TextOVSR: Text-Guided Real-World Opera Video Super-Resolution

Hua Chang[1] Xin Xu[1,2,*] Wei Liu[1] Jiayi Wu[1] Kui Jiang[3,*] Fei Ma[4] Qi Tian[5]
[1]School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan, China
[2]Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System
[3]Harbin Institute of Technology Zhengzhou Research Institute, Zhengzhou, China
[4]Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ)
[5]Huawei Technologies Ltd.
{changhua, xuxin, liuwei, wuaddone}@wust.edu.cn
jiangkui@hit.edu.cn, mafei@gml.ac.cn, tian.qi1@huawei.com

Abstract

Many classic opera videos exhibit poor visual quality due to the limitations of early filming equipment and long-term degradation during storage. Although real-world video super-resolution (RWVSR) has achieved significant advances in recent years, directly applying existing methods to degraded opera videos remains challenging. The difficulties are twofold. First, accurately modeling real-world degradations is complex: simplistic combinations of classical degradation kernels fail to capture the authentic noise distribution, while methods that extract real noise patches from external datasets are prone to style mismatches that introduce visual artifacts. Second, current RWVSR methods, which rely solely on degraded image features, struggle to reconstruct realistic and detailed textures due to a lack of high-level semantic guidance. To address these issues, we propose a Text-guided Dual-Branch Opera Video Super-Resolution (TextOVSR) network, which introduces two types of textual prompts to guide the super-resolution process. Specifically, degradation-descriptive text, derived from the degradation process, is incorporated into the negative branch to constrain the solution space. Simultaneously, content-descriptive text is incorporated into a positive branch and our proposed Text-Enhanced Discriminator (TED) to provide semantic guidance for enhanced texture reconstruction. Furthermore, we design a Degradation-Robust Feature Fusion (DRF) module to facilitate cross-modal feature fusion while suppressing degradation interference. Experiments on our OperaLQ benchmark show that TextOVSR outperforms state-of-the-art methods both qualitatively and quantitatively. The code is available at https://github.com/ChangHua0/TextOVSR.

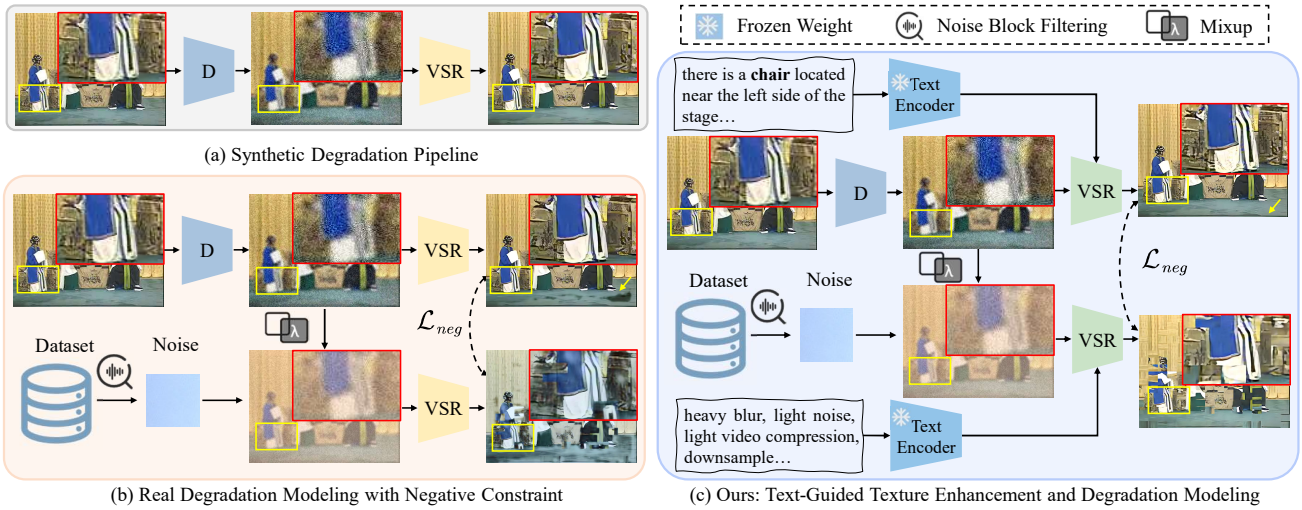
*Corresponding authors

1. Introduction

Many classic opera videos exhibit poor visual quality due to limitations in early filming equipment and complex degradation processes during storage and transmission. Video super-resolution (VSR) has advanced rapidly in recent years [6, 20], with the primary goal of restoring high-resolution (HR) frames from their low-resolution (LR) counterparts, typically generated by known kernels (e.g., bicubic). However, low-quality videos in the real-world are not the product of simple downsampling; they are affected by a complex combination of unknown degradations, including sensor noise, compression artifacts, and transmission losses [36]. Consequently, traditional VSR methods, trained under idealized degradation assumptions, suffer from a significant domain gap and demonstrate poor generalization when applied to real degraded videos [7, 31].

Real-world video super-resolution (RWVSR) has been proposed to address this domain shift [1, 25]. A central focus of this field is expanding the realism and diversity of the degradation space. For instance, Real-ESRGAN [36] employs a high-order degradation pipeline that combines simple kernels to synthesize complex, low-quality images, as shown in Figure 1(a). A key limitation of this approach is its confinement to a limited, synthetic degradation space, which often results in out-of-distribution noise. More recently, NegVSR [31] circumvented this by extracting real noise patches from external datasets to simulate authentic degradations (Figure 1(b)). However, this method is highly dependent on the stylistic similarity between the external and target datasets; a mismatch can introduce severe artifacts, as indicated by the yellow arrows in Figure 1. Furthermore, most existing RWVSR methods rely solely on degraded image features, which lack the high-level semantic information necessary for reconstructing realistic textures.

