditioning

Model Overview. We build our model on a pretrained Wan2.1 video diffusion backbone (Wan et al., 2025) and adapt it for one-step latent-space restoration. Although this framework is efficient, its prediction remains fully anchored to the upsampled LR latent. GraCo-VSR follows a granularity-aligned conditioning design, which introduces explicit reliability estimation and injects complementary conditions at their natural granularities. The model comprises Clip-Level Dynamics Conditioning (CLDC) and a Reliability-Aware Texture Encoder (RATE), which jointly mitigate the influence of degradation-corrupted signals and enable more reliable one-step restoration.

Clip-Level Dynamics Conditioning. Clip-Level Dynamics Conditioning (CLDC) aggregates flow information across adjacent frames into a global temporal descriptor for one-step restoration. Although RAFT is a powerful general-purpose estimator for online conditioning (Yang et al., 2024), it produces dense pixel-level motion fields. In contrast, our restoration model primarily requires the overall motion pattern of the clip. We therefore design a multi-scale flow-estimation network tailored to extract the essential motion cues at multiple spatial resolutions.

The network first uses a shared encoder to extract features at three progressively lower resolutions. It then predicts and refines the flow field from coarse to fine, jointly modeling global motion and local temporal variations. To mitigate GT-LR matching deviations, we introduce explicit supervision from a RAFT model (Wang et al., 2024d) that estimates optical flow from GT frame pairs.

To produce compact guidance from the dense flow map, we aggregate the distilled flow predictions into a single clip-level dynamics embedding. The embedding is injected into the DiT through two pathways. The motion token is concatenated with the empty-text token for cross-attention, while the motion modulation is incorporated into AdaLN through the timestep embedding.

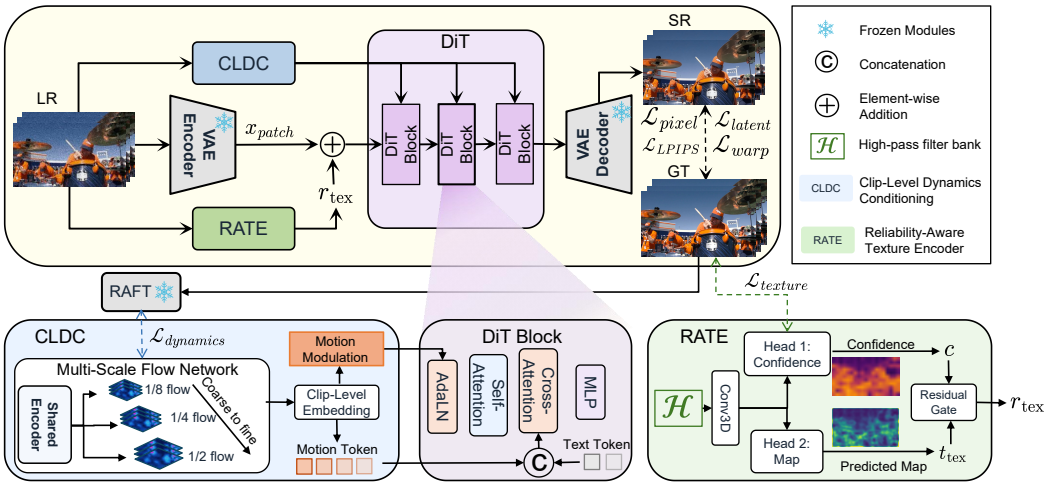

Figure 5: Overview of GraCo-VSR, comprising a DiT backbone, the Reliability-Aware Texture Encoder (RATE), and Clip-Level Dynamics Conditioning (CLDC). We first distill the Multi-Scale Flow Network (MSFN) from RAFT, then freeze MSFN and jointly train the remaining modules.

Reliability-Aware Texture Encoder. RATE generates a texture map and predicts confidence to suppress corrupted high-frequency responses. RATE applies fixed frequency filters to capture residual texture cues from LR videos. The filter bank combines spatial gradients, Laplacian responses, wavelet energy, and temporal differences in wavelet energy, providing complementary descriptions of local edges, fine textures, and their variation across frames. After channel-wise normalization, the frequency maps are processed by a 3D convolutional backbone. The resulting feature volume is adaptively pooled to the DiT patch grid to obtain an aligned texture feature.

Since LR textures are inherently unreliable in degraded regions, we adopt a dual-head reliability prediction design. One head estimates a frequency map, while the other predicts a pixel-wise confidence logit map. During training, the two heads are jointly constrained by the texture objective $\mathcal{L}_{\text{tex}}$ to align the predicted map and confidence with GT-derived targets. The predicted map is converted into texture tokens t_{tex}, while the confidence map provides the corresponding reliability weights C. A gate then produces the final residual results as follows

$$r_{\text{tex}} = \tanh(g_{\text{tex}})\,(C \odot t_{\text{tex}})\,, \tag{1}$$

where g_{tex} denotes the texture residual gate. The resulting r_{tex} is added to the initial patchified result x_{patch} before entering the DiT block to provide confidence-weighted local texture guidance. This weighting reduces the contribution of low-confidence texture cues before residual injection, while retaining stronger guidance from more reliable regions. The residual gate further regulates how strongly these weighted cues influence the DiT input.

Training Objective. Training proceeds in two phases. We first optimize the multi-scale flow network alone using the motion-distillation objective $\mathcal{L}_{\text{dynamics}}$ to learn robust motion representations, while keeping the remaining modules frozen. We then freeze the flow network and jointly optimize the DiT backbone, RATE, and the remaining CLDC modules with

$$\mathcal{L}_{\text{phase2}} = \mathcal{L}_{\text{main}} + \lambda_{\text{tex}}\mathcal{L}_{\text{tex}}. \tag{2}$$

where $\mathcal{L}_{\text{main}}$ supervises all the trainable components, while $\mathcal{L}_{\text{tex}}$ provides an auxiliary constraint for the texture-map and confidence predictions in RATE. The main objective $\mathcal{L}_{\text{main}}$ is formulated as

$$\mathcal{L}_{\text{main}} = \mathcal{L}_{\text{latent}} + \mathcal{L}_{\text{pixel}} + \mathcal{L}_{\text{LPIPS}} + \lambda_{\text{warp}}\mathcal{L}_{\text{warp}}. \tag{3}$$

We introduce $\mathcal{L}_{\text{latent}}$ to supervise the prediction in latent space, providing a stable optimization target for one-step restoration. $\mathcal{L}_{\text{pixel}}$ directly constrains the reconstruction in pixel space to preserve image content and structural fidelity, while $\mathcal{L}_{\text{LPIPS}}$ promotes perceptual similarity in deep feature space. To ensure consistency, we apply $\mathcal{L}_{\text{warp}}$ for suppressing motion-related flickering. Detailed formulations of $\mathcal{L}_{\text{dynamics}}$ and $\mathcal{L}_{\text{tex}}$ are provided in Appendix D.

