UNIMMVSR: A UNIFIED MULTI-MODAL FRAMEWORK FOR CASCADED VIDEO SUPER-RESOLUTION

Shian Du[1*], **Menghan Xia**[2†], **Chang Liu**[1], **Quande Liu**[3], **Xintao Wang**[3], **Pengfei Wan**[3], **Xiangyang Ji**[1†]
[1]Tsinghua University, [2]Huazhong University of Science and Technology,
[3]Kling Team, Kuaishou Technology
https://shiandu.github.io/UniMMVSR-website/

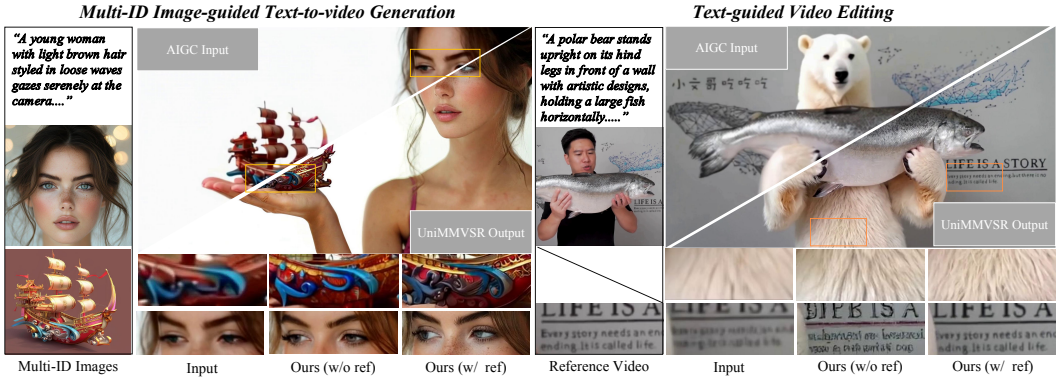

Figure 1: **UniMMVSR is a unified framework that supports video super-resolution with multi-modal input conditions.** By cooperating with the low-resolution multi-modal generative model, the proposed cascaded framework can effectively extend the controllable video generation to ultra-high-resolution (e.g., 4K) with high visual quality and subject consistency.

ABSTRACT

Cascaded video super-resolution has emerged as a promising technique for decoupling the computational burden associated with generating high-resolution videos using large foundation models. Existing studies, however, are largely confined to text-to-video tasks and fail to leverage additional generative conditions beyond text, which are crucial for ensuring fidelity in multi-modal video generation. We address this limitation by presenting UniMMVSR, the first unified generative video super-resolution framework to incorporate hybrid-modal conditions, including text, images, and videos. We conduct a comprehensive exploration of condition injection strategies, training schemes, and data mixture techniques within a latent video diffusion model. A key challenge was designing distinct data construction and condition utilization methods to enable the model to precisely utilize all condition types, given their varied correlations with the target video. Our experiments demonstrate that UniMMVSR significantly outperforms existing methods, producing videos with superior detail and a higher degree of conformity to multi-modal conditions. We also validate the feasibility of combining UniMMVSR with a base model to achieve multi-modal guided generation of 4K videos—a feat previously unattainable with existing techniques.

1 INTRODUCTION

Video generation foundation models (Seed, 2025; Wan et al., 2025; Hu et al., 2025) have made remarkable progress in synthesizing realistic videos, largely due to the scaling law of diffusion transformer architectures (Peebles & Xie, 2023b). Unfortunately, expanding model capacity typically

∗ This work was conducted during the author's internship at Kling Team, Kuaishou Technology.
† Corresponding author.

incurs a significant computational burden, a challenge particularly pronounced for high-resolution video generation (e.g., 2K, 4K, 8K), a growing trend in future applications. To resolve this dilemma, a stage-wise cascading paradigm, where a large-capacity base model generates a low-resolution video and subsequent lightweight super-resolution models synthesize the fine details, has emerged as a promising solution. Anyhow, existing research (Ho et al., 2023; Zhang et al., 2025a; Seed, 2025) on cascaded video super-resolution is limited to text-to-video task. A significant gap remains in understanding how super-resolution models can effectively use hybrid conditions—a crucial capability for maintaining generative fidelity in videos produced by multi-modal base models.

In this paper, we present the first unified latent diffusion framework for multi-modal video super-resolution, dubbed UniMMVSR. We focus on three common video generation tasks: text-to-video generation, multi-ID image-guided text-to-video generation, and text-guided video editing. For these tasks, our super-resolution model uses not only the low-resolution video but also text, ID images, and other videos as conditions. The main challenge is to integrate these diverse conditions into a single framework and to modulate these reference information in a compatible manner. This is crucial to ensure the model to use all conditions effectively, allowing it to generate vivid details that conform to the multi-modal guidance.

To achieve this, we conducted a thorough study on multi-modal condition injection, with a special focus on incorporating multiple ID images and reference videos. Our comparative analysis shows that token concatenation performs best among the baselines. Recognizing that the low-resolution video from a base model might not perfectly align with the multi-modal conditions, we improved the robustness of UniMMVSR in two ways: (i) We assign independent position embedding for condition tokens that are distinct from the noisy target video tokens. This encourages the model to use the references based on context and correlation even though their contents are pixel-aligned. (ii) We developed a custom training data pipeline that simulates the generation characteristics of base models using the SDEdit technique (Meng et al., 2021).

Our experiments prove that UniMMVSR is superior to existing baselines, especially in its visual fidelity to multi-modal references. Our ablation studies further validate the effectiveness of our key designs, offering a clear view of the advantages of our method. We also show the benefits of our unified training framework: high-quality training data can transfer across sub-tasks, which reduces the burden of collecting high-quality data for complex-modal tasks.

Our contributions are summarized as follows:

- We introduce UniMMVSR, the first multi-modal guided generative video super-resolution model built on a cascaded framework. Our model synthesizes vivid details while maintaining high fidelity to conditional references.
- We developed a unique SDEdit-based degradation pipeline to create synthetic training data for multi-modal video super-resolution. It enhances the model's robustness to discrepancies between low-resolution video inputs and multi-modal conditions.
- Our UniMMVSR framework demonstrates the ability to leverage high-quality training data across multiple tasks and can easily scale to ultra-high-resolution generation (e.g., 4K) with efficient computational overhead.

2 Related Works

2.1 Multi-modal Video Generation

With the advancement of video generative model, recent work has increasingly focused on enhancing the controllability of generated videos. (Huang et al., 2025; Chen et al., 2025; Yuan et al., 2025; He et al., 2024c; Hu, 2024; Lei et al., 2025; Ma et al., 2024b; Wei et al., 2024; Zhang et al., 2025b) introduced reference images to improve subject consistency in the output video, while methods such as (Chen et al., 2024b; Tu et al., 2025; Mou et al., 2024a; Ye et al., 2025; Liew et al., 2023) achieved mask-based or instruction-based video editing by incorporating referenced videos. Although these approaches demonstrate promising results for specific controllable tasks, they fail to generalize across diverse tasks, hindering the broader application of controllable video generation.

To establish a unified framework for controllable generation tasks, previous work (Ding et al., 2022; Ju et al., 2023) introduced multiple adapter modules to independently incorporate different reference conditions. This approach yielded poor performance while resulting in significant waste of model parameters. Consequently, recent methods such as FullDiT (Ju et al., 2025; Tan et al., 2025) leverage in-context condition mechanisms to flexibly combine multi-modal input condition signals through self-attention module, achieving multi-task controllable video generation in a unified framework.

However, the computational complexity of the self-attention module increases quadratically with the number of tokens, hindering the scalability of such methods to more tasks and higher resolutions. While FullDiT2 (He et al., 2025) optimizes computational overhead for reference conditions through kv cache, block skipping, and token selection techniques, it remains inapplicable to unified frameworks or high-resolution scenarios. To address this limitation, we propose the first unified cascaded framework for high-resolution multi-modal video generation. This approach effectively achieves controllable high-resolution video generation while faithfully preserving multiple input conditions.

2.2 Video Super-resolution

Most previous works (Chan et al., 2022b; Cao et al., 2021; Chan et al., 2021; 2022a) primarily focus on synthetic or real-world data by designing compositional synthetic degradation pipelines to model the degraded videos. With the widespread applications of video generation models, later approaches shift towards AI-generated data. Due to limited generative capabilities, previous methods tend to generate over-smooth results. Motivated by recent advances in diffusion models, several diffusion-based video super-resolution (VSR) methods (Wang et al., 2023c; Zhou et al., 2024; Yang et al., 2024a; He et al., 2024a; Li et al., 2025; Wang et al., 2025b;a) have been proposed, which show impressive performance and generate realistic details.

However, limited by recent video super-resolution framework, existing models can only take text prompt and input video as conditions, which hinders their applications towards controllable video generation tasks. Although producing fine details, due to the randomness of diffusion sampling process, it inevitably reduces the fidelity of the input video to multi-modal references, which further diminishes the controllability of the generated results. In this paper, for the first time, we design a generative video super-resolution framework that unifies the input of hybrid-modal conditions, which improves the visual quality of the input video while ensuring its fidelity to multi-modal conditions.

3 Method

We aim to achieve generative video super-resolution for AI-generated videos under hybrid-modal conditions, which synthesizes rich, vivid details and maintains high fidelity to various conditional inputs. It works in the scenario that multi-modal base model first generates a low-resolution video, which our UniMMVSR model then upscales using the original high-resolution reference conditions if they're available. The overview flowchart is depicted in Fig. 2. Specifically, we focus on three common video generation tasks: *Text-to-video*, *Text-to-video guided by multiple ID images*, and *Text-guided video editing*. To accomplish this, our super-resolution model incorporates diverse inputs—including low-resolution video, text, multiple ID images, and reference videos—in a compatible manner. We have tackled this challenge by exploring a unified condition injection mechanism, a custom training data pipeline, and a tailored training strategy to ensure the model effectively utilizes all multi-modal conditions.

3.1 Preliminaries

Our model is built upon a pretrained large-scale text-to-video latent diffusion model. It first pretrains an autoencoder that converts a video x into a low-dimensional latent z with an encoder $\mathcal{E}$ and reconstructs it with a decoder $\mathcal{D}$. The core of this framework is a conditional diffusion transformer that operates in the compressed latent space. Detailed architecture is presented in Supp. A.1.

