CONFIDENTIAL REVIEW COPY. DO NOT DISTRIBUTE.

How to Make MLLMs More Effective for Unified Video Generation and Editing?

Anonymous submission

Abstract

How can we make multimodal large language models (MLLMs) more effective for unified video generation and editing? MLLMs can interpret multimodal inputs and provide rich semantics for conditioning diffusion transformers (DiTs); existing approaches have explored diverse utilization mechanisms, yet no consensus has emerged on how to leverage them most effectively. To answer this question, we systematically dissect the MLLM–DiT interface from three perspectives: (i) is the visual information in MLLM inputs fully utilized by the DiT? (ii) how do MLLM-provided semantics coordinate with conditions from conventional text encoders such as T5? and (iii) what form of guidance do DiTs prefer from the MLLM: explicit plans or implicit features? Our analysis reveals that existing designs fall short on all three fronts. We therefore propose TANGO, a unified framework for video generation and editing that exploits source visual semantics via source-aware token routing, harmonizes MLLM and T5 conditioning through a two-stage training strategy, and delivers precisely the guidance that DiT generation requires via explicit target semantic planning. TANGO achieves strong performance across public video generation and editing benchmarks, validating the effectiveness of our framework design. This work provides new insights into how MLLMs can more effectively guide visual generation and offers a mechanism-driven perspective on multimodal generative model design.

1. Introduction

Video generation and editing are fundamental capabilities of multimodal generative models. While video generation aims to synthesize temporally coherent content from textual or visual conditions, video editing additionally requires interpreting user instructions, identifying relevant source content, and applying targeted modifications while preserving unrelated details. A unified framework for both tasks therefore demands effective integration of multimodal semantic reasoning and high-quality visual synthesis. A promising direction is to combine multimodal large language models (MLLMs), which can interpret complex instructions and visual inputs, with diffusion transformers (DiTs), which offer powerful video generation capabilities.

Recent MLLM–DiT approaches explore different mechanisms for transferring multimodal information into video generation. Bernini and Omni-Video 2 retain pretrained T5 conditioning alongside MLLM-derived semantic guidance [19, 33]. LoomVideo injects multi-level MLLM features into DiT blocks through cross-attention, replacing conventional text encoding with multimodal conditions [31]. UniVideo and OmniWeaving instead couple MLLM representations with video tokens through joint self-attention [25, 30]. These methods differ in two key respects: whether pretrained T5 conditioning is retained, and through which interface—cross-attention or joint self-attention—the MLLM features reach the DiT. Despite this architectural diversity, all share a common goal: enabling DiTs to fully utilize the visual and semantic information that MLLMs provide.

Incorporating MLLM features into a DiT, however, does not automatically guarantee their effective utilization. For MLLM conditioning to genuinely benefit generation, three requirements must hold: the source visual information encoded by the MLLM must actually reach the generation process; the MLLM semantics must complement, rather than be redundant with, conventional text-encoder conditions; and the guidance must take a form that the DiT can effectively exploit. Whether existing designs meet these requirements remains insufficiently understood, motivating three questions. First, is the visual information in MLLM inputs fully utilized by the DiT? Existing conditioning interfaces may provide only limited interaction between source visual representations and generated video tokens. Second, how can MLLM semantics and conventional text-encoder conditions complement each other? Directly juxtaposing MLLM features with pretrained T5 representations may create conditioning shortcuts that reduce the incentive to learn from the MLLM. Third, is directly transferring MLLM hidden states sufficient, or can explicit target semantic planning provide more effective guidance? Although recent studies have investigated MLLM–DiT fusion and semantic information transfer [7, 17], translating these observations into a coherent architecture remains an open challenge.

To answer these questions, we conduct a systematic mechanism analysis of five representative MLLM–DiT models, combining controlled condition ablations, attention