To achieve more effective degradation modeling and de-



(a) Synthetic Degradation Pipeline

(b) Real Degradation Modeling with Negative Constraint

(c) Ours: Text-Guided Texture Enhancement and Degradation Modeling

Figure 1. **Frameworks for real-world video super-resolution.** (a) The classical synthetic degradation pipeline (D) simulates real degradations for the VSR model. (b) Real degradation modeling extracts authentic noise from external datasets and applies a negative constraint ($\mathcal{L}_{neg}$) to enhance robustness. (c) Our proposed TextOVSR introduces text-guided priors to enrich image features and model diverse degradations in the feature space. λ controls the mixing ratio.

tail reconstruction, we propose TextOVSR, a dual-branch text-guided framework built upon NegVSR [31] (Figure 1(c)). We synthesize degraded inputs using a high-order degradation pipeline [7, 36]. A key innovation is the generation of textual prompts to guide the super-resolution. We employ a binning method to qualitatively describe the severity of each degradation component, concatenating these into a comprehensive degradation-descriptive text. This text is incorporated into the negative branch; after encoding by a CLIP text encoder [27], it guides the application of contrastive constraints to improve model robustness. To enhance texture reconstruction, the positive branch is guided by content-descriptive text, generated from pristine images using a multimodal large language model (MLLM) and fused with visual features via a novel Degradation-Robust Feature Fusion (DRF) module. The DRF module is designed to facilitate effective cross-modal fusion while suppressing interference from inherent degradations and style inconsistencies. Finally, we introduce a Text-Enhanced Discriminator (TED) that leverages high-level semantic cues from the content text to provide more accurate adversarial guidance, steering the generator toward photorealistic outputs. Extensive experiments on our self-constructed OperaLQ benchmark, comprising real-world degraded opera videos, demonstrate that TextOVSR achieves state-of-the-art performance in both qualitative and quantitative evaluations. In summary, our contributions are as follows:

- We propose TextOVSR, a text-guided dual-branch network for opera video super-resolution, leveraging content- and degradation-descriptive texts to improve texture reconstruction and handle complex degradations.
- We design a Degradation-Robust Feature Fusion (DRF) module that enables effective cross-modal fusion of visual and textual features while mitigating degradation-induced contamination.
- We introduce a Text-Enhanced Discriminator (TED) that leverages text semantics to improve discrimination and guide more realistic outputs.
- We construct and release OperaLQ, a benchmark of real-world degraded opera videos. On this dataset, our method achieves state-of-the-art performance across multiple image and video quality metrics.

2. Related Work

2.1. Video Super-Resolution

Video super-resolution (VSR) aims to reconstruct HR frames from their low-resolution LR counterparts by exploiting spatial and temporal correlations across frames [4]. Early approaches estimate optical flow for motion compensation to align neighboring frames before fusion [17, 29, 44], but motion estimation errors and occlusions often limit their accuracy. To overcome this, deformable convolution–based methods perform implicit alignment and aggregation in a data-driven manner [14, 33], and have been extended for intermediate-frame prediction [35] or joint reconstruction within recurrent frameworks [5, 6]. Recently, attention-driven architectures have been introduced to capture long-range temporal dependencies and enhance global feature interaction [3, 19, 20, 30]. Despite these advances, most VSR models are trained under simplified synthetic degradations (e.g., bicubic downsampling), leading to poor generalization to real-world scenarios with complex noise, compression, and motion blur.

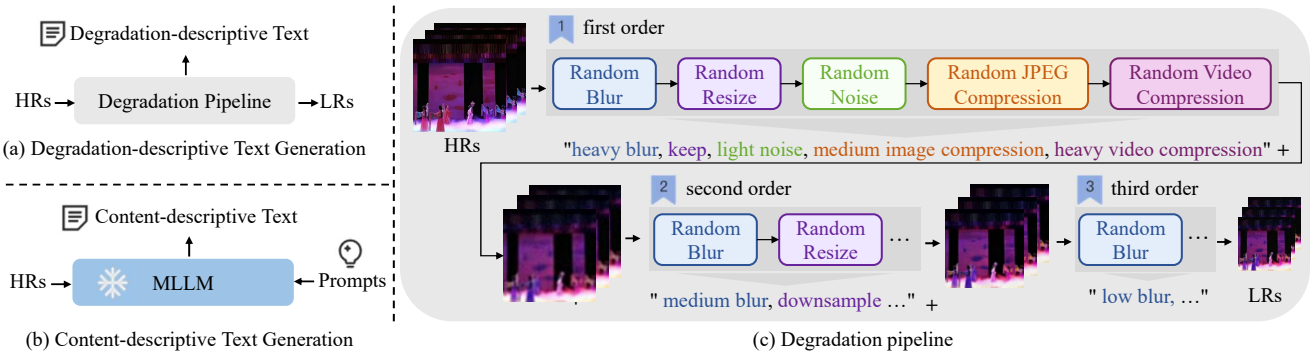


Figure 2. **Generation process of degradation- and content-descriptive texts.** Degradation-descriptive text is generated according to different intensity levels in the high-order degradation pipeline, while content-descriptive text is produced directly from high-resolution inputs (HRs) using a multimodal large language model (MLLM), rather than from degraded low-resolution videos (LRs).

2.2. Real-World Video Super-Resolution

To address complex and unknown degradations in real-world videos, recent studies have explored constructing realistic LR–HR pairs through either physical acquisition or synthetic degradation modeling. Data collection–based approaches such as RealVSR [46] employ synchronized multi-camera setups to capture paired sequences, but their dependence on specialized hardware limits scalability and general applicability. In contrast, degradation simulation–based methods focus on learning or expanding the degradation space. DBVSR [25] models blur kernels via convolutional learning, while AnimeSR [41] broadens the degradation domain using diverse synthetic operators. Real-ESRGAN [36] integrates multiple known kernels into a high-order degradation pipeline to better approximate real-world scenarios. NegVSR [31] further enhances realism by sampling noise patterns directly from real data. Beyond degradation modeling, several works incorporate degradation correction into the network architecture. RealBasicVSR [7] employs a dynamic cleaning module to suppress artifact propagation, while FastRealVSR [42] fuses sharpening and blur-kernel filtering for efficient compensation during hidden-state interaction. RealViformer [49] further highlights that channel attention can mitigate redundant information and enhance robustness against residual artifacts. Despite these advancements, existing methods still face two key challenges: the degradation space remains incomplete or domain-specific, and the reliance on image-only features limits the recovery of high-quality videos with clear structures and natural textures.