During training, given a LR-HR paired data (z_{HR}, z_{LR}) and multiple conditions C, isotropic gaussian noise is added to generate corresponding noise latent $z_t = (1-t)z_{HR} + t\epsilon$, where $\epsilon \in \mathcal{N}(0, I)$. With the formation of flow matching, it trains a network $\mu_\theta(z_t, t, C)$ to predict the

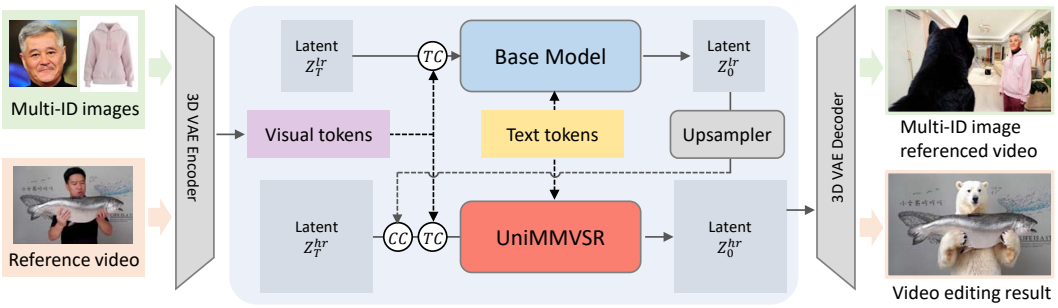

Figure 2: Overview of UniMMVSR in the context of a cascaded generation framework. Upsampler denotes the sequential operations of VAE decoding, upscaling via bilinear interpolation, and VAE encoding. *TC* and *CC* denote token concatenation and channel concatenation respectively. Texts are encoded by text encoder and then injected via cross-attention layers, which are omit for simplicity.

velocity $v = z_{HR} - \epsilon$. Then, the network μ_θ is optimized by minimizing the mean squared error loss $\mathcal{L}$ between the ground truth velocity and the model prediction:

$$\mathcal{L} = \mathbb{E}_{z_{HR},\epsilon,t,C}\|v - \mu_\theta(z_t, t, C)\|. \tag{1}$$

During inference, it first samples noise $\epsilon \in \mathcal{N}(0, I)$, then denoises it by a pre-defined ODE solver with a discrete set of N timesteps to generate clean latent z_{HR}. The final output x_{HR} is obtained by projecting z_{HR} to the pixel space using pre-trained decoder $\mathcal{D}$.

3.2 Unified Conditioning Framework

Our UniMMVSR model processes low-resolution input video with three types of reference conditions: text prompts, multiple ID images and reference videos. Since UniMMVSR is adapted from a pre-trained text-to-video model, it inherits the original text-prompt conditioning design. We mainly discuss the interaction strategy of conditional visual tokens and input video tokens.

Low-resolution video via channel concatenation. Since the low-resolution (LR) video and the target high-resolution (HR) video have pixel-aligned tempo-spatial correspondences, we use channel concatenation to directly incorporate the information of basic structure. To match their spatial sizes, we first upscale the LR video in pixel space and then encode it into a latent representation. This makes it ready for channel concatenation with the noisy HR latent[1]. During the inference phase, we first decode the LR latent generated by the base model. We then perform pixel interpolation to target resolution and finally encode it back into the latent space. This process ensures the LR latent has the same size as the noisy HR latent without corrupting the original information.

Visual references via token concatenation. We refer to multi-ID images and reference videos as visual references. Following the successful practice of recent in-context conditioning methods (Tan et al., 2025; Ju et al., 2025), we integrate the target video tokens with the conditional tokens using token concatenation. This strategy makes it feasible to incorporate general conditional modalities—whether they are spatially aligned or not—into the generation process, as they interact with the target video tokens in each layer's attention modules. Specifically, in each transformer block, noisy video tokens and visual references are processed in parallel through 2D self-attention, 2D cross-attention, and a feedforward network to preserve the alignment between the text and video modalities. For the 3D self-attention module, all tokens are treated as a single unified sequence and processed together, which ensures a bidirectional flow of information between the target video and visual reference tokens. Finally, all reference tokens are truncated from the transformer output to match the input shape.

Separated conditional RoPE. Before concatenation with the target video tokens, these conditional tokens are assigned with position encoding, where we adopt Rotary Position Embedding

[1]Note that, most VAE latent spaces do not support simple interpolation, which can cause significant structural distortion, as shown in (Xie et al., 2025a)

(RoPE) (Su et al., 2024). For multi-ID images, it is natural to assign an individual range of RoPE that are distinct from that of target video tokens, since no direct spatial correspondence exists between them. For reference video, since the LR video generated by our base model is not perfectly pixel-aligned with it, we also assign a separate range of RoPE for thest conditional tokens, so as to encourage our model to utilize it based on context and correlation rather than direct copy-and-paste. Specifically, we assign indices 0 to $f-1$ for noisy token, n_i to $n_i + k_i$ for i-th reference tokens, where f denotes the number of frames, n_i and k_i denotes the start index and length of i-th reference token.

3.3 Degradation Pipeline

Multi-modal conditional video generation model requires not only the vividness of generated content but also the conformity to provided conditions. Accordingly, our UniMMVSR model also needs to achieve high-quality details and maintains fidelity to the multi-modal conditions. Given a high-resolution video, it is of vital importance to design a degradation pipeline to process it into low-resolution video that simulates the degradation pattern of the base models' output. Specifically, the degradation pattern of multi-modal base model can be categorized into two scenarios: (i) At low resolution, high-frequency details in training data are lost due to resize operations. The base model can only generate basic structures that align with the semantic content of the text prompt, which lacks fine details and textures. (ii) For some challenging cases, the low-resolution output maintains a low fidelity to the visual references due to the sub-optimal controllability of the base model to harmonize text prompt and visual references. Therefore, the identity in LR video typically exhibits distortion in local structure and has low visual quality.

To tackle these two scenarios, a custom degradation pipeline needs to be designed to simulate the artifacts and distortions generated by the base model. However, traditional degradation pipelines (Wang et al., 2021; Chan et al., 2022b) constructed solely based on synthetic degradation factors (such as noise, blur, video compression, etc.) cannot be fully adapted to these scenarios since the resulting local structure of the LR video is strictly aligned with the HR video, failing to simulate the insufficient reference response in low-resolution output. To simulate the degradation scenario (ii), we recognized that **it is equivalent to the results obtained by base model using only text condition**. Therefore, based on the sdedit method (Meng et al., 2021), we constructed compatible high-frequency degradation features using inference result from the text-to-video base model, termed **SDEdit Degradation**.

Specifically, we downsample HR video to a resolution directly achievable by the text-to-video base model. The resized video x is encoded into latent space via a pre-trained 3D VAE encoder, and noise is added by the forward process of the diffusion model for k steps, where the step value k is randomly sampled from $[K_1, K_2]$ and K_2 denotes the maximum threshold to retain the main structure of the input video. Subsequently, we perform k steps of denoising process on the noisy latent using the base model, decoding the result via the 3D VAE decoder to obtain the LR video. After sdedit degradation, we apply synthetic degradation factors to the output x' to construct the final LR video. The degradation pipeline and samples are presented in Supp. A.4.

3.4 Training Strategy

Training Order. To train a unified model with hybrid conditions, the training order of subtasks is essential due to the varied difficulty. Instead of generating by text prompt only, multi-ID image-guided text-to-video generation and text-guided video editing tasks tend to synthesize high-fidelity textures and details by utilizing visual conditions and text prompt together, thus resulting in faster convergence speed than text-to-video generation task as shown in Fig. 12. Thus, we perform a difficult-to-easy training strategy, aiming to learn the difficult task first, and then effectively adapt to easier tasks.

Starting from a pre-trained text-to-video (T2V) model weight, we first train 21-frames text-to-video generation task independently in the first stage. In the second stage, we train 21-frames text-to-video generation and multi-ID image-guided text-to-video generation tasks together with a probability $0.6 : 0.4$, aiming to retain the ability to generate high-definition details from text. Next, all tasks are trained at 21 frames together with a probability $0.5 : 0.3 : 0.2$. Finally, we extend the frame length to 77 (5 seconds) while keeping the probability unchanged.

Table 1: Quantitative Evaluation of UniMMVSR on all three tasks. **Bold** and underlined indicate the best and second-best results, respectively. ↑ indicates higher is better; ↓ indicates lower is better.

Method	Visual Quality				Subject Consistency		Video Alignment		
	MUSIQ↑	CLIP-IQA↑	QAlign↑	DOVER↑	CLIP-I↑	DINO-I↑	PSNR↑	SSIM↑	LPIPS↓
Text-to-video Generation									
Base 512×512	30.996	0.246	3.741	0.594	-	-	-	-	-
Base 1080P	46.645	0.306	4.246	0.749	-	-	-	-	-
VEnhancer	**57.171**	0.367	4.214	0.733	-	-	-	-	-
STAR	56.904	0.369	4.435	0.769	-	-	-	-	-
SeedVR	55.596	**0.379**	4.396	**0.778**	-	-	-	-	-
Ours (single)	56.146	0.366	**4.535**	0.771	-	-	-	-	-
Ours (unified)	56.418	0.371	4.500	**0.778**	-	-	-	-	-
Text-guided Video Editing									
Base 512×512	35.073	0.234	3.615	0.400	-	-	30.191	0.699	0.364
Base 1080P	53.616	0.383	4.247	0.634	-	-	29.383	0.582	0.358
Ref Video	54.249	0.365	4.131	0.571	-	-	-	-	-
VEnhancer	57.036	0.380	4.013	0.590	-	-	28.417	0.571	0.489
STAR	56.802	0.397	4.264	0.608	-	-	29.421	0.631	0.397
SeedVR	57.820	0.370	4.183	0.635	-	-	29.535	0.597	0.413
Ours (no ref)	**59.119**	**0.399**	4.289	**0.648**	-	-	29.615	0.581	0.429
Ours (single)	53.388	0.348	4.302	0.597	-	-	**31.905**	**0.723**	**0.276**
Ours (unified)	53.245	0.344	**4.305**	0.597	-	-	31.556	0.713	0.282
Multi-ID Image-guided Text-to-video Generation									
Base 512×512	29.314	0.255	3.149	0.433	0.692	0.538	-	-	-
Base 1080P	46.780	0.345	4.092	0.662	0.691	0.507	-	-	-
VEnhancer	60.656	**0.469**	4.149	0.707	0.671	0.533	-	-	-
STAR	58.810	0.449	4.282	**0.763**	0.696	0.546	-	-	-
SeedVR	54.491	0.419	3.960	0.708	0.693	0.543	-	-	-
Ours (no ref)	60.947	0.445	4.385	0.742	0.693	0.543	-	-	-
Ours (single)	61.357	0.446	4.414	0.743	**0.728**	**0.566**	-	-	-
Ours (unified)	**62.248**	0.465	**4.428**	0.745	0.726	**0.566**	-	-	-