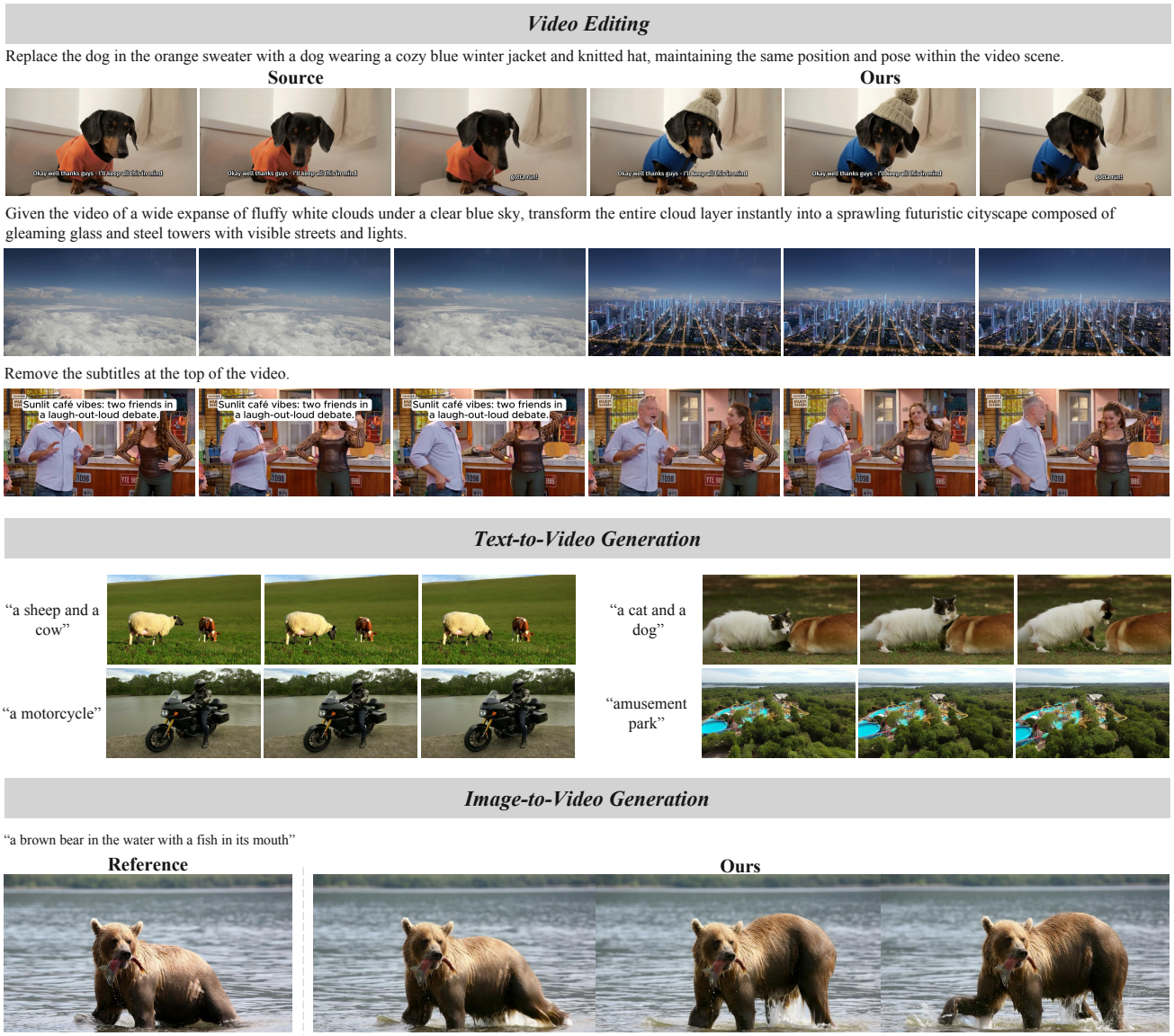


Figure 1. Qualitative results of TANGO for instruction-guided video editing, text-to-video generation, and image-to-video generation. The I2V example includes its reference image.

analysis, representation alignment, and planning comparisons. Our analysis yields three findings: source visual information is insufficiently utilized, MLLM and T5 conditions are partially redundant, and directly transferred hidden states provide only implicit guidance. Each finding translates into a dedicated mechanism in TANGO. Source-aware token routing lets source visual tokens interact with video tokens through joint self-attention, ensuring that visual information reaches the generation process. Progressive two-stage T5 conditioning first establishes an MLLM-driven backbone without T5, then concatenates projected T5 features with MLLM semantic features for shared DiT cross-attention while the backbone remains frozen. Explicit target semantic planning supervises the MLLM to predict the visual features of the desired outcome, giving the DiT well-defined, target-oriented guidance. Concretely, Stage I jointly trains the MLLM and DiT with language modeling, target visual semantic regression, and diffusion objectives; Stage II freezes this backbone and trains only the T5 projector that maps T5 features into the DiT conditioning space.

We evaluate TANGO on text-to-video generation, image-to-video generation, and instruction-guided video editing, using VBench and VBench++ to assess generation quality and semantic consistency [11, 12] and OpenVE-

Bench and IVE-Bench to assess editing performance and instruction adherence [4, 9]. TANGO achieves strong performance across these benchmarks, and ablations on the two-stage training strategy further quantify the contribution of progressive T5 conditioning.

Our contributions are summarized as follows:

- We present a systematic mechanism study of five representative MLLM–DiT models, examining source visual utilization, MLLM–T5 conditioning relationships, and the form of generation guidance, and identifying three limitations in existing designs: underutilized source visual information, redundant conditioning, and implicit guidance.
- Guided by these findings, we develop TANGO, a unified framework for video generation and editing featuring source-aware token routing, progressive two-stage T5 conditioning, and explicit target semantic planning.
- We evaluate TANGO across text-to-video generation, image-to-video generation, and instruction-guided video editing on VBench, VBench++, OpenVE-Bench, and IVE-Bench, demonstrating strong performance and quantifying the contribution of the two-stage training strategy through ablations.

2. Related Work

2.1. Unified Video Generation and Editing

Recent advances have demonstrated remarkable progress in video generation [2, 8, 10, 13, 28, 35] and video editing [1, 5, 16, 18, 20, 21, 24, 29, 36, 37]. As these two directions have advanced, recent work has increasingly explored unified frameworks that support both generation and editing. We focus on these unified approaches and on how MLLM representations can effectively condition video diffusion models.

Unified video generation and editing aim to support text-to-video generation, image-to-video generation, and instruction-guided editing within a shared framework. VACE and UNIC pursue this goal through flexible conditioning and joint attention mechanisms [14, 34]. Other approaches, including Omni-Video, VINO, and UniVideo, investigate multimodal semantic conditioning for video diffusion models [3, 27, 30]. Despite these advances, effectively utilizing the source visual information and semantic guidance available during generation remains challenging.

2.2. MLLM–DiT Conditioning and Fusion

Recent approaches integrate MLLMs with DiTs through diverse conditioning mechanisms. Bernini and Omni-Video 2 combine MLLM-derived semantic guidance with pretrained T5 conditioning [19, 33], while LoomVideo injects multi-level MLLM features through cross-attention [31]; UniVideo and OmniWeaving instead adopt joint self-attention for multimodal integration [25, 30]. Related studies, including MetaQueries and UniVid, further investigate semantic interfaces and cross-modal conditioning [23, 26]. These methods vary along two dimensions—whether pretrained T5 conditioning is retained, and whether MLLM features enter the DiT through cross-attention or joint self-attention. However, how these design choices shape source visual utilization, MLLM–T5 conditioning relationships, and target-oriented semantic planning remains insufficiently understood.