Dataset	Metric	RealBasicVSR	RealViFormer	VEnhancer	STAR	FlashVSR	PS-SR	**GraCo-VSR (ours)**
UDM10	PSNR ↑	24.0000	24.6400	21.3134	23.9995	23.4229	23.3345	**25.2523**
	SSIM ↑	0.6968	0.6947	0.6803	0.6983	0.6951	0.7186	**0.7263**
	LPIPS ↓	0.3636	0.3681	0.4335	0.4129	0.2956	0.3075	**0.2558**
	DISTS ↓	0.2012	0.2039	0.2305	0.2213	**0.1564**	0.1600	0.1673
	CLIP-IQA ↑	0.4377	0.4157	0.2852	0.2410	0.5146	0.4120	**0.5197**
	MUSIQ ↑	58.3230	57.9000	37.2538	36.6531	**65.2129**	62.1139	62.0742
	MANIQA ↑	0.2829	0.2878	0.2120	0.2122	**0.3767**	0.3320	0.3153
	DOVER ↑	0.7343	0.7303	0.4686	0.4780	0.7657	0.7566	**0.7707**
YouHQ40	PSNR ↑	22.0202	**23.6600**	19.7806	22.5653	21.8712	22.1063	23.2283
	SSIM ↑	0.5839	**0.6334**	0.5911	0.6281	0.5818	0.6045	0.6266
	LPIPS ↓	0.4048	0.3662	0.4742	0.4655	0.2991	0.3570	**0.2670**
	DISTS ↓	0.1891	0.1940	0.2140	0.2361	**0.1354**	0.1614	0.1680
	CLIP-IQA ↑	0.4729	0.4462	0.3309	0.2762	**0.5071**	0.4283	0.4922
	MUSIQ ↑	62.3136	63.1800	42.0159	33.1669	**68.8096**	63.6817	63.5833
	MANIQA ↑	0.2973	0.2903	0.2442	0.2346	**0.3833**	0.3434	0.3120
	DOVER ↑	0.8485	0.8464	0.6933	0.5968	0.8728	0.8510	**0.8763**
RealVSR	PSNR ↑	22.0039	**22.1600**	15.5017	17.7302	19.7673	20.8142	19.8605
	SSIM ↑	**0.7163**	0.6955	0.3934	0.5420	0.5632	0.6649	0.6258
	LPIPS ↓	0.2044	0.2023	0.3838	0.3151	0.2125	0.2261	**0.1973**
	DISTS ↓	**0.1018**	0.1138	0.1734	0.1667	0.1083	0.1182	0.1053
	CLIP-IQA ↑	0.3538	0.3108	0.4249	0.4724	0.4083	0.2986	**0.6054**
	MUSIQ ↑	70.2996	66.6800	72.3089	70.3364	71.9875	66.5488	**74.8212**
	MANIQA ↑	0.3802	0.3341	0.4078	0.3888	**0.4236**	0.3542	0.4161
	DOVER ↑	0.7369	0.7191	0.7731	0.7033	0.7590	0.7566	**0.8203**
REDS30	PSNR ↑	21.3318	21.7705	13.3273	21.5222	20.3816	21.2262	**21.6932**
	SSIM ↑	0.5201	0.5397	0.4094	0.5477	0.5033	**0.5605**	0.5432
	LPIPS ↓	0.3964	0.3732	0.7798	0.5631	0.3478	0.4060	**0.3345**
	DISTS ↓	0.1950	0.1893	0.4003	0.2635	**0.1600**	0.1880	0.1969
	CLIP-IQA ↑	0.3233	**0.3836**	0.2427	0.1866	0.3502	0.2858	0.3552
	MUSIQ ↑	**61.0596**	60.8387	23.8483	31.3451	60.9516	55.5433	55.5800
	MANIQA ↑	0.2868	0.2823	0.1636	0.1639	**0.2910**	0.2360	0.2388
	DOVER ↑	0.5812	0.5896	0.2485	0.3595	**0.6874**	0.5864	0.6243
SCOPE-test	PSNR ↑	25.1819	25.6788	20.8188	23.5462	24.1059	23.9382	**25.7405**
	SSIM ↑	0.7202	0.7278	0.6174	0.6827	0.7050	0.7267	**0.7427**
	LPIPS ↓	0.2916	0.2756	0.3903	0.3948	0.2215	0.2645	**0.2056**
	DISTS ↓	0.1707	0.1675	0.1830	0.1969	**0.1193**	0.1445	0.1406
	CLIP-IQA ↑	0.4274	0.3829	0.4113	0.3009	0.4715	0.3724	**0.4767**
	MUSIQ ↑	64.3304	62.9521	61.9197	44.8322	**68.5916**	64.0642	65.8728
	MANIQA ↑	0.3140	0.2931	0.3030	0.2288	**0.3916**	0.3165	0.3253
	DOVER ↑	0.7520	0.7187	0.7658	0.5828	**0.8024**	0.7522	0.7717

Table 1: Quantitative results on synthetic (UDM10 (Yi et al., 2019), YouHQ40 (Zhou et al., 2024), SCOPE-test) and real-world datasets (RealVSR (Yang et al., 2021), REDS30 (Nah et al., 2019)). The best and second-best results are colored in **red** and blue, respectively.

4 Experiments

4.1 Settings

Datasets. For training, we utilize our SCOPE-VSR dataset, which contains 27,351 high-quality videos. The LR-GT pairs are synthesized following the RealBasicVSR degradation pipeline. For evaluation, we apply both synthetic and real-world datasets. The synthetic datasets include UDM10 (Yi et al., 2019), YouHQ40 (Zhou et al., 2024), and our SCOPE-test dataset. SCOPE-test is constructed using the same SCOPE pipeline and contains 40 high-quality videos. These videos are not retained in SCOPE-VSR due to category-level sampling during semantic rebalancing. Therefore, their exclusion is determined by the balancing procedure rather than visual quality. We carefully select representative videos in each category to ensure diversity in the test set, and apply the same degradation pipeline in training to generate the LR-GT pairs. For real-world evaluation, we use RealVSR (Yang et al., 2021) and REDS30 (Nah et al., 2019).

Evaluation Metrics. We adopt a range of evaluation metrics to assess model performance comprehensively. To measure fidelity, we apply PSNR and SSIM (Wang et al., 2004) to evaluate pixel-wise similarity with GT. We also use LPIPS (Zhang et al., 2018) and DISTS (Ding et al., 2020) to assess the overall visual similarity based on GT. For perceptual quality, we introduce non-reference IQA metrics including CLIPIQA (Wang et al., 2023), MUSIQ (Ke et al., 2021), and MANIQA (Yang et al., 2022). In addition, we evaluate overall video quality using DOVER (Wu et al., 2023).