Reference Augmentation. For the multi-ID image-guided text-to-video generation task, most testing scenarios include cross-pair data, where the perspective, orientation, and position of the low-resolution output and ID images exhibit greater discrepancies than training datasets. For the text-guided video editing task, although the HR video and reference video are strictly pixel-aligned for non-editing area during training, the low-resolution output exhibits a certain degree of error compared with reference video. Since we only construct synthetic datasets for these two tasks (stated in Sec. A.2), directly utilizing synthetic reference conditions leads to a train-test gap, thereby compromising performance on the test set. To mitigate this issue, we design a reference augmentation technique to narrow this gap. Specifically, we apply several image-related transformations to simulate the cross-pair test scenarios of multi-ID image-guided text-to-video generation task. For the text-guided video editing task, we randomly shift the start frame of reference video, aiming to encourage the model to learn a more robust context-injection mechanism rather than directly copying the pixels from reference video.

4 Experiments

4.1 Experimental Settings

4.1.1 Implementation Details

Our model is trained on NVIDIA H800 GPUs with a total batch size of 32. AdamW (Loshchilov, 2017) is used as the optimizer with a learning rate of 10^{-4}. The text prompt is randomly replaced by a null prompt with 10% probability. To enhance the robustness of our model to different degradation scenarios, we utilize the noise augmentation technique by injecting noise into the input latent using a diffuse process. The noise timestep is randomly sampled from 200 to 600 to preserve the main structure. We have additionally encoded noise timestep as a micro condition for the model. We use a pretrained T2V model to provide initialization weight. During inference, we perform 50 PNDM (Liu et al., 2022b) sampling steps with independent classifier-free guidance as stated in Sec. A.3. The guidance scale s_{txt} and s_{ref} are set to 3.0 and 1.0 respectively, with reference guidance threshold $N_{ref} = 15$. We have also used timestep shift (Esser et al., 2024b) with shift value 1.0.

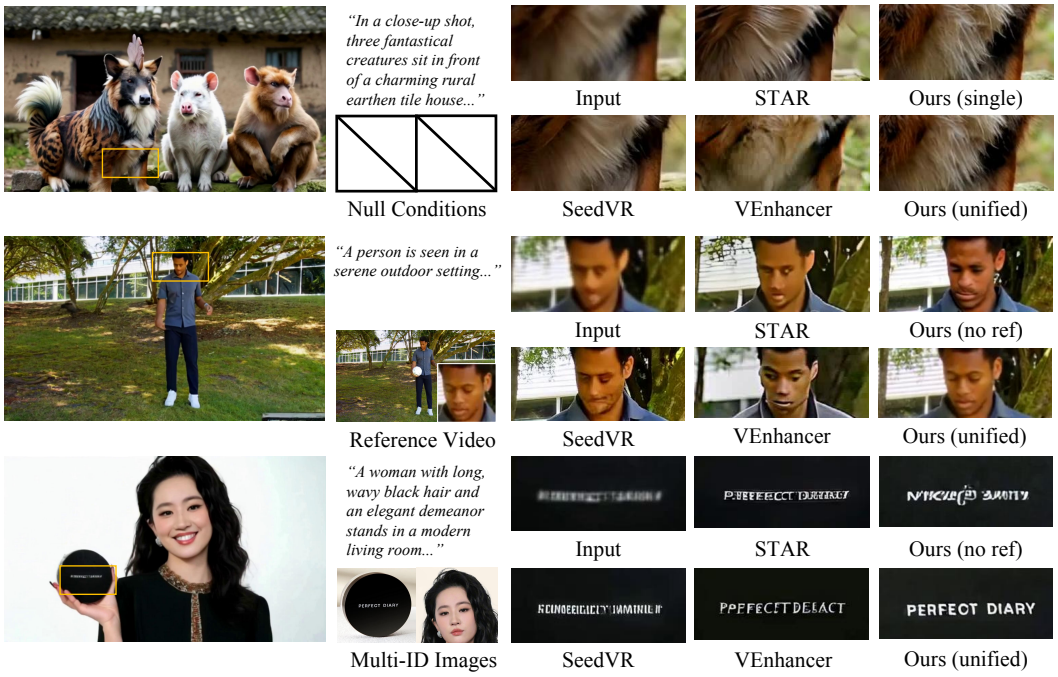

Figure 3: Qualitative comparisons on *text-to-video generation*, *text-guided video editing* and *multi-ID image-guided text-to-video generation* tasks from top to bottom. **(Zoom-in for best view)**

4.1.2 Testing Settings

Baseline Methods. To evaluate the effect of our cascaded framework, we compare with the end-to-end results of our base model (both 512×512 and 1080P). Since there is limited work on multi-modal VSR tasks, we also compare UniMMVSR with state-of-the-art VSR methods VEnhancer (He et al., 2024a), STAR (Xie et al., 2025b) and SeedVR (Wang et al., 2025b;a).

Evaluation Metrics. For visual quality, we conduct our evaluation of commonly used visual quality metrics MUSIQ (Ke et al., 2021), CLIP-IQA (Wang et al., 2023a), Q-Align (Wu et al., 2023b) and DOVER (Wu et al., 2023a). For multi-ID image-guided text-to-video generation task, we utilize DINO-I (Caron et al., 2021) and CLIP-I (Radford et al., 2021) to assess the fidelity to multiple ID images. We additionally use PSNR, SSIM and LPIPS (Zhang et al., 2018) to evaluate alignment of non-editing area with the reference video for text-guided video editing task.

4.2 Quantitative Comparison

Quantitative comparisons are shown in Tab. 1. Results show that although the UniMMVSR integrates multiple conditions, it still achieves state-of-the-art performance on controlling metrics compared with base model, previous VSR methods and our method without reference conditions, thereby validating the effectiveness of our method. For visual quality, UniMMVSR obtains the best QAlign&DOVER scores on text-to-video generation task and MUSIQ&QAlign scores on multi-ID image-guided text-to-video generation task, indicating its high perceptual quality. For text-guided video editing task, it is worth noting that our method maintains high pixel-level fidelity and structural similarity to the reference video for non-editing area, thus achieving similar metric values to the reference video. Even so, our approach remains competitive, achieving the best QAlign score. Furthermore, on multi-ID image-guided text-to-video generation task, our unified model exhibits high perceptual quality than our single-task model, indicating that complex-modal tasks can benefit from high-quality text-to-video data, effectively lowers the barrier to collect high-quality reference-video paired data. More comprehensive results can be seen in Supp. A.5.2.

4.3 Qualitative Comparison

Fig. 3 shows visual results on all three tasks. For text-to-video generation, both our single-task and unified model effectively remove existing degradation patterns and generate fine details like the dog's fur, while other approaches produce blurred details. For text-guided video editing and

Table 2: Ablation Study of UniMMVSR components on the multi-ID image-guided text-to-video generation Task. We analyze the impact of each component by visual quality and controlling metrics.

Ablation	Variant	MUSIQ↑	CLIP-IQA↑	QAlign↑	DOVER↑	CLIP-I↑	DINO-I↑
Ours	-	62.248	0.465	4.428	0.745	0.726	0.566
Architecture Design	full channel-concat	61.146	0.461	4.399	0.748	0.690	0.546
	full token-concat	61.974	0.464	4.442	0.739	0.728	0.565
Degradation Effect	synthetic degradation only	62.541	0.458	4.408	0.749	0.717	0.561
	sdedit degradation only	59.697	0.437	4.357	0.726	0.730	0.564
Training Order	full training	62.199	0.460	4.322	0.745	0.716	0.553
	easy-to-difficult	61.706	0.445	4.326	0.736	0.717	0.556

Figure 4: Visual Comparisons of single-task and unified model. **Zoom-in for best view.**

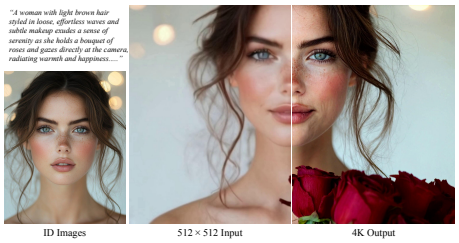

Figure 5: Qualitative results of 4K multi-ID image-guided text-to-video generation.

multi-ID image-guided text-to-video generation, UniMMVSR successfully leverages ID images and reference videos to generate high-fidelity textures and details, such as the facial structure of the man and the words "perfect diary" on the box. More results can be found in Supp. A.5.3.

4.4 Ablation Study

Due to space limit, we provide qualitative evaluation in Supp. A.5.4.

Architecture Design. In Tab. 2, we compare UniMMVSR with two architecture designs: full channel-concat and full token-concat. The former represents concatenating input video and reference tokens along channel dimension, while the latter represents concatenating along sequence dimension. As can be seen, full channel-concat method results in severe performance degradation in controlling metrics (0.690 vs 0.726 for CLIP-I and 0.546 vs 0.565 for DINO-I), which shows that it faces difficulties in reference injection. For full token-concat, while it achieves comparable performance, it results in nearly $2\times$ computational burden due to the quadratic computation complexity.

Degradation Effect. Since our degradation pipeline comprises both synthetic and sdedit degradations, we perform ablation study to investigate the effectiveness of each component. As can be seen in Tab. 2, although only using synthetic degradation obtains similar visual quality metrics, it leads to poorer quality in controlling metrics, which confirms that sdedit degradation successfully simulates the degradation scenario of base model results. Next, to validate the necessity of traditional synthetic degradation pipeline, we use sdedit degradation only to construct LR data. While it achieves comparable controlling metrics, it shows a decline in all visual quality metrics, demonstrating the effectiveness of synthetic degradation in detail synthesis.

