UniVideo: Unified Understanding, Generation, and Editing for Videos

Cong Wei[1,2*] **Quande Liu**[2†] **Zixuan Ye**[2] **Qiulin Wang**[2] **Xintao Wang**[2]
Pengfei Wan[2] **Kun Gai**[2] **Wenhu Chen**[1†]
[1] University of Waterloo [2] Kling Team, Kuaishou Technology

Website: https://congwei1230.github.io/UniVideo/

Code: https://github.com/KlingTeam/UniVideo

Abstract

Unified multimodal models have shown promising results in multimodal content understanding and generation but remain largely limited to the image domain. In this work, we present `UniVideo`, a versatile framework that extends unified modeling to the video domain. `UniVideo` adopts a dual-stream design, combining a Multimodal Large Language Model (MLLM) for visual understanding with a Multimodal DiT (MMDiT) for visual generation. This design preserves the MLLM's original text generation capabilities, enables accurate interpretation of complex multimodal instructions, and maintains visual consistency in the generated content. Built on this architecture, `UniVideo` unifies diverse video generation and editing tasks under a single multimodal instruction paradigm and is jointly trained across them. Extensive experiments demonstrate that `UniVideo` matches or surpasses state-of-the-art task-specific baselines in visual understanding, text/image-to-video generation, in-context video generation and in-context video editing. Notably, the unified design of `UniVideo` enables two forms of generalization. First, `UniVideo` supports task composition, such as combining editing with style transfer, by integrating multiple capabilities within a single instruction. Second, even without explicit training on free-form video editing, `UniVideo` transfers its editing capability from large-scale image editing data to this setting, handling unseen instructions, such as changing the environment or altering materials within a video. Beyond these core capabilities, `UniVideo` also supports visual-prompt-based video generation, where the MLLM interprets visual prompts and guides the MMDiT during synthesis. To foster future research, we released our model and code at https://github.com/KlingTeam/UniVideo.

1 Introduction

A long-term goal of multimodal AI assistants is to build models that can seamlessly **understand** diverse inputs across modalities and **generate** outputs in kind, enabling natural communication through language, images, and video demonstrations.

Recent advances in unified models suggest that this vision is increasingly attainable. Prior work (Shi et al., 2024a; Pan et al., 2025; Sun et al., 2023; Team, 2024; Tong et al., 2024; Wang et al., 2024b; Deng et al., 2025; Wu et al., 2025b; Ma et al., 2025b; Xie et al., 2024; 2025; Zhou et al., 2024) has demonstrated promising results in text–image understanding and generation by jointly optimizing these capabilities within unified systems. More recently, models such as Google Nano banana and GPT-image-1 have pushed this paradigm further by integrating computer vision, image

[*]Work done during an internship at Kling Team, Kuaishou Technology.
[†]Corresponding authors.

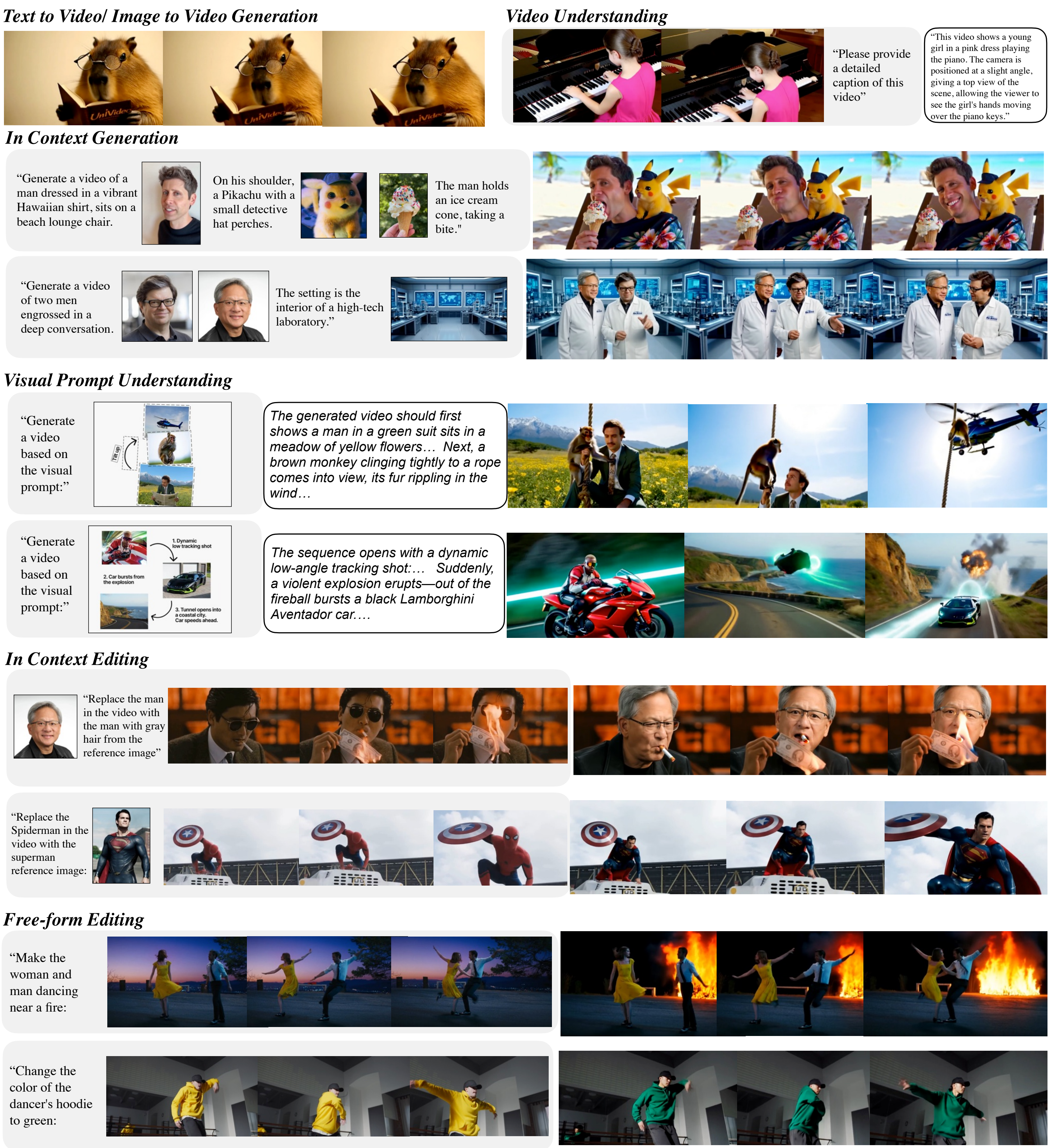


Figure 1: `UniVideo` is a unified system that can **understand** multi-modal instructions and **generate** multi-modal content. More videos are available on project website.

manipulation, and multimodal reasoning into a single framework, marking a shift from specialized single-modality generators toward powerful unified systems.

Despite this progress, unified understanding–generation models remain limited to text and image (Lin et al., 2025; Wu et al., 2025c), leaving video largely underexplored. Existing video generation models primarily address a single text-to-video task and rely on text encoders to process instructions (Wan et al., 2025; Ju et al., 2025b; Polyak et al., 2024; Kong et al., 2024), restricting their ability to understand and reason over multimodal instructions (Hu et al., 2024). Meanwhile, video editing methods typically employ task-specific modules or pipelines (Ku et al., 2024; Jiang et al., 2025; Ye et al., 2025b), which makes it difficult to scale across diverse tasks. Consequently, due to the lack of unified modeling, advanced capabilities such as multimodal prompting, in-context video generation, and sophisticated free-form editing remain beyond the reach of any single model.

Motivated by these limitations, we present `UniVideo` —a unified framework for understanding, generation, and editing in the video domain. `UniVideo` bridges this gap by enabling multimodal instruction following and delivering robust performance across diverse video tasks.

To build `UniVideo`, we propose a two-stream design, where an MLLM serves as the *understanding branch* and an MMDiT backbone (Esser et al., 2024) serves as the *generation branch*. While prior work such as Qwen-Image (Wu et al., 2025a) explores a similar idea in the image domain, our model generalizes this design to video. Both streams now receive image and video instructions: the understanding branch through a semantic encoder, and the generation branch through VAE-based encoders. In contrast, prior unified models such as GPT-image-1 (Lin et al., 2025) rely exclusively on semantic encoders, which often struggle to capture fine-grained visual details. Similarly, bottlenecked approaches using learnable query tokens (Tong et al., 2024; Pan et al., 2025) compress inputs into a fixed set of tokens, creating a severe capacity bottleneck when instructions contain videos. As a result, both approaches fall short in supporting in-context video generation. Our design preserves the multimodal reasoning capabilities of the MLLM while enabling the model to handle diverse video tasks with multimodal inputs. Moreover, it ensures cross-stream consistency, which is crucial for precise editing and for maintaining subject identity in in-context generation.

