OMNIWEAVING: TOWARDS UNIFIED VIDEO GENERATION WITH FREE-FORM COMPOSITION AND REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

While proprietary systems such as Seedance have achieved remarkable success in omni-capable video generation, the academic research community lags far behind: most of its models remain heavily fragmented, and the few efforts toward unified video generation still struggle to seamlessly integrate diverse tasks within a single framework. To bridge this gap, we propose OmniWeaving, an omni-level video generation model featuring powerful multimodal composition and reasoning-informed capabilities. By leveraging a massive-scale pretraining dataset that encompasses diverse compositional and reasoning-augmented scenarios, OmniWeaving learns to temporally bind interleaved text, multi-image, and video inputs while acting as an intelligent agent to infer complex user intentions for sophisticated video creation. Furthermore, we introduce IntelligentVBench, the first comprehensive benchmark designed to rigorously assess next-level intelligent unified video generation. Extensive experiments demonstrate that OmniWeaving achieves SoTA performance among open-source academic unified models. Project page: `https://anonymous.4open.science/r/omniweaving-24E4`.

1 INTRODUCTION

As unified visual generation comes into full bloom in the image domain on both the proprietary (OpenAI et al., 2024; Google, 2025a) and academic (Deng et al., 2025a; Wu et al., 2025c) fronts, video generation is now surfacing the same demand for unification, with the first answers coming from industry. Recently, models such as the proprietary Seedance-2.0 (Seed, 2026) and MiniMax-H3 (MiniMax, 2026) have drastically reshaped the landscape, dictating that next-generation models must be genuinely "omni-capable" through a synergy of two foundational capabilities: **(1) multimodal composition** requires systems to allow for the seamless spatio-temporal binding of free-form interleaved text, images, and video inputs. **(2) abstract reasoning** mandates that models act as active agents, capable of inferring complex user intentions and mastering the underlying semantic logic of a scene.

However, academic progress in unified video generation significantly lags behind industrial systems, revealing a substantial capability gap. First, the current landscape of video generation remains dominated by fragmented approaches (Wu et al., 2025a; Wan et al., 2025; Kong et al., 2024) narrowly tailored for text-to-video, image-to-video synthesis or video-to-video synthesis (He et al., 2025), relying on task-specific modules that impede scaling and integration. Furthermore, while recent academic models such as VACE (Jiang et al., 2025), UniVideo (Wei et al., 2025), and VINO (Chen et al., 2026b) attempt to unify video generation tasks, they either focus on basic task combinations or fail to leverage deep visual understanding to drive unified generation, and thus still struggle with multimodal composition and reasoning-informed video synthesis.

We argue that bridging this substantial capability gap of unified video models relies on three key drivers. **First**, the model architecture must integrate both visual comprehension and generation into a single framework to explicitly activate abstract reasoning, evolving models from passive renderers into "thinking-guided" generators. **Second**, a transition toward free-form, multi-task pretraining is essential to move beyond rigid text-video pairs and capture the intricate semantic relationships across diverse modalities. **Finally**, since existing benchmarks are largely limited to simplistic tasks with

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

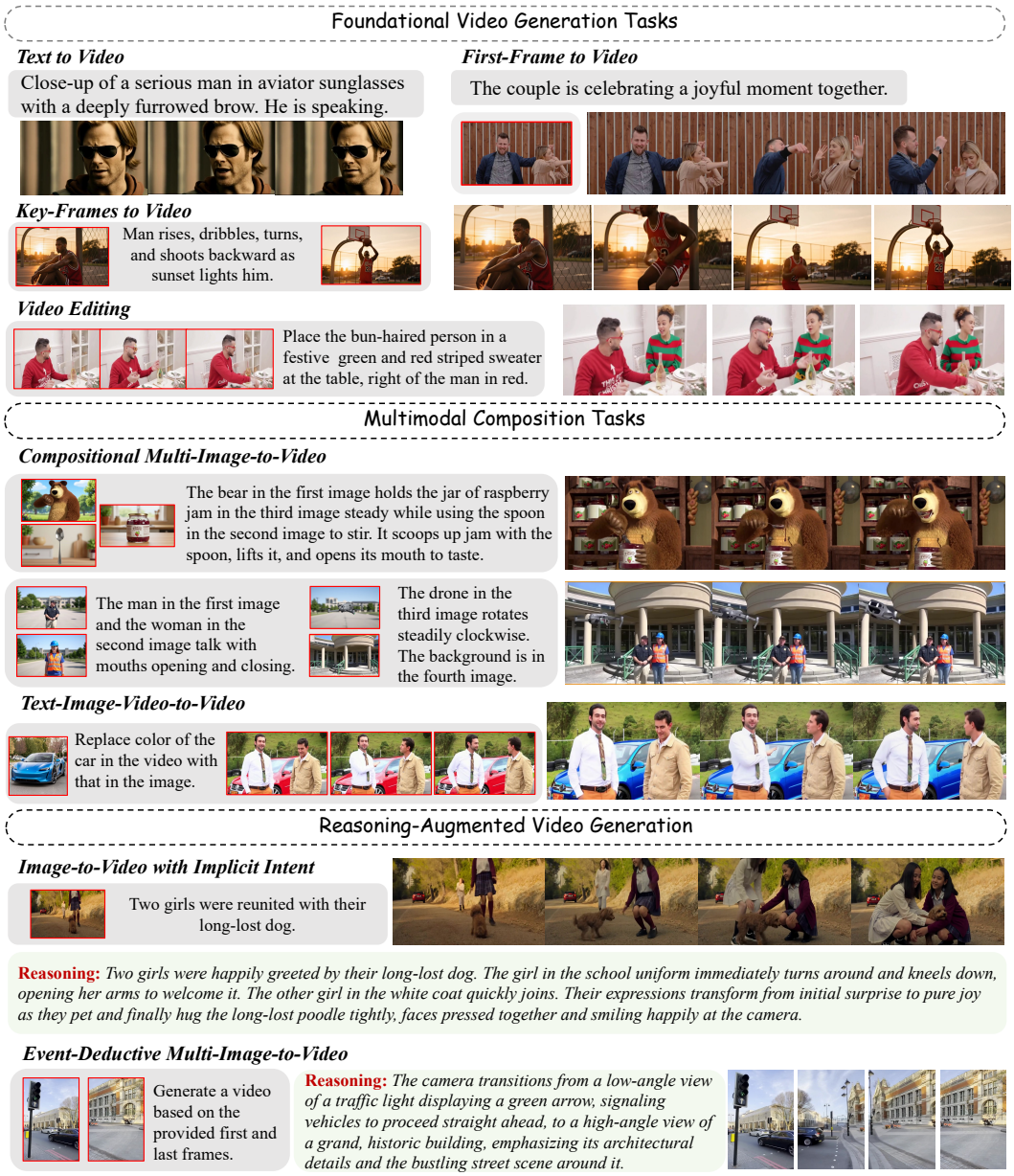

Figure 1: Showcase of OmniWeaving across diverse video generation scenarios, such as foundational tasks, multimodal composition tasks, and reasoning-augmented scenarios.

monolithic input formats, the community requires a more complex and comprehensive evaluation suite to foster the development of truly "omni-capable" video systems.

To address these challenges, we propose **OmniWeaving**, an omni-level video generation framework capable of both multimodal composition and abstract reasoning. Based on a unified architecture integrating visual comprehension and generation, we introduce a massive-scale training dataset that spans a broad spectrum of scenarios and diverse input formats, including both multimodal composition and reasoning-augmented tasks. Through a meticulous three-stage training strategy, as shown in Fig. 1, OmniWeaving could adeptly handle diverse video generation scenarios, effectively "weaving" free-form text, image, and video inputs into a coherent spatio-temporal narrative.

To rigorously evaluate unified video generation across heterogeneous and free-form inputs, we further introduce **IntelligentVBench**, specifically designed to assess abstract reasoning and free-form composition across four distinct video generation tasks, with a "VLM-as-a-judge" evalua-

tion paradigm. Extensive experiments demonstrate that OmniWeaving achieves SoTA performance among existing academic unified models. Overall, the main contributions are three-fold:

- We propose OmniWeaving, a unified framework that seamlessly integrates visual understanding to achieve omni-level video generation from free-form inputs with strong composition and reasoning capabilities.
- We introduce a massive-scale dataset, spanning a broad spectrum of generative scenarios including both composition and reasoning training tasks.
- We present IntelligentVBench, the first benchmark dedicated to measuring multimodal composition and abstract reasoning in unified video generation.

2 Related Work

Unified Video Generation. Industrial systems such as Seedance-2.0 (Seed, 2026) and MiniMax-H3 (MiniMax, 2026) already span a broad spectrum of video generation tasks within a single model, yet neither discloses its training data or recipe, leaving the research community without a reproducible path toward such capability. Although academic attempts at unification—VACE (Jiang et al., 2025), UniVideo (Wei et al., 2025), and VINO (Chen et al., 2026b)—broaden the range of tasks within a single model, they stop short of leveraging deep visual understanding for cohesive integration of multi-task capabilities. Thus, OmniWeaving aims to explore optimized architectural, data, and training paradigms, offering a robust reference for unified video synthesis.

Video Generation Benchmarks. Existing benchmarks fail to capture the true capabilities of unified video generation models on two axes. They lack **complexity**: most are confined either to single-shot tasks with rigid input formats (*e.g.*, VBench (Huang et al., 2024), VBench++ (Huang et al., 2025)) or to a narrow editing domain (*e.g.*, TGVE+ (Singer et al., 2024), OpenVE-Bench (He et al., 2025)), and even multi-task efforts like VACE-Bench (Jiang et al., 2025) still keep inputs inflexible. They also lack **comprehensiveness**: they prioritize basic rendering in simple scenes while neglecting advanced abilities like composition and reasoning. Their evaluation reliance on small, specialized models further yields unreliable judgments of whether a complex user request is actually satisfied. IntelligentVBench targets both gaps with free-form multimodal inputs, explicit reasoning/composition evaluation, a VLM-as-a-Judge paradigm for robust, human-aligned scoring.

