TRaM-VSR: Importance-Aware Token Routing and Merging for One-Step Diffusion Video Super-Resolution

Sicheng Gao[1,2], Zhuyun Zhou[2], Yixuan Liu[1], Tong Shen[1], Zongwei Wu[2⋆], and Radu Timofte[2]

[1] Advanced Micro Devices Inc., China
[2] Computer Vision Lab, CAIDAS & IFI, University of Würzburg, Germany

Abstract. Video super-resolution (VSR) using large-scale Diffusion Transformer (DiT) priors achieves exceptional perceptual quality but is often impractical due to the quadratic computational cost of processing dense spatio-temporal token sequences. Existing efficiency-oriented methods risk irreversible detail loss and temporal flickering, a vulnerability especially pronounced in one-step diffusion models. To address this, we propose TRaM-VSR, a Token Routing and Merging framework for adaptive token allocation, leveraging both context-aware video priors and network-level priors. First, token importance is estimated by fusing motion-sensitive temporal cues with semantic text similarity, isolating dynamic objects and structural boundaries. Next, this importance is further calibrated and adjusted by an offline planner to guide routing across optimally grouped network blocks. Technically, within each routed group, structurally critical tokens are processed in a high-fidelity local stream, while less informative tokens are aggregated into a compact global stream, both modulated by network depth and aligned with the multigranular nature of diffusion models. Extensive experiments show that TRaM-VSR accelerates inference significantly while preserving state-of-the-art reconstruction quality and robust temporal consistency. The code is available at https://github.com/Ree1s/TRaM-VSR.

1 Introduction

Video super-resolution (VSR) aims to recover high-resolution (HR) videos from low-quality (LQ) inputs while preserving temporal consistency across frames. Recently, diffusion-based approaches have achieved remarkable progress, largely driven by Diffusion Transformers (DiTs) and large-scale text-to-video priors [1, 10, 16, 22, 24, 28]. Leveraging powerful generative priors, recent VSR works [13, 32, 44] achieve impressive performance in super-resolution. However, these models incur substantial computational cost, as dense spatio-temporal tokens must be processed through deep transformer stacks with quadratic attention complexity.

To improve efficiency, a line of research explores lightweight generation pipelines, among which one-step diffusion has emerged as a promising paradigm [6, 18, 25].

⋆ Corresponding author.

Fig. 1: Motivation of TRaM-VSR. ToMeSD [3] (top right) yields scattered heatmaps that miss dynamic objects and prompt semantics. Instead, TRaM-VSR exploits the coarse-to-fine nature of DiTs: our offline planner locates safe compressible depths (Intervals 1 and 2), where our semantic-temporal scoring (bottom right) precisely highlights critical tokens. Notably, this scoring forms complementary focus regions across intervals, synergistically preserving both global motion and fine details. More discussion and analysis can be found in Section 6.

By collapsing the iterative denoising process into a single forward pass, these approaches substantially reduce inference latency. However, this efficiency often comes at the cost of temporal stability [6, 25]. Without the progressive refinement provided by multiple diffusion steps, generation errors cannot be gradually corrected and may accumulate over time, leading to artifacts such as shadow flickering and temporally inconsistent textures across frames. This issue becomes particularly pronounced when spatio-temporal dependencies are not sufficiently modeled within the one-step diffusion process.

However, effectively modeling spatio-temporal cues is challenging, as video frames can be affected by complex motion patterns arising from both camera egomotion and independently moving objects. When the learning model has limited capacity, as in the case of one-step diffusion, uniformly processing all tokens becomes difficult and can introduce significant errors. To improve efficiency, some studies [2, 9, 11, 19] have explored token merging or token dropping strategies. However, these accelerations are typically applied in a uniform manner [3, 4], as shown in Figure 1. Consequently, the modeling capacity is not allocated where it is most needed, leading to degraded temporal coherence. This observation motivates us to introduce a more context-aware perspective.

In this paper, we propose **TRaM-VSR**, a *Token Routing and Merging* framework for efficient video super-resolution. The core idea is to perform adaptive

token allocation by jointly leveraging context-aware *video priors* and *network-level priors*. For the video context, we learn a token importance map that captures both spatial and temporal relationships across frames, allowing the model to rank tokens according to their relevance for high-fidelity reconstruction. This selection is further guided by input text prompts, ensuring semantic alignment. Consequently, tokens corresponding to dynamic objects, structural boundaries, and temporally informative regions should receive higher importance, while less informative regions can be processed with reduced computational effort.

We then correlate the video-derived token importance with depth-aware contextual cues from the backbone network. In the diffusion-based architecture [22], each block of the DiT backbone functions as an independent denoising operator, exhibiting distinct behaviors at different depths. Inspired by complex systems [26], we hypothesize that organizing these operators at multiple granularities can induce emergent modeling capabilities that surpass the capabilities of individual blocks. Accordingly, we group DiT blocks into sets of varying granularities, with each group producing an independent dropping score that serves as a calibration prior, dependent solely on the compositional structure and depth of the network. By combining the video-derived token importance with the corresponding network-level dropping prior, our framework achieves holistic adaptive token routing and dropping throughout the transformer hierarchy. Extensive evaluation on VSR benchmarks shows that our method can achieve high-fidelity output at a very efficient scale. Our contributions are summarized as follows:

- We introduce a novel semantic–temporal importance scoring mechanism. Compared to generic compression metrics, our method effectively isolates motion-critical and prompt-relevant tokens, avoiding the temporal flickering induced by motion-blind pruning.
- We explicitly leverage the coarse-to-fine depth dynamics of DiTs to design a deterministic offline planner, precisely localizing optimal routing intervals to prevent catastrophic detail loss in one-step VSR.
- We design a highly efficient two-stream token pathway featuring explicit identity restoration. Extensive experiments demonstrate that TRaM-VSR significantly accelerates inference while preserving state-of-the-art quality and temporal consistency.