2.3. Conditioning Analysis and Semantic Planning

Recent studies examine how representations and semantics are transferred across the MLLM–DiT interface. Beyond Text Conditioning analyzes intermediate representations and their integration into video diffusion models [7], while What Semantics Survive the Connector? investigates semantic information loss during VLM-to-DiT alignment [17]. Bernini further explores explicit target semantic planning through visual representation prediction [19], and representation similarity measures such as CKA provide tools for analyzing feature relationships [15]. Building on these studies, we systematically investigate source visual utilization, MLLM–T5 conditioning relationships, and target-oriented semantic planning, which motivate the three mechanisms in TANGO: source-aware token routing, progressive two-stage T5 conditioning, and explicit target semantic planning.

3. Diagnosing the MLLM–DiT Interface

Current unified video generation and editing methods typically pair an MLLM with a DiT, but how the DiT actually uses these conditions remains underexplored. We therefore ask three questions that operationalize the requirements introduced in Section 1: (RQ1) is the visual information in MLLM inputs fully utilized by the DiT? (RQ2) are MLLM and T5 conditions complementary or redundant? and (RQ3) does explicit target planning provide more effective guidance than directly transferred hidden states? To answer them, we survey recent MLLM–DiT models for video generation and editing and organize them by their conditioning interface (cross-attention or joint self-attention) and their use of T5, as illustrated in Figure 2.

3.1. RQ1: Is the Visual Information in MLLM Inputs Fully Utilized by the DiT?

3.1.1. How Does Visual Input Affect Generation?

We first examine how the MLLM's visual inputs influence DiT generation. The visual input is an image in image-to-video (I2V) generation and a video in video-to-video (V2V) editing. We zero the MLLM's visual input embeddings, keep all other conditions unchanged, and compare the gen-

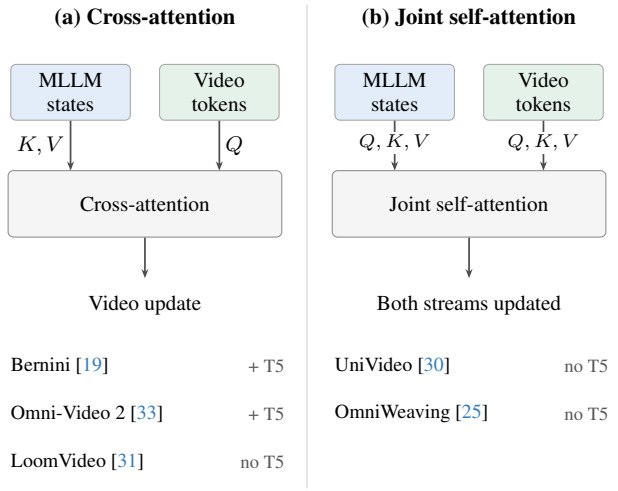


Figure 2. **MLLM–DiT conditioning interfaces.** Models are grouped by attention type and use of T5.

Table 1. **Output change after visual-input ablation.**

Model	LPIPS	CLIP-I
Bernini	**0.1491**	**0.9725**
Omni-Video 2	0.2157	0.9327
LoomVideo	0.4644	0.8402
UniVideo	0.7369	0.5339
OmniWeaving	0.5121	0.7431

erated outputs, measuring LPIPS and CLIP-I between corresponding frames of the original and the intervened outputs. Lower LPIPS and higher CLIP-I indicate greater similarity between the two videos. Table 1 reports the results across the five models.

Across the five models, visual-input ablation changes the generated video far less in the cross-attention group (Bernini, Omni-Video 2, and LoomVideo) than in the joint-attention group (UniVideo and OmniWeaving), suggesting that cross-attention designs exploit the MLLM's visual input only weakly.

3.1.2. Where Does the DiT Read Visual Information?

To investigate why cross-attention models respond weakly to visual-input ablation, we inspect the attention that the DiT pays to MLLM conditions during generation. Figure 3 compares Bernini and OmniWeaving as representatives of the cross-attention and joint self-attention designs, respectively.

In Bernini, visual-position tokens receive roughly 3,000× less attention per token than text-position tokens (0.0005% versus 1.47%), consistent with weak direct reading of visual states. Our qualitative observation of Omni-

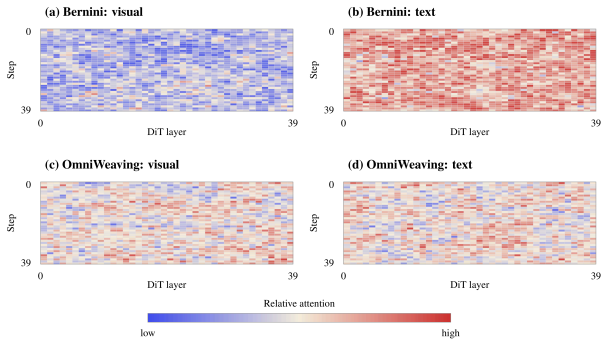


Figure 3. Attention to MLLM conditions in Bernini and OmniWeaving.

Weaving suggests stronger reading of visual positions under joint self-attention. These observations support the hypothesis that joint self-attention better preserves the DiT's access to source visual information, motivating the source-aware token routing in TANGO.

3.2. RQ2: Are MLLM and T5 Conditions Complementary or Redundant?

To examine whether the MLLM and the T5 encoder convey complementary or redundant information, we compare the condition vectors each provides to the DiT, using R^2 for linear predictability and centered kernel alignment (CKA) for representation similarity [15]. As shown in Table 2, Omni-Video 2 exhibits the strongest correspondence ($R^2 = 0.607$, CKA $= 0.829$), indicating substantial redundancy between its MLLM condition and the original instruction encoding. LoomVideo, UniVideo, and OmniWeaving show much weaker correspondence, suggesting more distinct condition spaces. Bernini lies in between in linear predictability ($R^2 = 0.343$), while its low CKA (0.112) indicates that linear predictability and sample-wise structural similarity capture different aspects of alignment. Feature similarity alone, however, does not imply that the two conditions are interchangeable.