Based on this unified architecture, we train `UniVideo` across a wide spectrum of tasks, including text-to-image, text-to-video, image-to-video, in-context video generation, in-context video editing, and image editing. As a unified system, `UniVideo` not only understands multimodal instructions and distinguishes between tasks but also achieves improvements over state-of-the-art task-specific methods. Thanks to unified training, `UniVideo` generalizes to novel task compositions unseen during training, such as deleting one identity while swapping another within a single instruction. More importantly, although `UniVideo` is not trained on free-form video editing data, it demonstrates generalization ability transfer from image editing to free-form video editing (e.g., changing object materials or modifying weather conditions), highlighting the effectiveness of our unified video understanding and generation framework.

Furthermore, `UniVideo` retains the strong visual understanding and text generation capability of its underlying frozen MLLM. By leveraging the MLLM's autoregressive reasoning and language generation abilities, `UniVideo` can effectively interpret ambiguous and complex multimodal instructions that require joint vision–language understanding, such as turning visual prompting inputs into in-context video generation tasks or image-to-video generation tasks. Since `UniVideo`'s text generation ability originates from a frozen MLLM, `UniVideo` should be regarded as *a post-trained unified multimodal generative system*(Wu et al., 2025c; Pan et al., 2025) capable of producing images, videos, and text, rather than a unified model trained from scratch(Ma et al., 2025b; Deng et al., 2025).

Our key contributions are:
1) We introduce `UniVideo`, a multimodal generative model that unifies understanding, generation, and editing of videos within a single framework. To build `UniVideo`, we propose a dual-stream architecture that combines the multimodal reasoning capabilities of the MLLM with the generation strengths of the MMDiT. Unlike prior task-specific or modality-restricted approaches, `UniVideo` can interpret multimodal instructions, distinguish between diverse tasks, and achieve state-of-the-art performance across a wide range of benchmarks.
2) We systematically study the key design choices that enable this unified framework, including connector architectures, generator designs, and multimodal conditioning strategies, and provide em-

pirical evidence for their effectiveness.
3) We demonstrate that UniVideo generalizes to unseen tasks and novel task compositions without ad hoc designs, highlighting the benefits of a unified framework.

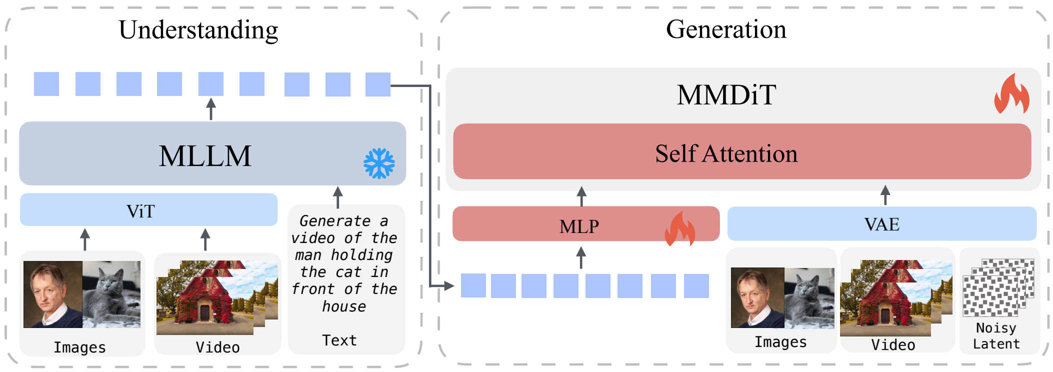


Figure 2: **Model architecture.** UniVideo is a dual-stream model consisting of an MLLM for understanding and an MMDiT module for generation. While prior work such as Qwen-Image and OmniGen2, explores a similar idea in the image domain, our model generalizes this design to video.

2 Method

2.1 Model Architecture

As demonstrated in Figure 2, UniVideo consists of two main components: a multimodal large language model (MLLM) and a multimodal DiT (MM-DiT). The MLLM handles visual–textual understanding, taking text, image, and video inputs and optionally producing text responses. The MM-DiT focuses on visual generation with two branches: one incorporates high-level semantic information from the MLLM, while the other integrates fine-grained reconstruction signals from a VAE. Specifically, we extract the last-layer hidden states of the MLLM, which encode rich semantic features of the multimodal input. These are aligned to the input space of the MM-DiT via a trainable connector and fed into its understanding stream. In parallel, visual signals are encoded by the VAE and passed into the MM-DiT generation stream to preserve fine details. This design enables strong semantic grounding together with high-fidelity visual detail, which is especially important for video editing and identity-preserving generation. We provide a model design analysis in subsection 3.2.

2.2 Unifying Multiple Tasks

We unify diverse multimodal tasks through natural language instructions, as illustrated in Figure 1. For text-to-video (T2V), the text input is processed by the MLLM, while the noisy video is fed into the MM-DiT. For image-to-video (I2V), both the image and text are processed by the MLLM, whereas the image and noisy video are provided to the MM-DiT. For in-context video generation (MultiID2V) and in-context video editing (ID-V2V), multiple visual conditions are often available, such as several reference images together with a reference video. Each visual signal is encoded with the VAE, padded to a uniform shape, concatenated along the temporal axis, and then processed with self-attention. Unlike prior approaches that introduce task-specific bias embeddings (Ye et al., 2025b) or context adapter modules (Jiang et al., 2025), we avoid task-specific customization. To help the MM-DiT distinguish between condition latents and noisy video latents, we apply 3D positional embeddings, which preserve the spatial indices across frames while incrementing only the temporal dimension. In practice, we find this strategy more effective than Qwen2-VL's MRoPE (Wang et al., 2024a), which offsets all axes whenever a new visual input is introduced.

2.3 Understanding Visual Prompt

UniVideo leverages its MLLM branch to interpret unconventional or hand-crafted prompts, as illustrated in Figure 3 and Figure 9. For example, users may provide an input image with man-

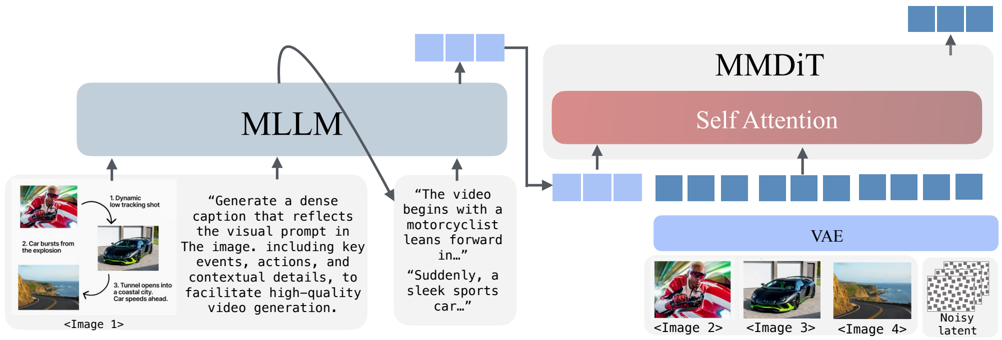


Figure 3: `UniVideo` leverages the MLLM stream to understand and interpret user intent from complex multimodal prompts that cannot be handled by the DiT alone. For example, users can provide diagrams or visual annotations to guide video generation without writing dense textual prompts.

Table 1: Training hyperparameters across different stages. Stage 1: Connector alignment, Stage 2: Fine-tuning, Stage 3: Multi-task training.