Implementation Details. We adopt Wan2.1-1.3B (Wan et al., 2025) as our backbone, with the SR timestep fixed at 799. Training is conducted on 2 NVIDIA A6000 GPUs. The AdamW optimizer (Loshchilov & Hutter, 2017) is adopted in both phases with a batch size of 1. We first train the

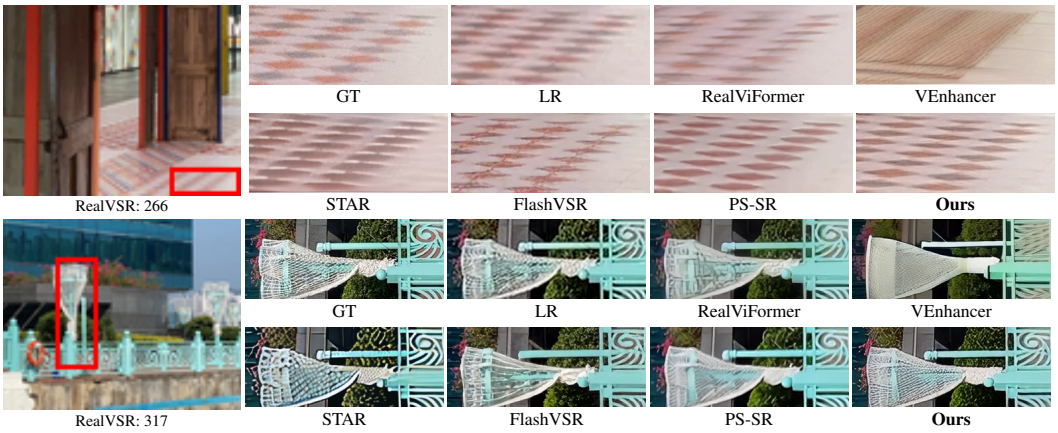

Figure 6: Visual comparisons on the real-world dataset (*i.e.*, RealVSR).

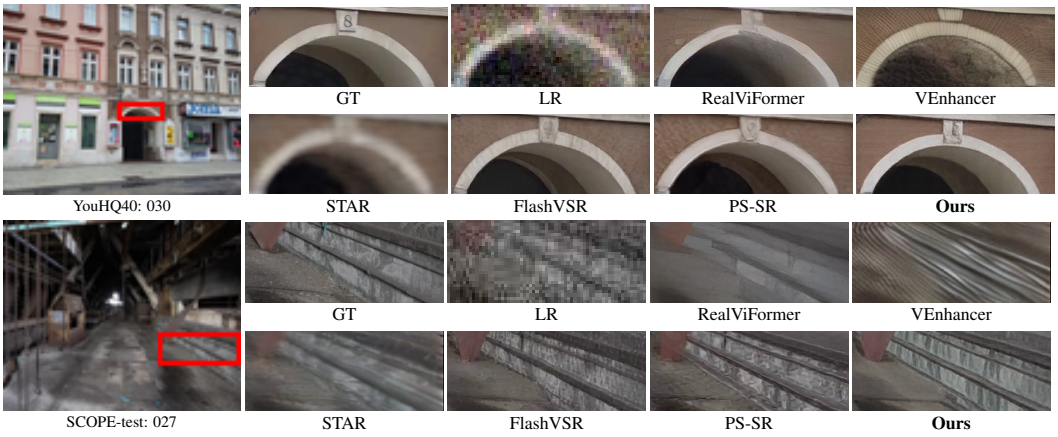

Figure 7: Visual comparisons on synthetic datasets (*i.e.*, YouHQ40 and SCOPE-test).

flow network for 14,500 steps at a learning rate of 1×10^{-4}. Then we freeze the network and jointly train the backbone and remaining modules in conditioning branches. In the joint training phase, the model is trained for 10,000 steps at a learning rate of 5×10^{-5} with $\lambda_{\text{warp}} = 0.1$ in Eq. (3).

4.2 Quantitative and Qualitative Comparisons

For comprehensive evaluation, we compare our model with a range of state-of-the-art VSR methods, including RealBasicVSR (Chan et al., 2022b), RealViFormer (Zhang & Yao, 2024), VEnhancer (He et al., 2024), STAR (Xie et al., 2025), FlashVSR (Zhuang et al., 2026) and PS-SR (Wu et al., 2026).

Quantitative Comparison. The quantitative results are summarized in Tab. 1. Our model consistently ranks among the top-performing methods across the evaluated benchmarks, demonstrating robust performance under both synthetic and real-world degradations. The improvements are particularly evident in perceptual quality, while reconstruction fidelity remains competitive. Although pixel-wise fidelity is not uniformly superior on every benchmark, the overall results indicate a balanced restoration capability across diverse scenarios.

Qualitative Comparison. We present the qualitative results on the real-world RealVSR dataset and the synthetic YouHQ40 and SCOPE-test datasets in Figs. 6 and 7. Across these examples, our method preserves clearer local structures and more coherent textures. In Fig. 6, our model preserves the edges of leaves more clearly, while several competing results either smooth these details or introduce distorted leaf patterns. Similarly, our model exhibits superior performance in the floor patterns in Fig. 7. These comparisons demonstrate the superior visual restoration capability of our method across diverse content. Additional qualitative results are provided in the Appendix E.

Pipeline	baseline	+quality	+foreground	+balance
PSNR ↑	25.2336	25.2225	25.2802	**25.2944**
SSIM ↑	0.7421	0.7474	0.7479	**0.7495**
LPIPS ↓	0.2728	0.2586	0.2594	**0.2493**
DISTS ↓	0.1803	0.1650	0.1643	**0.1570**
CLIP-IQA ↑	0.4217	0.4799	0.4815	**0.5018**
DOVER ↑	0.6815	0.7426	0.7406	**0.7653**

(a) Data curation pipeline.

Threshold	0	0.03	0.06
PSNR ↑	25.2323	**25.3705**	24.8917
SSIM ↑	0.7471	**0.7482**	0.7326
LPIPS ↓	0.2600	**0.2512**	0.2625
DISTS ↓	0.1624	**0.1554**	0.1664
CLIP-IQA ↑	0.5221	**0.5286**	0.5257
DOVER ↑	0.7494	0.7502	**0.7744**

(b) Texture frequency threshold.

Table 2: Ablation on the SCOPE data curation pipeline on UDM10. (a) Ablation on dataset curation stages. (b) Ablation on texture frequency threshold.

Component	PSNR ↑	SSIM ↑	LPIPS ↓	DISTS ↓	CLIP-IQA ↑	DOVER ↑
baseline	25.1312	0.7209	0.2690	0.1781	0.5040	0.7528
+CLDC	25.1172	**0.7298**	0.2623	0.1716	0.518	0.7541
+RATE	25.1099	0.7247	0.2636	0.1731	0.5152	0.7571
+CLDC+RATE	**25.2523**	0.7263	**0.2558**	**0.1673**	**0.5197**	**0.7707**

Table 3: Ablation on the GraCo-VSR architecture components on UDM10.

4.3 Ablation Study

We assess the effectiveness of the SCOPE dataset curation pipeline and the GraCo-VSR architecture components. All experiments are conducted on UDM10, with results summarized in Tabs. 2 and 3.