2.3. Text-guided Video Super-Resolution

Recently, text-guided image super-resolution (ISR) has gained increasing attention [37]. With powerful generative priors, text-to-image (T2I) diffusion models have shown strong potential for real-world SR [40, 45], yet their multi-step denoising leads to high computational cost. To address this, several studies adopt knowledge distillation to achieve single-step diffusion [9, 12, 39], effectively reducing complexity while maintaining perceptual quality. CLIP-SR [13] further integrates textual semantics into the reconstruction network to enhance fine texture recovery. Building on these advances, recent works extend generative priors to video SR, exploring different strategies to alleviate fidelity degradation and temporal inconsistency caused by diffusion randomness [18, 50]. STAR [43] further employs a text-to-video (T2V) diffusion model to enforce temporal coherence during generation. While diffusion-based approaches improve texture realism, their substantial computational overhead [47] still limits real-world deployment. Different from these diffusion-based methods, our approach embeds multiple types of textual prompts into a classical RWVSR framework, enabling effective degradation modeling and texture enhancement in a lightweight and robust manner.

3. Method

3.1. Description Text Generation

To provide textual guidance, we first produce degradation-descriptive text while synthesizing low-resolution (LR) videos using a high-order degradation pipeline [36], as illustrated in Figure 2(a) and (c). Each stage of the pipeline includes operations such as blur, resize, noise, JPEG compression, and video compression. Following PromptSR [11], degradations are categorized into three intensity levels: light, medium, and heavy, yielding descriptions such as “light blur”. For higher-order degradations, descriptions from successive stages are concatenated to form a comprehensive high-order degradation description. For frame-level semantic guidance, a multi-modal large language model (MLLM) generates content-descriptive text for each frame, as shown in Figure 2(b). This text captures the key visual semantics

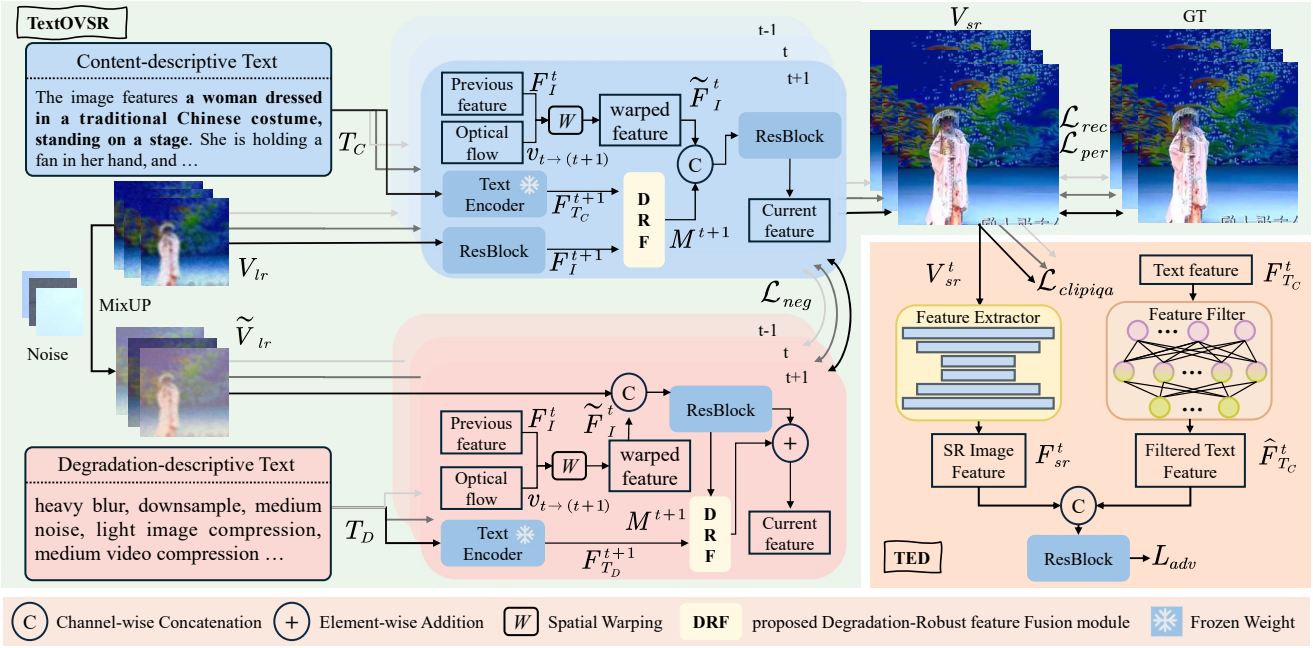


Figure 3. **The proposed TextOVSR network and TED adopt a two-stage training scheme.** In the first stage, only TextOVSR is trained. The positive branch (blue) takes content-descriptive text (T_C) and degraded videos (V_{lr}) as input, while the negative branch (red) takes degradation-descriptive text (T_D) and mixed-noise videos ($\widetilde{V}_{lr}$). Text features are extracted using a frozen CLIP encoder and fused with image features through the proposed DRF module. In the second stage, TextOVSR serves as the generator and TED as the discriminator. Adversarial training refines texture realism by selecting reliable textual features and integrating them with reconstructed image features. Here, $t-1$, t, and $t+1$ denote three consecutive frames, with detailed propagation described in Section 3.2.

of each frame, providing high-level guidance for super-resolution. To improve efficiency and facilitate video-level degradation modeling, text is shared across consecutive frames in batches, with a batch size of seven to match the format of the original dataset. During inference, only the positive branch is used. The inputs consist of the degraded video and its frame-wise content-descriptive text, while degradation-descriptive text is no longer required. The MLLM is used to generate content text for each test frame. The intensity levels of different degradation kernels and the MLLM prompts are provided in Appendix Section 1.