Hyperparameters	**Stages**		
	Stage 1 (Connector Alignment)	**Stage 2** (Fine-tuning)	**Stage 3** (Multi-task)
Learning rate	1×10^{-4}	2.0×10^{-5}	2.0×10^{-5}
LR scheduler	Constant	Constant	Constant
Weight decay	0.0	0.0	0.0
Gradient norm clip	1.0	1.0	1.0
Optimizer	AdamW ($\beta_1 = 0.9, \beta_2 = 0.95, \epsilon = 1.0 \times 10^{-15}$)		
Warm-up steps	50	50	50
Training steps	15K	5K	15K
EMA ratio	-	0.9999	0.9999
Gen resolution (min, max)	(240, 480)	(480, 854)	(480, 854)
Gen frames (min, max)	(1, 1)	(1, 129)	(1, 129)
Und resolution (min, max)	(240, 480)	(480, 854)	(480, 854)
Und frames (min, max)	(1, 1)	(1, 4)	(1, 4)
Diffusion timestep shift	5.0	5.0	5.0

ual annotations, which the MLLM translates into a structured plan and dense prompt tokens that guide video generation. Unlike agent-based approaches that invoke multiple downstream generators, `UniVideo` offers a more simplified design: the MMDiT directly integrates embeddings from the dense prompt tokens produced by the MLLM. This integration effectively turns visual prompting into in-context video generation.

2.4 Training Strategy

Stage 1. Connector alignment between MLLM and MMDiT. In this stage, we train only the MLP connector while keeping both the MLLM and MMDiT frozen. Training is performed on pretraining samples across text-to-image (T2I) and text-to-video (T2V) generation tasks, as well as an image-reconstruction task in which only images from the text-to-image dataset are fed into the MLLM and the MMDiT reconstructs the image using visual features from the MLLM. After this stage, `UniVideo` can generate images and videos conditioned on text or image inputs from the MLLM.

Stage 2. Fine-tuning MMDiT on T2I and T2V. In this stage, we keep the MLLM frozen and fine-tune the connector and MMDiT on small scale high-quality T2I and T2V samples. After this stage, `UniVideo` achieves performance comparable to the MMDiT backbone that uses its own text encoder.

Stage 3. Multi-task Training. Finally, we extend training to include in-context generation (multi-ID-to-video), in-context video editing (modifying the input video based on a reference image, such as ID swapping, ID addition, ID deletion, or style transfer), image editing and image-to-video tasks, alongside the previous text-to-image (T2I) and text-to-video (T2V) tasks. We keep the MLLM frozen and only train the connector and MMDiT. This stage enables `UniVideo` to unify a broad range of video generation and editing tasks under multimodal instruction. Details of training setting is provided in Table 1.

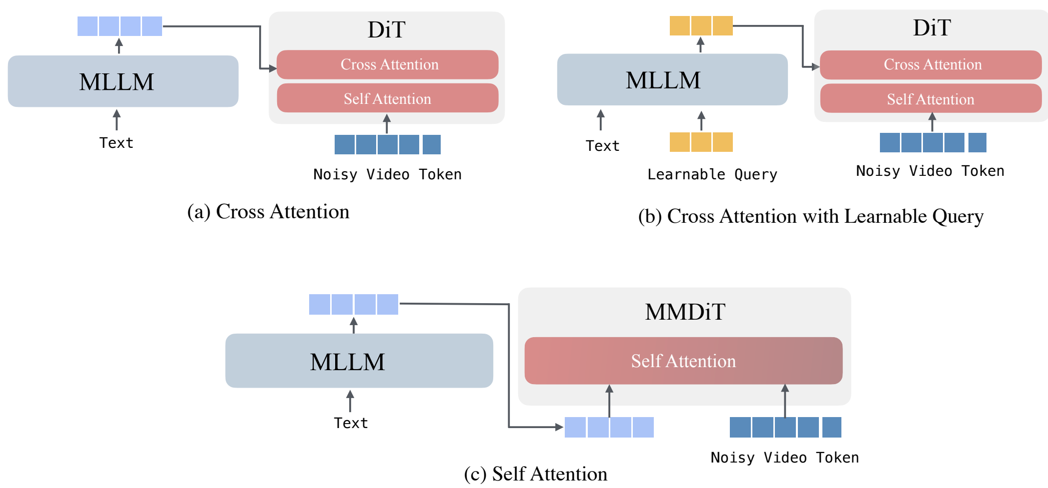


Figure 4: **Three design choices for aligning the MLLM with the diffusion generator in Stage 1 training.** We keep the MLLM fixed and vary the connector and DiT architecture across three variants: (a) the DiT uses cross-attention for text conditioning, where we replace its original text encoder with an MLP layer that aligns the final hidden states from the MLLM; (b) building upon (a), we introduce a learnable query design and extract the final hidden states from these learnable queries; and (c) our UniVideo architecture employs an MMDiT design that leverages self-attention for text conditioning.

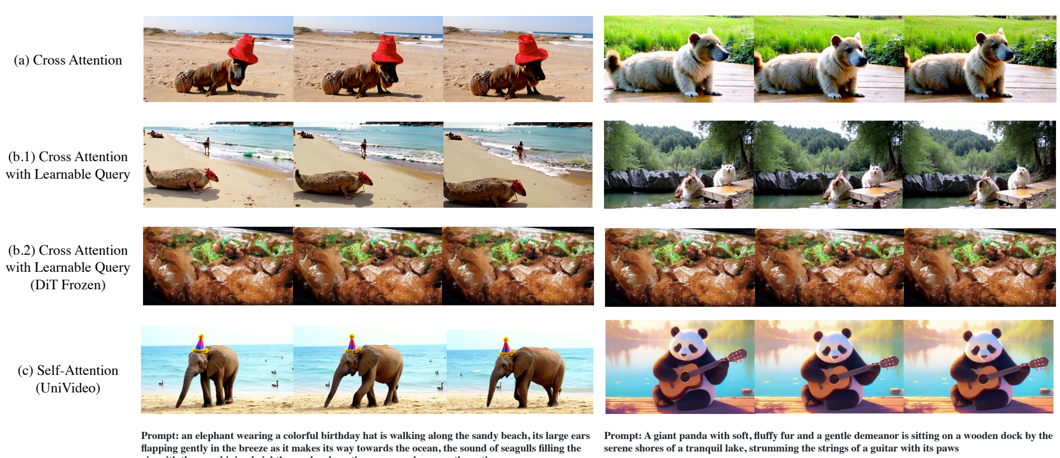
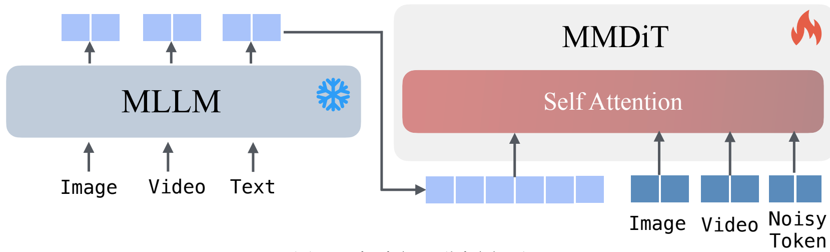
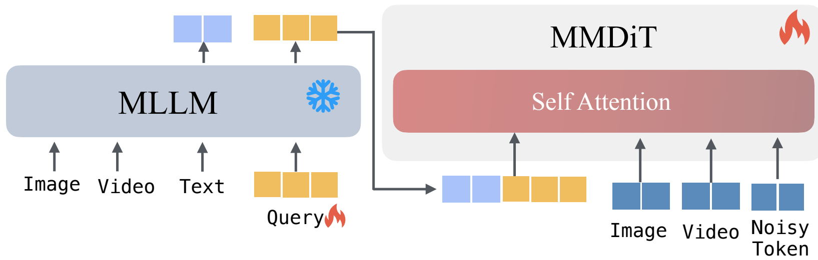


Figure 5: **Qualitative comparison of design choices for aligning the MLLM with the diffusion generator in Stage 1 training.** In all settings, the MLLM is kept frozen. (a) *Cross-Attention DiT:* we train the MLP connector and DiT; (b.1) *Cross-Attention DiT with Learnable Query:* following (Pan et al., 2025), we train the learnable query tokens, MLP connector, and DiT; (b.2) similar to (b.1), but the DiT is frozen while only the learnable query tokens and MLP connector are trained; (c) *UniVideo (MMDiT):* only the MLP connector is trained, with all other components frozen. All variants are trained for 15K steps. Among all variants, *UniVideo (MMDiT)* demonstrates the best prompt alignment.



(a) UniVideo (hidden)



(b) UniVideo (query)

Figure 6: **Two design variants for multimodal conditioning in Stage 3 multi-task training.** We study two multimodal conditioning strategies: (a) extracting the final hidden states of image, video, and text tokens produced by the MLLM; and (b) adopting a learnable query design and using the final hidden states of the learnable queries together with the final hidden states of the text tokens.