Training Order. To form a unified model, we have implemented three different training strategies: difficult-to-easy, easy-to-difficult and full training. Specifically, difficult-to-easy means training in the order: text-to-video $\rightarrow$multi-ID image-guided$\rightarrow$video editing, whereas easy-to-difficult denotes the reverse order. Full training represents training all tasks together from the pre-trained T2V model weight. In Tab. 2, we demonstrate the validness of our proposed training strategy by comparing with other strategies above. By utilizing a difficult-to-easy training order, UniMMVSR successfully adapts to multiple tasks while maintaining the performance on previous tasks.

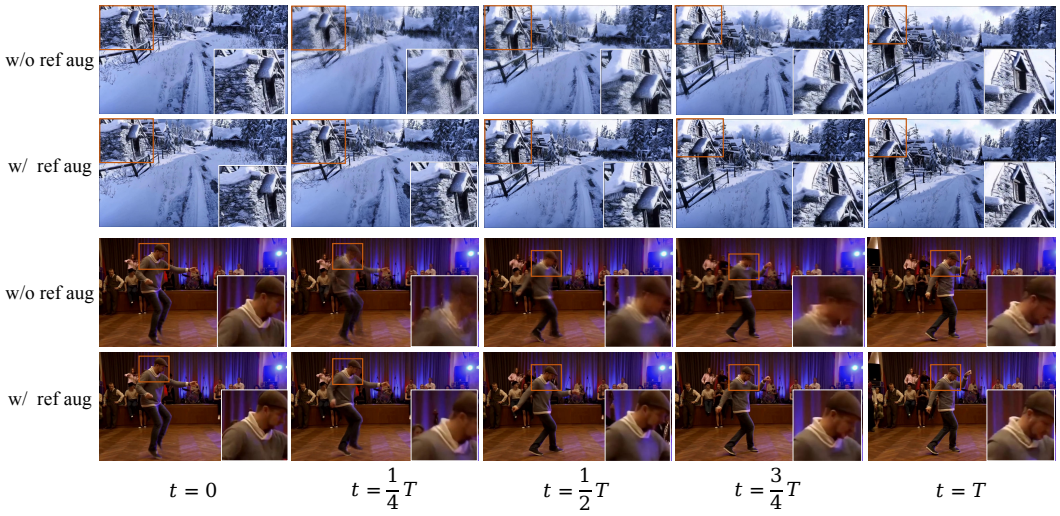

Figure 6: Visual Comparisons of reference augmentation. **Zoom-in for best view.**

4.5 Discussion

Importance of Reference Augmentation. Fig. 6 shows visual comparisons on text-guided video editing task. As can be seen in the second frame in the upper part and the middle three frames in the lower part, training without reference augmentation tend to produce unstable structure, which results in temporal jitter in some frames. After using reference augmentation, UniMMVSR learns to preserve the basic structure of the input video and avoids direct replication of the reference video, which mitigates the conflict of the basic structure between reference video and LR video.

High-quality Data Transfer across Sub-tasks. We perform an ablation study by training a single-task model on multi-ID image-guided text-to-video generation task without quality filtering. The model is compared with a unified model mix-trained on a high-quality text-to-video generation dataset. The results are shown in Fig. 4. As can be seen, the unified model generates more natural details such as teeth structure and facial expression, which demonstrates that the high resolution training data can transfer across sub-tasks.

Resolution Scaling Ability of Cascaded Model. Due to the scarcity of ultra-high-resolution (UHR) reference-video paired dataset and quadratic computational complexity, it is difficult to directly train a high-resolution controllable video generative model. By decoupling the process as low-resolution basic structure generation and high-frequency detail synthesis, UniMMVSR successfully generates 4K videos under multi-modal guidance. As shown in Fig. 5 and Supp. A.5.5, our method not only generates vivid details, but also preserves information in reference conditions.

5 Conclusion

In this paper, we introduce UniMMVSR, the first multi-modal guided generative video super-resolution model built on a cascaded framework. By treating the visual references as a unified sequence and processing them via the 3D self-attention module, UniMMVSR effectively synthesizes vivid details while maintaining high fidelity to conditional references. To enhance the model's robustness to discrepancies between low-resolution video inputs and multi-modal conditions, we develop a unique degradation pipeline based on sdedit method, which simulates the insufficient reference response in low-resolution output. Furthermore, we design a tailored training strategy to form a unified model, and demonstrate that high-quality training data can transfer across sub-tasks, which reduces the burden of collecting high-quality data for complex-modal tasks. The proposed cascaded framework shows its resolution scaling ability, which achieves multi-modal guided 4K video generation for the first time.

REFERENCES

Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), International Conference on Machine Learning (ICML), 2023.

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.

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

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In Proceedings of the IEEE/CVF international conference on computer vision, pp. 9650–9660, 2021.

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

Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 5972–5981, 2022a.

Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Investigating tradeoffs in real-world video super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5962–5971, 2022b.

Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. arXiv preprint arXiv:2310.19512, 2023.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7310–7320, 2024a.

Qihua Chen, Yue Ma, Hongfa Wang, Junkun Yuan, Wenzhe Zhao, Qi Tian, Hongmei Wang, Shaobo Min, Qifeng Chen, and Wei Liu. Follow-your-canvas: Higher-resolution video outpainting with extensive content generation. arXiv preprint arXiv:2409.01055, 2024b.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13320–13331, 2024c.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Yuwei Fang, Kwot Sin Lee, Ivan Skorokhodov, Kfir Aberman, Jun-Yan Zhu, Ming-Hsuan Yang, and Sergey Tulyakov. Multi-subject open-set personalization in video generation. In Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 6099–6110, 2025.

Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 1290–1299, 2022.

Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim M Alabdulmohsin, et al. Patch n'pack: Navit, a vision transformer for any aspect ratio and resolution. Advances in Neural Information Processing Systems, 36, 2024.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. Delta tuning: A comprehensive study of parameter efficient methods for pre-trained language models. arXiv preprint arXiv:2203.06904, 2022.

Shian Du, Xiaotian Cheng, Qi Qian, Henglu Wei, Yi Xu, and Xiangyang Ji. Efficient personalized text-to-image generation by leveraging textual subspace. arXiv preprint arXiv:2407.00608, 2024.

Shian Du, Menghan Xia, Chang Liu, Xintao Wang, Jing Wang, Pengfei Wan, Di Zhang, and Xiangyang Ji. Patchvsr: Breaking video diffusion resolution limits with patch-wise video super-resolution. In Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 17799–17809, 2025.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In International Conference on Machine Learning (ICML), 2024a.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In Forty-first International Conference on Machine Learning, 2024b.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618, 2022.

Zhicheng Geng, Luming Liang, Tianyu Ding, and Ilya Zharkov. Rstt: Real-time spatial temporal transformer for space-time video super-resolution. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 17441–17451, 2022.

Lanqing Guo, Yingqing He, Haoxin Chen, Menghan Xia, Xiaodong Cun, Yufei Wang, Siyu Huang, Yong Zhang, Xintao Wang, Qifeng Chen, Ying Shan, and Bihan Wen. Make a cheap scaling: A self-cascade diffusion model for higher-resolution adaptation. In European Conference on Computer Vision (ECCV), 2024.

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, 2024a.

Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, et al. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. arXiv preprint arXiv:2406.15252, 2024b.

Xuanhua He, Quande Liu, Shengju Qian, Xin Wang, Tao Hu, Ke Cao, Keyu Yan, and Jie Zhang. Id-animator: Zero-shot identity-preserving human video generation. arXiv preprint arXiv:2404.15275, 2024c.

Xuanhua He, Quande Liu, Zixuan Ye, Weicai Ye, Qiulin Wang, Xintao Wang, Qifeng Chen, Pengfei Wan, Di Zhang, and Kun Gai. Fulldit2: Efficient in-context conditioning for video diffusion transformers. arXiv preprint arXiv:2506.04213, 2025.

Yingqing He, Shaoshu Yang, Haoxin Chen, Xiaodong Cun, Menghan Xia, Yong Zhang, Xintao Wang, Ran He, Qifeng Chen, and Ying Shan. Scalecrafter: Tuning-free higher-resolution visual generation with diffusion models. In International Conference on Learning Representations (ICLR), 2024d.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626, 2022.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303, 2023.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In International Conference on Learning Representations (ICLR), 2022.

Jiahao Hu, Tianxiong Zhong, Xuebo Wang, Boyuan Jiang, Xingye Tian, Fei Yang, Pengfei Wan, and Di Zhang. Vivid-10m: A dataset and baseline for versatile and interactive video local editing. arXiv preprint arXiv:2411.15260, 2024.

Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8153–8163, 2024.

Teng Hu, Zhentao Yu, Zhengguang Zhou, Sen Liang, Yuan Zhou, Qin Lin, and Qinglin Lu. Hunyuancustom: A multimodal-driven architecture for customized video generation. arXiv preprint arXiv:2505.04512, 2025.

Yuzhou Huang, Ziyang Yuan, Quande Liu, Qiulin Wang, Xintao Wang, Ruimao Zhang, Pengfei Wan, Di Zhang, and Kun Gai. Conceptmaster: Multi-concept video customization on diffusion transformer models without test-time tuning. arXiv preprint arXiv:2501.04698, 2025.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 21807–21818, 2024.

Takashi Isobe, Xu Jia, Shuhang Gu, Songjiang Li, Shengjin Wang, and Qi Tian. Video super-resolution with recurrent structure-detail network. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16, pp. 645–660. Springer, 2020.

Mehran Jeelani, Noshaba Cheema, Klaus Illgner-Fehns, Philipp Slusallek, Sunil Jaiswal, et al. Expanding synthetic real-world degradations for blind video super resolution. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 1199–1208, 2023.

Xuan Ju, Ailing Zeng, Chenchen Zhao, Jianan Wang, Lei Zhang, and Qiang Xu. Humansd: A native skeleton-guided diffusion model for human image generation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 15988–15998, 2023.