3.2. TextOVSR

The proposed Text-guided Dual-Branch Opera Video Super-Resolution (TextOVSR) network comprises two branches, as illustrated in Figure 3. The positive branch (blue) takes content-descriptive text and degraded low-resolution videos as input, producing the super-resolution output. The negative branch (red) takes degradation-descriptive text and low-resolution videos with noise as input. During training, outputs from both branches are used to compute the negative loss ($\mathcal{L}_{neg}$), enhancing the positive branch's robustness to real-world noise. During inference, only the positive branch is employed. Both branches are built on the BasicVSR architecture [4], with text features extracted via the CLIP text encoder [27] and fused with image features using the Degradation-Robust feature Fusion (DRF) module. The key distinction lies in the timing of image-text fusion. In the positive branch, text features are fused early, before deep feature extraction, enhancing frame feature expressiveness and mitigating error propagation. In the negative branch, text features are fused after deep feature extraction, allowing degradation descriptions to model real-world noise at the feature level. Ablation results on the fusion position are provided in Section 5.4. TextOVSR is trained in two stages. In the first stage, TextOVSR is trained alone. In the second stage, the trained model serves as a generator, and the Text-Enhanced Discriminator (TED) is introduced to further improve reconstruction quality. Detailed training procedures are described in Section 4.2.

Both the positive and negative branches adopt a bidirectional propagation mechanism. Taking the forward propagation from time step t to $t+1$ in the positive branch as an example, the process is as follows. The degraded video $V_{lr} \in \mathbb{R}^{n\times c\times h\times w}$ is first processed by a residual module to extract the feature of the frame at time step $t+1$, denoted as F_I^{t+1}. The pre-trained CLIP text encoder then encodes the content description T_C of each frame into text feature $F_{T_C}^{t+1}$, which is fused with the corresponding frame feature via the DRF module to obtain the fused feature M^{t+1}. Meanwhile,

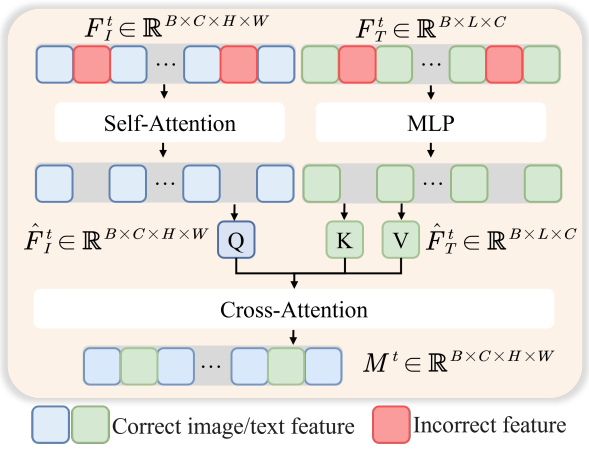


Figure 4. **The proposed Degradation-Robust Feature Fusion (DRF) module.** F_I^{t+1} and F_T^{t+1} denote the image and text feature vectors, respectively, and M^t represents the fused feature.

the propagated feature F_I^t and optical flow $v_{t\to t+1}$ are used to generate the aligned feature $\widetilde{F}_I^t$ through spatial warping. Finally, the fused feature and the aligned temporal feature are concatenated along the channel dimension, and a residual module produces the final feature for time step $t+1$.

3.3. Degradation-Robust feature Fusion Module

Input features from both positive and negative branches may contain unreliable information. In the positive branch, low-resolution inputs degraded through various processes inherently include distortions and erroneous features. Direct reliance on these features amplifies errors during temporal propagation, leading to blurred details and artifacts in reconstruction results [7, 42]. To better simulate real-world degradation, noise extracted from external datasets is introduced in the negative branch. However, such noise often exhibits style mismatches. For example, noise from everyday videos may not match the stylistic characteristics of opera videos, resulting in abnormal artifacts, as as indicated by the yellow arrows in Figure 1. Furthermore, content descriptions generated by the MLLM may also contain inaccuracies, introducing an additional source of uncertainty.

To address inaccuracies in the input features, we design a Degradation-Robust Feature Fusion (DRF) module for image-text fusion, as illustrated in Figure 4. The extracted frame features F_I^{t+1} and text features F_T^{t+1} are first processed by multi-head self-attention and linear layers, respectively, to amplify reliable information while suppressing noise and erroneous features. The filtered image feature $\hat{F}_I^t$ is then used to generate the Query (Q), and the filtered text feature $\hat{F}_T^t$ is used to generate the Key (K) and Value (V). Finally, a multi-head cross-attention mechanism computes the fused feature M^{t+1}.

3.4. Text-Enhanced Discriminator

The content-descriptive text contains high-level semantic information, which can enhance discrimination accuracy [32]. To leverage this, we introduce textual features into a standard UNet-based discriminator, forming the Text-Enhanced Discriminator (TED), as illustrated in Figure 3. Specifically, the inputs include the super-resolved frame V_{sr}^t and corresponding content description text feature $F_{T_C}^t$. A standard UNet extracts the image feature F_{sr}^t, while a feature filter emphasizes the effective textual features $\hat{F}_{T_C}^t$. The extracted image feature and filtered text feature are then concatenated along the channel dimension and processed by a residual module to compute the adversarial loss. This design fully exploits the high-level semantics from the content descriptions while mitigating the influence of inaccurate features, leading to more precise discrimination.