3 Model: OmniWeaving

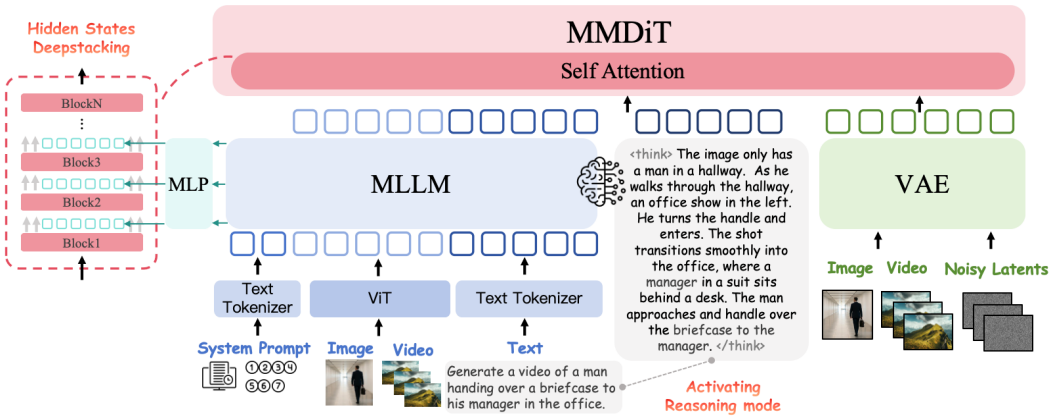

Figure 2: OmniWeaving couples an MLLM for multimodal understanding with an MMDiT for generation. We further activate the MLLM's thinking mode and introduce hidden states DeepStacking.

In this section, we introduce OmniWeaving, a unified framework for video generation and editing conditioned on free-form text, image, and video inputs, bringing advanced reasoning and compositional capabilities. We detail the model architecture, training tasks, and training strategy in Sec. 3.1, 3.2, and 3.3, respectively. More details are given in Appendix A.

3.1 Model Architecture

As shown in Fig. 2, OmniWeaving integrates multimodal understanding and generation in one architecture of three components. The **Multimodal Large Language Model (MLLM)** $\mathcal{M}$, with L layers, parses a free-form multimodal input $\mathcal{X}$ into semantic hidden states $\mathbf{H}^{(l)} = \mathcal{M}^{(l)}(\mathcal{X})$. The **Variational Autoencoder (VAE)** serves as a visual tokenizer, compressing visual inputs into low-level decodable latents $\mathbf{z}$. The **Multimodal Diffusion Transformer (MMDiT)** is the diffusion backbone that consumes both: its conditioning branch encodes the MLLM's last-layer states $\mathbf{H}^{(L)}$ through an MLP connector ϕ_0, while its generative branch fuses $\mathbf{z}$ with noise to synthesize the video. On this basis we add two mechanisms tailored to advanced reasoning and composition.

(1) Activating Thinking Mode of the MLLM: Direct MLLM encoding of interleaved visual-text inputs often yields semantic ambiguity due to weak intra-correlations and unclear video creation intents. We therefore elevate the MLLM from a passive feature extractor to an active reasoner: given $\mathcal{X}$ and a reasoning-guidance prompt p, it autoregressively generates a reasoning trace $\mathbf{r} \sim \mathcal{M}(\cdot \mid \mathcal{X}, p)$ that culminates in an explicit specification of the target video. The MLLM hidden states of the concatenated sequence $\phi_0(\mathcal{M}^{(L)}([\mathcal{X}; \mathbf{r}]))$ are then fed to the MMDiT conditioning branch, effectively bridging the cognitive gap between abstract user intent and pixel-level generation.

(2) Hidden States DeepStacking: Compositional video generation involving multiple subjects or intricate scenes often relies on both low- and high-level semantic representations, whereas the last-layer states $\mathbf{H}^{(L)}$ alone are biased toward high-level abstraction. As shown in Fig. 2, we therefore follow the DeepStacking mechanism of Qwen3-VL (Bai et al., 2025b) and additionally extract states from intermediate MLLM layers $\{\mathbf{H}^{(l_1)}, \ldots, \mathbf{H}^{(l_K)}\}$. We add them through layer-specific MLP connectors ϕ_k to the layer-k states $\mathbf{C}^{(k)}$ of the first K MMDiT conditioning layers: $\mathbf{C}^{(k)} \leftarrow \mathbf{C}^{(k)} + \phi_k(\mathbf{H}^{(l_k)})$, effectively injecting multi-granular semantic guidance into the generative process.

3.2 Training Tasks

Conventional text-video pairs cannot supervise the in-context reasoning and composition demanded by interleaved text, image, and video inputs. We therefore construct large-scale vision-text interleaved data guided by two principles: broad coverage of multimodal scenarios, and a hierarchical build-up of capability, realized as the three task categories below. See more details in Appendix B.

Foundational tasks cover text-to-image and text-to-video synthesis, instruction-guided video editing, and key-frame-to-video generation from either an initial frame or a first–last frame pair. **Multimodal composition tasks** require integrating subjects or scenes from multiple reference images, and optionally a source video, into a single coherent video that preserves every given identity. **Reasoning-augmented tasks** provide ambiguous inputs that do not fully specify the target video, such as a vague text query or key-frames with no stated link between them. Beyond the ground-truth video, each sample carries a reasoning trace that explicitly bridges the ambiguous input to the target, providing the next-token-prediction supervision for the MLLM's thinking mode.

3.3 Training Strategy

We adopt a progressive three-stage training paradigm that elevates OmniWeaving from a passive renderer to an intelligent generalist.

Stage 1: Representation Alignment. Since our generative backbone is a text-to-video expert that has never been exposed to multimodal hidden states, the objective of this stage is representation alignment between the MLLM and MMDiT. We restrict training to Text-to-Image (T2I), Text-to-Video (T2V), and Image-to-Video (I2V), finetuning only the MMDiT and the MLP connectors.

Stage 2: Multi-Task Free-Form Pretraining. We then scale to complex, heterogeneous inputs, covering every task of Sec. 3.2 except the reasoning-augmented ones, again keeping the MLLM frozen and finetuning only the MMDiT and the MLP connector. This yields a generalist that unifies diverse video generation tasks under free-form multimodal instructions.

Stage 3: Reasoning-Augmented Fine-Tuning. Finally, we introduce the reasoning-augmented tasks, co-trained with a curated subset from Stage 2, and unfreeze the MLLM for end-to-end optimization with the MMDiT. Supervising the reasoning traces with a next-token-prediction loss

Table 1: Compare IntelligentVBench with existing video generation benchmarks.

Benchmark	#Size	Multi-Task	Interleaved Input	Input modality			Ability to Measure			VLM-as-a-Judge
				Image	Multi-Images	Video	Foundation	Composition	Reasoning	
VBench Huang et al. (2024)	946	✗	✗	✗	✗	✗	✓	✗	✗	✗
VBench++ Huang et al. (2025)	1000	✓	✗	✓	✗	✗	✓	✗	✗	✗
OpenVE-Bench He et al. (2025)	431	✗	✗	✗	✗	✓	✓	✓	✗	✓
TGVE+ Singer et al. (2024)	1417	✗	✗	✗	✗	✓	✓	✓	✗	✗
OpenS2V-Eval Yuan et al. (2025)	360	✗	✓	✓	✓	✗	✓	✓	✗	✗
VACE-Bench Jiang et al. (2025)	696	✓	✓	✓	✓	✓	✓	✓	✗	✗
IntelligentVBench	1030	✓	✓	✓	✓	✓	✓	✓	✓	✓

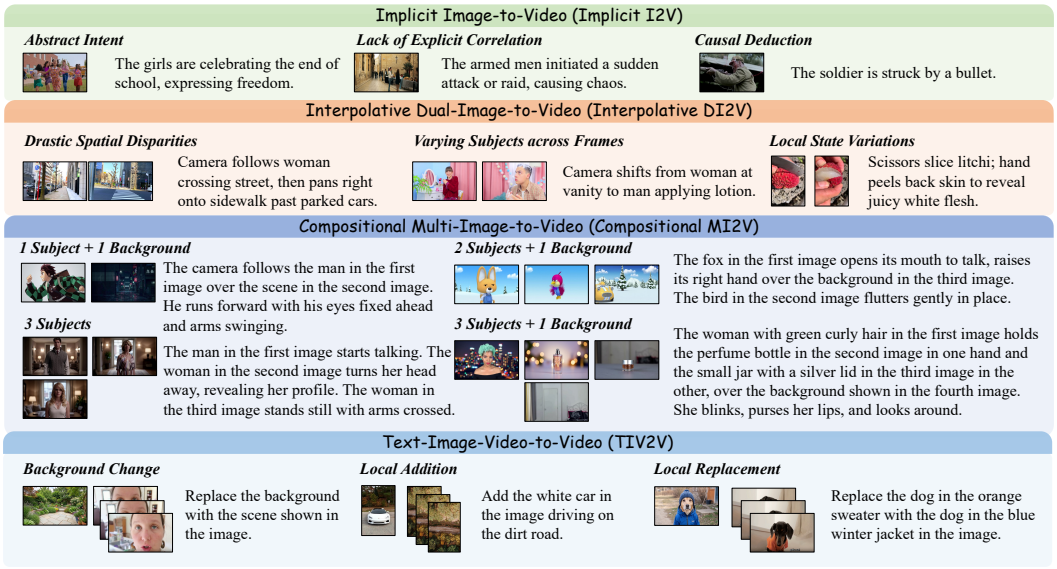

Figure 3: Examples of each task type in IntelligentVBench.

alongside the diffusion loss activates the "comprehend-then-generate" mode. Consequently, OmniWeaving evolves into an integrated reasoner and generator with higher intelligence.