Dataset Curation Pipeline. We evaluate the effect of each stage in SCOPE using a 0.1M subset of OpenVid-HD. Results are reported in Tab. 2a. Starting from the raw subset (*baseline*), we first apply metadata filtering and perceptual quality screening (+*quality*), which substantially improve the perceptual metrics. Foreground-sensitive mining (+*foreground*) further improves structural fidelity and content-faithful reconstruction, which demonstrates the effectiveness of prioritizing data with multi-level motion and fine-grained visual details. Finally, we apply semantic rebalancing (+*balance*), which yields the best overall performance, confirming the benefit of balancing semantic coverage while preserving high-quality content.

Texture Frequency Threshold. We examine the texture frequency threshold in the second stage of SCOPE, including a zero-threshold setting that disables texture filtering (Table 2b). Our default threshold of 0.03 yields the strongest reference-based results, indicating more accurate recovery of overall structure. Although higher thresholds are favored by DOVER, they penalize on faithful reconstructions with poorer fidelity. Thus, we retain 0.03 as our default threshold.

Architecture Components. We further evaluate the contribution of CLDC and RATE in our GraCo-VSR model. We use the Wan2.1 backbone (Wan et al., 2025) adapted for VSR as our baseline. The training data and configurations are kept identical to the main experiment. As shown in Tab. 3, adding CLDC improves structural fidelity and perceptual similarity by providing robust global dynamics guidance for recovering temporally coherent structures. RATE independently also improves perceptual similarity and visual quality, showing its contribution to reducing distorted textures during restoration. Combining the two components produces the best overall performance across nearly all metrics.

5 Conclusion

In this paper, we propose SCOPE, a structured curation pipeline for open-domain and perceptually enriched videos. By introducing semantic rebalancing, SCOPE emphasizes both restoration-relevant training quality and broad semantic coverage. Following this pipeline, we construct SCOPE-VSR, a high-quality, diverse, and balanced dataset from large-scale, multiple-source video collections. Additionally, we propose GraCo-VSR, which introduces GT supervision and reliability modeling at complementary granularities to mitigate degradation-corrupted motion and texture guidance. Extensive experiments demonstrate the effectiveness of our designs in both dataset and model architecture.

AI USE STATEMENT

Generative AI tools were used solely to assist with language editing and writing refinement. All AI-assisted text was carefully reviewed and revised by the authors, who take full responsibility for the final content of this paper.

ETHICS STATEMENT

This work does not involve human-subject experiments or the collection of private or sensitive personal information. We are not aware of any specific ethical concerns arising from the methodology or experiments presented in this paper.

REPRODUCIBILITY STATEMENT

The dataset curation procedure and dataset details are provided in Section 3.1 and Appendix C, while the benchmark and evaluation settings are described in Section 4.1. Model training objectives and implementation details are reported in Sections 3.2 and 4.1 and Appendix D. The efficiency evaluation protocol and results are provided in Appendix B. We will release our code and relevant artifacts upon acceptance.

REFERENCES

Hongyu An, Xinfeng Zhang, Shijie Zhao, Li Zhang, and Ruiqin Xiong. Spatio-temporal distortion aware omnidirectional video super-resolution. In *AAAI*, 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

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 *CVPR*, 2021.

Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *CVPR*, 2022a.

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

Zheng Chen, Bowen Chai, Rongjun Gao, Mingtao Nie, Xi Li, Bingnan Duan, Jianping Fang, Xiaohong Liu, Linghe Kong, and Yulun Zhang. Dvface: Spatio-temporal dual-prior diffusion for video face restoration. *arXiv preprint arXiv:2604.14560*, 2026a.

Zheng Chen, Ruofan Yang, Jin Han, Dehua Song, Zichen Zou, Chunming He, Yong Guo, and Yulun Zhang. Diffst: Spatiotemporal-aware diffusion for real-world space-time video super-resolution. *arXiv preprint arXiv:2605.13182*, 2026b.

Zheng Chen, Zichen Zou, Kewei Zhang, Xiongfei Su, Xin Yuan, Yong Guo, and Yulun Zhang. Dove: Efficient one-step diffusion model for real-world video super-resolution. In *NeurIPS*, 2026c.

Ali Diba, Mohsen Fayyaz, Vivek Sharma, Manohar Paluri, Jürgen Gall, Rainer Stiefelhagen, and Luc Van Gool. Large scale holistic video understanding. In *ECCV*, 2020.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *TPAMI*, 2020.

Marie Farge et al. Wavelet transforms and their applications to turbulence. *Annual review of fluid mechanics*, 1992.

Dario Fuoli, Shuhang Gu, and Radu Timofte. Efficient video super-resolution through recurrent latent space propagation. In *ICCVW*, 2019.

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

Younghyun Jo, Seoung Wug Oh, Jaeyeon Kang, and Seon Joo Kim. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In *CVPR*, 2018.

Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. Miradata: A large-scale video dataset with long durations and structured captions. In *NeurIPS*, 2024.

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

Wenbo Li, Xin Tao, Taian Guo, Lu Qi, Jiangbo Lu, and Jiaya Jia. Mucan: Multi-correspondence aggregation network for video super-resolution. In *ECCV*, 2020.

Xiaohui Li, Yihao Liu, Shuo Cao, Ziyan Chen, Shaobin Zhuang, Xiangyu Chen, Yinan He, Yi Wang, and Yu Qiao. Diffvsr: Revealing an effective recipe for taming robust video super-resolution against complex degradations. In *ICCV*, 2025.

Kun Liu, Qi Liu, Xinchen Liu, Jie Li, Yongdong Zhang, Jiebo Luo, Xiaodong He, and Wu Liu. Hoigen-1m: A large-scale dataset for human-object interaction video generation. In *CVPR*, 2025.

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

Alice Lucas, Santiago Lopez-Tapia, Rafael Molina, and Aggelos K Katsaggelos. Generative adversarial networks and perceptual losses for video super-resolution. *TIP*, 2019.

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 *CVPRW*, 2019.

Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. In *ICLR*, 2025.

Xuebin Qin, Hang Dai, Xiaobin Hu, Deng-Ping Fan, Ling Shao, and Luc Van Gool. Highly accurate dichotomous image segmentation. In *ECCV*, 2022.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022.

Tomás Soucek and Jakub Lokoc. Transnet v2: An effective deep network architecture for fast shot transition detection. In *ACM MM*, 2024.

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.

Boyang Wang, Fengyu Yang, Xihang Yu, Chao Zhang, and Hanbin Zhao. Apisr: anime production inspired real-world anime super-resolution. In *CVPR*, 2024a.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI*, 2023.

Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *IJCV*, 2024b.

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

Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *CVPRW*, 2019.

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

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *ECCV*, 2024c.

Yihan Wang, Lahav Lipson, and Jia Deng. Sea-raft: Simple, efficient, accurate raft for optical flow. In *ECCV*, 2024d.

Zhenzhi Wang, Yixuan Li, Yanhong Zeng, Youqing Fang, Yuwei Guo, Wenran Liu, Jing Tan, Kai Chen, Tianfan Xue, Bo Dai, et al. Humanvid: Demystifying training data for camera-controllable human image animation. In *NeurIPS*, 2024e.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 2004.