3 Experiments

In this section, we first describe the implementation details of `UniVideo` in subsection 3.1. Next, we discuss the design choices for aligning the MLLM and the Diffusion generator in subsection 3.2. Then, We present the main results in subsection 3.3. We conduct a comprehensive benchmark of `UniVideo` with SoTA methods across a broad spectrum of video understanding and generation tasks. Our results show that `UniVideo`'s strong unified capabilities across all settings. Next, we demonstrate the zero shot generalization ability of `UniVideo` and analysis the visual prompt understanding ability in subsection 3.4. Finally, we validate the design choices of `UniVideo` through ablation studies in subsection 3.5.

3.1 Implementation Details

We adopt qwen2.5VL-7B (Bai et al., 2025) as the MLLM backbone and HunyuanVideo-T2V-13B (Kong et al., 2024) as the MMDiT backbone. The original HunyuanVideo use two text encoders; we remove them and instead use qwen2.5VL as the unified multi-modal embedder. To align feature dimensions between qwen2.5VL and HunyuanVideo, we apply an MLP with a $4\times$ expansion.

3.2 Model Design

Our model design study addresses the following question: *What is the most effective approach for aligning a pretrained MLLM with a diffusion generator during Stage 1 training?*

We investigate three design choices for aligning the pretrained MLLM with the diffusion generator in Stage 1. Throughout this stage, the MLLM remains frozen, while we vary the connector and DiT architectures across three variants as shown in Figure 4.

(a) *Cross-attention DiT.* The first variant adopts a cross-attention–based DiT for text conditioning, where we replace its original text encoder with an MLP connector that projects the final hidden states from the MLLM into the DiT text embedding space. Both the MLP and DiT are trained.

(b) *Cross-attention DiT with Learnable query.* Building upon (a), we use a *learnable query* mechanism following Pan et al. (2025). Specifically, we extract the final hidden states of learnable queries from the MLLM, which are then passed through an MLP layer and used to replace the original text conditioning in the DiT's cross-attention module. We test two variants: (1) jointly training the learnable queries, MLP layer, and DiT (as in Pan et al. (2025)); and (2) training only the learnable queries and MLP while keeping the DiT frozen.

(c) *`UniVideo` architecture.* The main difference in this variant lies in its use of MMDiT, which employs self-attention for joint text–video interaction instead of cross-attention. We replace MMDiT's original text encoder with an MLP connector that projects the final hidden states from the MLLM into the MMDiT's text embedding space. Only the MLP layer is trained, while both the MLLM and MMDiT remain frozen.

For the cross-attention variants, we use an internal model with an architecture similar to (Wan et al., 2025), originally based on a T5 text encoder(Raffel et al., 2020), which we replace with Qwen2.5-VL. For UniVideo, we follow the implementation details described in subsection 3.1. All variants are trained for 15K steps, and the qualitative results are presented in Figure 5.

Our findings show that the cross-attention variants require unfreezing the DiT generator to achieve effective alignment with the MLLM, as evidenced by the comparison between (b.2) and (b.1). Nevertheless, even after unfreezing, variants (a) and (b.1) exhibit limited text-following ability—particularly for compositional object prompts. In contrast, the `UniVideo` architecture achieves efficient and robust alignment by training only the MLP connector.

We study two `UniVideo` variants in Stage 3 training. The MLLM is kept frozen, while we vary the connector design across two variants, as illustrated in Figure 6.

(a) ***`UniVideo` (hidden).*** In this variant, we extract the final-layer hidden states of all image, video, and text tokens produced by the MLLM. These token representations are used as inputs to the understanding branch of MMDiT. During training, only the MLP connector and the MMDiT are updated, while the MLLM remains frozen.

(b) ***`UniVideo` (query).*** This variant adopts a *learnable query* mechanism following Pan et al. (2025). Specifically, we extract the final hidden states of the learnable query tokens from the MLLM, together with the final hidden states of the text tokens. In this setting, we train the learnable queries, the MLP connector, and the MMDiT, while keeping the MLLM frozen.

`UniVideo` (query) can be more computationally efficient at inference time due to its fixed number of query tokens. By default, we use 512 learnable queries, which reduces the number of conditioning tokens compared to using all the multimodal hiddens from the MLLM. This efficiency gain is particularly beneficial for tasks where video inputs dominate the MMDiT conditioning, such as video editing.

During training, however, the *`UniVideo` (query)* variant requires backpropagation through the MLLM computation graph in order to optimize the learnable queries, which incurs additional memory overhead. In contrast, the *`UniVideo` (hidden)* variant does not require gradient flow through the MLLM and is therefore more memory-efficient during training.

Unless otherwise specified, we report results using the *`UniVideo` (hidden)* variant throughout this paper.

3.3 Main Results

3.3.1 Visual Understanding and Generation

`UniVideo` 's visual understanding is powered by a frozen pretrained MLLM. Freezing the MLLM preserves its strong native understanding ability and prevents performance degradation from joint training with generative tasks. As shown in Table 2, `UniVideo` achieves competitive scores of 83.5 on MMBench (Liu et al., 2024d), 58.6 on MMMU(Yue et al., 2024), and 66.6 on MM-Vet(Yu et al., 2023) for understanding tasks. At the same time, it retains strong generation ability, supporting both I2V and T2V within a single unified model. In contrast, baseline models rely on different variants for different tasks, whereas `UniVideo` reaches performance comparable to the HunyuanVideo backbone on the VBench(Huang et al., 2024) benchmarks.

Table 2: Quantitative comparison on **Visual Understanding and Video Generation**. Best results are shown in **bold**, and second-best are underlined. *We report understanding task results for `UniVideo` using the MLLM component — Qwen-2.5VL-7B results.

Model	Understanding			Video Generation
	MMB	MMMU	MM-Vet	Vbench T2V
Video Understanding Model				
LLaVA-1.5(Liu et al., 2024a)	36.4	**67.8**	36.3	×
LLaVA-NeXT(Liu et al., 2024b)	79.3	51.1	57.4	×
Video Generation Model				
CogVideoX(T2V/I2V)	×	×	×	81.61
I2VGen-XL	×	×	×	×
HunyuanVideo(T2V/I2V)	×	×	×	83.24
Step-Video-(T2V/TI2V)	×	×	×	81.83
Wan2.1(T2V/I2V)	×	×	×	**84.70**
Unified Understanding & Generation Model				
Emu3 (Wang et al., 2024b)	58.5	31.6	37.2	80.96
TokenFlow-XL (Qu et al., 2025)	76.8	43.2	48.2	×
Janus (Wu et al., 2025b)	69.4	30.5	34.3	×
JanusFlow (Ma et al., 2025b)	74.9	29.3	30.9	×
OmniGen2 (Wu et al., 2025c)	79.1	53.1	61.8	×
Show-o (Xie et al., 2024)	-	26.7	-	×
BAGEL (Deng et al., 2025)	**85.0**	55.3	**67.2**	×
Show-o2 (Xie et al., 2025)	79.3	48.9	56.6	81.34
UniVideo *	83.5	58.6	66.6	82.58

Figure 7: **Qualitative comparison** of `UniVideo` with SoTA Task Specific Experts on **In Context Generation** and **In Context Editing** tasks.

3.3.2 In-Context Video Generation

Benchmark: Following FullDiT (Ju et al., 2025b) and OmniGen2 (Wu et al., 2025c), we construct a test set covering both single-ID and multi-ID video generation scenarios. In the single-ID setting, a subject may have multiple reference images (e.g., different viewpoints of a person or object). In the multi-ID setting, the references include 2–4 distinct identities. Details are provided in the Appendix.

Metrics: We conduct both human evaluations and automatic metric assessments. For human evaluation, we follow the protocols of Instruct-Imagen (Hu et al., 2024) and OmniGen2 (Wu et al.,

Table 3: Quantitative comparison on **In-Context Generation**. Human evaluation includes Subject Consistency (SC), Prompt Following (PF), and Overall Video Quality (VQ). Automatic metrics measure video quality in terms of Smoothness and Aesthetics. Best results are shown in **bold**, and second-best are <u>underlined</u>. `UniVideo` achieves superior or competitive performance across all metrics compared to the SoTA methods and commercial models and in particular be the best for SC.