Xuan Ju, Weicai Ye, Quande Liu, Qiulin Wang, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, and Qiang Xu. Fulldit: Multi-task video generative foundation model with full attention. arXiv preprint arXiv:2503.19907, 2025.

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

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In International Conference on Learning Representations (ICLR), 2014.

Guojun Lei, Chi Wang, Rong Zhang, Yikai Wang, Hong Li, and Weiwei Xu. Animateanything: Consistent and controllable animation for video generation. In Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 27946–27956, 2025.

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, pp. arXiv–2501, 2025.

Jun Hao Liew, Hanshu Yan, Jianfeng Zhang, Zhongcong Xu, and Jiashi Feng. Magicedit: High-fidelity and temporally coherent video editing. arXiv preprint arXiv:2308.14749, 2023.

Han Lin, Jaemin Cho, Abhay Zala, and Mohit Bansal. Ctrl-adapter: An efficient and versatile framework for adapting diverse controls to any diffusion model. arXiv preprint arXiv:2404.09967, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In International Conference on Learning Representations (ICLR), 2023.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. Advances in neural information processing systems, 36, 2024a.

Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. arXiv preprint arXiv:2202.09778, 2022b.

Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. arXiv preprint arXiv:2402.17177, 2024b.

I Loshchilov. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101, 2017.

Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. arXiv preprint arXiv:2401.03048, 2024a.

Ze Ma, Daquan Zhou, Xue-She Wang, Chun-Hsiao Yeh, Xiuyu Li, Huanrui Yang, Zhen Dong, Kurt Keutzer, and Jiashi Feng. Magic-me: Identity-specific video customized diffusion. In European Conference on Computer Vision, pp. 19–37. Springer, 2024b.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. arXiv preprint arXiv:2108.01073, 2021.

Chong Mou, Mingdeng Cao, Xintao Wang, Zhaoyang Zhang, Ying Shan, and Jian Zhang. Revideo: Remake a video with motion and content control. Advances in Neural Information Processing Systems, 37:18481–18505, 2024a.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pp. 4296–4304, 2024b.

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

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. arXiv preprint arXiv:2407.02371, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4195–4205, 2023a.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4195–4205, 2023b.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952, 2023.

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 International conference on machine learning, pp. 8748–8763. PmLR, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research, 21(140):1–67, 2020.

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

ByteDance Seed. Seedance 1.0: Exploring the boundaries of video generation models. arXiv preprint arXiv:2506.09113, 2025.

Luohe Shi, Hongyi Zhang, Yao Yao, Zuchao Li, and Hai Zhao. Keep the cost down: A review on methods to optimize llm's kv-cache consumption. arXiv preprint arXiv:2407.18003, 2024.

Shuwei Shi, Jinjin Gu, Liangbin Xie, Xintao Wang, Yujiu Yang, and Chao Dong. Rethinking alignment in video super-resolution transformers. Advances in Neural Information Processing Systems, 35:36081–36093, 2022.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. Neurocomputing, 568:127063, 2024.

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

Yuanpeng Tu, Hao Luo, Xi Chen, Sihui Ji, Xiang Bai, and Hengshuang Zhao. Videoanydoor: High-fidelity video object insertion with precise motion control. In Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers, pp. 1–11, 2025.

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

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

Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin C. K. Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. International Journal of Computer Vision, 2024.

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

Jianyi Wang, Zhijie Lin, Meng Wei, Yang Zhao, Ceyuan Yang, Chen Change Loy, and Lu Jiang. Seedvr: Seeding infinity in diffusion transformer towards generic video restoration. In Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 2161–2172, 2025b.

Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. arXiv preprint arXiv:2308.06571, 2023b.

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

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

Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. arXiv preprint arXiv:2309.15103, 2023c.

Yujie Wei, Shiwei Zhang, Zhiwu Qing, Hangjie Yuan, Zhiheng Liu, Yu Liu, Yingya Zhang, Jingren Zhou, and Hongming Shan. Dreamvideo: Composing your dream videos with customized subject and motion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6537–6549, 2024.

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

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. arXiv preprint arXiv:2312.17090, 2023b.

Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7623–7633, 2023c.

Liangbin Xie, Yu Li, Shian Du, Menghan Xia, Xintao Wang, Fanghua Yu, Ziyan Chen, Pengfei Wan, Jiantao Zhou, and Chao Dong. Simplegvr: A simple baseline for latent-cascaded video super-resolution. arXiv preprint arXiv:2506.19838, 2025a.

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, 2025b.

Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In European Conference on Computer Vision, pp. 399–417. Springer, 2025.

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

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

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

Zixuan Ye, Huijuan Huang, Xintao Wang, Pengfei Wan, Di Zhang, and Wenhan Luo. Stylemaster: Stylize your video with artistic generation and translation. In Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 2630–2640, 2025.

Shenghai Yuan, Jinfa Huang, Xianyi He, Yunyang Ge, Yujun Shi, Liuhan Chen, Jiebo Luo, and Li Yuan. Identity-preserving text-to-video generation by frequency decomposition. In Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 12978–12988, 2025.

Ailing Zeng, Yuhang Yang, Weidong Chen, and Wei Liu. The dawn of video generation: Preliminary explorations with sora-like models. arXiv preprint arXiv:2410.05227, 2024.

David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. International Journal of Computer Vision, pp. 1–15, 2024.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3836–3847, 2023a.

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

Shilong Zhang, Wenbo Li, Shoufa Chen, Chongjian Ge, Peize Sun, Yida Zhang, Yi Jiang, Zehuan Yuan, Binyue Peng, and Ping Luo. Flashvideo: Flowing fidelity to detail for efficient high-resolution video generation. arXiv preprint arXiv:2502.05179, 2025a.

Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. arXiv preprint arXiv:2311.04145, 2023b.

Yuechen Zhang, Yaoyang Liu, Bin Xia, Bohao Peng, Zexin Yan, Eric Lo, and Jiaya Jia. Magic mirror: Id-preserved video generation in video diffusion transformers. arXiv preprint arXiv:2501.03931, 2025b.

Shangchen Zhou, Peiqing Yang, Jianyi Wang, Yihang Luo, and Chen Change Loy. Upscale-a-video: Temporal-consistent diffusion model for real-world video super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2535–2545, 2024.

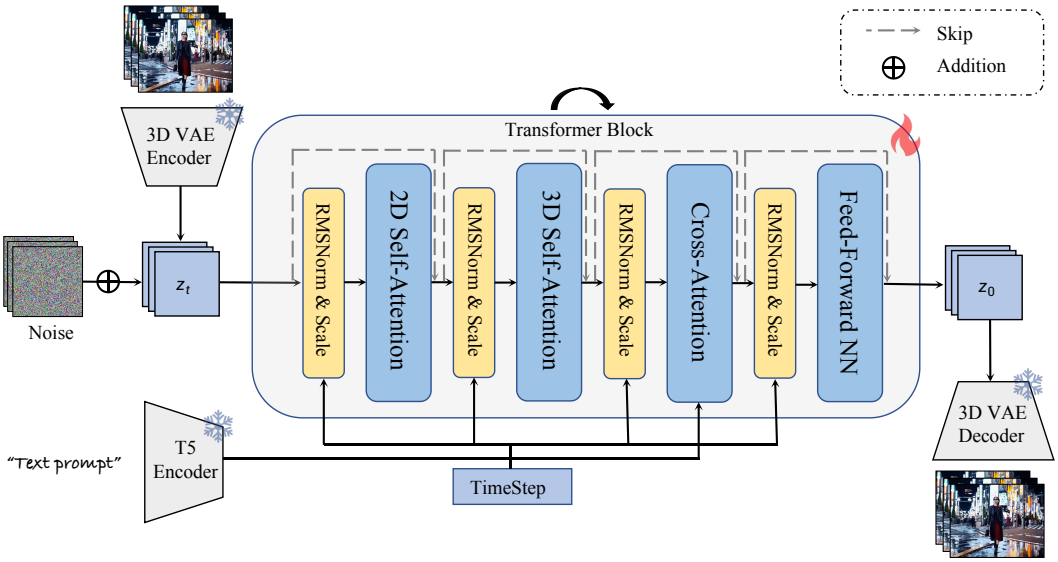

Figure 7: An overview of the architecture of our base model.

A Appendix

The content in the appendix is categorized as follows:

- Base Model.
- Training Datasets.
- Inference Technique.
- SDEdit Degradation.
- More Results.
  - Training Convergence Speed.
  - Quantitative Comparisons.
  - Qualitative Comparisons.
  - Ablation Study.
  - 4K Results.

A.1 Base Model

The architecture of our base model is shown in Fig. 7. Our UniMMVSR is built upon a pre-trained DiT-based video diffusion model, which comprises four main components: spatial self-attention (SSA), spatial cross-attention (SCA), temporal self-attention (TSA) and feed-forward network (FFN). Text prompts, time step and other micro conditions (aspect ratio, FPS, etc) are injected via the modulation mechanism (Peebles & Xie, 2023b). The pre-trained model is trained on 77-frames 512×512 resolution high-quality video data with diverse aspect ratio using NaViT Dehghani et al. (2024).

A.2 Training Datasets

Text-to-video Generation. We train our model using 840K self-collected high-quality video-text pairs, with each clip processed to 5 seconds and 1080P resolution. The dataset is constructed by applying several IQA/VQA methods (Wu et al., 2023b; Wang et al., 2023a; Ke et al., 2021; Wu et al., 2023a) to filter out low-quality data from 5M raw videos. The text prompts are all captioned using LLAVA captioner (Liu et al., 2024a), and encoded by T5 text encoder (Raffel et al., 2020) with no more than 512 tokens.