Aiqiu Wu, Zhaofan Qiu, Ting Yao, and Tao Mei. Ps-sr: Pseudo-single-step video super-resolution via speculative diffusion. In *CVPR*, 2026.

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 *ICCV*, 2023.

Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. In *NeurIPS*, 2024a.

Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *CVPR*, 2024b.

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

Yiran Xu, Taesung Park, Richard Zhang, Yang Zhou, Eli Shechtman, Feng Liu, Jia-Bin Huang, and Difan Liu. Videogigagan: Towards detail-rich video super-resolution. In *CVPR*, 2025.

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 *CVPRW*, 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 *ICCV*, 2021.

Xi Yang, Chenhang He, Jianqi Ma, and Lei Zhang. Motion-guided latent diffusion for temporally consistent real-world video super-resolution. In *ECCV*, 2024.

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 *ICCV*, 2019.

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

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

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 *CVPR*, 2024.

Junhao Zhuang, Shi Guo, Xin Cai, Xiaohui Li, Yihao Liu, Chun Yuan, and Tianfan Xue. Flashvsr: Towards real-time diffusion-based streaming video super resolution. In *CVPR*, 2026.

Yang Zou, Xingyue Zhu, Kaiqi Han, Jun Ma, Xingyuan Li, Zhiying Jiang, and Jinyuan Liu. Hatir: Heat-aware diffusion for turbulent infrared video super-resolution. In *AAAI*, 2026.

Training data	YouHQ40						REDS30					
	PSNR ↑	LPIPS ↓	CLIP-IQA ↑	MUSIQ ↑	MANIQA ↑	DOVER ↑	PSNR ↑	LPIPS ↓	CLIP-IQA ↑	MUSIQ ↑	MANIQA ↑	DOVER ↑
HQ-VSR	**23.9396**	0.3702	0.3433	47.3635	0.2528	0.7527	**22.4721**	0.5086	0.1994	34.2060	0.1721	0.3735
SCOPE-VSR	23.5259	**0.3585**	**0.4073**	**53.1609**	**0.2810**	**0.7929**	22.3664	**0.4800**	**0.2208**	**38.0918**	**0.2023**	**0.4362**

Table 4: Quantitative comparisons on synthetic datasets YouHQ40 and real-world dataset REDS30. DOVE is trained with the LoRA fine-tuning configuration. The best result is in **bold**.

Metric	FlashVSR	PS-SR	STAR	VEnhancer	**GraCo-VSR**
Params (B)	1.780	2.210	2.493	2.497	**1.554**
MACs (T)	144.666	**37.119**	3856.056	3955.055	349.060
Runtime (s)	**6.58**	37.99	187.57	342.87	13.39

Table 5: Complexity comparison.

A Ablation on Datasets

We further evaluate the performance of our dataset on DOVE (Chen et al., 2026c). Due to hardware limitation, we adopt the two-stage LoRA scheme to fine-tune DOVE separately on SCOPE-VSR and HQ-VSR. Following the original configuration, we set both the LoRA rank and scaling factor to 64. Stage 1 and stage 2 are trained for 10,000 and 500 steps, respectively. As shown in Tab. 4, SCOPE-VSR achieves better overall perceptual quality than HQ-VSR while remaining competitive in fidelity. These results demonstrate the generalization ability of our datasets on different methods, supporting the effectiveness of our multi-stage dataset pipeline.

B Complexity Comparison

We compare the computational cost of different VSR methods in terms of parameter count, computation complexity (MACs), and running time. Running time is measured on a single NVIDIA A6000 GPU. We use video of 32 frames with a resolution of 1280×720. As shown in Tab. 5, GraCo-VSR has the smallest parameter count and a comparably low running time. Overall, these results indicate that our model maintains a balanced computational profile, offering practical inference efficiency without relying on excessive model capacity or computational overhead.

C Pipeline Details

In this section, we provide the implementation details our data curation pipeline, SCOPE. Our pipeline are mainly composed of 4 steps.

C.1 Meta Filtering

To ensure sufficient spatial information for subsequent processing, we first discard videos with low resolution and number of frames. The resolution and frame count thresholds are set as 720p and 100, respectively. Abrupt changes within one clip can lead to inconsistency in spatial details and temporal information. Therefore, we detect and split shot transitions using TransNetV2 (Soucek & Lokoc, 2024) with a decision threshold of 0.5. 2 frames on either side of the detected boundary are removed to exclude unstable transition content. Among the continuous segments produced from a source video, we retain the segment with the largest number of frames.

C.2 Perceptual Quality Screening

We further assess candidate videos from three complementary perspectives: aesthetic quality, perceptual quality, and high-frequency texture richness. We use the LAION aesthetic predictor (Schuhmann et al., 2022) to estimate the aesthetic score of each video. For efficient perceptual-quality assessment, we uniformly sample at most 16 frames from each video, and evaluate them with CLIP-IQA (Wang et al., 2023). To quantify spatial detail richness, we convert the sampled frame to grayscale, and normalize its intensity to $[0, 1]$. We then apply a single-level 2D biorthogonal wavelet

Table 6: Semantic fields used for video annotation and rebalancing.

Field	Type and vocabulary	Role
primary_scene	indoor, outdoor, mixed, or unknown	Supplemental
scene_type	38 fine-grained scene classes, including unknown	**First-level index**
primary_subject	person, vehicle, animal, landscape, object, food, text/screen, or other	**Second-level index**
event_category	14 event classes	Supplemental
domain_category	6 content-domain classes	Supplemental
object_categories	45 object classes; at most 8 labels	Supplemental
action_categories	33 action classes; at most 5 labels	Supplemental
descriptive_tags	3-8 free-form descriptive tags	Display and retrieval

Table 7: Representative semantic classification results.

Example	Brief caption	Scene / subject / event	Objects / actions
Presentation	A man presents on a whiteboard about grass-fed butter benefits.	room / person / education or presentation	person, text / talking, gesturing
Food	A bowl of seafood pasta with mussels, clams, and shrimp.	restaurant / food / food	food, table / displaying, no obvious action
Daily life	Two men in patterned shirts stand indoors with colorful decorations.	room / person / daily life	person, clothing, plant / standing, talking, gesturing

transform with the bior1.3 wavelet (Farge et al., 1992). Let HL, LH, and HH denote the three high-frequency subbands. The frame-level texture score s_{tex} is defined as

$$s_{\text{tex}} = \frac{\left\| \sqrt{HL^2 + LH^2 + HH^2} \right\|}{\sqrt{N}}, \tag{4}$$

where N is the number of coefficients in one high-frequency subband. Thresholds of aesthetic, CLIP-IQA, and texture-frequency scores are set as 5.0, 0.4, and 0.03, respectively.

C.3 Foreground-Sensitive Mining

To begin with, we use SEA-RAFT (Wang et al., 2024d) to estimate the optical flow map. To make the motion magnitude comparable under different input, we normalize the flow magnitude by the video diagonal length, and calculate the average global flow magnitude, denoted as f_{global}. Videos with $f_{global} < 0.55$ are treated as nearly static and removed.