Model	Human Eval Score			Automatic Video Quality Score	
Single Reference Generation					
	SC↑	PF↑	VQ↑	Smoothness↑	Aesthetic↑
VACE	0.31	0.65	0.42	0.922	5.426
Kling1.6	<u>0.68</u>	**0.95**	<u>0.88</u>	<u>0.938</u>	**5.896**
Pika2.2	0.45	0.43	0.15	0.928	5.125
`UniVideo`	**0.88**	<u>0.93</u>	**0.95**	**0.943**	<u>5.740</u>
Multi Reference (≥ 2) Generation					
	SC↑	PF↑	VQ↑	Smoothness↑	Aesthetic↑
VACE	0.48	<u>0.53</u>	0.48	0.53	<u>5.941</u>
Kling1.6	<u>0.73</u>	0.45	**0.95**	0.916	6.034
Pika2.2	0.71	0.48	0.43	0.898	5.176
`UniVideo`	**0.81**	**0.75**	<u>0.85</u>	**0.942**	**6.128**

2025c) to perform a systematic study. Each sample is rated by at least three annotators on (i) subject consistency (SC), (ii) prompt following (PF), and (iii) overall video quality (VQ). Scores in each category are drawn from $\{0, 0.5, 1\}$, where 0 indicates inconsistency or extremely poor quality, and 1 indicates full consistency or high quality. For automatic evaluation, we adopt three metrics from VBench (Huang et al., 2024): smoothness, and aesthetics.

Baselines: We compare `UniVideo` with the state-of-the-art open-source model VACE, given the scarcity of video models capable of in-context generation. We also include commercial baselines such as Pika2.2 and Kling1.6.

Results: Quantitative comparisons are presented in Table 3. `UniVideo` achieves superior or competitive performance across all metrics compared to the baselines. Additional results are shown in Figure 7, and more examples are available on our project website. Notably, baseline models often struggle with complex instructions involving multiple identities (e.g., when the number of reference images is 4), whereas `UniVideo` can accurately follow instructions while preserving identity.

3.3.3 In-Context Video Editing

Benchmark: Following UNIC (Ye et al., 2025b), we construct a test set covering four editing types: swap, delete, addition, and style transfer. Each example consists of a source video and a reference image, together with a natural language instruction. Further details are provided in the Appendix.

Metrics: We adopt the evaluation protocol of UNIC (Ye et al., 2025b) and conduct automatic metric assessments. Specifically, we use CLIP-I and DINO-I to measure identity consistency, and CLIP-Score to measure prompt following.

Baselines: We compare `UniVideo` with state-of-the-art task-specific expert models, including UNIC, AnyV2V, and VideoPainter. We also evaluate against commercial models such as Pika2.2 and Kling1.6. **Note** that all baseline models require explicit mask inputs to localize editing regions and guide generation, whereas `UniVideo` operates without masks.

Results: Quantitative comparisons are presented in Table 4. Although `UniVideo` is evaluated under the more challenging mask-free setting, it still achieves superior or competitive performance across all metrics compared to the baselines. Additional results are shown in Figure 7, and further examples are provided on our project website. `UniVideo` can accurately follow instructions while preserving the identity of the reference images.

Table 4: Quantitative comparison with task-specific expert models on **In-Context Video Editing**. Our model is the **only mask-free approach**, capable of performing edits solely based on instructions without requiring explicit mask inputs to indicate editing regions. Despite this more challenging setting, it achieves superior or competitive performance across all metrics compared to state-of-the-art task-specific expert baselines. Best scores are shown in **bold**, and second-best are <u>underlined</u>.

Model	In Context Insert				
	Identity		Alignment	Video Quality	
	CLIP-I↑	DINO-I↑	CLIP-score↑	Smoothness↑	Aesthetic↑
VACE	0.513	0.105	0.103	0.947	5.693
UNIC	0.598	0.245	0.216	<u>0.961</u>	5.627
Kling1.6	0.632	0.287	0.246	**0.993**	<u>5.798</u>
Pika2.2	<u>0.692</u>	**0.399**	<u>0.253</u>	0.951	5.591
UniVideo *(Mask Free)*	**0.693**	<u>0.398</u>	**0.259**	0.943	**6.031**
Model	**In Context Swap**				
	Identity		Alignment	Video Quality	
	CLIP-I↑	DINO-I↑	CLIP-score↑	Smoothness↑	Aesthetic↑
VACE	0.703	0.391	0.218	0.960	5.961
UNIC	<u>0.725</u>	<u>0.429</u>	<u>0.242</u>	0.971	<u>6.056</u>
Kling1.6	0.707	**0.437**	0.211	**0.995**	6.042
Pika2.2	0.704	0.406	0.211	0.967	5.097
AnyV2V	0.605	0.229	0.218	0.917	4.842
UniVideo *(Mask Free)*	**0.728**	0.427	**0.244**	<u>0.973</u>	**6.190**
Model	**In Context Delete**				
	Video Reconstruction		Alignment	Video Quality	
	PSNR↑	RefVideo-CLIP↑	CLIP-score↑	Smoothness↑	Aesthetic↑
VACE	<u>20.601</u>	0.874	0.206	0.968	**5.637**
UNIC	19.171	0.817	**0.217**	0.970	5.493
Kling1.6	15.476	<u>0.888</u>	0.208	**0.998**	4.965
AnyV2V	19.504	0.869	0.205	0.964	5.325
VideoPainter	**22.987**	**0.920**	0.212	0.957	5.403
UniVideo *(Mask Free)*	17.980	<u>0.888</u>	<u>0.214</u>	<u>0.971</u>	<u>5.498</u>
Model	**In Context Stylization**				
	Style & Content		Alignment	Video Quality	
	CSD-Score↑	ArtFID↓	CLIP-score↑	Smoothness↑	Aesthetic↑
AnyV2V	0.207	43.299	0.195	0.937	4.640
StyleMaster	**0.306**	38.213	0.188	<u>0.952</u>	<u>5.121</u>
UNIC	0.197	**36.198**	<u>0.215</u>	0.932	5.045
UniVideo *(Mask Free)*	<u>0.228</u>	<u>37.877</u>	**0.226**	**0.963**	**6.281**

3.4 Model Analysis

3.4.1 Zero Shot Generalization

We observed two type of generalization ability of UniVideo. Although the training data of UniVideo does not include general free-form video editing tasks, it transfers this ability from diverse image editing data and in-context video editing data (limited to ID deletion, swapping, addition, and stylization) to the video domain, enabling it to handle free-form video editing instructions(e.g., changing material or environment). Surprisingly, we find that UniVideo can perform tasks such as changing materials of character. We also observe that UniVideo is capable of handling task compositions. It can combine in-context editing with style transfer, or perform multiple edits simultaneously (e.g., deleting one identity while adding another). Demonstrations in Figure 8.

3.4.2 Visual Prompt Understanding

We demonstrate the results of visual prompting with UniVideo in Figure 9. We consider two types of visual prompts. In the first setting, users draw reference images and story plans on a canvas. Here, the model can interpret the plan and generate corresponding videos. In the second setting, annotations are drawn directly on an input image, which the model treats as an I2V task; in this case, UniVideo can interpret the motion or new events described by the visual prompt. These results highlight the advantages of UniVideo in handling complex multimodal instructions. Although the

Figure 8: **Zero-Shot Generalization**. We demonstrate two type of generalization. (i) `UniVideo` was not trained on General Free-form Video Editing data. It transfers this ability from diverse image editing data to the video domain through joint training with in-context video generation and editing data (limited to ID deletion, swapping, addition, and stylization), enabling it to handle previously unseen video editing instructions. (ii) `UniVideo` can also generalize to novel task compositions, even though it was not explicitly trained on such compositions.

Figure 9: **Qualitative results of `UniVideo` with visual prompt inputs.** We illustrate two types of visual prompts: in the first three examples, annotations are drawn on a canvas, while in the last example, the annotation is drawn directly on an input image.

qualitative results are obtained in a zero-shot setting, future end-to-end training on task-specific data may further improve performance.