4 BENCHMARK: INTELLIGENTVBENCH

As stated in Sec. 2, traditional benchmarks fail to capture the true capabilities of unified video generation models. We therefore introduce IntelligentVBench, the first benchmark built around abstract reasoning and free-form composition for next-level intelligent video generation. A comparison between IntelligentVBench and existing benchmarks is shown in Table 1.

Task Design. Our benchmark rests on two principles. First, inputs are **heterogeneous and free-form**: images, text, and video interleaved without a fixed template, with images serving as spatial anchors (references) or temporal anchors (key-frames). Second, evaluation targets **composition and reasoning**: the former integrates elements from separate inputs into one coherent video, the latter infers what they leave ambiguous. Accordingly, IntelligentVBench comprises 1,030 cases across the four tasks detailed below (Fig. 3), the first two probing reasoning and the latter two composition.

Implicit I2V (Image-to-Video) keeps the standard I2V setup with an initial frame, but the text conveys abstract intent rather than concrete actions (e.g., stating only a premise requiring causal deduction ungrounded in the image), forcing the model to unfold events under high ambiguity.

Interpolative DI2V (Dual-Image-to-Video) requires synthesizing a temporally coherent video conditioned on two boundary frames, reconstructing what happens in between, such as the path of a widely displaced camera, or the simultaneous activity of several subjects.

Compositional MI2V (Multi-Image-to-Video) provides 1–4 reference images of subjects or backgrounds plus a prompt that may introduce additional objects, and asks the model to integrate these disparate elements into a single coherent video while preserving every given identity.

Table 2: Main results of the Implicit I2V, Interpolative DI2V, and TIV2V tasks in IntelligentVBench. The best results are in **bold** for specialized- and unified-models, with the second best underlined. Acad. indicates academic models. #Params reports only the generation parameters.

Model	#Params	Acad.	Implicit I2V					Interpolative DI2V					TIV2V				
			$\mathcal{IF}\uparrow$	$\mathcal{CP}\uparrow$	$\mathcal{VQ}\uparrow$	MIN	AVG	$\mathcal{IF}\uparrow$	$\mathcal{CP}\uparrow$	$\mathcal{VQ}\uparrow$	MIN	AVG	$\mathcal{IF}\uparrow$	$\mathcal{CP}\uparrow$	$\mathcal{VQ}\uparrow$	MIN	AVG
Specialized Video Generation Models																	
CogVideoX-I2V	5B	✓	3.53	3.74	2.90	2.68	3.39	-	-	-	-	-	-	-	-	-	-
Wan2.1-I2V	14B	✗	3.76	3.33	3.34	2.78	3.48	-	-	-	-	-	-	-	-	-	-
Wan2.1-FLF2V	14B	✗	-	-	-	-	-	4.48	4.28	4.49	3.98	4.42	-	-	-	-	-
HunyuanVideo	13B	✗	3.37	**4.17**	**3.87**	3.00	3.80	-	-	-	-	-	-	-	-	-	-
HunyuanVideo1.5	8.3B	✗	3.98	3.44	3.67	3.12	3.70	-	-	-	-	-	-	-	-	-	-
Wan2.2-I2V	14B	✗	**4.21**	3.70	3.68	**3.30**	**3.86**	-	-	-	-	-	-	-	-	-	-
Unified Video Generation Models																	
MiniMax-H3	33B	✗	**4.55**	**3.88**	**4.12**	**3.60**	**4.19**	4.71	4.37	4.51	4.10	4.53	**4.36**	**4.21**	**4.03**	**3.71**	**4.20**
VACE	14B	✓	3.66	3.22	3.32	2.76	3.40	4.01	3.61	3.96	3.32	3.86	1.46	1.42	1.71	1.27	1.53
VINO	13B	✓	3.70	2.64	2.96	2.28	3.10	2.38	1.73	3.29	1.65	2.47	2.86	2.90	2.52	2.11	2.76
UniVideo(query)	13B	✓	3.66	3.37	3.16	2.77	3.39	2.31	1.85	3.27	1.76	2.48	3.22	3.91	3.26	2.66	3.46
UniVideo(hidden)	13B	✓	3.74	3.82	3.28	2.95	3.61	2.36	1.98	3.40	1.84	2.58	3.13	4.01	2.93	2.56	3.36
Ours (w/o think)	8.3B	✓	4.05	3.75	3.36	3.11	3.72	4.60	4.22	**4.52**	3.99	4.45	4.00	4.04	3.65	3.31	3.89
Ours (think)	8.3B	✓	4.33	3.82	3.63	3.34	3.93	**4.74**	**4.38**	4.50	**4.11**	**4.54**	-	-	-	-	-

Table 3: Main results of Compositional MI2V task in IntelligentVBench. We define 3 subcategories based on the number of subjects presented in the 1–4 input images.

Model	#Params	Acad.	1 Subject (w/ or w/o BKG)					2 Subjects (w/ or w/o BKG)					3 Subjects (w/ or w/o BKG)				
			$\mathcal{IF}\uparrow$	$\mathcal{CP}\uparrow$	$\mathcal{VQ}\uparrow$	MIN	AVG	$\mathcal{IF}\uparrow$	$\mathcal{CP}\uparrow$	$\mathcal{VQ}\uparrow$	MIN	AVG	$\mathcal{IF}\uparrow$	$\mathcal{CP}\uparrow$	$\mathcal{VQ}\uparrow$	MIN	AVG
Specialized Video Generation Models																	
SkyReels-A2	14B	✗	**3.51**	**4.08**	4.46	**3.24**	**4.02**	3.22	3.76	4.37	2.97	3.78	1.64	1.76	2.50	1.56	1.97
SkyReels-V3	14B	✗	3.46	3.71	**4.65**	2.98	3.94	**3.28**	**3.84**	**4.44**	**3.04**	**3.86**	**2.59**	**3.10**	4.30	**2.37**	**3.33**
MAGREF	14B	✓	3.15	2.48	4.32	2.18	3.32	3.04	2.81	4.33	2.44	3.39	2.50	2.21	**4.46**	2.07	3.06
Phantom	14B	✓	3.21	2.95	4.29	2.47	3.48	2.88	3.42	4.38	2.62	3.56	2.36	2.79	4.21	2.20	3.12
Unified Video Generation Models																	
MiniMax-H3	33B	✗	**4.44**	**4.75**	**4.78**	**4.27**	**4.66**	**4.40**	**4.75**	**4.75**	**4.24**	**4.63**	**4.12**	**4.48**	**4.65**	**3.88**	**4.42**
VACE-Wan2.1	14B	✓	3.88	4.48	4.68	3.62	4.35	3.31	4.03	4.51	3.08	3.95	2.60	3.03	4.40	2.50	3.34
VINO	13B	✓	3.72	4.22	4.46	3.42	4.13	3.56	4.34	4.58	3.39	4.16	2.63	2.97	4.24	2.53	3.28
UniVideo(query)	13B	✓	3.35	3.90	4.41	3.01	3.89	2.98	3.73	4.18	2.76	3.63	2.30	2.50	3.89	2.20	2.89
UniVideo(hidden)	13B	✓	3.33	4.18	4.38	3.08	3.97	3.22	4.12	4.36	3.09	3.90	2.31	2.83	3.94	2.29	3.03
Ours	8.3B	✓	4.35	4.53	4.58	4.01	4.49	4.08	4.22	4.52	3.61	4.27	3.53	4.01	4.54	3.26	4.03

TIV2V (Text-Image-Video-to-Video) targets cross-modal compositional editing: given a source video, reference images, and an instruction, the model must inject the designated concepts into the source video while preserving the temporal consistency and visual fidelity of non-target regions.

Evaluation Metrics. Following the "VLM-as-a-Judge" (Zheng et al., 2023) paradigm, we employ Gemini2.5-Pro (Comanici et al., 2025) to rate every generated video on a 1–5 scale along three metrics: **Instruction Following** ($\mathcal{IF}$) evaluates whether the generated video accurately adheres to the prompt requests. **Condition Preserving** ($\mathcal{CP}$) evaluates whether the generated video faithfully retains the provided visual references (e.g., anchoring key-frames and multi-subject identities). **Overall Visual Quality** ($\mathcal{VQ}$) assesses the general aesthetic quality of the video such as temporal consistency and physical naturalness. We further report two overall scores averaged over the N cases of each task, $\textbf{AVG} = \frac{1}{N}\sum_i(\mathcal{IF}_i+\mathcal{CP}_i+\mathcal{VQ}_i)/3$ and $\textbf{MIN} = \frac{1}{N}\sum_i \min(\mathcal{IF}_i, \mathcal{CP}_i, \mathcal{VQ}_i)$, where subscript i denotes the scores of the i-th case. See more details in Appendix C.

5 Experiments

5.1 Implementation Details and Evaluation Setup

We implement OmniWeaving by integrating Qwen2.5-VL (Bai et al., 2025c) as the MLLM and HunyuanVideo-1.5 (Wu et al., 2025a) as the generative engine. For our DeepStacking mechanism, we extract hidden states from the 8th, 16th, and 24th MLLM layers and project them via trainable MLPs into the initial three conditioning layers of the MMDiT. During training, we use 400 H20 GPUs with a batch size of 400, focusing on generating 480p videos. More details are in Appendix D.1. To assess model performance, we conduct zero-shot evaluations on both IntelligentVBench and existing benchmarks, with the compared baselines introduced in Appendix D.2.