The strong alignment of Omni-Video 2 is also condition-dependent: replacing the original instruction with a target caption reduces the correspondence to $R^2 = 0.134$ and CKA $= 0.256$. These observations suggest that MLLM–T5 relationships vary across models and conditions, and that naively juxtaposing the two conditions risks conditioning shortcuts rather than genuine complementarity.

RQ1 and RQ2 together motivate two design choices in TANGO: source-aware token routing, which lets source visual states interact with video tokens through a dedicated joint-attention route while the remaining MLLM states are injected via cross-attention, and progressive two-stage T5 conditioning, which establishes the MLLM–DiT pathway before T5 and adds projected T5 features to the MLLM

Table 2. **MLLM–T5 feature similarity.** The upper two models retain T5 conditioning, whereas the lower three do not. R^2 measures linear predictability and CKA measures sample-wise representation similarity; neither metric alone establishes functional interchangeability.

Model	R^2 (mean $\pm$ std.)	CKA
Bernini	0.343 ± 0.042	0.112
Omni-Video 2	0.607 ± 0.036	0.829
LoomVideo	0.020 ± 0.015	0.210
UniVideo	0.099 ± 0.020	0.185
OmniWeaving	0.063 ± 0.023	0.232

condition sequence for shared cross-attention after the backbone has been learned.

3.3. RQ3: Does Explicit Target Planning Provide More Effective Guidance?

To test whether explicit planning improves the MLLM condition, we train two variants from the same baseline for the same number of steps: one predicts target-ViT features, and the other projects MLLM latent states through an MLP. In both variants, the resulting MLLM condition is concatenated with T5 features for DiT cross-attention. We compare editing performance on OpenVE-Bench [9] (Table 3).

RQ3 demonstrates the benefit of explicit planning in this comparison: the planning variant achieves stronger editing performance than direct latent projection, showing that explicitly specified, target-oriented semantics provide more effective guidance than implicitly transferred features. TANGO therefore adopts explicit target semantic planning, supervising the MLLM to predict the visual features of the desired outcome.

Table 3. **Ablation study of target semantic planning on OpenVE-Bench.** Higher scores indicate better performance.

Configuration	Overall ↑	Local Add ↑	Subtitle Edit ↑
w/o target semantic planning	3.94	3.61	3.76
TANGO	**4.12**	**4.10**	**4.78**

Taken together, these findings motivate TANGO's progression from independent conditioning to complementary fusion. Stage 1 learns source-aware routing and target semantic planning without T5, while Stage 2 freezes this backbone and concatenates projected T5 features with the MLLM semantic condition for cross-attention.

4. Method

Guided by the analyses in Section 3, we propose TANGO, a unified framework for video generation and editing built on source-aware token routing, explicit target semantic planning, and progressive T5 conditioning. TANGO is trained in two stages: Stage I jointly trains the MLLM and DiT to establish an MLLM-driven generation backbone without T5, and Stage II freezes this backbone and learns a T5 projector whose output is concatenated with the MLLM semantic condition for DiT cross-attention. Sections 4.1 and 4.2 detail the two stages, and Section 4.3 describes how the resulting framework supports unified video generation and editing. An overview of TANGO is shown in Figure 4.

4.1. Stage I: MLLM-Driven Generation Backbone

Stage I establishes an MLLM-driven generation backbone without T5. Given a textual instruction and an optional source image or video, the MLLM produces semantic conditions and target-oriented representations that guide DiT generation. Two complementary designs, motivated by RQ1 and RQ3 respectively, improve source visual utilization and target semantic guidance.

Source-Aware Token Routing. Motivated by RQ1, we separate source visual tokens from other MLLM features and route them through different attention mechanisms. Specifically, source visual tokens H_{src} interact directly with video tokens X_l through joint self-attention, while the remaining semantic features H_{sem} are injected through cross-attention:

$$\begin{aligned} \widetilde{X}_l &= \text{JointSelfAttn}_l(X_l, H_{\text{src}}), \\ X_{l+1} &= \text{CrossAttn}_l(\widetilde{X}_l, H_{\text{sem}}). \end{aligned} \tag{1}$$

This separation enables direct interaction between source visual information and the tokens being generated, while cross-attention preserves a dedicated channel for semantic guidance.

Explicit Target Semantic Planning. Motivated by RQ3, we further supervise the MLLM to explicitly predict the desired visual outcome rather than relying solely on directly projected hidden states. A planning head g_ϕ predicts the visual features of the target video, which are aligned with features that a pretrained ViT encoder E_{ViT} extracts from the ground-truth video V_{tgt}:

$$\mathcal{L}_{\text{plan}} = \mathcal{L}_{\text{reg}}\left(g_\phi(H_{\text{MLLM}}), E_{\text{ViT}}(V_{\text{tgt}})\right). \tag{2}$$

where $\mathcal{L}_{\text{reg}}$ denotes a regression loss. This objective encourages the MLLM to encode target-oriented semantics, giving the DiT well-defined guidance aligned with the generation target.

Joint Training. We jointly optimize the MLLM and DiT using language modeling, target semantic planning, and dif-