3.5. Objective Functions

To comprehensively train the proposed TextOVSR network, we adopt a two-stage training strategy following RealBasicVSR [7] and NegVSR [31]. In the first stage, TextOVSR is trained with the reconstruction loss ($\mathcal{L}_{rec}$) and the negative loss ($\mathcal{L}_{neg}$), formulated as:

$$\mathcal{L}_{stage1} = \mathcal{L}_{rec}\left(V_{sr}^t, V_{GT}^t\right) + \alpha\mathcal{L}_{neg}\left(V_{sr}^t, \hat{V}_{sr}^t\right), \quad (1)$$

where V_{sr}^t and $\hat{V}_{sr}^t$ denote the outputs of the positive and negative branches, respectively. The loss weight α for $\mathcal{L}_{neg}$ is set to 0.5 following previous work [31].

In the second stage, TextOVSR serves as the generator, while the proposed TED acts as the discriminator. In addition to the reconstruction and negative losses from the first stage, the generator is further optimized with a perceptual loss ($\mathcal{L}_{per}$) [15] and a CLIPIQA loss [34] to enhance perceptual quality. The CLIPIQA loss is defined as:

$$\mathcal{L}_{clipiqa} = 1 - \mathcal{R}\left(V_{sr}^t\right), \quad (2)$$

where $\mathcal{R}$ denotes the CLIP-IQA model. To further improve detail recovery, the adversarial loss is computed using TED:

$$\mathcal{L}_{adv} = -\mathbb{E}_{\left(\mathcal{H}\left(V_{lr}^t, T_C^t\right)\sim P_g\right)} log\left(TED\left(V_{sr}^t, F_{T_C}^t\right)\right), \quad (3)$$

where TED denotes the proposed text-enhanced discriminator, and $F_{T_C}^t$ represents the text features.The overall objective for stage two is:

$$\mathcal{L}_{stage2} = \mathcal{L}_{stage1} + \mathcal{L}_{per}\left(V_{sr}^t, V_{GT}^t\right) + \beta\mathcal{L}_{clipiqa} + \mathcal{L}_{adv}, \quad (4)$$

where the hyperparameter β is used to adjust the weight of the $\mathcal{L}_{clipiqa}$, which is set to 0.5 during training.

	Bicubic	DBVSR[25]	RealBasicVSR[7]	FTVSR[26]	Self-BlindVSR[1]	RealViformer[49]	NegVSR[31]	BVSR-IK[51]	Ours
Params(M)	-	36.3	4.9	45.8	139.5	8.5	3.4	6.5	5.7
Runtimes(ms)	-	411.4	81.2	825.4	985.6	87.9	63.8	145.3	195.3
FLOPs(G)	-	1159.2	376.9	2417.7	2518.0	213.0	292.0	317.4	309.6
NRQM↑	2.9193	3.9296	5.1708	3.0731	3.0474	5.1894	5.7761	3.6682	**5.8184**
MUSIQ↑	27.0694	38.8246	48.3548	28.5172	31.6829	52.3948	**58.6386**	39.6988	58.3033
CLIPIQA+↑	0.4212	0.3529	0.3494	0.2917	0.3147	0.3774	0.3990	0.3855	**0.5667**
TOPIQ↑	0.2125	0.2625	0.3556	0.2095	0.2134	0.3669	0.4354	0.2607	**0.4636**
BRISQUE↓	61.3265	56.2738	41.7475	53.7206	61.8668	39.3883	33.5291	57.9673	**33.3799**
NIQE↓	7.5734	6.0774	4.4300	7.0509	6.9784	4.4347	4.0756	6.8368	**3.5139**
ILNIQE↓	37.9113	29.2702	31.1341	35.4652	34.0444	32.7323	32.4800	33.7224	**30.0242**
PI↓	7.3350	6.2133	4.7243	7.0915	7.0388	4.7117	4.2232	6.6727	**3.9913**
DOVER↑	8.6303	15.9180	33.4799	9.5426	11.8134	39.4318	40.6763	16.7777	**45.0415**

Table 1. **Quantitative comparison with existing methods.** Best and second-best results are highlighted. Model parameters (Params), runtime, and FLOPs are evaluated on the same device. ↑ denotes higher values are better; ↓ denotes lower values are better.

Figure 5. **OperaLQ Dataset.** Our OperaLQ dataset consists of real degraded opera videos with varying content and resolutions.

4. Experiments

4.1. Datasets and Metrics

Dataset. For training, we use the Chinese Opera Video Clips (COVC) dataset introduced in MambaOVSR [8]. Each 7-frame clip is split into individual frames for reconstruction. Unlike RealBasicVSR [7], which applies online degradations, we pre-generate degraded inputs using the RealESRGAN [36] pipeline to ensure consistent degradations across epochs and accurate alignment with the degradation-descriptive text for each frame. As described in Section 3.1, content-descriptive texts are produced by an MLLM, while degradation-descriptive texts are obtained through a binning strategy. Detailed parameter settings and classification criteria are provided in the appendix. During preprocessing, the ground-truth (GT) frames and their degraded counterparts are randomly cropped to 256×256. The degraded frames are then downsampled to 64×64 using bicubic interpolation. Random horizontal flipping is applied for augmentation. For evaluation, we construct a real-world benchmark named OperaLQ, consisting of 50 opera videos with 100 frames each. As shown in Figure 5, the videos are collected from diverse sources to cover a broad range of real-world degradations, including various motion patterns, resolutions, and scene complexities.

Metric. Since the test set contains real-world degraded videos without ground-truth references, reference-based metrics are strictly inapplicable. To comprehensively and more robustly evaluate overall real-world video super-resolution performance, we employ a suite of reference-free image and video quality metrics. For image-level evaluation, we adopt NRQM [22], MUSIQ [16], CLIPIQA+ [34], TOPIQ [10], BRISQUE [23], NIQE [24], ILNIQE [48], and PI [2] to measure perceptual quality and naturalness. For video-level evaluation, we use DOVER [38] to assess overall temporal and perceptual consistency.