2 Related Work

Diffusion Priors for Video Restoration: Diffusion models, particularly those leveraging large-scale text-to-video (T2V) priors, have recently dominated video super-resolution (VSR) due to their robust perceptual quality and highly flexible conditioning [22, 35]. Systems such as Upscale-A-Video [44], SeedVR [32], and VEnhancer [13] demonstrate how adapting these rich spatio-temporal priors significantly improves real-world restoration by handling complex motion and severe degradations [37–39]. To address the inherent latency of iterative sampling, recent advancements have introduced one-step diffusion VSR models, notably DOVE [6], One-Step VSR [25], OS-DiffVSR [18], and UltraVSR [20]. Despite

drastically reducing the number of sampling steps, these architectures still inherit the massive spatial and temporal token footprints of their foundational DiT backbones. Consequently, shifting the efficiency bottleneck from temporal iteration to spatial token computation remains an open challenge.

Efficiency in Diffusion Models: Improving the inference efficiency of diffusion models has inspired numerous architectural and system-level accelerations. Traditional efforts primarily target sampling dynamics, reducing the number of denoising iterations through advanced ODE/SDE solvers [8, 14] or distillation techniques like Latent Consistency Models (LCM) [21]. Other orthogonal directions include attention scheduling [29] and hardware-aware system optimizations [23]. While these methods effectively accelerate multi-step models, they are fundamentally complementary to token-level routing. For one-step VSR frameworks, where sampling-step reduction is already saturated, alleviating the internal token complexity is not trivial.

Token Compression in DiT Backbones: Token compression directly reduces sequence length to mitigate the computational overhead of self-attention. Methods like Token Merging (ToMe) [2, 3] and its video extension, VVidToMe [19], aggregate redundant tokens, while various pruning adaptations selectively discard them to improve diffusion efficiency [4, 9, 11, 12]. However, applying naive token compression to DiT-based VSR often degrades high-frequency textures and temporal consistency, particularly when token selection ignores motion dynamics or is applied uniformly across all network depths. Unlike existing literature that relies on static or depth-agnostic compression, TRaM-VSR introduces a dynamic, importance-aware routing mechanism. By coupling risk-guided interval planning with a motion-sensitive TCG saliency cue, our method strictly preserves the generative priors of DiT backbones while maximizing inference throughput. Importantly, TCG is exclusively utilized for routing and does not alter the underlying denoising trajectory.

3 Preliminaries

We operate in the latent space of a pretrained Variational Autoencoder (VAE) to process a low-quality (LQ) video $\mathbf{x}^{\mathrm{lq}} \in \mathbb{R}^{B\times 3\times F\times H\times W}$ and its high-quality (HQ) counterpart $\mathbf{x}^{\mathrm{hq}}$. The VAE encoder $\mathcal{E}$ maps these videos to their respective latent representations:

$$\mathbf{z}^{\mathrm{lq}} = \mathcal{E}(\mathbf{x}^{\mathrm{lq}}), \qquad \mathbf{z}^{\mathrm{hq}} = \mathcal{E}(\mathbf{x}^{\mathrm{hq}}). \tag{1}$$

Following the standard latent diffusion formulation, a clean latent $\mathbf{z}_0$ is perturbed by Gaussian noise. At a fixed super-resolution timestep t, the noised latent is

$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{2}$$

where $\bar{\alpha}_t \in (0, 1)$ is the cumulative noise level determined by the noise schedule.

Recent one-step VSR frameworks [6, 18, 25] train the denoiser to predict the velocity target $\mathbf{v}_\theta$. Conditioned on the noised latent $\mathbf{z}_t$, a text prompt $\mathbf{c}$, and the

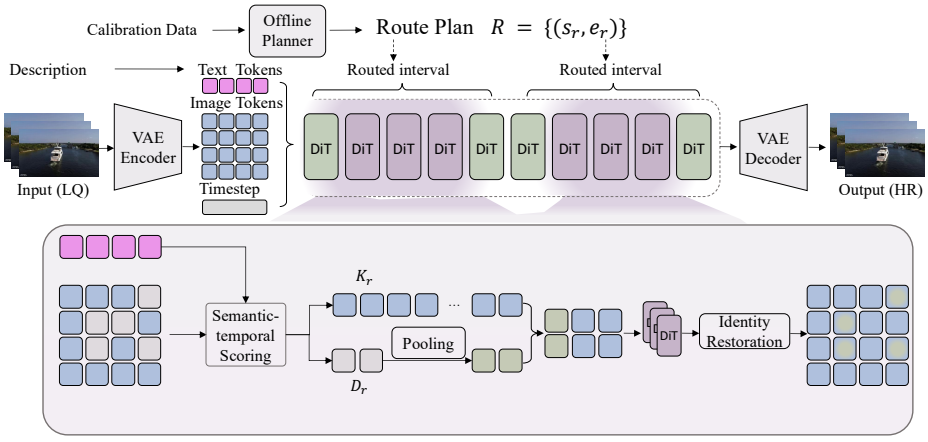

Fig. 2: Overview of the proposed TRaM-VSR pipeline. Video frames are first analyzed to estimate token importance. This video-aware prior is then combined with network-level priors, organizing tokens across multiple block granularities.

timestep t, the network directly estimates the clean latent in a single forward pass:

$$\hat{\mathbf{z}}_0 = \sqrt{\bar{\alpha}_t}\,\mathbf{z}_t - \sqrt{1-\bar{\alpha}_t}\,\mathbf{v}_\theta(\mathbf{z}_t, \mathbf{c}, t). \tag{3}$$

The final restored video is then decoded via the VAE decoder as $\hat{\mathbf{x}} = \mathcal{D}(\hat{\mathbf{z}}_0)$.

4 Methodology

4.1 Overview

Our goal is to reduce the token-computation cost of a one-step DiT denoiser without sacrificing restoration quality. As illustrated in Figure 2, TRaM-VSR confines token compaction to explicitly selected depth intervals and maintains full-resolution processing elsewhere. Each routed interval is specified by its start layer and end layer; routing begins at the start layer, compact processing is applied within the interval, and full-resolution tokens are seamlessly restored at the end layer.

TRaM-VSR comprises three core components: (1) risk-guided offline planning for route intervals (Sec. 4.2), (2) importance-aware token routing at the route entry (Sec. 4.3), and (3) a two-stream route-and-merge process inside routed intervals with identity restoration at the route exit (Sec. 4.4). To reduce train–test mismatch, we follow a standard two-stage one-step training protocol and keep the routing mechanism enabled during training [6, 18, 25].