3.5 Ablation Study

Our ablation studies address the following two questions: (i) *Does multi-task learning enhance performance compared with single-task learning?* (ii) *Is our two-branch design effective? Specifically, should visual embeddings be streamed to both the MLLM and MMDiT branches?*

We conduct human evaluations on In-Context Video Editing and In-Context Video Generation, using the same evaluation protocol as in subsubsection 3.3.2. (i) To study multi-task learning, we compare

UniVideo with a single-task baseline. The single-task baseline shares the same architecture as UniVideo but requires an independent model for each task and has access only to task-specific data. Results in Table 5 demonstrate the effectiveness of multi-task learning, especially for the editing task, where UniVideo benefits from large-scale image editing data during joint learning. (ii) To evaluate the impact of streaming visual inputs, we compare UniVideo with a variant that share the same architecture: - **w/o visual for MMDiT:** visual inputs are fed only to the MLLM branch. As shown in Table 6, feeding visual inputs exclusively to the MLLM results in a dramatic drop in identity preservation.

Table 5: Ablation study on UniVideo and single-task model across different in-context tasks.

		Single-task model			UniVideo		
		PF↑	SC↑	VQ↑	PF↑	SC↑	VQ↑
In-context generation	singleid	0.85	0.83	0.93	0.93	0.88	0.95
	multiid	0.75	0.79	0.73	0.75	0.81	0.85
In-context editing	insert	0.81	0.85	0.86	0.92	0.92	0.91
	swap	0.53	0.78	0.68	0.91	0.85	0.85
	delete	0.32	0.42	0.89	0.52	0.58	0.92
	stylization	0.56	0.43	0.63	0.79	0.64	0.64
Average		0.64	0.67	0.79	0.80 (+0.16)	0.78 (+0.11)	0.85 (+0.06)

Table 6: Ablation study on UniVideo and UniVideo w/o Visual for MMDIT.

		UniVideo			UniVideo w/o Visual for MMDIT		
		PF↑	SC↑	VQ↑	PF↑	SC↑	VQ↑
In-context generation	singleid	0.93	0.88	0.95	0.75	0.32	0.86
	multiid	0.75	0.81	0.85	0.81	0.23	0.83
In-context editing	insert	0.92	0.92	0.91	0.68	0.18	0.75
	swap	0.91	0.85	0.85	0.63	0.15	0.62
	delete	0.52	0.58	0.92	0.21	0.13	0.63
	stylization	0.79	0.64	0.64	0.86	0.11	0.57
Average		0.80	0.78	0.85	0.66	0.18	0.71

4 Related Work

Unified Multimodal Understanding and Generation. Recent progress in multimodal generation has been driven primarily by the text and image domains. Autoregressive models such as LlamaGen, Chameleon, Emu2, and Emu3(Sun et al., 2024a; Team, 2024; Sun et al., 2024b; Wang et al., 2024b) adopt discrete token prediction. Hybrid approaches like Show-o, Transfusion, and DreamLLM (Xie et al., 2024; Zhou et al., 2024; Dong et al., 2023) integrate autoregression with diffusion for image synthesis. Regression- or instruction-tuning–based methods, including SEED-X, Janus, MetaMorph, Next-gpt and OmniGen2 (Ge et al., 2024; Wu et al., 2025b; Gupta et al., 2022; Wu et al., 2024; 2025c), adapt LLMs for image feature prediction and controllable generation. Efficiency-oriented designs such as LMFusion and MetaQueries (Shi et al., 2024a; Pan et al., 2025) freeze MLLMs and add lightweight modules or learnable queries, while large-scale pretraining efforts like Show-o2, BLIP3-o, MoGao, and BAGEL (Xie et al., 2025; Chen et al., 2025a; Liao et al., 2025; Deng et al., 2025) demonstrate strong generalization on interleaved multimodal data. Despite these advances, most works remain centered on image understanding and generation. In contrast, we move beyond the image domain by presenting a unified video model. The most related works to ours are Omni-Video and UniVid (Tan et al., 2025; Luo et al., 2025), which primarily focus on the basic text-to-video generation task. However, these approaches do not investigate the potential benefits of a unified architecture—such as how unification can enhance compositional generalization in tasks like in-context editing and in-context generation. In contrast, our work explicitly demonstrates that a unified framework leads to stronger generalization to unseen tasks, highlighting the advantages of architectural unification across diverse understanding and generation scenarios.

Image/Video Generation and Editing. Diffusion models have achieved remarkable success in high-fidelity image synthesis, with systems like Stable Diffusion, DALL·E, and Imagen(Rombach et al., 2022; Podell et al., 2023; Esser et al., 2024; Ramesh et al., 2021; Saharia et al., 2022) establishing strong text-to-image capabilities and recent video diffusion models(Blattmann et al., 2023b; Polyak et al., 2024; Chen et al., 2025c; 2023; Yang et al., 2024; Blattmann et al., 2023a; Kong et al., 2024; Brooks et al., 2024; Ma et al., 2025a) enabling scalable video generation. To improve controllability, models including ControlNet, T2I-Adapter(Zhang et al., 2023b; Mou et al., 2024) introduce external condition modules, while editing frameworks like InstructPix2Pix, EMU-Edit (Brooks et al., 2023; Sheynin et al., 2024) support instruction-driven refinement. Recently, unified image generation has emerged, with OmniGen, OmniControl, and UniReal (Xiao et al., 2025; Tan et al., 2024; Chen et al., 2025d) expanding from generation to reference-guided editing. General editing methods (Wei et al., 2024; Zhao et al., 2024; Liu et al., 2025b; Shi et al., 2024b; Zhang et al., 2023a) further highlight this trend. In contrast, the video domain remains dominated by single-task frameworks such as Video-P2P, MagicEdit, MotionCtrl (Liu et al., 2024c; Liew et al., 2023; Wang et al., 2024c; Liu et al., 2025a). Attempts at unification include AnyV2V(Ku et al., 2024), which requires task-specific pipelines, EditVerse(Ju et al., 2025a), which can not perform visual understanding task. VACE(Jiang et al., 2025), which relies on heavy adapter designs, FullDiT(Ju et al., 2025b), which supports multi-condition video generation but lacks editing, and UNIC(Ye et al., 2025b), which unifies tasks but depends on task-specific condition bias, limiting scalability. Yet, compared to images, unified and flexible video generation and editing remains far less explored. Our work bridges this gap by unifying diverse video tasks under a multimodal instruction framework. We provide the model capabilities comparison in Table 7.

5 Conclusion

We introduce `UniVideo`, a unified multimodal generative model for video understanding, generation, and editing. By integrating an MLLM for semantic understanding with an MMDiT for generation, `UniVideo` combines strong multimodal reasoning with fine-grained visual consistency. It can interpret multimodal instructions and handle diverse tasks effectively. Our experiments show that `UniVideo` not only matches or outperforms task-specific baselines across text/image-to-video, video editing, and in-context generation, but also generalizes to unseen tasks and novel task compositions—capabilities that specialized pipelines struggle to achieve. Beyond robust performance, `UniVideo` can also support visual prompting understanding, underscoring the advantages of unified modeling over fragmented approaches. Looking forward, `UniVideo` opens new directions for multimodal research, advancing us toward assistants that can naturally communicate through language, images, and video.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

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

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22563–22575, 2023b.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18392–18402, 2023.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1:8, 2024.

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.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025a.

Junying Chen, Zhenyang Cai, Pengcheng Chen, Shunian Chen, Ke Ji, Xidong Wang, Yunjin Yang, and Benyou Wang. Sharegpt-4o-image: Aligning multimodal models with gpt-4o-level image generation. *arXiv preprint arXiv:2506.18095*, 2025b.

Shoufa Chen, Chongjian Ge, Yuqi Zhang, Yida Zhang, Fengda Zhu, Hao Yang, Hongxiang Hao, Hui Wu, Zhichao Lai, Yifei Hu, Ting-Che Lin, Shilong Zhang, Fu Li, Chuan Li, Xing Wang, Yanghua Peng, Peize Sun, Ping Luo, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Goku: Flow based video generative foundation models. *arXiv preprint arXiv:2502.04896*, 2025c.

Xi Chen, Zhifei Zhang, He Zhang, Yuqian Zhou, Soo Ye Kim, Qing Liu, Yijun Li, Jianming Zhang, Nanxuan Zhao, Yilin Wang, et al. Unireal: Universal image generation and editing via learning real-world dynamics. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12501–12511, 2025d.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. Dreamllm: Synergistic multimodal comprehension and creation. *arXiv preprint arXiv:2309.11499*, 2023.

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*, 2024.

Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*, 2024.

Agrim Gupta, Linxi Fan, Surya Ganguli, and Li Fei-Fei. Metamorph: Learning universal controllers with transformers. *arXiv preprint arXiv:2203.11931*, 2022.