4.2. Implementation Details

We use the pre-trained SPyNet [28] to estimate optical flow between adjacent frames, with its weights frozen during training. The content-descriptive texts are generated by LLaVA [21], and their textual embeddings are extracted using the CLIP text encoder [27] with a ViT-L/14@336px backbone. The training process is divided into two stages. In the first stage, the TextOVSR model is trained for 100K iterations using the Adam optimizer with a learning rate of 1×10^{-4} and the loss function $\mathcal{L}_{stage1}$. In the second stage, the model is fine-tuned within a GAN framework to enhance fine details and perceptual realism, using a reduced learning rate of 5×10^{-5} and the loss function $\mathcal{L}_{stage2}$.

4.3. Comparison with State-of-the-Arts

We compare the proposed method with all publicly reproducible real-world video super-resolution (VSR) approaches, including Self-BlindVSR [1], NegVSR [31], Re-

Figure 6. **Qualitative comparison with other methods.** (Zoom-in for best view.)

Variants	baseline	DRF				TED	NRQM↑	CLIPIQA+↑	TOPIQ↑	NIQE↓	BRISQUE↓
		w/o T, N	w/o T	w T_D	w T_D&T_C						
1	✓	✗	✗	✗	✗	✗	5.7761	0.3990	0.4354	4.0756	33.5291
2	✓	✓	✗	✗	✗	✗	5.4949	0.5462	0.4436	3.7303	38.3618
3	✓	✗	✗	✓	✗	✗	5.4610	0.5471	0.4483	3.6396	39.3594
4	✓	✗	✓	✗	✗	✗	5.5697	0.5507	0.4523	3.6627	38.1570
5	✓	✗	✗	✗	✓	✗	5.6838	**0.5667**	0.4636	3.5139	35.7444
6(Ours)	✓	✗	✗	✗	✓	✓	**5.8184**	0.5659	**0.4733**	**3.4291**	**33.3799**

Table 2. **Comparison of quantitative results with different components of our method on the OperaLQ dataset.** w/o T, N denotes enhancing only the negative branch propagation, while w/o T denotes enhancing both positive and negative branches simultaneously. w T_D and w T_D&T_C denote using only degradation-descriptive text and using both degradation-descriptive and content-descriptive texts.

Figure 7. **Qualitative comparison of different component ablation studies.** (Zoom-in for best view.)

alViformer [49], FTVSR [26], RealBasicVSR [7], and DBVSR [25]. In addition, we evaluate against BVSR-IK [51], a state-of-the-art framework that jointly performs video super-resolution and deblurring.

The quantitative results on the OperaLQ dataset are shown in Table 1. Compared with existing methods, TextOVSR achieves superior performance. Specifically, it attains the best results on multiple image quality metrics (e.g., CLIPIQA+, TOPIQ, NIQE) and achieves the highest score on the video metric DOVER. Qualitative results in Figure 6 show that TextOVSR better suppresses blur and restores fine details, especially in text and facial regions, yielding clearer and more faithful visual results than other methods.

5. Analysis and Discussions

5.1. Ablation Study

We conducted ablation studies on OperaLQ to evaluate each component. Using NegVSR [31] as the baseline (Variant 1), we added the DRF module to enhance dual-branch feature propagation with textual information. Four variants were then created to analyze individual contributions: enhancing only the negative branch, fusing only degradation text, enhancing both branches, and fusing both texts (Variants 2–5). Adding TED to Variant 5 yielded Variant 6.

Quantitative and qualitative results are shown in Table 2 and Figure 7. Variant 2 enhances the negative branch via the DRF module, effectively suppressing artifacts caused by style-inconsistent noise in Variant 1 (as indicated by arrows) and improving CLIPIQA+ by 0.1472. Variant 3 further incorporates degradation-descriptive text, yielding additional improvements in CLIPIQA+. Variant 4 strengthens both positive and negative branches, reducing erroneous features in the positive branch and producing sharper contours. Variant 5 integrates degradation- and content-descriptive texts, significantly enhancing texture representation. Finally, the introduction of TED to guide texture reconstruction further boosts overall reconstruction quality.

	NRQM↑	CLIPIQA+↑	TOPIQ↑	NIQE↓
Caption	5.5552	**0.5718**	0.4602	3.6206
Text	**5.6838**	0.5667	**0.4636**	**3.5139**

Table 3. **Quantitative comparison of coarse-grained and fine-grained textual descriptions**, where Caption denotes coarse-grained textual descriptions and Text denotes fine-grained ones.

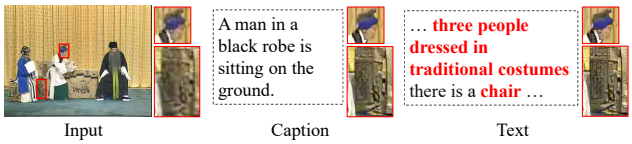


Figure 8. **Qualitative comparison of coarse-grained and fine-grained textual descriptions.** (Zoom-in for best view.)

	NRQM↑	CLIPIQA+↑	TOPIQ↑	NIQE↓
CLIP	5.0118	0.3081	0.2860	5.3520
UNet	5.6838	**0.5667**	0.4636	3.5139
TED	**5.8184**	0.5659	**0.4733**	**3.4291**

Table 4. **Quantitative comparison of different discriminators.**

5.2. Impacts of Coarse-Grained and Fine-Grained Content-descriptive text

To study the effect of textual granularity on super-resolution, we generated two types of content descriptions for the OperaLQ test set: Caption (coarse-grained) and Text (fine-grained). Quantitative results (Table 3) show that fine-grained descriptions improve overall reconstruction with only a minor CLIPIQA+ drop (0.0051). Qualitative results (Figure 8) indicate that fine-grained texts better guide recovery of fine structures, producing clearer and more realistic textures (e.g., facial and chair regions).