We then employ DIS (Qin et al., 2022) to extract foreground regions with a visually salient subject. To reduce computation, DIS is employed once every four frames, and foreground masks for the remaining frames are propagated with the SEA-RAFT flow. We use a mask probability threshold of 0.50 and a temporal decay factor of 0.90 during propagation to suppress temporal flickering. The raw masks are further refined with geometric and morphological constraints so that they better align with the semantic subject.

With the DIS-refined mask, we measure the foreground flow magnitude and texture score following the previous stages. We additionally compute the mean foreground-area ratio over all valid frames. A video is labeled *rich* only when it meets with the following thresholds: mean foreground-area ratio at least 0.20, foreground-flow score at least 1.50, and foreground-texture score at least 0.06. Videos that fail the threshold are labeled *others*. The *rich* videos are of higher priority in semantic rebalancing procedure.

C.4 Semantic Rebalancing

Semantic rebalancing starts from the high-quality subset produced by foreground-sensitive mining rather than selecting again from all candidates. The initial subset contains all *rich* videos and is then filled to the target size N with *other* videos in descending order of global flow magnitude. The remaining *other* videos are retained in the candidate pool for subsequent exchanges.

To obtain the semantic labels used for rebalancing, we employ Qwen3-VL-2B-Instruct (Bai et al., 2025) to jointly generate a caption and classify every video in the initial subset and candidate pool. The prompt provides a closed candidate set for each field and requires the model to rely only on visible evidence. As summarized in Table 6, `scene_type` and `primary_subject` serve as the two

hierarchical balancing indices, whereas the remaining fields provide complementary annotations. The following prompt is adopted:

Prompt Template for Semantic Annotation

Analyze this video for balancing a real-world video super-resolution dataset.
Return exactly one compact, single-line JSON object and stop immediately after its final }. No Markdown, explanation, repeated keys, null values, or copied candidate lists. Keep the complete response under 180 tokens. Describe only visible evidence. Do not infer identity, profession, emotion, intent, or a hidden story.
brief_caption: concise English caption, at most 20 words, including the main subject, scene, and visible action when clear.
primary_scene: ONE JSON STRING from [PRIMARY_SCENES].
scene_type: ONE JSON STRING from [SCENE_TYPES].
primary_subject: ONE JSON STRING from [PRIMARY_SUBJECTS]; NEVER an array.
object_categories: ONE JSON ARRAY containing zero to eight labels from [OBJECT_CATEGORIES]. Never inventory every visible object.
action_categories: ONE JSON ARRAY containing zero to five labels from [ACTION_CATEGORIES].
event_category: ONE JSON STRING from [EVENT_CATEGORIES].
domain_category: ONE JSON STRING from [DOMAIN_CATEGORIES].
descriptive_tags: three to eight short lowercase English tags.
If evidence is insufficient, use unknown, other, no_obvious_action, or synthetic_or_uncertain as appropriate. Output every key exactly once. Use this exact key order: brief_caption, primary_scene, scene_type, primary_subject, event_category, domain_category, object_categories, action_categories, descriptive_tags.

Table 7 presents representative normalized outputs spanning presentation, food, and daily-life content. Each example highlights the hierarchy formed by scene_type and primary_subject, while the event, object, and action labels supply additional semantic context. All displayed labels have been normalized to the closed vocabulary so that quota statistics are computed over a stable and compact label space.

Based on the normalized labels, we first allocate the target count across scene_type classes and then allocate each scene quota across its primary_subject classes. For a category c with n_c available samples and a target subset size N, its continuous quota is

$$q_c = N \frac{\sqrt{n_c}}{\sum_j \sqrt{n_j}}. \tag{5}$$

Square-root weighting places the allocation between uniform sampling and the natural data distribution, reducing the dominance of head classes without unrealistically amplifying extremely rare classes. Each continuous quota is floored and capped by category availability, while every nonempty category receives at least one sample when the budget permits. Applying the same procedure within each scene yields the final target $q_{c,s}$ for every (scene_type, primary_subject) cell.

We define a semantic cell whose current count is below its target as a *deficit* cell and one whose count is above its target as a *surplus* cell. Each exchange introduces one reserve-pool sample from a deficit cell and removes one selected sample from a surplus cell, thereby keeping the subset size exactly equal to N.

To preserve the high-value samples found by foreground-sensitive mining, we introduce the following two constraints when exchanging: (i) within a *surplus* cell, non-rich samples are removed before rich samples. (ii) let R_0 and R denote the numbers of rich videos in the initial and current subsets, respectively. Every exchange must satisfy $R \geq \lceil 0.60 R_0 \rceil$. If removing a rich sample would violate this bound, that pairing is skipped in favor of another feasible exchange. The procedure terminates when all achievable quotas are met or when the reserve pool contains no deficit-surplus pairing that satisfies the rich-retention constraint.

D Architecture and Training Details

This section details the GraCo-VSR architecture and its training procedure. We first describe the multi-scale flow network in CLDC and its standalone distillation, followed by the objectives used in the main training stage.

D.1 Motion Distillation

The multi-scale flow network is distilled independently before the main training phase. A frozen SEA-RAFT (Wang et al., 2024d) model serves as the teacher and estimates optical flow from GT frame pairs, whereas the multi-scale flow network predicts the corresponding flow from LR inputs. The network first downsamples each input to $1/8$, $1/4$, and $1/2$ of its original resolution and processes these inputs with a shared encoder built primarily from 3D convolutions. At each scale, a scale-specific fusion module aggregates inter-frame information, and a flow head maps the resulting motion features to a flow estimate. Estimation begins at the $1/8$ scale to capture the principal motion direction and displacement range. The resulting coarse flow is then propagated to the $1/4$ and $1/2$ scales, where it is refined through residual corrections conditioned on higher-resolution features. The subsequent motion-conditioning module primarily uses the $1/4$-scale flow, which balances global motion representation and spatial detail.

At each output scale, we resize the teacher flow to the spatial resolution of the prediction before evaluating the dynamic loss as follows:

$$\mathcal{L}_{\text{dynamics}} = \mathcal{L}_{\text{flow}} + 0.01\mathcal{L}_{\text{dir}} + 0.2\mathcal{L}_{\text{grad}} + 0.1\mathcal{L}_{\text{global}}. \tag{6}$$

Both $\mathcal{L}_{\text{flow}}$ and $\mathcal{L}_{\text{grad}}$ use the Charbonnier penalty for robust regression. Specifically, $\mathcal{L}_{\text{flow}}$ matches the orthogonal components of the predicted and teacher flows, assigning larger weights to regions with greater teacher-flow magnitude. $\mathcal{L}_{\text{grad}}$ matches the first-order differences of flow, thereby preserving motion boundaries and local deformations. The direction term $\mathcal{L}_{\text{dir}}$ normalizes the flows into unit vectors and minimizes their cosine distance. We exclude flows whose teacher motion magnitude is below 10^{-3}, preventing small component-wise errors from obscuring an incorrect motion direction. In addition, $\mathcal{L}_{\text{global}}$ is designed to align the output on statistics level with a Smooth L1 loss. Together, these four objectives transfer complementary cues ranging from local displacement values to clip-level dynamics, yielding a stable motion representation for subsequent conditioning.