4.2 Risk-Guided Route Interval Planning

This stage determines a small set of non-overlapping routed intervals, each specified by its start layer s_r and end layer e_r. The token drop ratio is treated as a fixed hyper-parameter ρ shared across intervals.

An offline analysis on the uncompressed baseline model computes two layer-wise signals: a local update magnitude Δ_ℓ and an end-to-end perturbation shift S_ℓ. Both signals are normalized by their layer-wise means to enable cross-depth comparison. The perturbation shift S_ℓ is measured as the deviation in the final restored output when routing is simulated exclusively at layer ℓ (averaged over a small calibration set), which captures how local perturbations propagate to the final prediction.

With these signals, the routing risk proxy is defined as

$$rss_\ell = \hat{\Delta}_\ell \cdot \hat{S}_\ell, \tag{4}$$

and mapped to a smooth routing gate

$$gate_\ell = \sigma\left(\frac{\tau - rss_\ell}{s}\right), \qquad gate_\ell \in (0,1), \tag{5}$$

where $\sigma(\cdot)$ is the sigmoid, τ is a risk threshold, and s controls smoothness. Lower-risk layers yield larger $gate_\ell$.

We score each candidate window $\Omega = [s, s+W-1]$ by its average gate:

$$\text{Score}(\Omega) = \frac{1}{W}\sum_{\ell\in\Omega} \text{gate}_\ell. \tag{6}$$

We then greedily select R non-overlapping windows in descending order of $\text{Score}(\Omega)$,

4.3 Importance-Aware Token Routing

Given a planned interval (s_r, e_r) from Sec. 4.2, the routing mechanism dictates which image tokens to preserve at the route entry, as summarized in Algorithm 1. Let $\mathbf{T}^\ell \in \mathbb{R}^{B\times L\times C}$ and $\mathbf{H}^\ell \in \mathbb{R}^{B\times N\times C}$ be text and image tokens at layer ℓ. Token compaction is applied strictly to image tokens $\mathbf{H}^\ell$, while text tokens remain entirely untouched. Temporal curvature is used only as a lightweight saliency cue for token ranking under a fixed routing budget; it does not modify the denoising trajectory.

Text-similarity importance. Let $\bar{\mathbf{t}}_b = \frac{1}{L}\sum_{j=1}^{L} \mathbf{T}^\ell_{b,j}$ be the mean text embedding. For token n:

$$s^{\text{text}}_{b,n} = \text{Norm}\left(\frac{\langle \mathbf{H}^\ell_{b,n}, \bar{\mathbf{t}}_b\rangle}{\|\mathbf{H}^\ell_{b,n}\|_2 \|\bar{\mathbf{t}}_b\|_2 + \epsilon}\right), \tag{7}$$

where $\text{Norm}(\cdot)$ denotes per-sample min–max normalization over image tokens.

Temporal Curvature Guidance (TCG) importance. We reshape image tokens as $\mathbf{X} \in \mathbb{R}^{B\times F_g\times P\times C}$, where F_g is the number of frame groups and P is the number of tokens per group. Define temporal velocity and curvature:

$$\mathbf{v}_{b,f,p} = \mathbf{X}_{b,f+1,p} - \mathbf{X}_{b,f,p}, \tag{8}$$

$$\kappa_{b,f,p} = \arccos\left(\frac{\langle \mathbf{v}_{b,f,p}, \mathbf{v}_{b,f+1,p}\rangle}{\|\mathbf{v}_{b,f,p}\|_2\|\mathbf{v}_{b,f+1,p}\|_2 + \epsilon}\right), \tag{9}$$

where the cosine term is clamped to $[-1, 1]$ prior to the arccos operation for numerical stability. We reshape/broadcast κ back to the flattened token order ($N = F_gP$) and normalize to obtain $s^{\text{tcg}} \in \mathbb{R}^{B\times N}$.

Semantic–temporal importance and keep set. We fuse cues by normalized weighted averaging:

$$s_{b,n} = \text{Norm}\left(\frac{w_t\, s^{\text{text}}_{b,n} + w_p\, s^{\text{tcg}}_{b,n}}{w_t + w_p}\right). \tag{10}$$

Given the fixed drop ratio ρ, the keep count is $K_r = \max(1, N - \lfloor \rho N \rfloor)$. We select the keep/drop sets by top-K_r:

$$\mathcal{K}_r = \text{TopK}(s, K_r), \qquad \mathcal{D}_r = \{1, \ldots, N\} \setminus \mathcal{K}_r. \tag{11}$$

4.4 Two-Stream Token Merge and Identity Restoration

Inside a routed interval, computation proceeds via a two-stream architecture (see Algorithm 1): a local stream for kept tokens and a compact global stream formed by merging dropped tokens.

Route entry (merge). Kept tokens are gathered as

$$\mathbf{H}_{\text{loc}} = \text{Gather}(\mathbf{H}^{\ell}, \mathcal{K}_r). \tag{12}$$

Let $M_r = |\mathcal{D}_r|$ and global ratio γ. The number of global tokens is bounded by

$$G_r = \min\Big(M_r,\ g_{\max},\ \max(g_{\min}, \lfloor \gamma M_r \rfloor)\Big). \tag{13}$$

We sort $\mathcal{D}_r$ and partition it into G_r groups $\{\mathcal{D}_{r,g}\}_{g=1}^{G_r}$. Each group is merged by mean pooling:

$$\mathbf{g}_{b,g} = \frac{1}{|\mathcal{D}_{r,g}|} \sum_{n\in\mathcal{D}_{r,g}} \mathbf{H}^{\ell}_{b,n}. \tag{14}$$

The working tokens are

$$\mathbf{H}^{\ell}_{\text{work}} = [\mathbf{H}_{\text{loc}}; \mathbf{G}], \quad \mathbf{G} = [\mathbf{g}_1, \ldots, \mathbf{g}_{G_r}]. \tag{15}$$

For rotary embeddings, local tokens use routed indices, while each global token uses averaged rotary features within its group.