Hexiang Hu, Kelvin CK Chan, Yu-Chuan Su, Wenhu Chen, Yandong Li, Kihyuk Sohn, Yang Zhao, Xue Ben, Boqing Gong, William Cohen, et al. Instruct-imagen: Image generation with multi-modal instruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4754–4763, 2024.

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.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. *arXiv preprint arXiv:2503.07598*, 2025.

Xuan Ju, Tianyu Wang, Yuqian Zhou, He Zhang, Qing Liu, Nanxuan Zhao, Zhifei Zhang, Yijun Li, Yuanhao Cai, Shaoteng Liu, Daniil Pakhomov, Zhe Lin, Soo Ye Kim, and Qiang Xu. Editverse: Unifying image and video editing and generation with in-context learning. *arXiv preprint arXiv:2509.20360*, 2025a. URL `https://arxiv.org/abs/2509.20360`.

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*, 2025b.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Max Ku, Cong Wei, Weiming Ren, Harry Yang, and Wenhu Chen. Anyv2v: A tuning-free framework for any video-to-video editing tasks. *arXiv preprint arXiv:2403.14468*, 2024.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Chao Liao, Liyang Liu, Xun Wang, Zhengxiong Luo, Xinyu Zhang, Wenliang Zhao, Jie Wu, Liang Li, Zhi Tian, and Weilin Huang. Mogao: An omni foundation model for interleaved multi-modal generation. *arXiv preprint arXiv:2505.05472*, 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.

Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, et al. Uniworld: High-resolution semantic encoders for unified visual understanding and generation. *arXiv preprint arXiv:2506.03147*, 2025.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024a.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llavanext: Improved reasoning, ocr, and world knowledge, 2024b.

Lijie Liu, Tianxiang Ma, Bingchuan Li, Zhuowei Chen, Jiawei Liu, Gen Li, Siyu Zhou, Qian He, and Xinglong Wu. Phantom: Subject-consistent video generation via cross-modal alignment. *arXiv preprint arXiv:2502.11079*, 2025a.

Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8599–8608, 2024c.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025b.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024d.

Jiabin Luo, Junhui Lin, Zeyu Zhang, Biao Wu, Meng Fang, Ling Chen, and Hao Tang. Univid: The open-source unified video model. *arXiv preprint arXiv:2509.24200*, 2025.

Guoqing Ma, Haoyang Huang, Kun Yan, Liangyu Chen, Nan Duan, Shengming Yin, Changyi Wan, Ranchen Ming, Xiaoniu Song, Xing Chen, et al. Step-video-t2v technical report: The practice, challenges, and future of video foundation model. *arXiv preprint arXiv:2502.10248*, 2025a.

Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, et al. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7739–7751, 2025b.

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, 2024.

Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, et al. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025.

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.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K Du, Zehuan Yuan, and Xinglong Wu. Tokenflow: Unified image tokenizer for multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2545–2555, 2025.

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.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pp. 8821–8831. Pmlr, 2021.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

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

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8871–8879, 2024.

Weijia Shi, Xiaochuang Han, Chunting Zhou, Weixin Liang, Xi Victoria Lin, Luke Zettlemoyer, and Lili Yu. Lmfusion: Adapting pretrained language models for multimodal generation. *arXiv preprint arXiv:2412.15188*, 2024a.

Yichun Shi, Peng Wang, and Weilin Huang. Seededit: Align image re-generation to image editing. *arXiv preprint arXiv:2411.06686*, 2024b.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024a.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Emu: Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*, 2023.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14398–14409, 2024b.

Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. *arXiv preprint arXiv:2411.15098*, 2024.

Zhiyu Tan, Hao Yang, Luozheng Qin, Jia Gong, Mengping Yang, and Hao Li. Omni-video: Democratizing unified video understanding and generation. *arXiv preprint arXiv:2507.06119*, 2025.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Shengbang Tong, David Fan, Jiachen Zhu, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. Metamorph: Multimodal understanding and generation via instruction tuning. *arXiv preprint arXiv:2412.14164*, 2024.

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

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024b.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024c.

Cong Wei, Zheyang Xiong, Weiming Ren, Xeron Du, Ge Zhang, and Wenhu Chen. Omniedit: Building image editing generalist models through specialist supervision. In *The Thirteenth International Conference on Learning Representations*, 2024.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025a.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12966–12977, 2025b.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025c.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. In *Forty-first International Conference on Machine Learning*, 2024.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13294–13304, 2025.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *arXiv preprint arXiv:2506.15564*, 2025.

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*, 2024.

Yang Ye, Xianyi He, Zongjian Li, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, and Li Yuan. Imgedit: A unified image editing dataset and benchmark. *arXiv preprint arXiv:2505.20275*, 2025a.

Zixuan Ye, Xuanhua He, Quande Liu, Qiulin Wang, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, Qifeng Chen, and Wenhan Luo. Unic: Unified in-context video editing. *arXiv preprint arXiv:2506.04216*, 2025b.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36:31428–31449, 2023a.

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, 2023b.

Haozhe Zhao, Xiaojian Shawn Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. Ultraedit: Instruction-based fine-grained image editing at scale. *Advances in Neural Information Processing Systems*, 37:3058–3093, 2024.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024.

A Appendix

Appendix contains the following sections:

- Training Details
- Limitation and Future Work
- Training Dataset Construction
- Evaluation Benchmark

B Training Details

We adopt qwen2.5VL-7B (Bai et al., 2025) as the MLLM backbone and HunyuanVideo-T2V-13B (Kong et al., 2024) as the MMDiT backbone. The original HunyuanVideo also uses CLIP as its text encoder; we remove it and instead employ qwen2.5VL as the unified multimodal embedder. The released HunyuanVideo checkpoint is a CFG-distilled model, whose distillation embeddings we discard to simplify the training. To align feature dimensions between qwen2.5VL and HunyuanVideo, we apply an MLP with a $4\times$ expansion. We report training configurations, hyperparameters in Table 1

C Limitation and Future Work

Our model is trained on diverse tasks with multimodal instructions. While we do not observe task confusion, it sometimes fails to strictly follow editing instructions, occasionally over-editing unrelated regions. Due to backbone limitations, the model also struggles to fully preserve the motion of original videos, indicating the need for stronger video backbones. Moreover, although `UniVideo` generalizes to free-form video editing, its success rate remains lower than in image editing, underscoring the greater difficulty of video editing. Future work could explore large-scale video editing datasets and improved backbones for motion fidelity. Additionally, as `UniVideo` represents an assembled multimodal generative system capable of producing images, videos, and text, future work could aim to develop a native multimodal video model trained end-to-end.

D Training Dataset Construction

This section details the construction of our datasets.

D.1 ID-related Tasks

For in-context video generation, which requires identity annotations, we follow the data creation pipeline of ConceptMaster (Huang et al., 2025), including fast elimination of unsuitable videos and fine-grained identity extraction. To generate training data for in-context video editing tasks such as deletion, swap, and insertion, we first apply SAM2 (Ravi et al., 2024) to obtain object segmentation masks from the source video. We then train a video inpainting model to remove the target object while preserving the original background, thereby creating the edited input clip.

D.2 Stylization

Following UNIC (Ye et al., 2025b), Text-to-Video (T2V) models can generate stylized videos with superior quality and stronger fidelity to a given reference style image. So instead of stylizing an existing real video, we first generate a high-quality stylized video using a T2V model. We then transform this stylized video into a realistic counterpart using a video ControlNet model.

D.3 Image and Video

We leverage state-of-the-art image editing models such as FLUX.1 Kontext (Labs et al., 2025) to create diverse image editing data. We also source open source data such as OmniEdit(Wei et al.,

Table 7: Model capabilities across understanding, generation, editing, and in-context generation. ✓indicates support; ✗indicates not supported. The last row is highlighted.

Model	Understanding	Image Gen.	Video Gen.	Image Edit.	Video Edit.	In-context Video Gen.
LLaVA-1.5	✓	✗	✗	✗	✗	✗
SD3-medium	✗	✓	✗	✗	✗	✗
FLUX.1-dev	✗	✓	✗	✗	✗	✗
QwenImage	✓	✓	✗	✓	✗	✗
HunyuanVideo	✗	✓	✗	✗	✗	✗
Show-o	✓	✓	✗	✗	✗	✗
Janus-Pro	✓	✓	✗	✓	✗	✗
Emu3	✓	✓	✗	✓	✗	✗
BLIP3-o	✓	✓	✗	✗	✗	✗
BAGEL	✓	✓	✗	✓	✗	✗
OmniGen2	✓	✓	✗	✗	✗	✗
VACE	✗	✓	✓	✗	✗	✓
UniVideo	✓	✓	✓	✓	✓	✓

2024), ImgEdit(Ye et al., 2025a), and ShareGPT-4o-Image(Chen et al., 2025b). For image and video generation tasks, we utilize internal datasets.