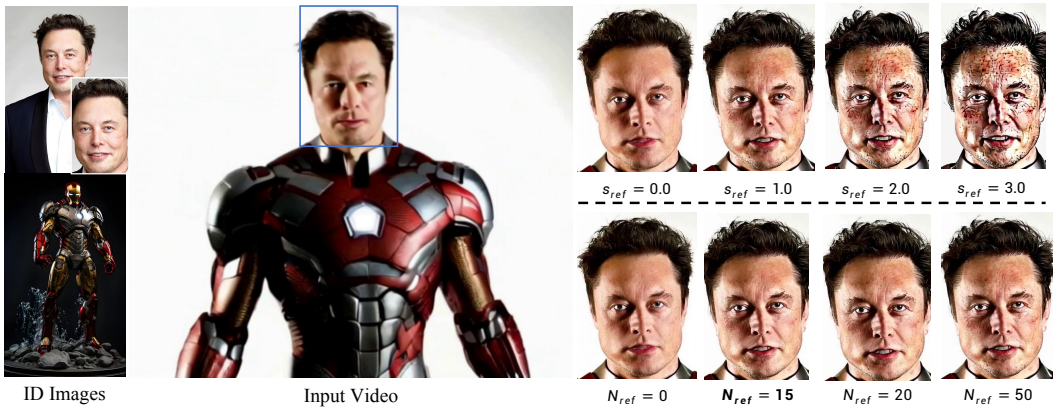

Figure 8: Qualitative comparisons of different inference settings. The text prompt is omitted.

Multi-ID Image-guided Text-to-video Generation. Since portrait-related images dominate the application of multi-ID image-guided text-to-video generation task, we collect around 1.5M videos from open-sourced movies and television series. We then apply the same data filtering as text-to-video generation task to obtain 480K high-quality samples for training. We randomly select a non-overlapping frame from the video clip to extract the reference image. Finally, we apply Mask2former method (Cheng et al., 2022) to identify and extract referenced images.

Text-guided Video Editing. Although inpainting-based datasets align better with test scenarios, model-generated reference videos naturally lack high-definition details required by our method. Thus, we follow the local-editing data pipeline (Hu et al., 2024) to preserve the high-frequency information in the non-editing area of the reference video, which results in 450K high-quality samples.

A.3 Inference Technique

UniMMVSR is trained on video data with either reference conditions or null conditions, and thus it can handle both scenarios. During inference, we apply independent classifier-free guidance (CFG) for each condition as:

$$\begin{aligned}\tilde{\epsilon}_\theta(z_t, t, c_{txt}, c_{ref}) = \epsilon_\theta(z_t, t, c_{txt}, c_{ref}) \\ + s_{txt} \cdot (\epsilon_\theta(z_t, t, c_{txt}, c_{ref}) - \epsilon_\theta(z_t, t, \phi_{txt}, c_{ref})) \\ + s_{ref} \cdot (\epsilon_\theta(z_t, t, c_{txt}, c_{ref}) - \epsilon_\theta(z_t, t, c_{txt}, \phi_{ref})),\end{aligned} \tag{2}$$

where c_{txt}, c_{ref} denote the condition of text prompt and reference, ϕ_{txt}, ϕ_{ref} denote the corresponding null conditions and s_{txt}, s_{ref} are the guidance scale. However, we find that simply increasing reference scale s_{ref} leads to over-sharpen results and even generates artifacts. Since our goal is to modify the local structure of the input video based on reference conditions and generate high-frequency details, we introduce reference guidance threshold (RGT) technique to only utilize reference condition for first N_{ref} steps as below:

$$\tilde{s}_{ref} = \begin{cases} s_{ref}, & n < N_{ref} \\ 0, & n \geq N_{ref} \end{cases} \tag{3}$$

We have compared the results of different inference settings in Fig. 8. As shown below, directly increasing reference guidance scale leads to over-sharpen details and even artifacts. By utilizing the proposed RGT technique ($s_{ref} = 1.0 \& N_{ref} = 15$), UniMMVSR enhances the guidance of the reference conditions, further strengthening the generalization on the cross-pair test set.

A.4 SDEdit Degradation

The degradation pipeline is shown in Fig. 9. We first perform sdedit degradation to modify the local structure of HR video using the sdedit method by our text-to-video base model. Afterwards, we

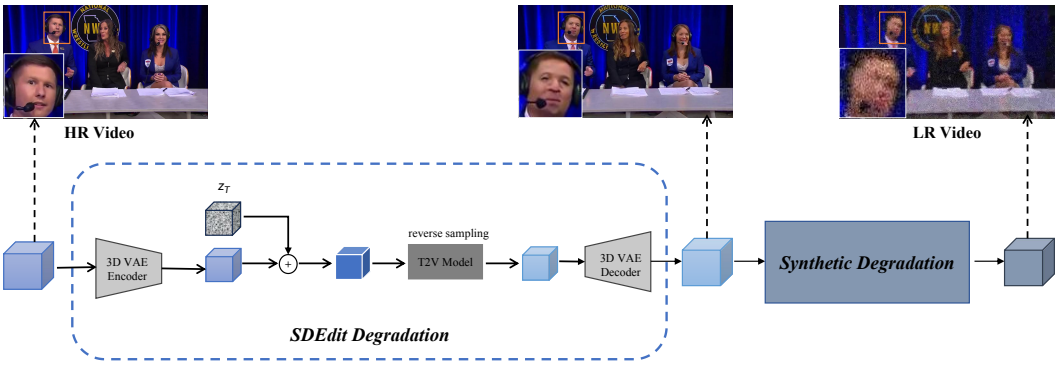

Figure 9: Degradation pipeline for UniMMVSR.

Table 3: Quantitative comparison of text-to-video generation task. **Bold** and underlined indicate the best and second-best results, respectively. ↑ indicates higher is better; ↓ indicates lower is better.

Text-to-video Generation									
Method	**Visual Quality**				**Subject Consistency**		**Video Alignment**		
	MUSIQ↑	**CLIP-IQA↑**	**QAlign↑**	**DOVER↑**	**CLIP-I↑**	**DINO-I↑**	**PSNR↑**	**SSIM↑**	**LPIPS↓**
Base 512×512	30.996	0.246	3.741	0.594	-	-	-	-	-
Base 1080P	46.645	0.306	4.246	0.749	-	-	-	-	-
VEnhancer-v1	**57.171**	0.367	4.214	0.733	-	-	-	-	-
VEnhancer-v2	44.603	0.364	4.091	0.693	-	-	-	-	-
STAR-light	56.904	0.369	4.435	0.769	-	-	-	-	-
STAR-heavy	55.294	0.361	**4.579**	**0.795**	-	-	-	-	-
SeedVR-7B	55.596	**0.379**	4.396	0.778	-	-	-	-	-
SeedVR-3B	54.310	0.375	4.281	0.762	-	-	-	-	-
SeedVR2-7B	48.490	0.310	4.292	0.763	-	-	-	-	-
SeedVR2-7B-sharp	47.013	0.300	4.234	0.753	-	-	-	-	-
SeedVR2-3B	50.604	0.328	4.318	0.783	-	-	-	-	-
Ours (single)	56.146	0.366	4.535	0.771	-	-	-	-	-
Ours (unified)	56.418	0.371	4.500	0.778	-	-	-	-	-

apply traditional synthetic degradation to introduce high-frequency degradation pattern. Light and heavy sdedit degradation samples are shown in Fig. 10 and 11 respectively.

A.5 More Results

A.5.1 Training Convergence Speed

We have shown the training loss curve of single-task model on text-to-video generation, multi-ID image-guided text-to-video generation and text-guided video editing tasks in Fig. 12. As can be seen, text-guided video editing task tends to converge faster at a lower loss value 0.18, while text-to-video generation task converges slowest, at around $3k$ steps.

A.5.2 Quantitative Comparisons

Full quantitative comparisons of text-to-video generation, multi-ID image-guided text-to-video generation and text-guided video editing tasks are shown in Tab. 3, 4 and 5 respectively.

A.5.3 Qualitative Comparisons

Additional qualitative comparisons of text-to-video generation, multi-ID image-guided text-to-video generation and text-guided video editing tasks are presented in Fig. 13, 14 and 15 respectively.

A.5.4 Ablation Study

Qualitative comparisons with different components are shown in Fig. 16. For architecture design, full channel-concat (Full CC) struggles to inject visual references. For full token-concat (Full TC),

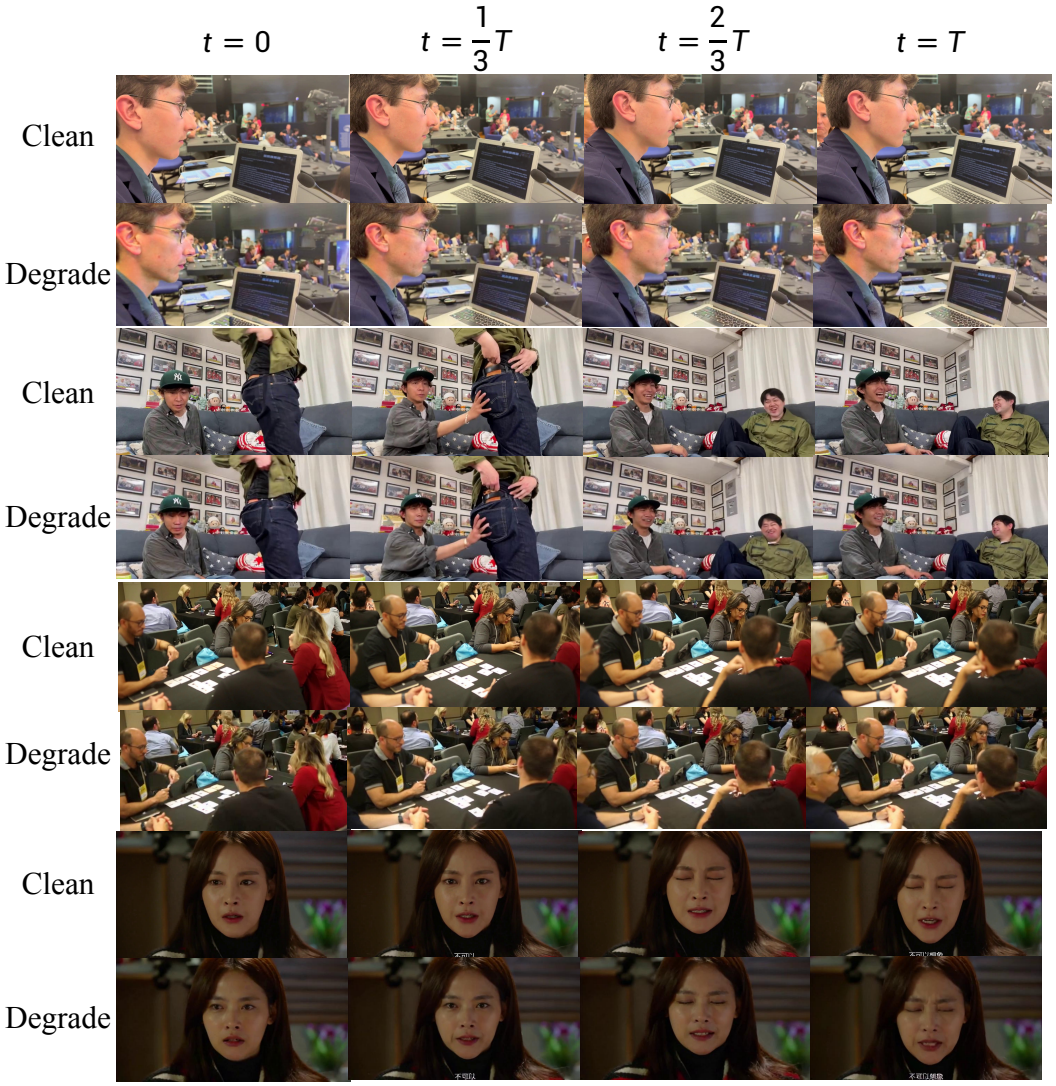

Figure 10: Samples of light sdedit degradation.