Algorithm 1 TRaM-VSR Route-and-Merge in a Routed Interval

Require: Full tokens $X \in \mathbb{R}^{N\times C}$, text tokens T, drop ratio ρ, global ratio γ, weights (w_t, w_p), routed blocks $\mathcal{F}$
Ensure: Restored full tokens $\hat{X} \in \mathbb{R}^{N\times C}$
Semantic–temporal scoring:
  $s^{text} \leftarrow \mathrm{TextSim}(X, T)$
  $s^{tcg} \leftarrow \mathrm{TCG}(X)$ ▷ time-group curvature cue
  $s \leftarrow \mathrm{Norm}(w_t s^{text} + w_p s^{tcg})$
Top-K keep:
  $K \leftarrow \max(1, N - \lfloor \rho N \rfloor)$
  $\mathcal{K}_r \leftarrow \mathrm{TopK}(s, K)$; $\mathcal{D}_r \leftarrow [1..N] \setminus \mathcal{K}_r$
Two-stream merge:
  $X_{loc} \leftarrow X[\mathcal{K}_r]$
  $X_{glob} \leftarrow \mathrm{GroupMean}(X[\mathcal{D}_r], \gamma)$
  $X_{work} \leftarrow [X_{loc}; X_{glob}]$
Routed computation: $Y_{work} \leftarrow \mathcal{F}(X_{work})$
Unmerge:
  $\hat{X} \leftarrow X$ ▷ freeze base for dropped tokens
  $\hat{X}[\mathcal{K}_r] \leftarrow Y_{work}[1{:}K]$ ▷ scatter kept tokens
return $\hat{X}$

Route exit (unmerge). Let $\mathbf{H}^{\ell'}_{\mathrm{work}} = [\mathbf{H}^{\ell'}_{\mathrm{loc}}; \mathbf{G}^{\ell'}]$ be the output at the end layer e_r. To reconstruct the full-resolution token grid, we bypass the global stream features and populate the dropped positions with their original features saved at the route entry (denoted as $\mathbf{H}_{\mathrm{snap}}$):

$$\tilde{\mathbf{H}}^{\ell'}_{b,n} = \begin{cases} \mathbf{H}^{\ell'}_{\mathrm{loc},b,\pi(n)}, & n \in \mathcal{K}_r, \\ \mathbf{H}_{\mathrm{snap},b,n}, & n \in \mathcal{D}_r, \end{cases} \tag{16}$$

where $\pi(\cdot)$ maps an original kept index to its local position. This mechanism reduces attention complexity from $\mathcal{O}(N^2)$ to approximately $\mathcal{O}((K_r+G_r)^2)$ inside routed layers.

Rationale. The local stream preserves high-fidelity tokens for detail synthesis, while the global stream retains low-cost context within routed layers. Identity restoration avoids injecting potentially noisy updates into dropped positions and yields stable behavior under aggressive routing.

5 Experiments

5.1 Experimental Settings

Datasets. Following standard VSR evaluation protocols, all experiments are conducted at a $\times 4$ upscaling factor. We evaluate our method on three synthetic benchmarks (UDM10 [27], SPMCS [41], and YouHQ40 [44]) and three real-world benchmarks (RealVSR [40], MVSR4x [33], and VideoLQ [5]). The synthetic datasets employ the same degradation pipeline used during training.

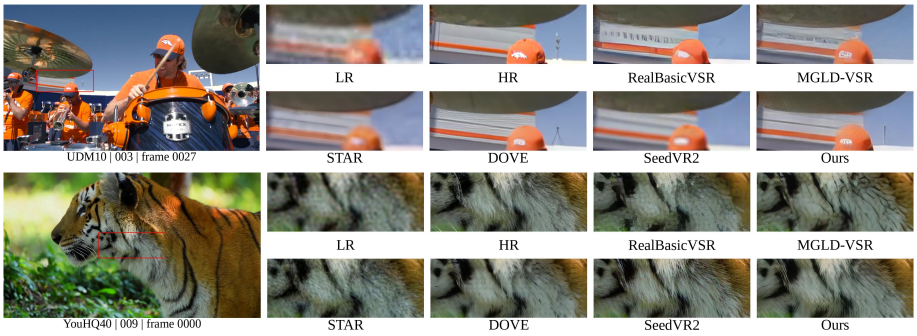

Fig. 3: Visual comparison on real-world UDM10 [27] (top) and synthetic YouHQ40 [44] (bottom). TRaM-VSR recovers substantially sharper fine textures and structurally coherent edges compared to recent baselines.

RealVSR and MVSR4x contain real-world LQ–HQ pairs captured in the wild, whereas VideoLQ consists of internet-sourced videos without ground-truth HR.
Evaluation Metrics. For datasets with HR references, we report standard fidelity metrics (PSNR and SSIM [36]) and perceptual full-reference metrics (LPIPS [43] and DISTS [7]). For no-reference perceptual evaluation, we employ CLIP-IQA [30] and MUSIQ [15]. To strictly assess temporal consistency, we measure the flow-warping error (E^*_{warp}) [17], where lower values indicate smoother temporal transitions. Unless otherwise stated, all metrics are computed in the RGB color space.
Implementation Details. TRaM-VSR is built upon a one-step latent DiT VSR architecture and is evaluated using a fixed one-step sampling scheme. To ensure fair evaluation, we maintain strictly identical degradation and testing protocols across all compared methods. By default, TRaM-VSR is configured with two routed intervals, an importance fusion mechanism combining text similarity and TCG cues, and the two-stream local–global token architecture with identity restoration at the route exits. At inference time, TRaM-VSR does not require any manually provided caption or prompt; all quantitative and visual results are obtained under this prompt-free setting.

5.2 Comparison with State-of-the-Art Methods

We compare TRaM-VSR against representative state-of-the-art image and video restoration baselines, including RealESRGAN [34], ResShift [42], RealBasicVSR [5], Upscale-A-Video [44], MGLD-VSR [39], STAR [38], and DOVE [6].
Quantitative Results. Quantitative comparisons on synthetic and real-world benchmarks are reported in Table 1. Overall, TRaM-VSR achieves a superior quality–efficiency balance under strictly one-step inference. On synthetic datasets, our framework consistently improves perceptual metrics (e.g., LPIPS and DISTS) while maintaining highly competitive pixel-wise fidelity (PSNR/SSIM).