5.3. Impacts of Different Discriminators

Table 4 and Figure 9 present quantitative and qualitative comparisons of different discriminator architectures. The UNet-based discriminator (UNet) remains the most commonly adopted architecture in previous works [7, 31]. GALIP [32] directly utilizes CLIP's image and text encoders for feature extraction and alignment (CLIP); however, its unfiltered alignment often leads to ambiguous reconstructions, particularly in fine-grained regions such as faces and sleeves. In contrast, our proposed TED employs a UNet for image feature extraction while selectively filtering textual features, producing more accurate and sharper reconstructions and delivering the best overall performance.

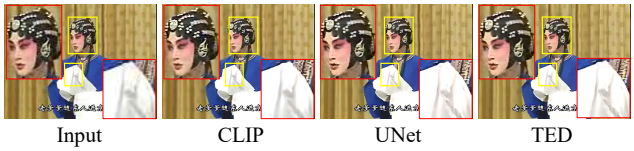


Figure 9. **Qualitative comparison of different Discriminators.** (Zoom-in for best view.)

Location	NRQM↑	CLIPIQA+↑	TOPIQ↑	NIQE↓
without	**5.7761**	0.3990	0.4354	4.0753
before	5.4280	0.5438	0.4316	**3.7274**
after	5.4949	**0.5462**	**0.4436**	3.7303

Table 5. **Quantitative comparison of DRF at different positions in the negative branch** (before and after image feature extraction). without indicates no DRF is added.

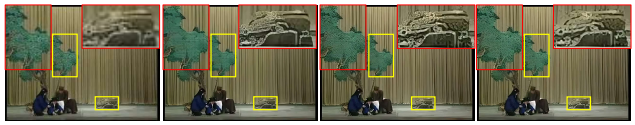


Figure 10. **Qualitative comparison of DRF at different positions in the negative branch.** (Zoom-in for best view.)

5.4. Impacts of DRF Module Location

In the positive branch, DRF is placed before image feature extraction to fuse image and text features effectively. In the negative branch, we test DRF placement before or after feature extraction. Qualitative results (Figure 10) show that pre-extraction fusion recovers richer details but adds noise, while post-extraction fusion suppresses out-of-distribution noise and improves quantitative metrics (Table 5).

6. Conclusion

In this work, we propose a novel Text-guided Dual-Branch Opera Video Super-Resolution (TextOVSR) network for real-world opera video super-resolution. Specifically, the negative branch of TextOVSR integrates degradation-descriptive text derived from the degradation process to broaden the degradation space. Simultaneously, to enhance texture reconstruction, content-descriptive text is incorporated into both the positive branch and the proposed TED. Furthermore, we design a DRF module that alleviates degradation-induced contamination while enabling effective cross-modal fusion of visual and textual features. On our constructed real-world benchmark, OperaLQ, TextOVSR outperforms existing state-of-the-art methods in both quantitative and qualitative evaluations.

7. Acknowledgments

This work was supported by the Natural Science Foundation of China (62376201, and 62501189), Hubei Provincial Science & Technology Talent Enterprise Services Program (2025DJB059), Hubei Provincial Special Fund for Central-Guided Local S&T Development (2025CSA017), and the Natural Science Foundation of Heilongjiang Province of China for Excellent Youth Project (YQ2024F006).

References

[1] Haoran Bai and Jinshan Pan. Self-supervised deep blind video super-resolution. *IEEE TPAMI*, 46(7):4641–4653, 2024. 1, 6, 7

[2] Yochai Blau, Roey Mechrez, Radu Timofte, Tomer Michaeli, and Lihi Zelnik-Manor. The 2018 pirm challenge on perceptual image super-resolution. In *ECCVW*, pages 0–0, 2018. 6

[3] Jiezhang Cao, Yawei Li, Kai Zhang, and Luc Van Gool. Video super-resolution transformer. *arXiv preprint arXiv:2106.06847*, 2021. 2

[4] 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*, pages 4947–4956, 2021. 2, 4

[5] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Understanding deformable alignment in video super-resolution. In *AAAI*, pages 973–981, 2021. 2

[6] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *CVPR*, pages 5972–5981, 2022. 1, 2

[7] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Investigating tradeoffs in real-world video super-resolution. In *CVPR*, pages 5962–5971, 2022. 1, 2, 3, 5, 6, 7, 8

[8] Hua Chang, Xin Xu, Wei Liu, Wei Wang, Xin Yuan, and Kui Jiang. Mambaovsr: Multiscale fusion with global motion modeling for chinese opera video super-resolution. *arXiv preprint arXiv:2511.06172*, 2025. 6

[9] Bin Chen, Gehui Li, Rongyuan Wu, Xindong Zhang, Jie Chen, Jian Zhang, and Lei Zhang. Adversarial diffusion compression for real-world image super-resolution. In *CVPR*, pages 28208–28220, 2025. 3

[10] Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. Topiq: A top-down approach from semantics to distortions for image quality assessment. *IEEE TIP*, 33:2404–2418, 2024. 6

[11] Zheng Chen, Yulun Zhang, Jinjin Gu, Xin Yuan, Linghe Kong, Guihai Chen, and Xiaokang Yang. Image super-resolution with text prompt diffusion. *arXiv preprint arXiv:2311.14282*, 2023. 3

[12] Linwei Dong, Qingnan Fan, Yihong Guo, Zhonghao Wang, Qi Zhang, Jinwei Chen, Yawei Luo, and Changqing Zou. Tsd-sr: One-step diffusion with target score distillation for real-world image super-resolution. In *CVPR*, pages 23174–23184, 2025. 3

[13] Bingwen Hu, Heng Liu, Zhedong Zheng, and Ping Liu. Clip-sr: Collaborative linguistic and image processing for super-resolution. *IEEE TMM*, 2025. 3

[14] 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*, pages 3224–3232, 2018. 2

[15] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711. Springer, 2016. 5

[16] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *ICCV*, pages 5148–5157, 2021. 6

[17] Tae Hyun Kim, Mehdi SM Sajjadi, Michael Hirsch, and Bernhard Scholkopf. Spatio-temporal transformer network for video restoration. In *ECCV*, pages 106–122, 2018. 2