5.2 Main Results on IntelligentVBench

We conduct zero-shot evaluations on IntelligentVBench for both open-source task-specific models and unified models. The primary results for Implicit I2V, Interpolative DI2V, and TIV2V are in Table 2, while the performance for Compositional MI2V is detailed in Table 3.

Table 4: Main results on OpenVE-Bench.

Model	#Params	Acad.	Global-Style↑	BKG-Change↑	Local-Change↑	Local-RM↑	Local-Add↑	Subtitle-Edit↑	Overall↑
Specialized Video Generation Models									
OmniVideo	1.3B	✓	1.11	1.18	1.14	1.14	1.36	1.00	1.16
InsViE	2B	✓	2.20	1.06	1.48	1.36	1.17	2.18	1.58
Lucy-Edit	5B	✗	2.27	1.57	**3.20**	1.75	**2.30**	1.61	2.12
ICVE	13B	✓	2.22	1.62	2.57	**2.51**	1.97	2.09	2.16
Ditto	14B	✓	**4.01**	1.68	2.03	1.53	1.41	2.81	2.25
OpenVE-Edit	5B	✓	3.16	**2.36**	2.98	1.85	2.15	**2.91**	**2.57**
Unified Video Generation Models									
VACE	14B	✓	1.49	1.55	2.07	1.46	1.26	1.48	1.55
VINO	13B	✓	**4.32**	2.37	3.65	3.19	2.71	2.94	3.19
UniVideo(query)	13B	✓	3.47	**2.47**	**3.86**	2.48	2.87	3.13	3.05
UniVideo(hidden)	13B	✓	3.55	2.42	3.67	2.89	**2.90**	2.99	3.07
Ours	8.3B	✓	4.16	2.32	3.68	**3.22**	2.41*	**3.54**	**3.22**

Comparing existing academic unified models against specialized models and MiniMax-H3 (MiniMax, 2026), an industrial open-weight model, we identify two notable shortcomings of current academic models. First, they exhibit significant performance imbalances across subtasks: VINO (Chen et al., 2026b) and VACE (Jiang et al., 2025) achieve competitive results on Compositional MI2V, yet fall considerably short on interpolative DI2V and TIV2V, respectively. Second, on individual subtasks, they still lag behind specialized models in reasoning-intensive scenarios, such as implicit I2V and interpolative DI2V. These indicate that existing academic models remain far from truly "omni-capable", and their capacity for reasoning-informed generation is still underdeveloped.

Compared to existing academic unified models, OmniWeaving achieves the following advantages: **(1) Synergistic Multi-Task Integration.** Across all four tasks, OmniWeaving consistently achieves SoTA performance among open-source academic unified models under both **MIN** and **AVG** overall metrics. **(2) Enhanced Multimodal Composition.** Whether processing complex multi-subject image combinations or heterogeneous tri-modal inputs (image, video, and text), OmniWeaving outperforms all academic baselines, showcasing superior compositional flexibility for video generation. **(3) Comprehension-Guided Generation.** For reasoning-related tasks like Implicit I2V and Interpolative DI2V, enabling the "thinking mode" of MLLM yields consistently superior results compared to direct generation. This confirms that our framework successfully unifies comprehension and generation, leveraging deep reasoning to elevate visual synthesis. *Of course, compared to the industrial MiniMax-H3, which benefits from substantially larger training scale and parameter count, OmniWeaving, as an academic model, still exhibits a notable performance gap.*

5.3 Main Results on Existing Benchmarks

We conduct zero-shot evaluations on VBench for text-to-video (T2V) generation and OpenVE-Bench for video-to-video (V2V) editing. We also evaluate on a concurrent benchmark, *i.e.*, VideoArgus-Bench (Zeng et al., 2026), which covers five categories of tasks: T2V, TI2V, subject-to-video, text-guided video editing, and subject-guided video editing.

As reported in Table 5, OmniWeaving outperforms existing unified frameworks on VBench, achieving performance comparable to specialized models like HunyuanVideo. Given that T2V data constitutes only about 10% of our training corpus, OmniWeaving successfully preserves its foundational generation capabilities without significant degradation.

Table 5: Main Results on VBench.

Model	#Params	Quality↑	Semantic↑	Total↑
Specialized Video Generation Models				
StepVideo	30B	84.46	71.28	81.83
CogVideoX	5B	83.05	**77.33**	81.91
HunyuanVideo	13B	85.09	75.82	83.24
Wan2.1	14B	**85.59**	76.11	**83.69**
Unified Video Generation Models				
VINO	13B	84.00	78.01	82.80
UniVideo	13B	-	-	82.58
Ours	8.3B	**84.37**	**78.03**	**83.10**

For V2V editing (Table 4), OmniWeaving achieves an average score that surpasses both specialized models and unified approaches such as UniVideo and VINO. Notably, our model exhibits well-balanced performance across global, local, and text-based editing. However, we observe a performance drop on the "local-add" subtask, which we attribute to its focus on animated content addition with little emphasis on natural integration, whereas our training prioritizes realistic object insertion.

The results for VideoArgus-Bench are shown in Table 9 in Appendix E.1. OmniWeaving handles all five VideoArgus-Bench tasks with a single model, outperforming the unified baselines VACE and UniVideo on every task, as well as several specialized models in T2V, TI2V, subject-to-video, and text-guided video editing.

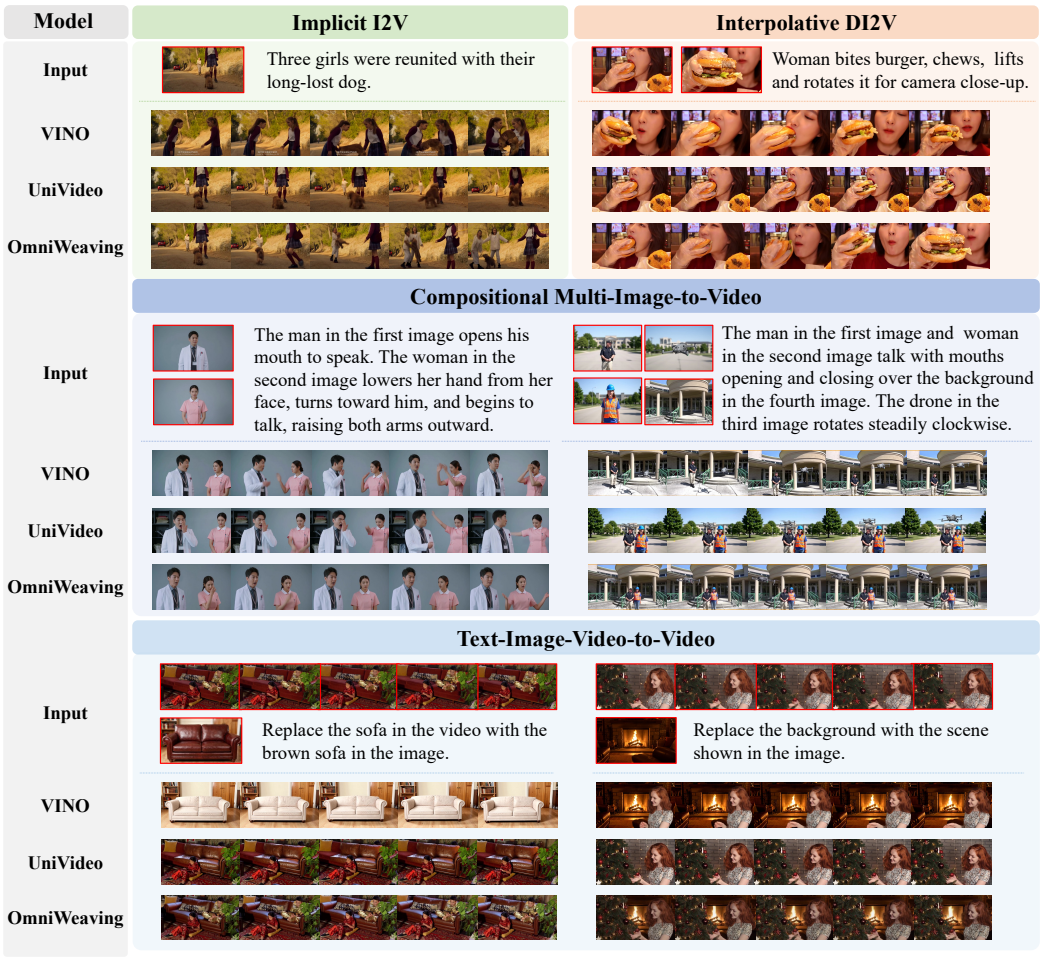

Figure 4: Qualitative comparison of VINO, UniVideo, and OmniWeaving.

5.4 Qualitative Comparisons

As illustrated in Fig. 4, we conduct a qualitative comparison of VINO, UniVideo, and OmniWeaving. In both Implicit I2V and Interpolative DI2V tasks, the two baselines frequently produce start or end frames misaligned with the given images. Also, in Implicit I2V, when required to generate three girls while the input image contains only one, both UniVideo and VINO fail to synthesize the correct number of subjects as specified by the prompt. Moreover, in Compositional MI2V, though both baselines can handle simple two-element compositions, they struggle to integrate four distinct elements. Finally, in TIV2V, the baselines often introduce unwanted modifications in unedited regions of the source video. In contrast, OmniWeaving effectively mitigates these issues, showing higher-quality generation across all cases. See more cases in Appendix E.3.