Crucially, on the RealVSR dataset, TRaM-VSR remains robust, yielding improved perceptual scores and strong temporal consistency. These results substantiate the effectiveness of our importance-aware routing and two-stream token processing for precision-critical video restoration.

Qualitative Results. Visual comparisons on real-world UDM10 and synthetic YouHQ40 are provided in Figure 3. TRaM-VSR synthesizes sharper structures and more faithful high-frequency textures, effectively avoiding the over-smoothing artifacts typical of purely regression-based restorers or aggressive token compaction. In particular, our method reliably preserves repetitive patterns and thin structural edges.

Efficacy of the Routing Strategy. To explicitly isolate the benefits of our framework, we compare TRaM-VSR against alternative acceleration strategies applied to the identical one-step DiT backbone: (i) uniform token merging and (ii) uniform token pruning (Figure 4). While naive merging and pruning reduce computational overhead, they consistently introduce characteristic artifacts such as texture over-smoothing, local contrast drift, and inconsistent edge transitions—flaws that are exacerbated by the skip-free nature of DiT architectures. In contrast, TRaM-VSR leverages semantic–temporal scoring to safeguard visually vital regions and preserves global context via the two-stream representation. Furthermore, identity restoration definitively prevents the irreversible detail loss observed in the baselines. Consequently, our method reconstructs sharper, more coherent textures, demonstrating a strictly more favorable quality–efficiency trade-off for real-world scenarios.

Temporal Consistency. We visualize temporal coherence using space–time slice inspection in Figure 5. While prior baselines exhibit noticeable structural wobbling and temporal fluctuations under complex degradations (reflected by significant slice errors), TRaM-VSR yields remarkably smooth temporal profiles with minimized error accumulation. This visual stability strongly aligns with the superior E^*_{warp} performance reported in Table 1.

5.3 Ablation Study

We conduct extensive ablations on the SPMCS dataset to validate the core architectural designs of TRaM-VSR.

Routing Structure. Table 2 compares random routing, single-stream importance-aware routing, and our proposed two-stream framework. Random routing severely degrades perceptual quality due to unstructured token removal. While implementing importance-aware routing improves selection, the single-stream variant (lacking a global stream) still suffers perceptually. This indicates that aggressively discarding tokens without retaining global context is profoundly detrimental to VSR. Our full two-stream design achieves the best overall perceptual quality (LPIPS and MUSIQ) while maintaining robust temporal consistency, demonstrating the necessity of jointly modeling localized high-fidelity details and compact global context within routed layers.

Importance Fusion. We ablate the components of our token-scoring mechanism in Table 3. Relying solely on text similarity provides semantic guidance

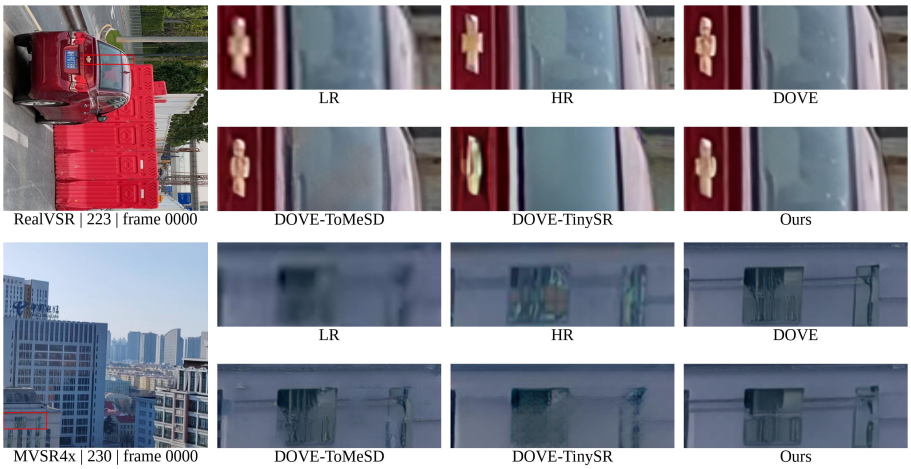

Fig. 4: Visual comparison of alternative acceleration strategies applied to the one-step DOVE backbone on RealVSR [40] (top) and MVSR4x [33] (bottom). Compared to uniform token merging (DOVE-ToMeSD) and token pruning (DOVE-TinySR), our TRaM-VSR effectively mitigates texture over-smoothing and strictly preserves structural integrity under high compression regimes.

but neglects motion-sensitive regions. Conversely, utilizing TCG scoring alone strictly enforces temporal dynamics (yielding the lowest E^*_{warp} error) but limits semantic synthesis. Fusing text similarity and TCG cues (at a 0.6/0.4 ratio) achieves the optimal balance, maximizing overall perceptual quality without severely compromising video-level fidelity. This confirms that semantic relevance and temporal dynamics offer complementary signals for identifying precision-critical tokens.

Offline Planner and Route Scheduling. The efficacy of our routed-interval selection is evaluated via the speed–quality analysis in Figure 6. Each plotted point represents a distinct interval configuration and drop ratio. The specific routing intervals identified by our offline planner consistently trace a superior quality–efficiency Pareto front compared to alternative heuristic placements at comparable speedups. This validates our calibration-guided scheduling approach. Furthermore, sweeping the drop ratio produces a smooth trade-off curve, confirming that ρ serves as a predictable and robust control knob for practical deployment.

Identity Restoration. Finally, we analyze the critical role of route-exit restoration. In standard skip-free DiT backbones, intermediate token compaction inevitably induces irreversible detail loss. By completely bypassing the global stream and reconstructing the full-resolution grid using untouched features saved at the interval entry, our identity restoration mechanism drastically improves structural robustness. This ensures that precision-critical details remain pristine

Fig. 5: Temporal consistency visualization on SPMCS by stacking the red vertical line across consecutive frames (left) and showing the corresponding temporal slices (right); the last row is the error map (red indicates larger error).

while fully retaining the computational benefits of the compacted routed layers (as evidenced by Table 2).