although it achieves comparable results, it largely sacrifices the inference efficiency. For degradation effect, sdedit-only and synthetic-only methods lack in generating vivid details and preserving input ID images respectively. For training order, both full training and easy-to-difficult paradigm show suboptimal results compared with our difficult-to-easy paradigm, which demonstrates the effectiveness of our training strategy.

A.5.5 4K Results

Additional 4K results of text-to-video generation, multi-ID image-guided text-to-video generation and text-guided video editing tasks are presented in Fig. 17, 18 and 19 respectively. The results show that the proposed cascaded framework excels at scaling resolution on all three controllable video generation tasks. We also present 4K videos in the supplementary material.

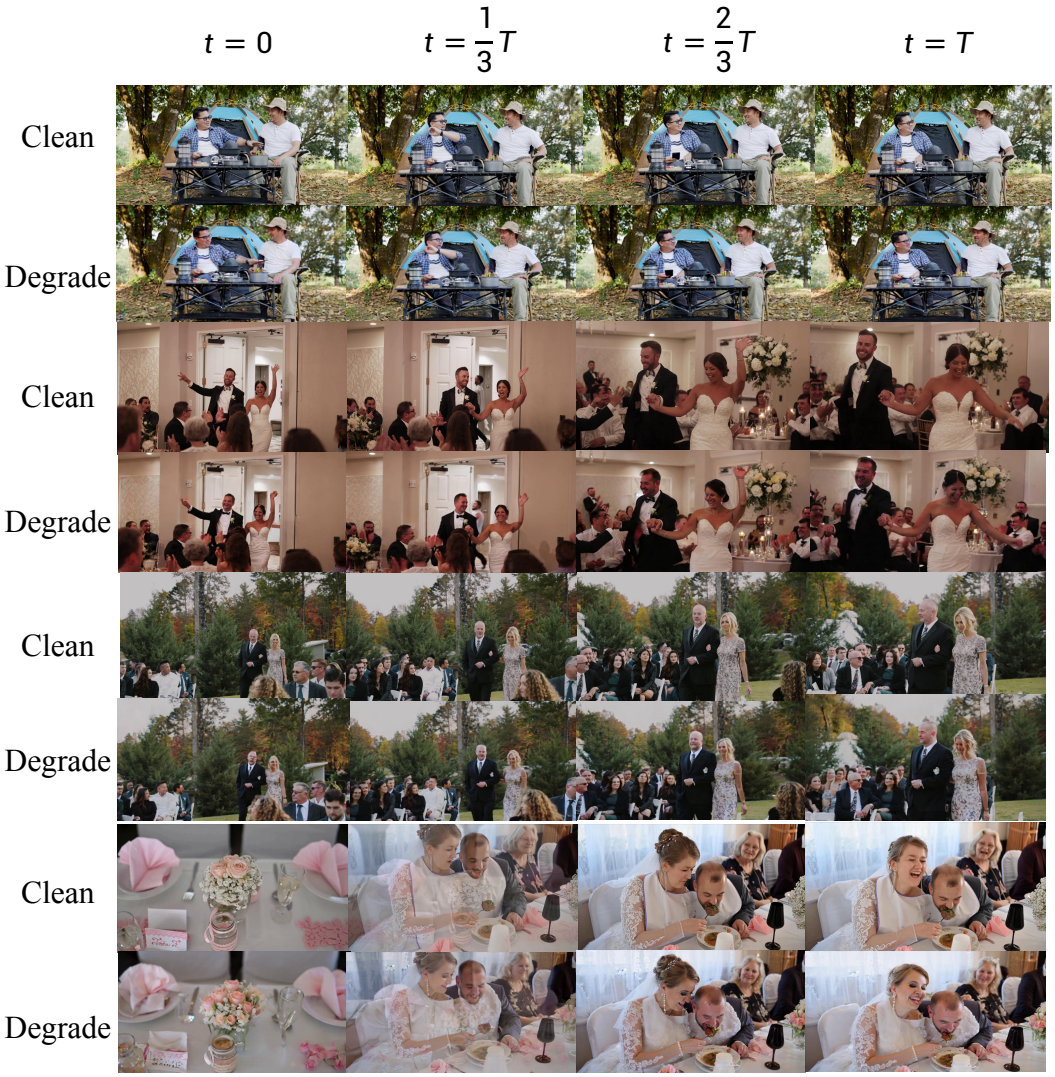

Figure 11: Samples of heavy sdedit degradation.

Table 4: Quantitative comparison of multi-ID image-guided text-to-video generation task. **Bold** and underlined indicate the best and second-best results, respectively. ↑ indicates higher is better; ↓ indicates lower is better.

Multi-ID Image-guided Text-to-video Generation									
Method	**Visual Quality**				**Subject Consistency**		**Video Alignment**		
	MUSIQ↑	**CLIP-IQA↑**	**QAlign↑**	**DOVER↑**	**CLIP-I↑**	**DINO-I↑**	**PSNR↑**	**SSIM↑**	**LPIPS↓**
Base 512×512	29.314	0.255	3.149	0.433	0.692	0.538	-	-	-
Base 1080P	46.780	0.345	4.092	0.662	0.691	0.507	-	-	-
VEnhancer-v1	60.656	**0.469**	4.149	0.707	0.671	0.533	-	-	-
VEnhancer-v2	43.776	0.422	3.860	0.628	0.690	0.538	-	-	-
STAR-light	58.810	0.449	4.282	**0.763**	0.696	0.546	-	-	-
STAR-heavy	54.446	0.399	4.223	0.721	0.695	0.547	-	-	-
SeedVR-7B	54.491	0.419	3.960	0.708	0.693	0.543	-	-	-
SeedVR-3B	53.943	0.416	3.845	0.689	0.696	0.544	-	-	-
SeedVR2-7B	49.220	0.344	3.814	0.664	0.689	0.543	-	-	-
SeedVR2-7B-sharp	46.718	0.332	3.751	0.639	0.691	0.545	-	-	-
SeedVR2-3B	51.169	0.368	3.850	0.690	0.694	0.544	-	-	-
Ours (no ref)	60.947	0.445	4.385	0.742	0.693	0.543	-	-	-
Ours (single)	61.357	0.446	4.414	0.743	**0.728**	**0.566**	-	-	-
Ours (unified)	**62.248**	0.465	**4.428**	0.745	0.726	**0.566**	-	-	-

Figure 12: Training loss curve of all three tasks.

Table 5: Quantitative comparison of text-guided video editing task. **Bold** and underlined indicate the best and second-best results, respectively. ↑ indicates higher is better; ↓ indicates lower is better.

Text-guided Video Editing									
Method	**Visual Quality**				**Subject Consistency**		**Video Alignment**		
	MUSIQ↑	**CLIP-IQA↑**	**QAlign↑**	**DOVER↑**	**CLIP-I↑**	**DINO-I↑**	**PSNR↑**	**SSIM↑**	**LPIPS↓**
Base 512×512	35.073	0.234	3.615	0.400	-	-	30.191	0.699	0.364
Base 1080P	53.616	0.383	4.247	0.634	-	-	29.383	0.582	0.358
Ref Video	54.249	0.365	4.131	0.571	-	-	-	-	-
VEnhancer-v1	57.036	0.380	4.013	0.590	-	-	28.417	0.571	0.489
VEnhancer-v2	48.084	0.353	3.959	0.557	-	-	28.712	0.628	0.410
STAR-light	56.802	0.397	4.264	0.608	-	-	29.421	0.631	0.397
STAR-heavy	56.207	0.378	4.259	0.599	-	-	29.442	0.648	0.371
SeedVR-7B	57.820	0.370	4.183	0.635	-	-	29.535	0.597	0.413
SeedVR-3B	55.326	0.360	4.048	0.628	-	-	29.338	0.588	0.416
SeedVR2-7B	54.046	0.361	4.087	0.610	-	-	29.600	0.614	0.367
SeedVR2-7B-sharp	52.723	0.359	4.010	0.579	-	-	29.563	0.619	0.362
SeedVR2-3B	54.310	0.355	4.099	0.597	-	-	29.326	0.593	0.382
Ours (no ref)	**59.119**	**0.399**	4.289	**0.648**	-	-	29.615	0.581	0.429
Ours (single)	53.388	0.348	4.302	0.597	-	-	**31.905**	**0.723**	**0.276**
Ours (unified)	53.245	0.344	**4.305**	0.597	-	-	31.556	0.713	0.282