E Evaluation Benchmark

E.1 Visual Understanding and Generation

For the **text-to-video generation task**, we use the prompt suite provided in VBench Huang et al. (2024), which contains 946 prompts covering 16 dimensions, including *subject consistency, background consistency, aesthetic quality, imaging quality, object class, multiple objects, color, spatial relationship, scene, temporal style, overall consistency, human action, temporal flickering, motion smoothness, dynamic degree, appearance style*.

E.2 In-Context Video Generation

For the in-context video generation, we construct a test set consisting of 20 cases, evenly split between single-ID and multi-ID scenarios. For each case, we collect ID images and carefully design prompts to ensure reasonable evaluation. As shown in Fig. 10, we build an ID pool with diverse images, ranging from cartoons to real-world subjects, including humans, animals, and common objects. We then select ID images from this pool and design appropriate prompts for them.

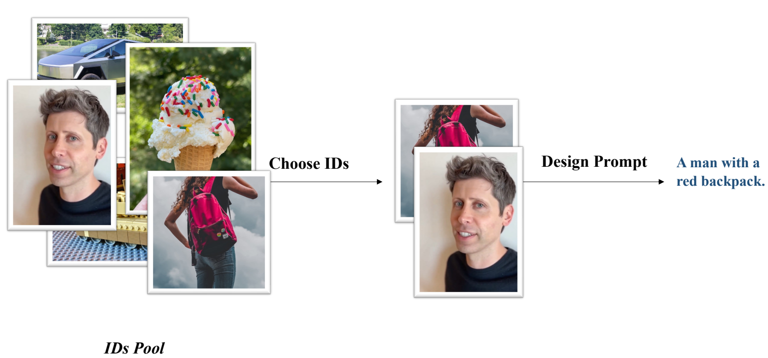


Figure 10: **Construction pipeline of in-context video generation test set.**

The single-ID examples are shown in Fig. 11. The single ID can have either one ID image, as shown by the cat example, or multiple shots of the same ID, as demonstrated by the human example.

As shown in Fig. 12, in the multiple-ID scenarios, the number of IDs in a case ranges from 2 to 4, with larger numbers leading to higher difficulty. Our prompts focus on the interaction between

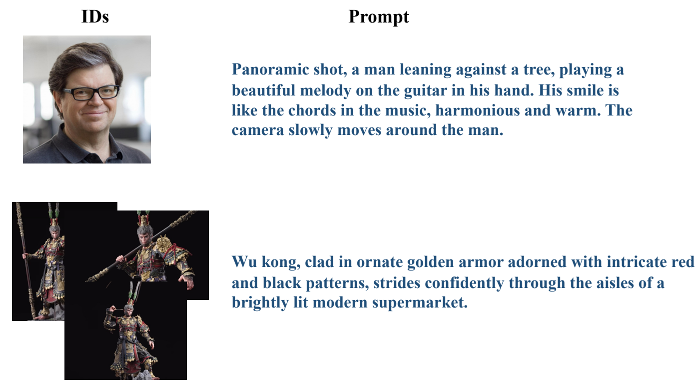


Figure 11: **Example of single-ID test case in in-context video generation test set.**

these ID images and describe the relationships among them. For example, in the first case, the prompt describes a woman sitting on the sofa beside the bag, which connects the woman, sofa, and bag provided in the ID images. In the second case, the relationship between the two characters is described as Psyduck riding Pikachu.

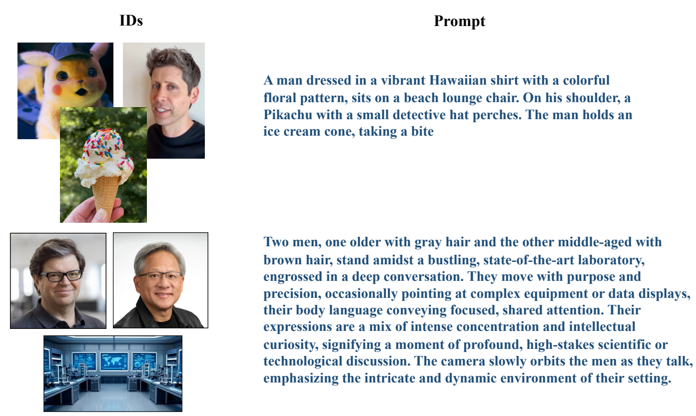


Figure 12: **Example of multi-ID test case in in-context video generation test set.**

E.3 In-Context Video Editing

For the in-context video editing, we evaluate on the UNICBench Ye et al. (2025b) across four tasks: ID Insertion, ID Swap, ID Deletion, and Stylization. Since our setting differs from other video editing models (which may require masks to indicate the edited area, while ours uses instructions instead), we demonstrate in detail how we derive our inputs from the existing video editing benchmark.

First, as shown in Fig. 13, for ID insertion, the elements in UNICBench consist of a reference video, reference ID, and a caption for the target video. The goal of ID insertion is to naturally integrate new objects or elements from the reference ID into the target video. Here we replace the caption with a more direct instruction.

For ID swap, the elements in UNICBench consist of a reference video, mask, reference ID, and a caption for the target video. The goal of ID swap is to replace specific elements in the target video with corresponding elements from the reference ID while preserving the original video's context and motion. In our setting, we don't need a mask to indicate the editing area; instead, we use a more

Figure 13: **Example of ID insertion test case.**

convenient instruction-based approach. For example, in Fig. 14, we simply use the instruction ”Use the man’s face in the reference image to replace the man’s face in the video.”

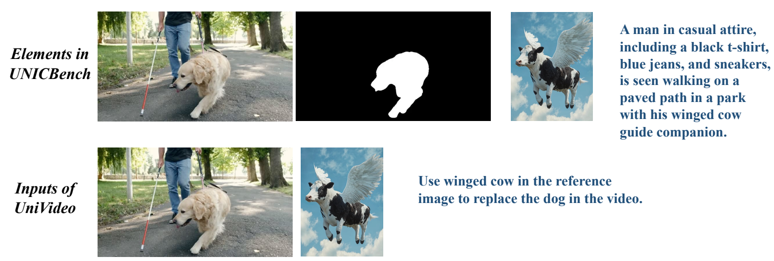


Figure 14: **Example of ID swap test case.**

For ID deletion, UNICBench provides a reference video, mask, and a caption for the target video. ID deletion aims to naturally remove specified objects or elements from the video while maintaining visual consistency and filling the removed areas with appropriate background content. While current video editing methods use masks to specify the object for removal, our approach simplifies this through text instructions. As demonstrated in Fig. 15, we use straightforward prompts such as ”Delete the computer in the video.”

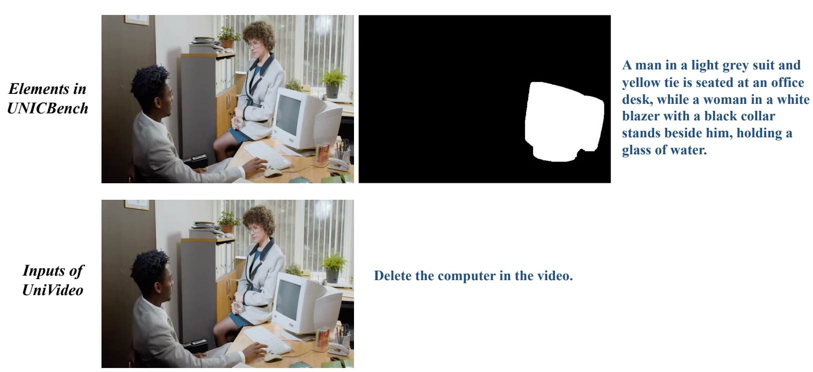


Figure 15: **Example of ID deletion test case.**

For stylization, the existing elements in UNICBench include a style reference image, target caption, and reference video. The purpose of stylization is to transform the visual appearance of the target video to match the artistic style of the reference image while preserving the original video’s content and motion dynamics. We standardize the instruction format to "Transform the video into the style of the reference image," as shown in Fig. 16.

Figure 16: **Example of stylization test case.**