6 Token Visualization and Discussion

In Figure 7, we visualize the token selection for the video examples shown in Figure 1. From left to right, we present: (a) token selection for the original video, (b) for the reversed video, and (c) for a static video where the first frame is repeated across all video frames, resulting in no changes.

Asymmetric Token Routing. Examining case (a) in detail, we observe that routing decisions differ between block intervals 1 and 2. More importantly, when the car is approaching (left), tokens corresponding to the car's shadow are dropped more aggressively, whereas when the car is leaving (right), the trend is reversed. We hypothesize that this behavior reflects the natural motion dynamics and the model's implicit world understanding of the car's movement. As the car approaches, its front shadow gradually expands. Although this introduces local pixel changes, the newly appearing shadow regions follow a highly regular and predictable pattern, progressively turning local areas into darker pixels and reducing their informative content. As a result, our method identifies these regions as statistically predictable and applies more aggressive dropping.

Interestingly, compared to shallow block segments, deeper blocks exhibit slightly higher activation on the same region. While this may appear to introduce additional computation, it actually reflects a meaningful hierarchical modeling strategy, which is further facilitated by our network-level prior. Although the shadow itself is predictable at the pixel level, it remains closely related to the object's geometric structure, illumination relationships, and temporal continuity. Consequently, the network-level prior encourages deeper layers, which operate at a more semantic and structural level, to retain limited attention on these regions

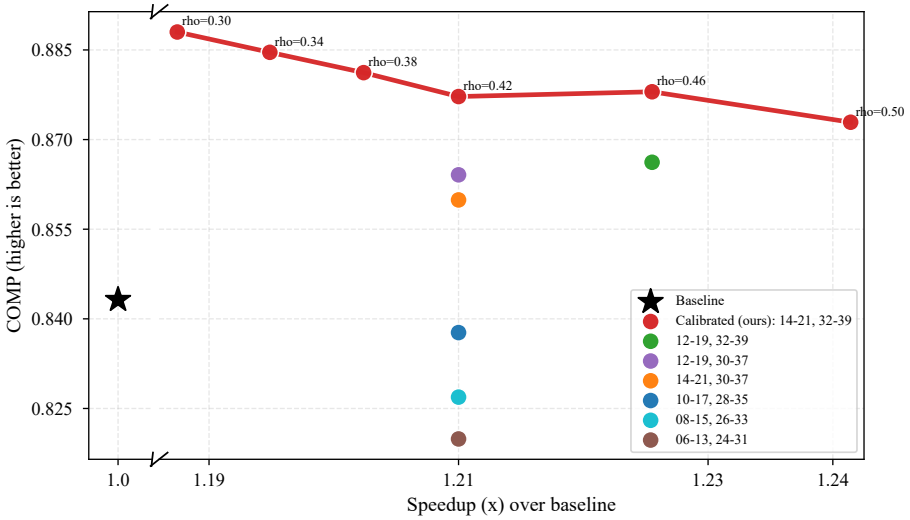

Fig. 6: Speed–quality trade-off under different routed intervals and drop ratios.

Fig. 7: Spatio-temporal routing visualization on the dynamic "Jeep" sequence.

in order to maintain geometric and temporal consistency between the shadow and the moving object.

Emerging Bidirectional Temporal Behavior. We further analyze the reversed video and visualize the same frames reordered as in (b). In this case, the car moves backward and the shadow continues to expand. Here we observe a similar activation trend as in (a). This observation is particularly notable because the baseline model [6] was originally trained only for single-direction temporal processing. However, after fine-tuning the baseline with our proposed network, our token adaptation mechanism enables the originally unidirectional model to effectively handle bidirectional temporal dynamics in a consistent way.

Without Motion. Finally, we manually construct a static video by repeating the first frame, as illustrated in (c), such that no motion exists throughout the sequence. This setting allows us to isolate the behavior of the network-level priors. Interestingly, when comparing (a-left) with (c), we observe that the deeper

Table 1: Quantitative comparison with state-of-the-art methods. Best and second-best results are marked in red and blue.

Dataset	Metric	Real-ESRGAN [34]	ResShift [42]	RealBasicVSR [5]	Upscale-A-Video [44]	MGLD-VSR [39]	STAR [38]	DOVE [6]	SeedVR2-7B [31]	Ours
UDM10	PSNR ↑	24.04	23.65	24.13	21.72	24.23	23.47	26.48	25.47	26.73
	SSIM ↑	0.7107	0.6016	0.6801	0.5913	0.6957	0.6804	0.7827	0.7582	0.7759
	LPIPS ↓	0.3877	0.5537	0.3908	0.4116	0.3272	0.4242	0.2696	0.2507	0.2678
	DISTS ↓	0.2184	0.2898	0.2067	0.2230	0.1677	0.2156	0.1492	0.1263	0.1450
	CLIP-IQA ↑	0.4189	0.4344	0.3494	0.4697	0.4557	0.2417	0.5107	0.2958	0.4547
	E^*_{warp} ↓	4.83	6.12	3.10	3.97	3.59	2.08	1.77	2.20	1.73
SPMCS	PSNR ↑	21.22	21.68	22.17	18.81	22.39	21.24	23.11	22.53	23.18
	SSIM ↑	0.5613	0.5153	0.5638	0.4113	0.5896	0.5441	0.6210	0.6187	0.6237
	LPIPS ↓	0.3721	0.4467	0.3662	0.4468	0.3263	0.5257	0.2888	0.2781	0.2883
	DISTS ↓	0.2220	0.2697	0.2164	0.2452	0.1960	0.2872	0.1713	0.1530	0.1613
	CLIP-IQA ↑	0.5238	0.5442	0.3513	0.5248	0.4348	0.2646	0.5690	0.4271	0.5588
	E^*_{warp} ↓	5.61	8.07	1.88	4.22	1.68	1.01	1.04	1.38	0.97
YouHQ40	PSNR ↑	22.82	23.32	22.39	19.62	23.17	22.64	24.30	23.30	24.38
	SSIM ↑	0.6337	0.6273	0.5895	0.4824	0.6194	0.6323	0.6740	0.6659	0.6856
	LPIPS ↓	0.3571	0.4211	0.4091	0.4268	0.3608	0.4600	0.2997	0.2705	0.2903
	DISTS ↓	0.1790	0.2159	0.1933	0.2012	0.1685	0.2287	0.1477	0.1117	0.1370
	CLIP-IQA ↑	0.4704	0.4633	0.3964	0.5258	0.4657	0.2739	0.4985	0.3321	0.4676
	E^*_{warp} ↓	5.91	5.75	3.08	6.84	3.45	2.21	2.05	4.03	1.70
RealVSR	PSNR ↑	20.85	20.81	22.12	20.29	22.02	17.43	22.32	22.09	22.37
	SSIM ↑	0.7105	0.6277	0.7163	0.5945	0.6774	0.5215	0.7301	0.7181	0.7288
	LPIPS ↓	0.2016	0.2312	0.1870	0.2671	0.2182	0.2943	0.1851	0.2014	0.1911
	DISTS ↓	0.1279	0.1435	0.0983	0.1425	0.1169	0.1599	0.0978	0.1146	0.1005
	CLIP-IQA ↑	0.7472	0.5553	0.2905	0.4855	0.4510	0.3641	0.5207	0.2872	0.5062
	E^*_{warp} ↓	6.32	9.55	4.45	6.25	3.16	9.88	3.52	3.54	3.23
MVSR4x	PSNR ↑	22.47	21.58	21.80	20.42	22.77	22.42	22.42	22.46	22.56
	SSIM ↑	0.7412	0.6473	0.7045	0.6117	0.7418	0.7421	0.7523	0.7426	0.7591
	LPIPS ↓	0.4534	0.5945	0.4235	0.4717	0.3568	0.4311	0.3476	0.3643	0.3373
	DISTS ↓	0.3021	0.3351	0.2498	0.2673	0.2245	0.2714	0.2363	0.2293	0.2245
	CLIP-IQA ↑	0.4396	0.5003	0.4118	0.6106	0.3769	0.2674	0.5453	0.2127	0.5076
	E^*_{warp} ↓	1.64	3.89	1.69	5.10	1.55	0.61	0.78	0.92	0.75