D.2 Main Training Phase and Objectives

During the main training phase, the multi-scale flow network remains frozen, while the DiT backbone, RATE, and remaining modules in CLDC are optimized end to end. The trainable components are jointly supervised by the main restoration objective in Eq. 3.

The additional supervision applied to RATE is detailed as follows. The map and confidence prediction branches within RATE are further constrained by a texture loss $\mathcal{L}_{\text{tex}}$. Given an input video, a fixed frequency bank extracts four high-frequency energy-map channels: Sobel gradient magnitude, Laplacian response, bior1.3 wavelet energy, and the temporal difference between adjacent wavelet-energy maps. We apply the same bank to the GT video and pool the resulting map $\widehat{\mathbf{M}}$ onto the RATE output grid. The predicted map $\mathbf{M}$ is supervised using a channel-weighted Smooth L1 loss:

$$\mathcal{L}_{\text{map}} = \frac{1}{N\sum_{c=1}^{4} w_c} \sum_{c=1}^{4} w_c \sum_{i=1}^{N} \rho\Big(M_{c,i} - \widehat{M}_{c,i}\Big), \qquad \mathbf{w} = (1, 1, 1, 0.25), \tag{7}$$

where ρ denotes the Smooth L1 penalty, w_c is the weight of four channels, and N is the number of spatiotemporal positions.

For confidence supervision, we first average the four channels of the GT and LR maps to obtain E^{GT} and E^{LR}, respectively. Using the sample-wise GT energy scale $s = \max(\text{mean}(E^{\text{GT}}), \epsilon)$, we define the texture-presence and LR-visibility terms as

$$P = \text{clip}\left(\frac{E^{\text{GT}}}{s}, 0, 1\right), \qquad V = \exp\left(-\left|\log\left(1 + \frac{E^{\text{LR}}}{s}\right) - \log\left(1 + \frac{E^{\text{GT}}}{s}\right)\right|\right). \tag{8}$$

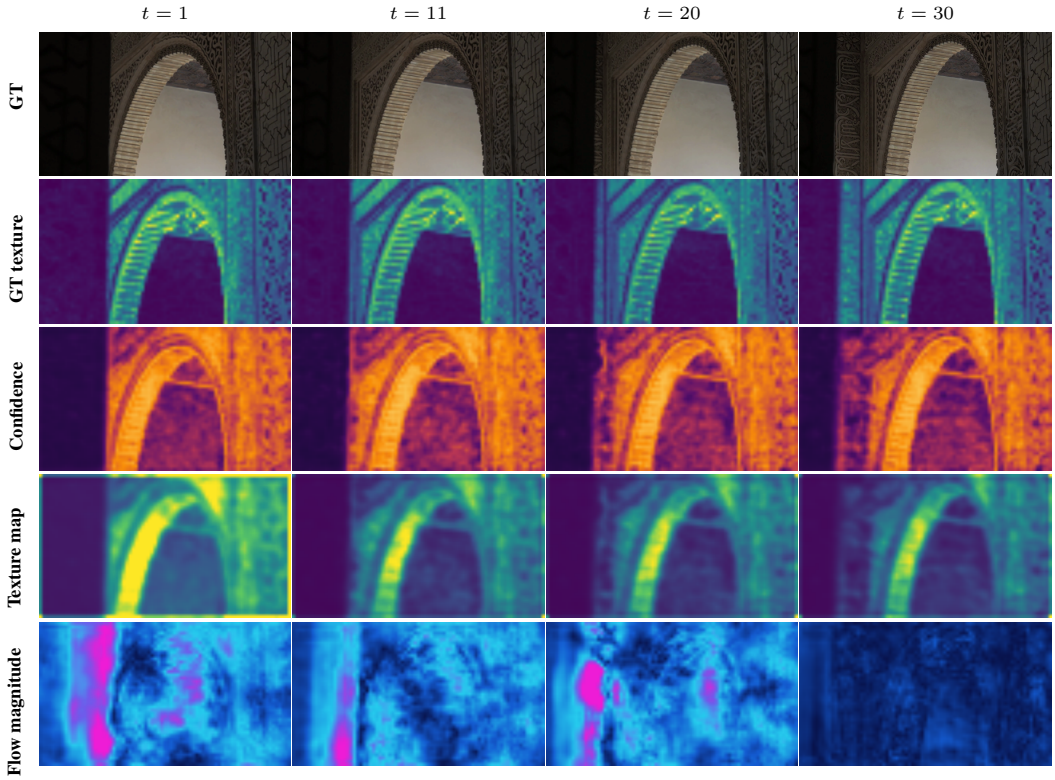

Figure 8: Visualization of motion and texture guidance for UDM10: 001 (Yi et al., 2019).

Their product, $C^* = PV$, serves as the confidence target. We then optimize the confidence logits $\widehat{C}$ using

$$\mathcal{L}_{\text{conf}} = \text{BCEWithLogits}\left(\widehat{C}, C^*\right). \tag{9}$$

The weights of $\mathcal{L}_{\text{map}}$ and $\mathcal{L}_{\text{conf}}$ are set to 1.0 and 0.05, respectively. The resulting texture loss is

$$\mathcal{L}_{\text{tex}} = \mathcal{L}_{\text{map}} + 0.05\mathcal{L}_{\text{conf}} \tag{10}$$

To complement the preceding descriptions of condition extraction and injection, we visualize the intermediate motion and texture guidance for representative videos from UDM10 (Yi et al., 2019) and RealVSR (Yang et al., 2021). We present four keyframes from each video in temporal order and show the texture predicted map, texture confidence maps, and flow magnitudes, respectively. These visualizations provide a direct view of the guidance signals supplied to the restoration model.

E Additional Qualitative Comparisons

We provide additional qualitative comparisons on UDM10 (Yi et al., 2019), YouHQ40 (Zhou et al., 2024), RealVSR (Yang et al., 2021), and SCOPE-test. Our method consistently produces clear and faithful static reconstructions across diverse synthetic and real-world scenes. As shown in Fig. 12, our result preserves more recognizable character contours and cleaner sign boundaries without introducing severe structural deformation. Similarly, in Fig. 10, our method retains the dense debris and splash-like details while avoiding the excessive smoothing or artificial repetitive patterns observed in other results. These comparisons demonstrate that our method restores natural texture and details with greater clarity and structural faithfulness across diverse visual content.