[18] Xiaohui Li, Yihao Liu, Shuo Cao, Ziyan Chen, Shaobin Zhuang, Xiangyu Chen, Yinan He, Yi Wang, and Yu Qiao. Diffvsr: Enhancing real-world video super-resolution with diffusion models for advanced visual quality and temporal consistency. *arXiv e-prints*, pages arXiv–2501, 2025. 3

[19] Jingyun Liang, Yuchen Fan, Xiaoyu Xiang, Rakesh Ranjan, Eddy Ilg, Simon Green, Jiezhang Cao, Kai Zhang, Radu Timofte, and Luc V Gool. Recurrent video restoration transformer with guided deformable attention. *NeurIPS*, 35:378–393, 2022. 2

[20] Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. Vrt: A video restoration transformer. *IEEE TIP*, 33:2171–2182, 2024. 1, 2

[21] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 36:34892–34916, 2023. 6

[22] Chao Ma, Chih-Yuan Yang, Xiaokang Yang, and Ming-Hsuan Yang. Learning a no-reference quality metric for single-image super-resolution. *Computer Vision and Image Understanding*, 158:1–16, 2017. 6

[23] Anish Mittal, Anush K Moorthy, and Alan C Bovik. Blind/referenceless image spatial quality evaluator. In *2011 conference record of the forty fifth asilomar conference on signals, systems and computers (ASILOMAR)*, pages 723–727. IEEE, 2011. 6

[24] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Sign. Process. Letters*, 20(3):209–212, 2012. 6

[25] Jinshan Pan, Haoran Bai, Jiangxin Dong, Jiawei Zhang, and Jinhui Tang. Deep blind video super-resolution. In *ICCV*, pages 4811–4820, 2021. 1, 3, 6, 7

[26] Zhongwei Qiu, Huan Yang, Jianlong Fu, Daochang Liu, Chang Xu, and Dongmei Fu. Learning degradation-robust spatiotemporal frequency-transformer for video super-resolution. *IEEE TPAMI*, 45(12):14888–14904, 2023. 6, 7

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 2, 4, 6

[28] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *CVPR*, pages 4161–4170, 2017. 6

[29] Mehdi SM Sajjadi, Raviteja Vemulapalli, and Matthew Brown. Frame-recurrent video super-resolution. In *CVPR*, pages 6626–6634, 2018. 2

[30] Shuwei Shi, Jinjin Gu, Liangbin Xie, Xintao Wang, Yujiu Yang, and Chao Dong. Rethinking alignment in video super-resolution transformers. *NeurIPS*, 35:36081–36093, 2022. 2

[31] Yexing Song, Meilin Wang, Zhijing Yang, Xiaoyu Xian, and Yukai Shi. Negvsr: Augmenting negatives for generalized noise modeling in real-world video super-resolution. In *AAAI*, pages 10705–10713, 2024. 1, 2, 3, 5, 6, 7, 8

[32] Ming Tao, Bing-Kun Bao, Hao Tang, and Changsheng Xu. Galip: Generative adversarial clips for text-to-image synthesis. In *CVPR*, pages 14214–14223, 2023. 5, 8

[33] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. Tdan: Temporally-deformable alignment network for video super-resolution. In *CVPR*, pages 3360–3369, 2020. 2

[34] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI*, pages 2555–2563, 2023. 5, 6

[35] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *CVPRW*, pages 0–0, 2019. 2

[36] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *ICCV*, pages 1905–1914, 2021. 1, 2, 3, 6

[37] Hongyang Wei, Shuaizheng Liu, Chun Yuan, and Lei Zhang. Perceive, understand and restore: Real-world image super-resolution with autoregressive multimodal generative models. *arXiv preprint arXiv:2503.11073*, 2025. 3

[38] 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*, pages 20144–20154, 2023. 6

[39] Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. *NeurIPS*, 37:92529–92553, 2024. 3

[40] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *CVPR*, pages 25456–25467, 2024. 3

[41] Yanze Wu, Xintao Wang, Gen Li, and Ying Shan. Animesr: Learning real-world super-resolution models for animation videos. *NeurIPS*, 35:11241–11252, 2022. 3

[42] Liangbin Xie, Xintao Wang, Shuwei Shi, Jinjin Gu, Chao Dong, and Ying Shan. Mitigating artifacts in real-world video super-resolution models. In *AAAI*, pages 2956–2964, 2023. 3, 5

[43] Rui Xie, Yinhong Liu, Penghao Zhou, Chen Zhao, Jun Zhou, Kai Zhang, Zhenyu Zhang, Jian Yang, Zhenheng Yang, and Ying Tai. Star: Spatial-temporal augmentation with text-to-video models for real-world video super-resolution. *arXiv preprint arXiv:2501.02976*, 2025. 3

[44] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *IJCV*, 127(8):1106–1125, 2019. 2

[45] Tao Yang, Rongyuan Wu, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. In *ECCV*, pages 74–91. Springer, 2024. 3

[46] 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*, pages 4781–4790, 2021. 3

[47] Xi Yang, Chenhang He, Jianqi Ma, and Lei Zhang. Motion-guided latent diffusion for temporally consistent real-world video super-resolution. In *ECCV*, pages 224–242. Springer, 2024. 3

[48] Lin Zhang, Lei Zhang, and Alan C Bovik. A feature-enriched completely blind image quality evaluator. *IEEE TIP*, 24(8): 2579–2591, 2015. 6

[49] Yuehan Zhang and Angela Yao. Realviformer: Investigating attention for real-world video super-resolution. In *ECCV*, pages 412–428. Springer, 2024. 3, 6, 7

[50] 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*, pages 2535–2545, 2024. 3

[51] Qiang Zhu, Yuxuan Jiang, Shuyuan Zhu, Fan Zhang, David Bull, and Bing Zeng. Blind video super-resolution based on implicit kernels. *arXiv preprint arXiv:2503.07856*, 2025. 6, 7