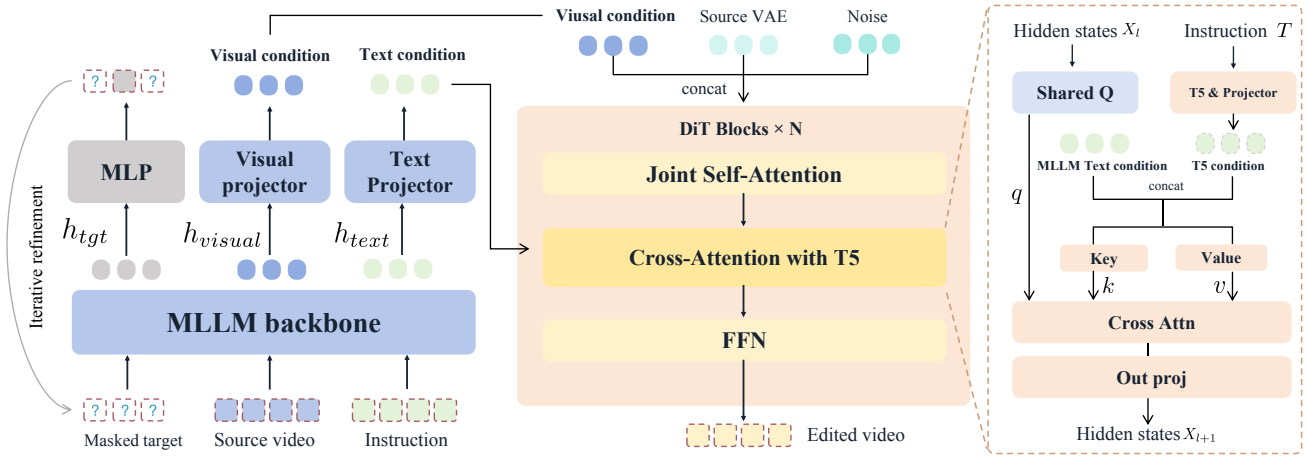


Figure 4. Overview of TANGO. Stage I learns source-aware routing and target semantic planning without T5. In Stage II, projected T5 features are concatenated with MLLM text conditions and read through the same DiT cross-attention module.

fusion objectives:

$$\mathcal{L}_{\text{Stage1}} = \lambda_{\text{LM}}\mathcal{L}_{\text{LM}} + \lambda_{\text{plan}}\mathcal{L}_{\text{plan}} + \lambda_{\text{diff}}\mathcal{L}_{\text{diff}}. \quad (3)$$

By removing T5 throughout Stage I, the backbone learns to generate directly from multimodal conditions, yielding an independent MLLM-driven foundation for the subsequent T5 enhancement.

4.2. Stage II: Concatenated T5 Conditioning

Stage I establishes an MLLM-driven generation backbone without T5. In Stage II, we add the pretrained T5 textual prior through the cross-attention pathway shown in Figure 4. A frozen T5 encoder extracts representations from the input instruction P, and a trainable projector P_ψ maps them into the DiT conditioning space:

$$H_{\text{T5}} = E_{\text{T5}}(P), \qquad \widehat{H}_{\text{T5}} = P_\psi(H_{\text{T5}}). \quad (4)$$

The projected T5 tokens are concatenated with the MLLM semantic tokens along the token dimension:

$$C = [H_{\text{sem}}; \widehat{H}_{\text{T5}}], \qquad X_{l+1} = \text{CrossAttn}_l(\widetilde{X}_l, C). \quad (5)$$

At each DiT layer, the video hidden states provide the queries, while the concatenated MLLM and T5 conditions provide keys and values to the same cross-attention module. This lets the DiT read both condition sources through the interface learned in Stage I.

During Stage II, the MLLM, target planning module, DiT backbone, and T5 encoder are frozen; only the T5 projector is optimized with the diffusion objective:

$$\mathcal{L}_{\text{Stage2}} = \mathcal{L}_{\text{diff}}. \quad (6)$$

This training sequence preserves the learned MLLM–DiT parameters while adding T5 information to the shared conditioning sequence.

4.3. Unified Video Generation and Editing

After the two-stage training, TANGO supports text-to-video generation, image-to-video generation, and instruction-guided video editing within a shared architecture. Given a textual instruction and an optional source image or video, the MLLM produces semantic conditions and target-oriented planning representations. Source visual tokens, when present, interact with video tokens through joint self-attention; the remaining MLLM semantic tokens are concatenated with projected T5 tokens and read through shared cross-attention. These conditions jointly guide the DiT to generate the target video.

5. Experiments

We evaluate TANGO on text-to-video (T2V) generation on VBench [11], image-to-video (I2V) generation on VBench++ [12], and instruction-guided video editing on OpenVE-Bench [9] and IVE-Bench [4]. An ablation study on OpenVE-Bench further quantifies the contribution of the two-stage training strategy.

5.1. Video Generation Evaluation

We assess T2V generation on VBench [11] and I2V generation on VBench++ [12], reporting overall scores together with representative metrics for visual quality, semantic alignment, and temporal consistency.

Text-to-Video Generation. Table 5 compares TANGO with recent T2V models on VBench. Following Bernini [19], we report the total score, quality score, semantic score, aesthetic quality, dynamic degree, object class, and overall consistency. TANGO reaches 84.30 in total score, exceeding the unified models Show-o2 (81.34) and

Table 4. Quantitative comparison and mechanism ablations on OpenVE-Bench. Baseline scores are reported using Gemini 2.5 Pro; TANGO should be ranked against these baselines only when evaluated with the same evaluator and protocol. Bold and underlined values denote the best and second-best results among benchmark methods, respectively; ties receive the same rank.

Method	Editing Categories								
	Overall ↑	Global Style ↑	Background Change ↑	Local Change ↑	Local Remove ↑	Local Add ↑	Subtitle Edit ↑	Creative Edit ↑	Camera Edit ↑
VACE-14B	1.57	1.49	1.55	2.07	1.46	1.26	1.48	1.47	1.62
OmniVideo	1.31	1.11	1.18	1.14	1.14	1.36	1.00	2.26	1.00
InsViE	1.53	2.20	1.06	1.48	1.36	1.17	2.18	2.02	1.09
Ditto	1.98	4.01	1.68	2.03	1.53	1.41	2.81	1.23	1.32
ICVE	2.07	2.22	1.62	2.57	2.51	1.97	2.09	2.41	1.11
Lucy-Edit	2.15	2.27	1.57	3.20	1.75	2.30	1.61	2.86	1.61
OpenVE-Edit	2.49	3.16	2.36	2.98	1.85	2.15	2.91	2.31	2.02
VINO	3.18	4.34	2.54	3.73	3.22	2.77	2.61	3.29	2.81
Bernini	3.96	4.29	3.58	**4.72**	**4.10**	3.35	3.55	3.84	**4.18**
Ablation variants									
w/o source-aware routing	3.98	4.29	3.73	3.84	4.21	3.97	3.65	4.43	3.79
w/o Stage 2 training	3.89	4.18	3.66	3.83	4.08	4.08	3.47	4.66	2.90
w/o target semantic planning	3.94	4.25	3.60	3.92	4.21	3.61	3.76	4.61	3.91
TANGO	**4.12**	**4.56**	**3.66**	3.77	3.32	**4.10**	**4.78**	**4.87**	3.90