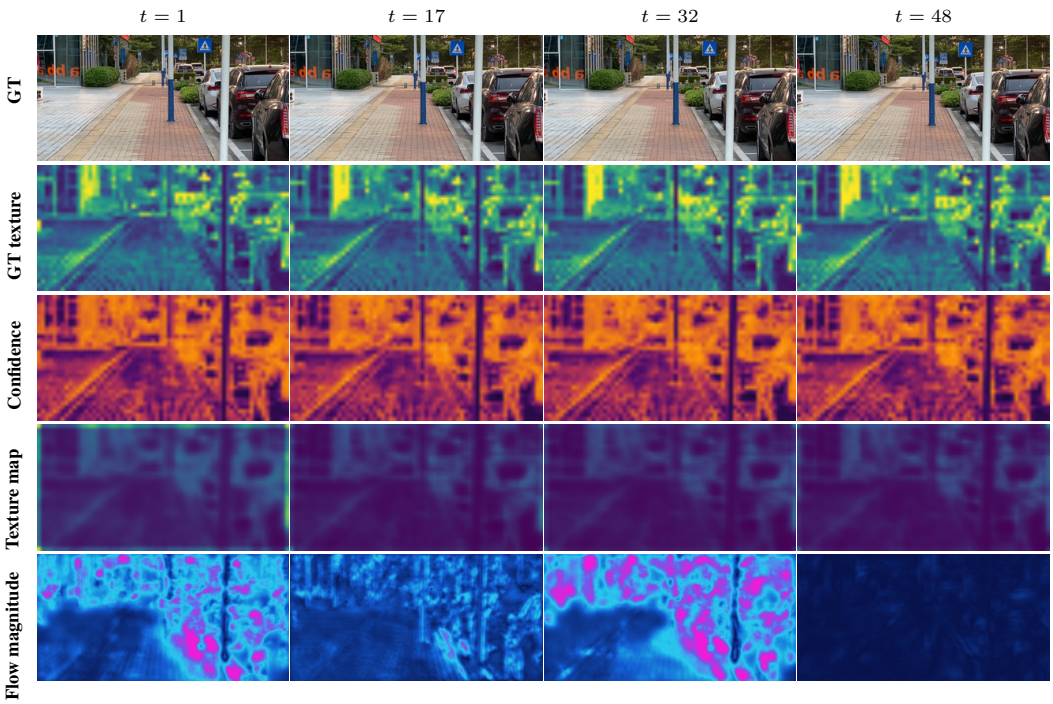

Figure 9: Visualization of motion and texture guidance for RealVSR: 090 (Yang et al., 2021).

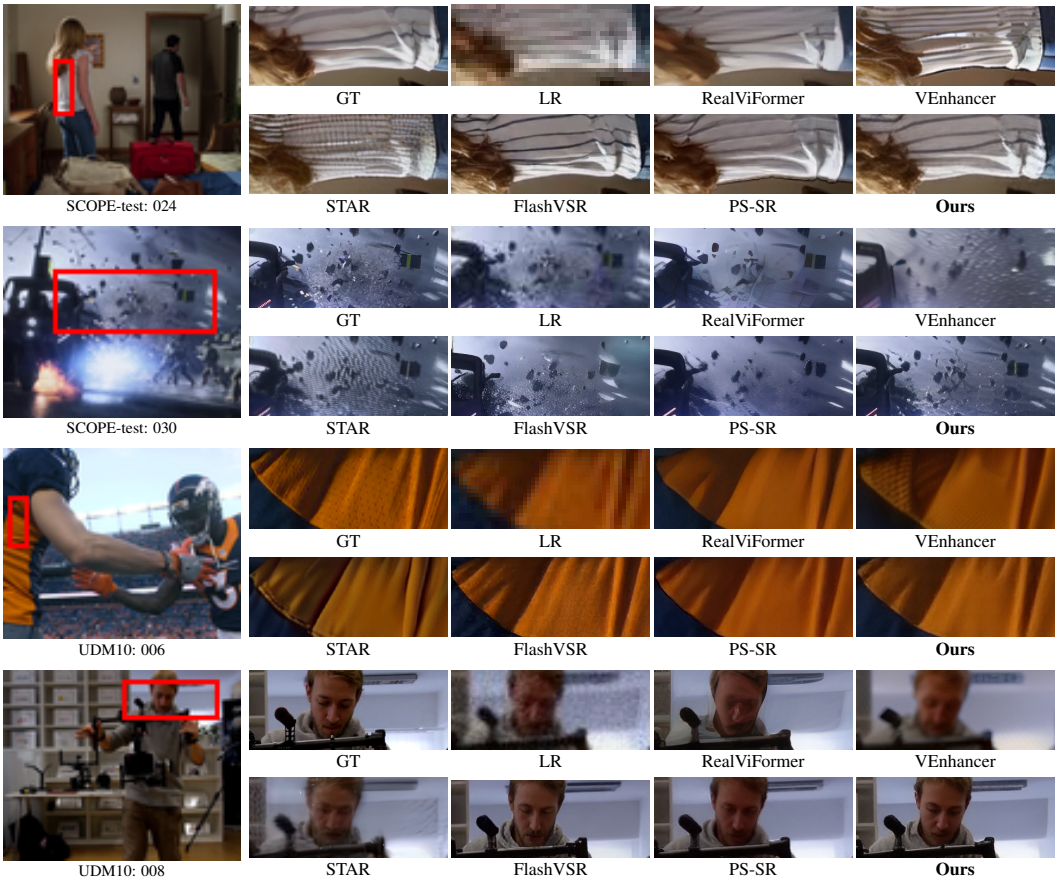

Figure 10: Additional visual comparisons on UDM10 (Yi et al., 2019) and SCOPE-test.



Figure 11: Additional visual comparisons on YouHQ40 (Zhou et al., 2024) and REDS30 (Nah et al., 2019).

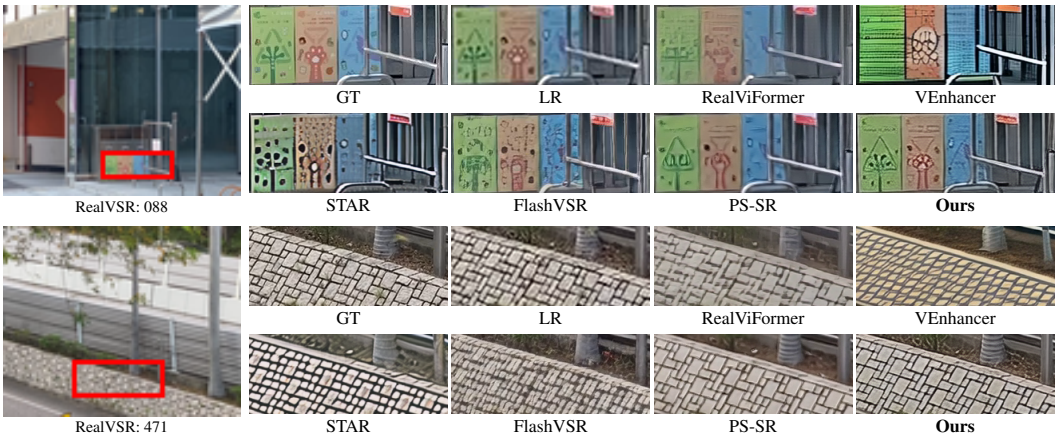

Figure 12: Additional visual comparisons on RealVSR (Yang et al., 2021).