5.5 In-Depth Analysis

Effect of Reasoning-Augmented Video Generation. As shown in Fig. 5(a), we evaluate the AVG performance of OmniWeaving and its precursor prior to the final training stage on Implicit I2V and Interpolative DI2V tasks, both with and without "think mode" activated. Notably, before training stage 3, enabling "think mode" severely degrades performance, revealing a lack of synergy between comprehension and generation. In contrast, after Reasoning-Augmented Fine-Tuning, "think mode" significantly boosts overall results, demonstrating that **enhanced reasoning effectively drives higher-quality video generation**. In Table 11 of Appendix E.4, we also enable the thinking mode for UniVideo and VINO, but find it likewise degrades their performance, further highlighting the **effectiveness of our Reasoning-Augmented Fine-Tuning**.

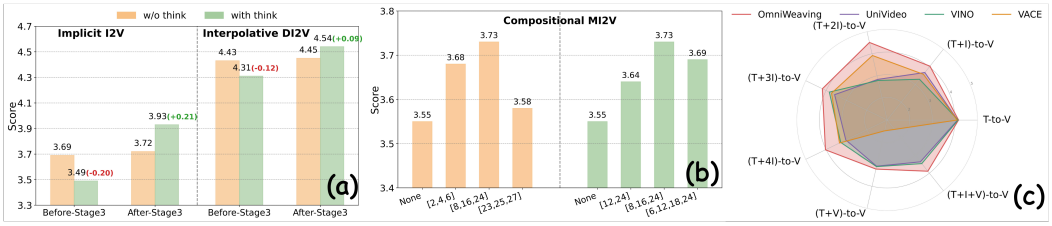

Figure 5: (a) AVG performance when enabling or disabling the thinking mode of OmniWeaving. (b) AVG performance with different DeepStacking strategies. (c) Performance visualization across different input formats for each unified video generation models.

Effect of DeepStacking. To evaluate the impact of DeepStacking, we fine-tune the Stage-1 model on a multimodal-composition subset with various layer-selection strategies. Fig. 5(b) compares their average performance on Compositional MI2V. When fixing the selection to three layers, sampling across a broad depth (e.g., [8, 16, 24]) outperforms concentrations in exclusively shallow ([2, 4, 6]) or deep ([23, 25, 27]) layers, with all configurations surpassing non-DeepStacking. Furthermore, this 3-layer setup achieves superior results compared to 2-layer ([12, 24]) or 4-layer ([6, 12, 18, 24]) alternatives. These findings suggest that integrating a balanced spectrum of semantic features—spanning low-level to high-level—optimizes compositional video generation.

Capability Visualization. Fig. 5(c) presents a radar chart visualizing the average performance scores across diverse input formats on established test cases for each unified model. It intuitively shows that OmniWeaving consistently outperforms baselines, exhibiting superior versatility and robustness when processing a wide variety of input formats.

Assessment of Evaluation Protocol. To validate the evaluation reliability for IntelligentVBench, we conducted a user study comparing human expert ratings with evaluations from several frontier VLMs, including Gemini2.5-Pro (Comanici et al., 2025), GPT-5 (OpenAI, 2025), Seed-1.6 (Guo et al., 2025), and Qwen3-VL (Bai et al., 2025b). By calculating the Pearson correlation coefficient across randomly sampled representative cases from each subtask, we assessed the alignment between automated scores and human judgment. As shown in Table 6, Gemini2.5-Pro consistently achieves the highest correlation across all metrics, demonstrating its superior proficiency in accurately reflecting model performance.

Table 6: Pearson correlation with human ratings for each VLM.

VLMs	$\mathcal{IF}\uparrow$	$\mathcal{CP}\uparrow$	$\mathcal{VQ}\uparrow$	AVG
Qwen3-VL-235B	0.66	0.54	0.59	0.60
Seed-1.6	0.72	0.63	0.67	0.67
GPT-5	0.75	0.72	0.71	0.73
Gemini2.5-Pro	**0.81**	**0.74**	**0.72**	**0.76**

Internal vs. External Prompt Rewriter. The thinking mode of the MLLM acts similarly to a prompt enhancer (PE). After three-stage training, we compare using the internal MLLM module and external LLMs (Qwen3-VL or Gemini2.5-Pro) as the PE. As shown in Table 7, the internal rewriter achieves higher AVG scores on both Implicit I2V and Interpolative DI2V. This suggests that an effective PE should be generation-aware (e.g., through end-to-end optimization with the MMDiT) rather than merely stronger in general understanding.

Table 7: AVG scores with internal and external prompt rewriters.

Rewriter	Implicit I2V	Interp. DI2V
None (w/o think)	3.72	4.45
Qwen3-VL	3.77	4.48
Gemini2.5-Pro	3.86	4.52
Internal MLLM(Ours)	**3.93**	**4.54**

6 Conclusion

In this paper, we introduce OmniWeaving, an omni-level video generation model featuring powerful multimodal composition and reasoning-informed generation capabilities. OmniWeaving can accommodate free-form, multimodal interleaved inputs to generate compliant videos. It is also capable of leveraging deep visual understanding to actively guide the generative process with a unified visual comprehension and generation framework. Furthermore, we propose IntelligentVBench, a comprehensive benchmark specifically designed to evaluate next-level intelligent video generation, containing diverse tasks. Extensive experiments show that OmniWeaving achieves SoTA performance across existing open-source academic unified models and even surpasses specialized models.

AI USE STATEMENT

We used large language models (LLMs) solely to aid or polish the writing of this manuscript, such as improving grammar, wording, and readability. And all text was reviewed and verified by the authors, who take full responsibility for the content of this paper. Separately, as part of our methodology, LLMs and multimodal LLMs serve as components of the training data construction pipeline for unified video generation (Appendix B.3 and Fig. 10). Additionally, we employed Gemini2.5-Pro as the evaluator for IntelligentVBench, enabling a more reliable assessment of model capabilities. Beyond the uses described above, we have not used generative AI tools to conceive research ideas, design the method or experiments, etc. All remaining work was carried out by humans.

ETHICS STATEMENT

This study does not raise any ethical concerns. The data used for training and evaluation come from existing datasets, publicly accessible sources, and generative models, and are used for research purposes only; the human evaluations are limited to rating generated videos. Like other video generation models, OmniWeaving could potentially be misused to create misleading content; this risk is not specific to our method, and we encourage its responsible use.

REPRODUCIBILITY STATEMENT

All results in this paper are reproducible. The project code is released at the anonymized link `https://anonymous.4open.science/r/omniweaving-24E4`, and the data processing pipeline is detailed in Appendix B.3. Additionally, the model weights and data will be open-sourced after the ICLR anonymization period.

REFERENCES

Qingyan Bai, Qiuyu Wang, Hao Ouyang, Yue Yu, Hanlin Wang, Wen Wang, Ka Leong Cheng, Shuailei Ma, Yanhong Zeng, Zichen Liu, et al. Scaling instruction-based video editing with a high-quality synthetic dataset. *arXiv preprint arXiv:2510.15742*, 2025a.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025b.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025c. URL `https://arxiv.org/abs/2502.13923`.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

Junsong Chen, Yuyang Zhao, Jincheng Yu, Ruihang Chu, Junyu Chen, Shuai Yang, Xianbang Wang, Yicheng Pan, Zhou Daquan, Huan Ling, et al. Sana-video: Efficient video generation with block linear diffusion transformer. In *International Conference on Learning Representations*, volume 2026, pp. 38973–38997, 2026a.

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*, 2026b.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

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*, 2025a.

Yufan Deng, Xun Guo, Yuanyang Yin, Jacob Zhiyuan Fang, Yiding Yang, Yizhi Wang, Shenghai Yuan, Angtian Wang, Bo Liu, Haibin Huang, et al. Magref: Masked guidance for any-reference video generation. *arXiv e-prints*, pp. arXiv–2505, 2025b.

Zhengcong Fei, Debang Li, Di Qiu, Jiahua Wang, Yikun Dou, Rui Wang, Jingtao Xu, Mingyuan Fan, Guibin Chen, Yang Li, et al. Skyreels-a2: Compose anything in video diffusion transformers. *arXiv preprint arXiv:2504.02436*, 2025.

Google. Introducing gemini 2.5 flash image, our state-of-the-art image model. `https://developers.googleblog.com/en/introducing-gemini-2-5-flash-image/?utm_source=chatgpt.com`, 2025a.

Google. Gemini ai video generator powered by veo 3.1. `https://gemini.google/overview/video-generation/`, 2025b.

Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

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.

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

Yinming Huang, Shuyuan Tu, Xi Yan, Zihan Yang, Jianhua Han, Xu Hang, Yu-Gang Jiang, and Zuxuan Wu. Va-judger: Reward modeling from human preference feedback for joint video-audio generation. *arXiv preprint arXiv:2608.18607*, 2026.

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.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, et al. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17191–17202, 2025.

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.

Black Forest Labs. Flux.2 - next generation image generation. `https://bfl.ai/models/flux-2`, 2025.

Debang Li, Zhengcong Fei, Tuanhui Li, Yikun Dou, Zheng Chen, Jiangping Yang, Mingyuan Fan, Jingtao Xu, Jiahua Wang, Baoxuan Gu, et al. Skyreels-v3 technique report. *arXiv preprint arXiv:2601.17323*, 2026.

Xinyao Liao, Xianfang Zeng, Ziye Song, Zhoujie Fu, Gang Yu, and Guosheng Lin. In-context learning with unpaired clips for instruction-based video editing. *arXiv preprint arXiv:2510.14648*, 2025.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, et al. Muon is scalable for llm training. *arXiv preprint arXiv:2502.16982*, 2025a.

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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14951–14961, 2025b.