Table 5. Quantitative comparison of text-to-video generation on VBench. Total, Quality, and Semantic scores are reported on the original 0–1 scale converted to percentages. Show-o2 and Gen-3 results are from [32]; TUNA results are from [22]. Higher scores indicate better performance; bold and underline mark the best and second-best full-model results in each column.

Method	Total ↑	Quality ↑	Semantic ↑	Aesthetic Quality ↑	Dynamic Degree ↑	Object Class ↑	Overall Consistency ↑
CogVideoX-5B	81.91	83.05	77.33	61.88	69.51	85.07	27.65
Gen-3	82.32	84.11	75.17	63.34	60.14	87.81	26.69
Show-o2	81.34	82.10	78.31	65.15	40.83	94.81	27.00
HunyuanVideo	83.24	85.09	75.82	60.36	70.83	86.10	26.44
Wan2.1-14B	83.69	**85.59**	76.11	66.07	65.46	86.28	25.91
VINO	83.17	83.69	81.08	**68.11**	55.56	91.17	27.00
Wan2.2-A14B	**84.79**	85.33	82.61	67.06	69.72	96.00	27.36
Bernini	84.46	85.20	81.49	65.32	**79.72**	95.08	27.73
TUNA	84.06	84.32	**83.04**	65.88	69.39	95.41	27.71
w/o source-aware routing	81.42	87.40	57.50	66.32	87.50	85.94	26.37
w/o Stage 2 training	62.46	68.10	39.90	52.14	25.00	60.94	21.78
w/o target semantic planning	81.20	83.23	73.09	67.07	37.50	97.66	25.80
TANGO	84.30	85.00	81.54	65.27	62.50	**100.00**	**28.30**

TUNA (84.06), and attains the highest object-class and overall-consistency scores among the listed full models.

The three ablation rows correspond to the official RQ1 variant, the Stage I model without T5 conditioning (RQ2), and direct latent projection instead of target semantic planning (RQ3), respectively. The reported TANGO results use Stage II with concatenated T5 conditioning.

Image-to-Video Generation. Table 6 reports I2V results on VBench++, covering the total score, quality score, I2V score, subject consistency, background consistency, and camera motion, which together measure visual quality and fidelity to the reference image. TANGO achieves the highest total, quality, subject-consistency, and camera-motion scores among the listed models, together with the second-highest background-consistency score.

The reported TANGO configuration uses Stage II with concatenated T5 conditioning and 3000/2000 inference steps.

5.2. Video Editing Evaluation

We evaluate instruction-guided video editing on OpenVE-Bench [9] and IVE-Bench [4], which provide complementary assessments of editing quality, instruction compliance, and source video preservation.

OpenVE-Bench. Table 4 reports the overall score and eight editing categories—global style transfer, background change, local change, local removal, local addition, subtitle editing, creative editing, and camera editing—with baseline results following Bernini [19]. TANGO achieves the highest overall score and leads in five of the eight categories.

IVE-Bench. Table 7 evaluates TANGO on the official short subset of IVE-Bench, reporting the total score and three dimensions: video quality, instruction compliance,

Table 6. Quantitative comparison of image-to-video generation on VBench++. Additional baseline results are from the URSA evaluation [6]. The TANGO values are reported on the original 0–1 scale converted to percentages. Higher scores indicate better performance; bold and underline mark the best and second-best results in each column.

Method	Overall Scores			Detailed Metrics		
	Total ↑	Quality ↑	I2V Score ↑	I2V Subject Consistency ↑	I2V Background Consistency ↑	Camera Motion ↑
DynamiCrafter	86.9	80.5	93.5	97.2	97.4	31.2
Cosmos	84.2	75.8	92.6	96.0	97.4	25.4
VideoMAR	84.8	75.6	94.0	97.9	98.4	21.6
CogVideoX	86.7	78.6	94.8	97.2	96.7	<u>67.7</u>
HunyuanVideo	86.8	78.5	<u>95.1</u>	<u>98.5</u>	97.4	49.9
Wan2.1	86.9	80.8	92.9	97.0	96.4	34.8
Step-Video	<u>88.4</u>	<u>81.2</u>	**95.5**	97.9	98.5	49.2
Pusa	87.3	79.8	94.8	97.6	**99.2**	29.5
URSA	86.2	79.8	92.6	96.1	96.5	37.6
TANGO	**88.80**	**82.60**	95.00	**98.70**	<u>98.93</u>	**87.50**

Table 7. Quantitative comparison on IVE-Bench (Short Subset). Higher scores indicate better performance. Values are displayed to four decimal places. Bold and underlined values denote the best and second-best reported values, respectively; ties receive the same rank.

Method	Total Score ↑	Video Quality ↑	Instruction Compliance ↑	Video Fidelity ↑
VACE	0.6261	<u>0.7983</u>	0.2542	**0.8258**
Omni-Video	0.5865	0.7820	0.4364	0.5411
InsV2V	0.6668	0.7958	0.3861	<u>0.8185</u>
Ditto	**0.6675**	0.7812	<u>0.4908</u>	0.7304
ICVE	0.6033	0.7125	<u>0.4535</u>	0.6437
Lucy-Edit-Dev	0.6353	**0.8209**	0.3394	0.7455
AnyV2V	0.5767	0.7271	0.4172	0.5859
TANGO	<u>0.6652</u>	0.7811	**0.4913**	0.7231

and video fidelity. TANGO is competitive in instruction compliance, but does not lead consistently across all three dimensions.