block segments exhibit substantially higher activation than the shallow ones in the same shadow regions. As motion cues vanish in this static setting, shadows are no longer predictable from temporal dynamics. Instead, we venture that in this case shadows become stable spatial structures tightly coupled with object geometry and illumination, carrying greater semantic significance. Consequently, deeper blocks tend to preserve these regions to maintain semantic and geometric consistency, leading to reduced dropping compared to the dynamic scenario.

7 Limitations and Future Work

Although TRaM-VSR improves temporal stability and efficiency, several limitations remain. First, one-step VSR may still show mild residual oscillation in challenging regions such as fine repetitive textures, shadows, and fast local motion. Second, our speedup primarily targets DiT token computation, so the end-to-end acceleration depends on the relative cost of VAE decoding and implementation details such as tiling. Third, the routed intervals are selected by offline

Table 2: Structural ablation on SPMCS: random routing vs. single-stream routing vs. two-stream local/global routing. Best results are marked in **bold**.

Variant	PSNR ↑	LPIPS ↓	MUSIQ ↑	E^*_{warp} ↓
Random route (single-stream)	23.13	0.3187	62.84	1.0030
Text+TCG (single-stream, no global)	23.08	0.3391	63.28	0.9839
Full TRaM-VSR (two-stream, with global)	**23.18**	**0.2883**	**67.27**	**0.9720**

Table 3: Ablation on SPMCS. The 0.6/0.4 fusion setting yields the best overall perceptual quality, while utilizing TCG-only achieves the lowest warp error. Best results are marked in **bold**.

Variant (SPMCS, inference)	PSNR ↑	LPIPS ↓	MUSIQ ↑	E^*_{warp} ↓
Text-only	23.09	0.3450	62.79	0.9295
PSG-only	22.98	0.3343	59.49	**0.9068**
Text+PSG (0.6/0.4)	**23.18**	**0.2883**	**67.27**	0.9720

calibration; while this setup is fixed across our benchmarks, new backbones or substantially different degradation regimes may benefit from recalibration. Future work will explore stronger temporal priors and adaptive routing calibration.

8 Conclusion

In this paper, we presented TRaM-VSR, an efficient Token Routing and Merging framework that addresses the prohibitive computational costs and temporal instability inherent in one-step DiT-based video super-resolution. Moving beyond uniform token compression, our approach dynamically allocates computational budgets by jointly leveraging context-aware video priors and multi-granular network-level priors. By fusing semantic-temporal importance scoring with depth-calibrated routing, TRaM-VSR explicitly preserves dynamic objects and structural boundaries within a high-fidelity local stream, while compactly aggregating redundant tokens. Extensive evaluations confirm that our framework significantly accelerates inference without compromising state-of-the-art perceptual quality and temporal coherence, offering a highly practical solution for real-world VSR deployment.

Acknowledgments. This work was supported by The Alexander von Humboldt Foundation.

References

1. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Esser, P.: Align your latents: High-resolution video synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)