MiniMax. Minimax h3. https://huggingface.co/MiniMaxAI/MiniMax-H3, 2026.

OpenAI. Introducing GPT-5. https://openai.com/index/introducing-gpt-5/, 2025. Accessed: 2026-03-05.

OpenAI, Aaron Hurst, Adam Lerer, et al. Gpt-4o system card. Technical report, OpenAI, 2024. URL https://arxiv.org/abs/2410.21276.

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.

ByteDance Seed. Seedance 2.0. https://seed.bytedance.com/en/seedance2_0, 2026.

Uriel Singer, Amit Zohar, Yuval Kirstain, Shelly Sheynin, Adam Polyak, Devi Parikh, and Yaniv Taigman. Video editing via factorized diffusion distillation. In *European Conference on Computer Vision*, pp. 450–466. Springer, 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.

DecartAI Team. Lucy edit: Open-weight text-guided video editing, 2025.

Kling Team. Kling video 3.0 omni model. https://kling.ai/quickstart/klingai-video-3-omni-model-user-guide, 2026.

Meituan LongCat Team, Xunliang Cai, Qilong Huang, Zhuoliang Kang, Hongyu Li, Shijun Liang, Liya Ma, Siyu Ren, Xiaoming Wei, Rixu Xie, et al. Longcat-video technical report, 2025. *URL https://arxiv. org/abs/2510.22200*, 2025.

Shuyuan Tu, Qi Dai, Zuxuan Wu, Zhi-Qi Cheng, Han Hu, and Yu-Gang Jiang. Implicit temporal modeling with learnable alignment for video recognition. In *ICCV*, 2023.

Shuyuan Tu, Qi Tian, Zihan Yang, Yue Wu, Xintong Han, Weijie Kong, Jiangfeng Xiong, Jian-Wei Zhang, Zhao Zhong, Liefeng Bo, et al. Baton: Explicit semantic blueprints for joint video-audio generation. *arXiv preprint arXiv:2605.25195*, 2026.

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.

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.

Bing Wu, Chang Zou, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Jack Peng, Jianbing Wu, Jiangfeng Xiong, Jie Jiang, et al. Hunyuanvideo 1.5 technical report. *arXiv preprint arXiv:2511.18870*, 2025a.

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*, 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.

Yuhui Wu, Liyi Chen, Ruibin Li, Shihao Wang, Chenxi Xie, and Lei Zhang. Insvie-1m: Effective instruction-based video editing with elaborate dataset construction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16692–16701, 2025d.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306, 2022.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 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.

Zihan Yang, Shuyuan Tu, Licheng Zhang, Qi Dai, Yu-Gang Jiang, and Zuxuan Wu. Arcflow: Unleashing 2-step text-to-image generation via high-precision non-linear flow distillation. *arXiv preprint arXiv:2602.09014*, 2026.

Shenghai Yuan, Xianyi He, Yufan Deng, Yang Ye, Jinfa Huang, Bin Lin, Jiebo Luo, and Li Yuan. Opens2v-nexus: A detailed benchmark and million-scale dataset for subject-to-video generation. *arXiv preprint arXiv:2505.20292*, 2025.

Ziyun Zeng, Zixuan Wang, Yongsheng Yu, Hang Hua, and Jiebo Luo. Videoargus: Agentic rubric-grounded unified evaluation for video generation and editing. *arXiv preprint arXiv:2608.05485*, 2026.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

APPENDIX OVERVIEW

In the Appendix, we present the following details. Project page: `https://anonymous.4open.science/r/omniweaving-24E4`.

- Model Details in Section A.
- Training Details in Section B.
- Benchmark Details in Section C.
- Experiment Setup in Section D.
- More Experimental Results in Section E.
- Future Work in Section F.

A MODEL DETAILS

Building upon the general architecture outlined in Sec. 3.1, we detail the specific design and implementation of our model by instantiating OmniWeaving with Qwen2.5-VL (Bai et al., 2025c) as the MLLM and HunyuanVideo-1.5 (Wu et al., 2025a) as the core generative engine. Because HunyuanVideo-1.5 utilizes a multilingual Glyph-ByT5 text encoder (Xue et al., 2022) for precise cross-lingual text rendering and a SigLIP (Zhai et al., 2023) encoder for robust vision-text semantic alignment, we retain both modules to preserve architectural consistency with the generative backbone. Specifically, in scenarios involving multiple visual inputs, the global SigLIP feature is computed as the mean embedding across all provided images. Similarly, for video inputs, we extract a representative set of frames and average their corresponding SigLIP embeddings to yield a unified semantic representation. Notably, both the Glyph-ByT5 and SigLIP encoders remain strictly frozen throughout the entire training process.

Furthermore, a key aspect of multitask training is that the representations provided by the MLLM should enable the MMDiT to understand exactly what task needs to be performed (Wu et al., 2025b). To this end, we set distinct MLLM system prompts for different tasks to facilitate better differentiation. The system prompts for each video generation task are shown in Fig. 6, Fig. 7, Fig. 8 and Fig. 9. For reasoning-augmented tasks, we append a specialized reasoning-guidance prompt to the user input (*i.e.*, *based on the provided image and text, please generate a detailed description that explicitly articulates the visual, temporal, and semantic characteristics of the target video*). This reasoning-guidance prompt explicitly directs the MLLM to perform reasoning, yielding an enhanced prompt that provides a granular description of the target video's spatial composition and temporal dynamics. Subsequently, the hidden states corresponding to this enhanced prompt are concatenated with those of the original user input; this combined representation is then fed into the MMDiT to effectively steer the generative process.

B TRAINING DETAILS

B.1 TRAINING DATA SOURCE

To construct a robust training corpus, we curate data from two complementary sources based on video provenance: real-world and synthetic domains. Real-world data encompasses a broad spectrum of visual content essential for capturing rich appearances, naturalistic motions, and complex scene dynamics; crucially, this anchors the generated videos to natural distributions and mitigates noticeable generative artifacts. However, because naturally occurring paired videos for highly conditioned tasks, such as video editing, are frequently sparse or inherently noisy, we incorporate synthetic data by leveraging off-the-shelf generation models (Wu et al., 2025a; Wan et al., 2025; Google, 2025b) to rapidly synthesize target videos aligned with specific input conditions. While relying exclusively on synthetic data tends to introduce pronounced artificial biases, combining these two domains creates a synergistic effect that perfectly balances natural realism with task-specific conditioning density. Overall, the resulting training corpus contains tens of millions of samples spanning all tasks in Appendix B.2.

System Prompt for Text-to-Video Generation Task

<|im_start|>system
You are a helpful assistant. Describe the video by detailing the following aspects: (1) The main content and theme of the video. (2) The color, shape, size, texture, quantity, text, and spatial relationships of the objects. (3) Actions, events, behaviors temporal relationships, physical movement changes of the objects. (4) Background environment, light, style and atmosphere. (5) camera angles, movements, and transitions used in the video.<|im_end|>
<|im_start|>user
<|user_text|><|im_end|>
<|im_start|>assistant

System Prompt for First-Frame-to-Video Generation Task

<|im_start|>system
You are a helpful assistant. Describe the key features of the input image (color, shape, size, texture, objects, background), then explain how the user's text instruction should alter the image to introduce motion and evolution over time. Generate a video using this image as the first frame that meets the user's requirements, ensuring the specified elements evolve or move in a way that fulfills the text description while maintaining consistency. <|im_end|>
<|im_start|>user
<|user_img|><|user_text|><|im_end|>
<|im_start|>assistant

Figure 6: System prompts for Text-to-Video and First-Frame-to-Video generation tasks, where **<|user_text|>** is the user input prompt and **<|user_img|>** is the user input image.

System Prompt for Key-Frames-to-Video Generation Task

<|im_start|>system
You are a helpful assistant. Given a text instruction and multiple images (greater than 1) as key-frames of the video, you need to analyze the visual trajectory required to transition from the start to the end across these key-frames. Determine how the elements in each key-frame must evolve, move, or transform to align with the subsequent key-frame based on the text instruction. Generate a video that seamlessly connects these key-frames, ensuring the motion and evolution between them fulfill the text description while maintaining temporal consistency. <|im_end|>
<|im_start|>user
<|key_frame_1|>...<|key_frame_n|><|user_text|><|im_end|>
<|im_start|>assistant

System Prompt for Video-to-Video Editing Task

<|im_start|>system
You are a helpful assistant. Given a text instruction and an input video, you need to analyze the visual content and temporal dynamics of the input video, and then explain how the user's text instruction should modify the video's visual style, objects, or scene composition. Generate an edited video that meets the user's requirements, ensuring the specified modifications are applied consistently across frames while preserving the original motion flow and coherence.<|im_end|>
<|im_start|>user
<|user_video|><|user_text|><|im_end|>
<|im_start|>assistant

Figure 7: System prompts for Key-Frame-to-Video generation and Video-to-Video editing tasks, where **<|user_text|>** is the user input prompt, **<|key_frame_i|>** is the user input image, and **<|user_video|>** is the user input video.