5.3. Ablation Study of the Three Mechanisms

Table 4 compares TANGO with variants without source-aware routing (RQ1), Stage II T5 conditioning (RQ2), or target semantic planning (RQ3) on OpenVE-Bench. The RQ2 variant is the Stage I MLLM-driven backbone; both it and TANGO use the same test samples and inference settings. Their comparison isolates the contribution of projected T5 features after the MLLM pathway has been established.

Stage II raises the overall score by 0.23. This gain is consistent with the pretrained textual prior complementing the MLLM condition through shared cross-attention.

6. Conclusion

In this work, we investigated how to make MLLMs more effective for unified video generation and editing. Through a systematic diagnosis of the MLLM–DiT interface, we found that existing designs underutilize source visual information, produce partially redundant MLLM and T5 conditions, and rely on implicitly transferred guidance. Guided by these findings, we presented TANGO, a unified framework that addresses each limitation with a dedicated mechanism: source-aware token routing, which ensures that source visual information reaches the generation process; progressive two-stage T5 conditioning, which concatenates projected T5 features with the established MLLM condition for shared cross-attention; and explicit target semantic planning, which supplies the DiT with well-defined, target-oriented guidance. Experiments on VBench, VBench++, OpenVE-Bench, and IVE-Bench demonstrate competitive performance across text-to-video generation, image-to-video generation, and instruction-guided video editing. On VBench++, TANGO achieves the highest total, quality, and camera-motion scores among the compared methods; on OpenVE-Bench, it attains the highest overall editing score. These results show how distinct roles for source evidence, target intent, and textual priors can strengthen a shared MLLM–DiT interface for generation and editing.

LLM Usage Statement

Generative AI tools, including ChatGPT (OpenAI), assisted with manuscript writing, language editing, and the presentation of technical content. The authors reviewed and verified the text and take full responsibility for the scientific claims, experimental results, and references.

References

[1] Qingyan Bai, Qiuyu Wang, Hao Ouyang, Yue Yu, Hanlin Wang, Wen Wang, Ka Leong Cheng, Shuailei Ma, Yanhong Zeng, Zichen Liu, Yinghao Xu, Yujun Shen, and Qifeng Chen. Scaling instruction-based video editing with a high-quality synthetic dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 37971–37981, 2026. 3

[2] Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, et al. SkyReels-V2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025. 3

[3] Junyi Chen, Tong He, Zhoujie Fu, Pengfei Wan, Kun Gai, and Weicai Ye. VINO: A unified visual generator with interleaved omnimodal context. *arXiv preprint arXiv:2601.02358*, 2026. 3

[4] Yinan Chen, Jiangning Zhang, Teng Hu, Yuxiang Zeng, Zhucun Xue, Qingdong He, Chengjie Wang, Yong Liu, Xiaobin Hu, and Shuicheng Yan. IVEBench: Modern benchmark suite for instruction-guided video editing assessment. *arXiv preprint arXiv:2510.11647*, 2025. ICLR 2026. 3, 6, 7

[5] Xiaoyan Cong, Haotian Yang, Angtian Wang, Yizhi Wang, Yiding Yang, Canyu Zhang, and Chongyang Ma. VIVA: VLM-guided instruction-based video editing with reward optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 34364–34374, 2026. 3

[6] Haoge Deng, Ting Pan, Fan Zhang, Yang Liu, Zhuoyan Luo, Yufeng Cui, Wenxuan Wang, Chunhua Shen, Shiguang Shan, Zhaoxiang Zhang, and Xinlong Wang. Uniform discrete diffusion with metric path for video generation. In *International Conference on Learning Representations*, 2026. 8

[7] Yanbo Ding, Yijia Fan, Caihua Shan, Yifan Yang, Yifei Shen, Weijie Wang, Xirui Hu, Dongsheng Li, Lili Qiu, Yuqing Yang, and Yali Wang. Beyond text conditioning: A systematic study of MLLM-DiT fusion for video generation. *arXiv preprint arXiv:2608.14043*, 2026. 1, 3

[8] Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, et al. LTX-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026. 3

[9] Haoyang He, Jie Wang, Jiangning Zhang, Zhucun Xue, Xingyuan Bu, Qiangpeng Yang, Shilei Wen, and Lei Xie. OpenVE-3M: A large-scale high-quality dataset for instruction-guided video editing. *arXiv preprint arXiv:2512.07826*, 2025. 3, 5, 6, 7

[10] Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. 3

[11] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 2, 6

[12] Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026. 2, 6

[13] Longbin Ji, Xiaoxiong Liu, Junyuan Shang, Shuohuan Wang, Yu Sun, Hua Wu, and Haifeng Wang. VideoAR: Autoregressive video generation via next-frame & scale prediction. *arXiv preprint arXiv:2601.05966*, 2026. 3

[14] 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. 3

[15] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, pages 3519–3529, 2019. 3, 4

[16] Sen Liang, Fengbin Guan, Youliang Zhang, Xin Li, and Zhibo Chen. CoT-Edit: Let CoT guide instruction video editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 37960–37970, 2026. 3

[17] Hangyu Lin, Chao Wen, Chengming Xu, Jianxiong Gao, Jiangning Zhang, Xiaobin Hu, and Yanwei Fu. What semantics survive the connector? diagnosing VLM-to-DiT alignment in video editing. *arXiv preprint arXiv:2605.20795*, 2026. 1, 3

[18] Yiqi Lin, Guoqiang Liang, Ziyun Zeng, Zechen Bai, Yanzhe Chen, and Mike Zheng Shou. Kiwi-Edit: Versatile video editing via instruction and reference guidance. *arXiv preprint arXiv:2603.02175*, 2026. 3