2. Bolya, D., Fu, C.Y., Dai, X., Zhang, P., Feichtenhofer, C., Hoffman, J.: Token merging: Your ViT but faster. In: International Conference on Learning Representations (ICLR) (2023)
3. Bolya, D., Hoffman, J.: Token merging for fast stable diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2023)
4. Castells, T., Song, H.K., Kim, B.K., Choi, S.: LD-Pruner: Efficient pruning of latent diffusion models using task-agnostic insights. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2024)
5. Chan, K.C., Zhou, S., Xu, X., Loy, C.C.: Investigating tradeoffs in real-world video super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5962–5971 (2022)
6. Chen, Z., Zou, Z., Zhang, K., Su, X., Yuan, X., Guo, Y., Zhang, Y.: Dove: Efficient one-step diffusion model for real-world video super-resolution (2025). https://doi.org/10.48550/arXiv.2505.16239
7. Ding, K., Ma, K., Wang, S., Simoncelli, E.P.: Image quality assessment: Unifying structure and texture similarity. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(5), 2567–2581 (2020)
8. Dockhorn, T., Vahdat, A., Kreis, K.: Genie: Higher-order denoising diffusion solvers. Advances in Neural Information Processing Systems **35**, 30150–30166 (2022)
9. Dong, L., Fan, Q., Yu, Y., Zhang, Q., Chen, J., Luo, Y., Zou, C.: TinySR: Pruning diffusion for real-world image super-resolution (2025)
10. Esser, P., Chiu, J., Atighehchian, P., Gritsenko, A., Fidler, S.: Structure and content-guided video synthesis with diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
11. Fang, G., Li, K., Ma, X., Wang, X.: Tinyfusion: Diffusion transformers learned shallow. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 18144–18154 (2025)
12. Fang, G., Ma, X., Wang, X.: Structural pruning for diffusion models. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)
13. He, J., Xue, T., Liu, D., Lin, X., Gao, P., Lin, D., Qiao, Y., Ouyang, W., Liu, Z.: VEnhancer: Generative space-time enhancement for video generation (2024), technical report (arXiv:2407.07667)
14. Karras, T., Aittala, M., Lehtinen, J., Hellsten, J., Aila, T., Laine, S.: Analyzing and improving the training dynamics of diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24174–24184 (2024)
15. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: MUSIQ: Multi-scale image quality transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5148–5157 (2021)
16. Khachatryan, L., Movsisyan, A., Tadevosyan, V., Navasardyan, R., Shi, Y., Sargsyan, S.: Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
17. Lai, W.S., Huang, J.B., Wang, O., Shechtman, E., Yumer, E., Yang, M.H.: Learning blind video temporal consistency. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 170–185 (2018)

18. Li, H., Tang, H., Han, J., Zhou, T., Cui, J., Xie, H., Chen, Y., Hu, J.: Os-diffvsr: Towards one-step latent diffusion model for high-detailed real-world video super-resolution (2025)
19. Li, X., Ma, C., Yang, X., Yang, M.H.: VidToMe: Video token merging for zero-shot video editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
20. Liu, Y., Pan, J., Li, Y., Dong, Q., Zhu, C., Guo, Y., Wang, F.: Ultravsr: Achieving ultra-realistic video super-resolution with efficient one-step diffusion space (2025)
21. Luo, S., Chen, Y., Sun, Y., Li, L., Li, J., Zhao, H.: Latent consistency models: Synthesizing high-resolution images with few-step inference (2023)
22. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
23. Shen, H., et al.: Efficient diffusion models: A survey. Transactions on Machine Learning Research (TMLR) (2025)
24. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. In: International Conference on Learning Representations (ICLR) (2023)
25. Sun, Y., Sun, L., Liu, S., Wu, R., Zhang, Z., Zhang, L.: One-step diffusion for detail-rich and temporally consistent video super-resolution (2025)
26. Tang, L., Jia, M., Wang, Q., Phoo, C.P., Hariharan, B.: Emergent correspondence from image diffusion. Advances in neural information processing systems **36**, 1363–1389 (2023)
27. Tao, X., Gao, H., Liao, R., Wang, J., Jia, J.: Detail-revealing deep video super-resolution. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 4472–4480 (2017)
28. Villegas, R., Babaeizadeh, M., Kindermans, P.J., Sajjadi, M.S., Zhang, H., Gu, M., Kumar, S., Tulyakov, S., Tagliasacchi, M., Dumoulin, V.: Phenaki: Variable length video generation from open domain textual descriptions. In: International Conference on Learning Representations (ICLR) (2023)
29. Wang, H., Liu, D., Kang, Y., Li, Y., Lin, Z., Jha, N.K., Liu, Y.: Attention-driven training-free efficiency enhancement of diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
30. Wang, J., Chan, K.C., Loy, C.C.: Exploring CLIP for assessing the look and feel of images. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 2555–2563 (2023)
31. Wang, J., Lin, S., Lin, Z., Ren, Y., Wei, M., Yue, Z., Zhou, S., Chen, H., Zhao, Y., Yang, C., et al.: Seedvr2: One-step video restoration via diffusion adversarial post-training. arXiv preprint arXiv:2506.05301 (2025)
32. Wang, J., Lin, Z., Wei, M., Zhao, Y., Yang, C., Loy, C.C., Jiang, L.: Seedvr: Seeding infinity in diffusion transformer towards generic video restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2161–2172 (2025)
33. Wang, R., Liu, X., Zhang, Z., Wu, X., Feng, C.M., Zhang, L., Zuo, W.: Benchmark dataset and effective inter-frame alignment for real-world video super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1168–1177 (2023)
34. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1905–1914 (2021)

35. Wang, Y., Chen, X., Ma, X., Zhou, S., Huang, Z., Wang, Y., Yang, C., He, Y., Yu, J., Yang, P., Guo, Y., Wu, T., Si, C., Jiang, Y., Chen, C., Loy, C.C., Dai, B., Lin, D., Qiao, Y., Liu, Z.: LAVIE: High-quality video generation with cascaded latent diffusion models. International Journal of Computer Vision (IJCV) (2024)
36. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004)
37. Wu, R., Yang, T., Sun, L., Zhang, Z., Li, S., Zhang, L., Qiao, Y., Dong, C.: SeeSR: Towards semantics-aware real-world image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
38. Xie, R., Liu, Y., Zhou, P., Zhao, C., Zhou, J., Zhang, K., Zhang, Z., Yang, J., Yang, Z., Tai, Y.: Star: Spatial-temporal augmentation with text-to-video models for real-world video super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17108–17118 (2025)
39. Yang, X., He, C., Ma, J., Zhang, L.: Motion-guided latent diffusion for temporally consistent real-world video super-resolution. In: European conference on computer vision. pp. 224–242. Springer (2024)
40. Yang, X., Xiang, W., Zeng, H., Zhang, L.: 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 (ICCV). pp. 4781–4790 (2021)
41. Yi, P., Wang, Z., Jiang, K., Jiang, J., Ma, J.: Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3106–3115 (2019)
42. Yue, Z., Wang, J., Loy, C.C.: Efficient diffusion model for image restoration by residual shifting. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(1), 116–130 (2025). https://doi.org/10.1109/TPAMI.2024.3461721
43. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 586–595 (2018)
44. Zhou, S., Wang, J., Chan, K.C., Loy, C.C.: 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 (CVPR) (2024)