System Prompt for Interleaved Text-and-Multi-Image-to-Video Generation Task

<|im_start|>system
You are a helpful assistant. Given text instructions and one or more input images, you need to explain how to extract and combine key information from the input images to construct a new image as the video's first frame, and then how the user's text instruction should alter the image to introduce motion and evolution over time. Generate a video that meets the user's requirements, ensuring the specified elements evolve or move in a way that fulfills the text description while maintaining consistency.<|im_end|>
<|im_start|>user
<|user_img_1|><|user_text_1|>...<|user_img_n|><|user_text_n|>
<|im_end|>
<|im_start|>assistant

System Prompt for Text-Image-Video-to-Video Generation Task

<|im_start|>system
You are a helpful assistant. Given a text instruction, one or more reference images, and an input video, you need to analyze the visual content and temporal dynamics of the input video, alongside the scene or subject characteristics of the reference images. Explain how the user's text instruction directs the application of the reference images' visual attributes onto the input video. Generate an edited video that meets the user's requirements, ensuring the specified modifications are applied consistently across frames while preserving the original motion flow and coherence.
<|im_end|>
<|im_start|>user
<|user_video|><|user_img_1|>...<|user_img_n|><|user_text|>
<|im_end|>
<|im_start|>assistant

Figure 8: System prompts for Interleaved Text-and-Multi-Image-to-Video and Text-Image-Video-to-Video generation tasks, where **<|user_img_i|>** is the user input images, and **<|user_video|>** is the user input video.

B.2 Training Tasks

To ensure our training tasks facilitate richer multimodal interactions, we establish two core design principles: comprehensive coverage of ***diverse multimodal scenarios*** and the systematic optimization of ***hierarchical model capabilities***. Accordingly, we structure our training framework around three primary competencies, each encompassing a diverse array of task formats.

Foundational Video Generation Tasks: This category integrates some foundational generation and editing tasks across three primary domains: **(a)** Text-to-image and text-to-video synthesis with text-video or text-image paired data; **(b)** Instruction-guided video-to-video editing for both local and global modifications, such as background replacement, style transfer, object manipulation (addition, removal, or replacement), and text rendering; and **(c)** Key-frame(s)-to-video generation, which synthesizes continuous temporal sequences either from a single initial frame or via interpolation across multiple key-frames.

Multimodal Composition Tasks. Multimodal composition requires extracting and integrating distinct subjects or scenes from diverse inputs to synthesize a coherent video without unnatural blending artifacts. We formulate two primary tasks: **(a)** Interleaved Text-and-Multi-Image-to-Video generation, where the inputs contain multiple reference images (capturing key visual elements such as subjects or scenes) interleaved with text, requiring the model to accurately compose these elements into a cohesive video sequence; and **(b)** Text-Image-Video-to-Video generation, where the inputs consist of three modalities (image, text, and video), requiring the model to seamlessly integrate target visual elements extracted from reference images into the temporal dynamics of a source video.

Reasoning-Augmented Tasks: When user inputs are ambiguous, reasoning is essential to decipher the intended video content. Accordingly, we construct a reasoning-augmented dataset encompassing three main tasks: **(a)** Text-to-Video generation, where the model is trained to deduce comprehensive

System Prompt for Reasoning-augmented Task (Example 1)

<|im_start|>system
You are a helpful assistant. Describe the key features of the input image (color, shape, size, texture, objects, background), then explain how the user's text instruction should alter the image to introduce motion and evolution over time. Note that the text instruction may implicitly express the intent of video generation. Generate a video using this image as the first frame that meets the user's requirements, ensuring the specified elements evolve or move in a way that fulfills the text description while maintaining consistency. <|im_end|>
<|im_start|>user
<|user_img|><|user_text|>
Based on the provided image and text, please generate a detailed description that explicitly articulates the visual, temporal, and semantic characteristics of the target video.<|im_end|>
<|im_start|>assistant

System Prompt for Reasoning-augmented Task (Example 2)

<|im_start|>system
You are a helpful assistant. Given a text instruction and multiple images (greater than 1) as key-frames of the video, you need to analyze the visual trajectory required to transition from the start to the end across these key-frames. Determine how the elements in each key-frame must evolve, move, or transform to align with the subsequent key-frame based on the text instruction. Note that the text instruction may implicitly express the intent of video generation. Generate a video that seamlessly connects these key-frames, ensuring the motion and evolution between them fulfill the text description while maintaining temporal consistency. <|im_end|>
<|im_start|>user
<|key_frame_1|>...<|key_frame_n|><|user_text|>
Based on the provided image and text, please generate a detailed description that explicitly articulates the visual, temporal, and semantic characteristics of the target video.<|im_end|>
<|im_start|>assistant

Figure 9: Two System prompt examples for reasoning-augmented video generation tasks.

descriptions from brief, ambiguous input text queries prior to synthesis; **(b)** Intent-Driven Image-to-Video generation, where the model learns to formulate a reasoning trace detailing the temporal progression when visual and textual inputs lack explicit linkage (*e.g.*, the text outlines abstract intents); and **(c)** Event-Deductive Multi-Image-to-Video generation, given several highly disparate reference images as the key-frames, the model is optimized to bridge disparate reference key-frames by first uncovering implicit temporal dynamics via providing transition descriptions for these key-frames, before generating temporally coherent videos.

B.3 Training Data Construction

To construct our training tasks, we employ a dual-pipeline data construction strategy: output-first and input-first. In the output-first pipeline, we curate a diverse array of real-world videos from sources such as YouTube, cinematic clips, live-stream excerpts, and social media platforms to serve as ground-truth target videos. Subsequently, an ensemble of auxiliary models is utilized to extract corresponding images or generate descriptive texts that act as task-specific inputs. Conversely, the input-first pipeline begins by formulating the input conditions, leveraging video generation models, augmented by various tool models, to synthesize the corresponding ground-truth videos. To facilitate both pipelines, we integrate a robust suite of models (Yang et al., 2025; Bai et al., 2025b; Carion et al., 2025; Labs, 2025; Huang et al., 2026; Tu et al., 2023; 2026; Yang et al., 2026; Wu et al., 2025a; Wan et al., 2025; Google, 2025b), such as Qwen3, Qwen3-VL, SAM3, FLUX2, and various video generation models, with Qwen3-VL additionally serving as an evaluator for data quality filtering. Next, we provide a detailed exposition of the training data construction pipeline.

Foundational Video Generation Tasks: Our primary training corpus for Text-to-Image (**T2I**), Text-to-Video (**T2V**), and Key-Frame(s)-to-Video generation (**I2V**) tasks is predominantly derived

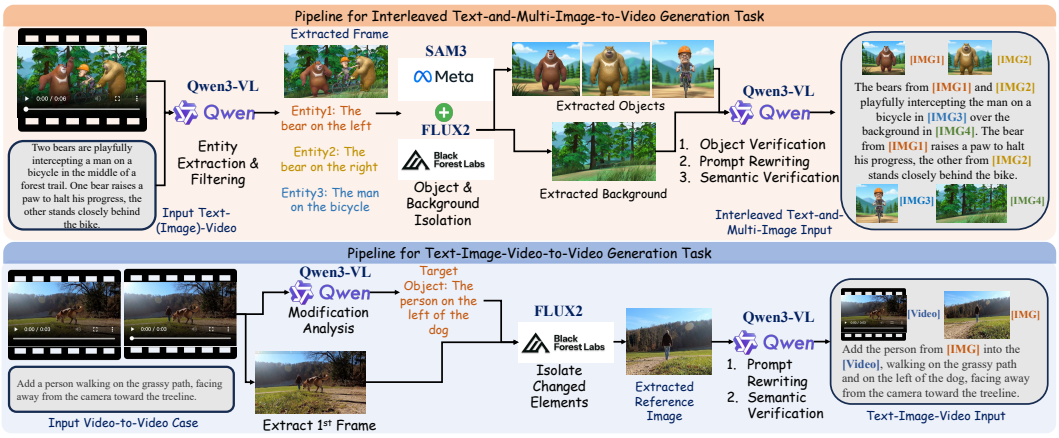

Figure 10: Training data construction pipeline for Multimodal Composition Tasks.

from extensive in-house datasets utilizing an output-first pipeline. Specifically, the videos are mainly collected from web-sourced clips, and we leverage Qwen3-VL-235B to generate high-quality textual annotations as the user input. To ensure the annotations align with specific task requirements, we design tailored prompting strategies: T2I and T2V prompts focus on visual semantic descriptions, whereas I2V prompts emphasize the dynamic transitions originating from the initial frame or the characterization of spatio-temporal offsets across multiple key-frames. Beyond this output-first paradigm, we also integrate an input-first strategy to produce a set of synthetic video data. Specifically, we first construct a carefully curated set of textual prompts and key-frames, and subsequently query Veo3 (Google, 2025b) for high-quality video synthesis.

Furthermore, the training task for Instruction-guided video-to-video editing (**V2V**) encompasses both global and local modifications. Global edits primarily focus on background transformations and style changes, while local edits include fine-grained operations such as object addition, removal, replacement, and text manipulation within the video. To construct a robust training corpus, we aggregate data from existing datasets, specifically OpenVE-3M (He et al., 2025) and Ditto (Bai et al., 2025a), and synthesize additional samples following their established pipelines. Finally, to guarantee high dataset fidelity, the entirety of this collected corpus undergoes rigorous quality filtration via Qwen3-VL-235B, ensuring the elimination of unsuccessful or low-quality edits.

Notably, we observe that existing pipelines for local object addition V2V frequently yield physical inconsistencies, where newly introduced objects appear detached from the scene's underlying geometry and lighting. To rectify this, we invert the local object removal process, treating the post-removal video as the source input and the original, unedited video as the ground-truth target. By adapting the corresponding instructions from "removal" to "addition", we successfully generate high-quality V2V training samples for local object addition, characterized by physical realism and seamless environmental integration.

Multimodal Composition Tasks. The Multimodal Composition Tasks encompass two primary sub-tasks: Interleaved Text-and-Multi-Image-to-Video generation and Text-Image-Video-to-Video generation. The respective data construction pipelines are detailed below, also shown in Figure 10.