[19] Chenchen Liu, Junyi Chen, Lei Li, Lu Chi, Mingzhen Sun, Zhuoying Li, Yi Fu, Ruoyu Guo, Yiheng Wu, Ge Bai, and Zehuan Yuan. Bernini: Latent semantic planning for video diffusion. *arXiv preprint arXiv:2605.22344*, 2026. 1, 3, 4, 6, 7

[20] Qi Liu, Gang Yue, Mingyu Yin, Lisai Zhang, Yidi Wu, Yaole Wang, Yaohui Wang, Chang Yao, Jingyuan Chen, and Lin Ma. TIDE: Task-isolated diffusion for unified video editing and generation. *arXiv preprint arXiv:2606.08260*, 2026. 3

[21] Xinyu Liu, Shihao Li, Weihong Lin, Xinlong Chen, Yang Shi, Yujin Han, Yiyang Cai, Yanghao Wang, Ruibin Yuan, Yuanxing Zhang, Pengfei Wan, Wenhan Luo, and Yike Guo. ReBind: Multi-reference video editing via structured instructions with explicit reference relationships. *arXiv preprint arXiv:2607.14681*, 2026. 3

[22] Zhiheng Liu, Weiming Ren, Haozhe Liu, Zijian Zhou, Shoufa Chen, Haonan Qiu, Xiaoke Huang, Zhaochong An, Fanny Yang, Aditya Patel, Viktar Atliha, Tony Ng, Xiao Han, Chuyan Zhu, Chenyang Zhang, Ding Liu, Juan-Manuel Perez-Rua, Sen He, Jürgen Schmidhuber, Wenhu Chen, Ping Luo, Wei Liu, Tao Xiang, Jonas Schult, and Yuren Cong. TUNA: Taming unified visual representations for native unified multimodal models. *arXiv preprint arXiv:2512.02014*, 2025. 7

[23] Jiabin Luo, Junhui Lin, Zeyu Zhang, Biao Wu, Meng Fang, Ling Chen, and Hao Tang. Univid: The open-source unified video model, 2025. 3

[24] Jinjie Mai, Chaoyang Wang, Gordon Guocheng Qian, Willi Menapace, Sergey Tulyakov, Bernard Ghanem, Peter Wonka, and Ashkan Mirzaei. EasyV2V: A high-quality instruction-based video editing framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 30435–30445, 2026. 3

[25] Kaihang Pan, Qi Tian, Jianwei Zhang, Weijie Kong, Jiangfeng Xiong, Yanxin Long, Shixue Zhang, Haiyi Qiu, Tan Wang, Zheqi Lv, Yue Wu, Liefeng Bo, Siliang Tang, and Zhao Zhong. Omniweaving: Towards unified video generation with free-form composition and reasoning. *arXiv preprint arXiv:2603.24458*, 2026. 1, 3, 4

[26] Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, Ji Hou, and Saining Xie. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025. 3

[27] Zhiyu Tan, Hao Yang, Luozheng Qin, Jia Gong, Mengping Yang, and Hao Li. Omni-video: Democratizing unified video understanding and generation, 2025. 3

[28] Team Wan, Ang Wang, Baole Ai, Chaojie Mao, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 3

[29] Shengzhi Wang, Yingkang Zhong, Jiangchuan Mu, Kai Wu, Mingliang Xiong, Wen Fang, Mingqing Liu, Hao Deng, Bin He, Gang Li, and Qingwen Liu. Align-a-video: Deterministic reward tuning of image diffusion models for consistent video editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2074–2083, 2025. 3

[30] Cong Wei, Quande Liu, Zixuan Ye, Qiulin Wang, Xintao Wang, Pengfei Wan, Kun Gai, and Wenhu Chen. Univideo: Unified understanding, generation, and editing for videos. *arXiv preprint arXiv:2510.08377*, 2025. ICLR 2026. 1, 3, 4

[31] Jianzong Wu, Hao Lian, Jiongfan Yang, Dachao Hao, Ye Tian, Yunhai Tong, Jingyuan Zhu, Biaolong Chen, Qiaosong Qi, Aixi Zhang, Wanggui He, Mushui Liu, Jinlong Liu, and Hao Jiang. Loomvideo: Unifying multimodal inputs into video generation and editing. *arXiv preprint arXiv:2606.06042*, 2026. 1, 3, 4

[32] Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. In *Advances in Neural Information Processing Systems*, 2025. 7

[33] Hao Yang, Zhiyu Tan, Jia Gong, Luozheng Qin, Hesen Chen, Xiaomeng Yang, Yuqing Sun, Yuetan Lin, Mengping Yang, and Hao Li. Omni-video 2: Scaling MLLM-conditioned diffusion for unified video generation and editing. *arXiv preprint arXiv:2602.08820*, 2026. 1, 3, 4

[34] 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, 2025. 3

[35] Zangwei Zheng, Xiangyu Peng, Yuxuan Lou, Chenhui Shen, Tom Young, et al. Open-Sora 2.0: Training a commercial-level video generation model in $200k. *arXiv preprint arXiv:2503.09642*, 2025. 3

[36] Donghao Zhou, Haoyang He, Fan Zhang, Hao Yang, Guisheng Liu, Xin Gao, Zhongwei Wan, Xingyuan Bu, Jie Wang, Qiangpeng Yang, Shilei Wen, Chi-Wing Fu, and Pheng-Ann Heng. ThinkV2V: Unleashing the reasoning capability of MLLMs for instruction-guided video editing. *arXiv preprint arXiv:2609.38541*, 2026. 3

[37] Yixuan Zhu, Haolin Wang, Shilin Ma, Wenliang Zhao, Yansong Tang, Lei Chen, and Jie Zhou. FADE: Frequency-aware diffusion model factorization for video editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28426–28435, 2025. 3