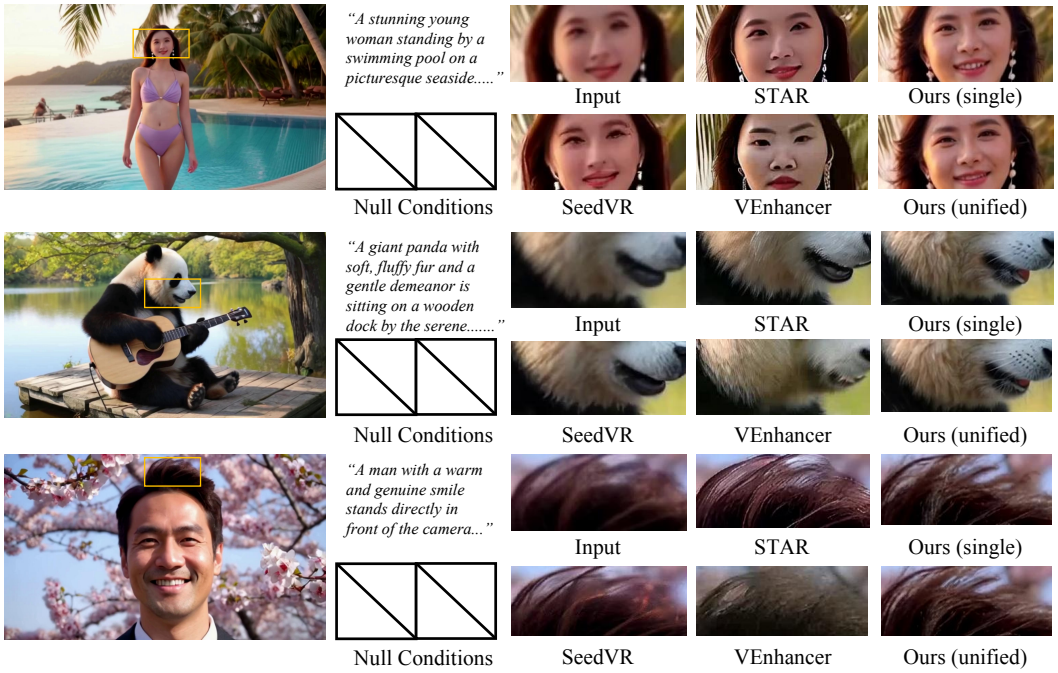

Figure 13: Qualitative comparisons on text-to-video generation task.

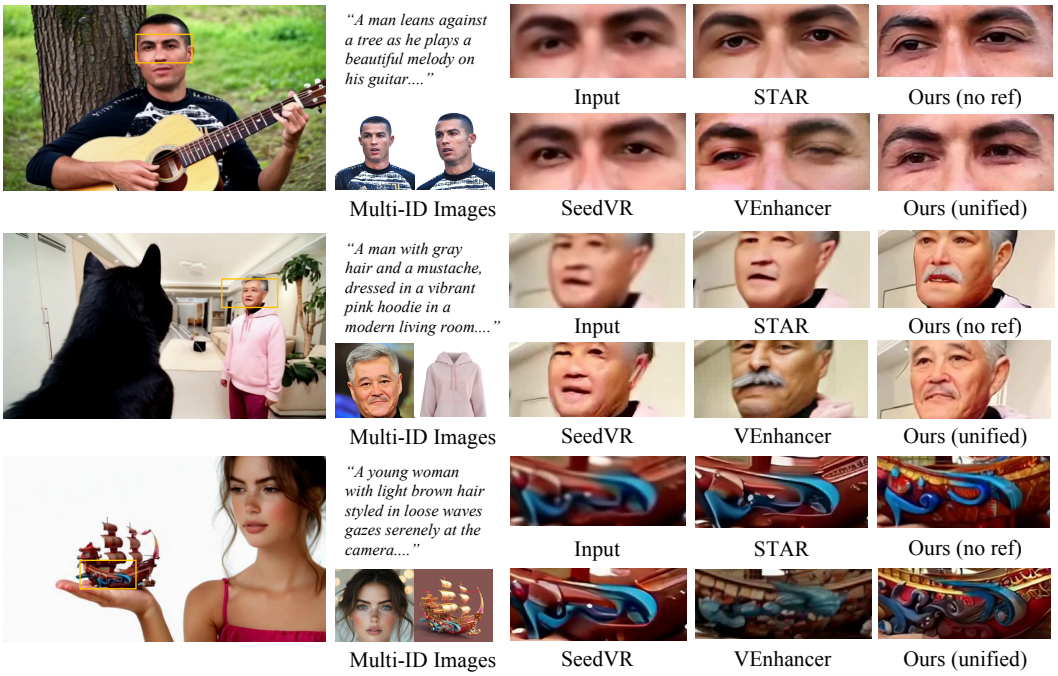

Figure 14: Qualitative comparisons on multi-ID image-guided text-to-video generation task.

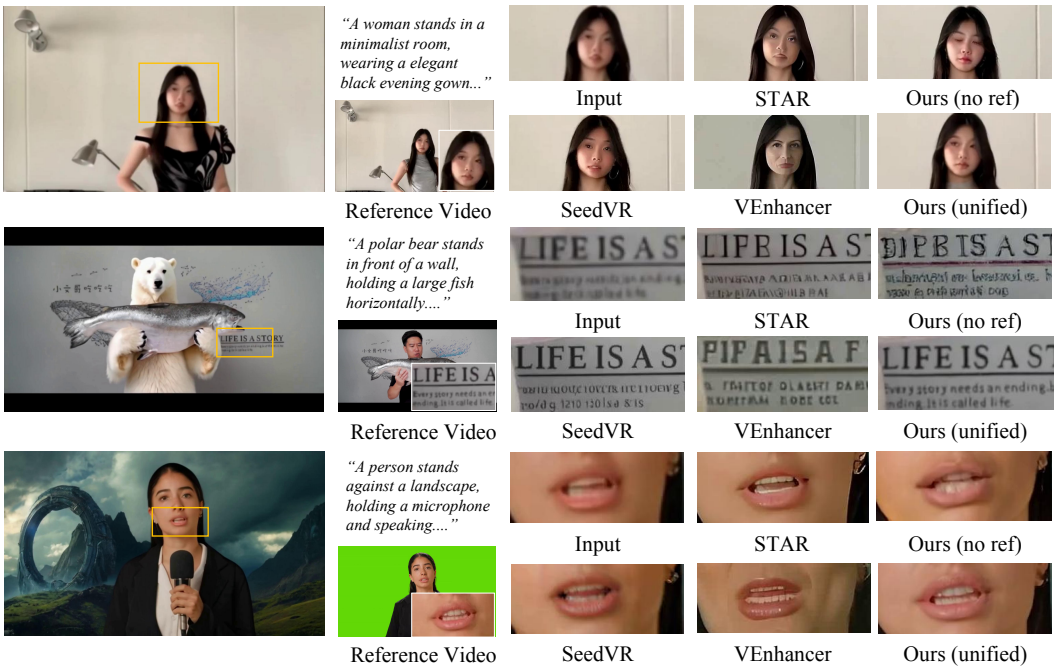

Figure 15: Qualitative comparisons on text-guided video editing task.

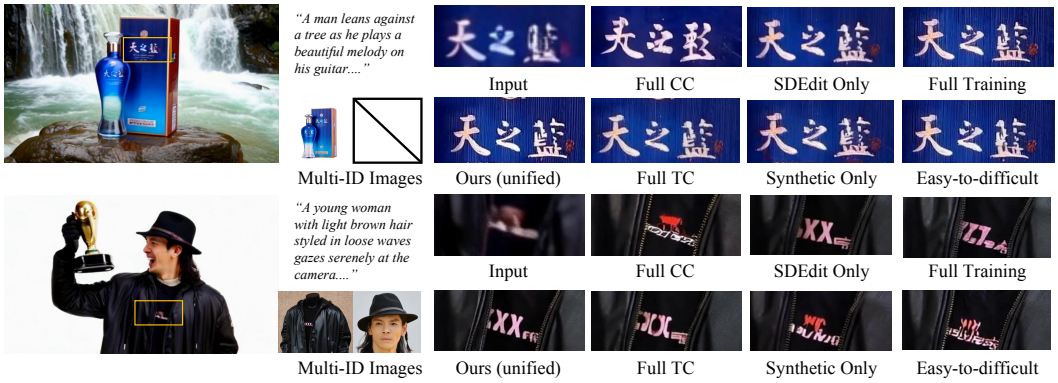

Figure 16: Qualitative comparisons with different components.

Figure 17: Additional 4K results on text-to-video generation task. **Zoom-in for best view.**

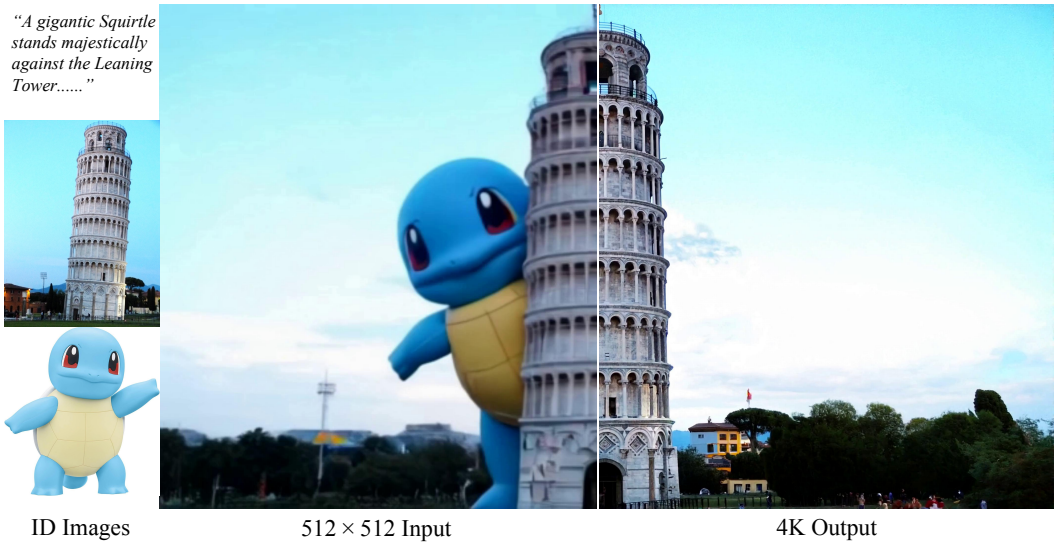

Figure 18: Additional 4K results on multi-ID image-guided text-to-video generation task. **Zoom-in for best view.**

Figure 19: Additional 4K results on text-guided video editing task. **Zoom-in for best view.**