Specifically, to facilitate **Interleaved Text-and-Multi-Image-to-Video generation**, we construct our dataset by systematically processing our image-to-video (I2V) task data. Given a detailed video caption and its corresponding key-frame, we employ Qwen3-VL-235B to identify moving objects, representing them as named entities enriched with precise descriptive modifiers to ensure accurate localization. Qwen3-VL also acts as a filter to discard entities that feature ambiguous descriptions, isolated body parts, or objects absent from the key-frame. Subsequently, for each validated entity, we utilize SAM3 to estimate its spatial location and temporal presence across the video sequence, followed by the application of FLUX2 to extract the corresponding object. To ensure appearance diversity and avoid pose duplication with the first frame, we leverage FLUX2 to extract these entity images from subsequent frames. Moreover, recognizing the potential extraction inac-

curacies of FLUX2, we formulate four distinct prompt variations for each target object, leveraging Qwen3-VL as a verifier to confirm the identity alignment between the subject in the extracted image and the specified entity within the key-frame. In addition to entity extraction, we also isolate the background from the first frame with FLUX2. We then instruct Qwen3-VL to synthesize a comprehensive, reorganized prompt that integrates the extracted objects and the background, thereby formulating the interleaved text-and-multi-image input. Ultimately, following a rigorous final verification by Qwen3-VL to ensure semantic consistency between the constructed interleaved input and the ground-truth video, we construct a large-scale, high-quality dataset tailored for Interleaved Text-and-Multi-Image-to-Video generation.

In contrast, we formulate the **Text-Image-Video-to-Video generation** task by repurposing existing video-to-video editing datasets. Given a source and target video pair, we first employ Qwen3-VL-235B, in conjunction with the original editing instruction, to analyze the specific modifications made in the target video, generating precise descriptive terms for the altered elements, such as a specific localized object or the overall video background. We then extract the first frame of the target video and apply FLUX2 to isolate the specific visual elements that have changed relative to the source video to serve as our reference images. Finally, we prompt Qwen3-VL a second time to rewrite the editing instructions and verify the semantic alignment between the input and output, thereby ensuring that the target elements detailed in the prompt are explicitly grounded in the extracted reference images.

Reasoning-Augmented Tasks. The Reasoning-Augmented Tasks comprise three primary subtasks: Text-to-Video generation, Intent-Driven Image-to-Video generation, and Event-Deductive Multi-Image-to-Video generation. In addition to standard user inputs and ground-truth videos, each task incorporates a reasoning trace that explicitly bridges the input to the corresponding output. Next, we detail the pipelines for constructing the training data across these three tasks.

For **Text-to-Video generation**, we initialize our pipeline with a collection of brief, ambiguous text queries, where a subset of these queries has already been accompanied by ground-truth videos. To process the queries without corresponding videos, we leverage Qwen3-30B to generate detailed prompts acting as query-guidance. These detailed prompts are subsequently fed into HunyuanVideo-1.5 (Wu et al., 2025a) to generate the corresponding target videos. Furthermore, for queries that already possess matching videos, we employ Qwen3-VL-235B to deeply analyze the video content and expand the initial brief query into a comprehensive, detailed prompt. Ultimately, this dual-pathway approach yields robust training triplets, comprising the initial query, the detailed prompt guidance, and the associated video. Therefore, it enables video generation models to ground video generation in sophisticated language-based reasoning.

To advance **Intent-Driven Image-to-Video generation**, we curate a dataset of cinematic text-video pairs wherein the text describes the action intent or behavioral premise driving the corresponding video scene. For each sample, we leverage Qwen3-VL-235B to synthesize a detailed description of the temporal motion unfolding from the initial frame. To ensure rigorous data quality, we further employ Qwen3-VL to strictly filter out instances where the generated motion lacks causal alignment with the driving intent. On this basis, the action intent or behavioral premise serves as the user query, and the generated motion description functions as an intermediate reasoning trace. It effectively enables the model to infer the reasoning trace prior to generating temporally coherent videos conditioned on both the first frame and the user query.

To facilitate **Event-Deductive Multi-Image-to-Video generation**, we process our collected video dataset by extracting key-frames and rigorously filtering them to ensure significant visual distinctions among key-frames. Then we leverage Qwen3-VL-235B to generate two distinct types of textual annotations: a concise, single-sentence prompt that outlines the underlying intent or premise of the video, and a detailed prompt that intricately describes the dynamic temporal transitions across the given key-frames. In formulating the training data, each instance incorporates the ordered key-frames and the corresponding ground-truth video, alongside a user query that either utilizes the concise prompt or relies solely on a generic instruction devoid of any explicit event cues, *i.e.*, "*generate a complete video based on the provided key-frames*". Crucially, the detailed prompt acts as an explicit reasoning trace, thereby conditioning the model to logically deduce the underlying event trajectory implied by the key-frames before synthesizing the final video.

B.4 Training Stage

Stage 1: Modality Alignment Training. As our generative backbone, *i.e.*, HunyuanVideo-1.5, is highly proficient as a text-to-video expert but lacks prior exposure to multimodal hidden states produced by our MLLM, the primary objective of this initial stage is modality alignment between MLLM and MMDiT. We restrict the training to the most fundamental tasks: Text-to-Image (T2I), Text-to-Video (T2V), and Image-to-Video (I2V) from a single first frame. We keep the MLLM parameters strictly frozen and exclusively finetune the MMDiT and the MLP connector with high-quality samples of these three tasks. After this stage, OmniWeaving achieves performance comparable to the MMDiT backbone that uses its own text encoder.

Stage 2: Multi-Task Free-Form Pretraining. Having aligned the foundational modalities, we scale the training to encompass complex, heterogeneous inputs. Specifically, this stage incorporates all tasks detailed in Sec. 3.2, excluding the Reasoning-Augmented Tasks. However, we observe that introducing training task involving video input such as video editing at the very onset of the training, impedes the model's learning efficiency on tasks that require processing interleaved image-text inputs. Thus, we further partition this training phase into two distinct sub-stages: we initially exclude tasks involving video inputs, later integrating them for joint mixed training in the second sub-stage. Throughout this training stage, the MLLM parameters still remain frozen to focus exclusively on fine-tuning the MMDiT. Ultimately, we empower OmniWeaving to unify diverse video generation and editing tasks via multimodal instructions, establishing it as a versatile generalist framework with robust free-form compositional capabilities.

Stage 3: Reasoning-Augmented Fine-Tuning. In the final stage, we further elevate OmniWeaving's cognitive capabilities by introducing the reasoning-augmented tasks, jointly co-training alongside a curated high-quality subset from Stage 2. Crucially, we unfreeze the MLLM in this phase for end-to-end optimization alongside the MMDiT. Beyond the standard diffusion loss, the reasoning traces within the augmented data also introduce a next-token-prediction loss specifically designed to enhance the MLLM's reasoning proficiency. This establishes a "comprehend-then-generate" paradigm. Specifically, when confronted with ambiguous inputs, the MLLM learns to activate a "thinking mode" to extract explicit requirements via visual comprehension, thereby better guiding the MMDiT to synthesize highly aligned videos. Consequently, OmniWeaving evolves into an integrated, proactive reasoner and generator, functioning as an intelligent agent adept at managing diverse video generation tasks.

C Benchmark Details

C.1 Design Principles

To comprehensively evaluate the complex capabilities required of unified video generation models, the design of IntelligentVBench is guided by two core principles, closely mirroring the demands of real-world human queries.

Heterogeneous and Free-Form Inputs. To better reflect the complexities of real-world human queries, our benchmark aims to employ test cases featuring free-form, interleaved vision-text inputs. Specifically, inputs should consist of diverse combinations of multiple images, text, and, when necessary, reference video contexts. Crucially, images within the prompt serve multifaceted roles: Some act as spatial anchors, providing specific objects intended for scene composition, while others serve as temporal anchors to delineate event progression. This design compels models to dynamically interpret unstructured, multi-modal conditions rather than over-fitting to rigid input templates.

Evaluating Composition and Reasoning: We consider composition and reasoning to be the two most critical capabilities for advanced intelligent video generation, thus structuring our evaluation around them. Specifically, for **(a) Composition**, given complex inputs comprising diverse subjects and scenes, we require the model to not only achieve accurate static spatial composition, but also to temporally synthesize the dynamic interactions among these entities and their interplay with contextual background elements throughout the generated frames. Furthermore, regarding **(b) Reasoning**, many cases within our benchmark cannot be resolved through direct explicit scene rendering; instead, the model may first need to deduce temporal correlations across significantly disparate ref-

Table 8: Training recipe of OmniWeaving.

	Stage1	Stage2	Stage3	
			MLLM	MMDiT
Hyperparameters				
Learning rate	2×10^{-5}	2.0×10^{-5}	3.0×10^{-5}	1.0×10^{-5}
LR scheduler	Constant	Constant	Cosine	Constant
Min Learning rate	2×10^{-5}	2.0×10^{-5}	2.0×10^{-6}	1.0×10^{-5}
Weight decay	0.01	0.01	0.05	0.01
Gradient norm clip	1.0	1.0	1.0	1.0
Optimizer	Muon	Muon	AdamW	Muon
GPUs	128× H20s	400× H20s	400× H20s	
Batch Size	128	400	400	
Loss weight (CE : MSE)	-	-	0.25 : 1	
Warm-up steps	250	1200	250	
Training steps	5K	50K	3k	
Resolution	480p	480p	480p	
Text dropout ratio	0.1	0.1	0.1	
Diffusion timestep shift	1.0 for image and 4.0 for video			
Data sampling ratio				
Text-to-Image	0.10	0.05 → 0.03	0.00	
Text-to-Video	0.40	0.15 → 0.07	0.05	
First-Frame-to-Video	0.50	0.15 → 0.06	0.05	
Key-Frames-to-Video	0.00	0.25 → 0.14	0.05	
Video-to-Video	0.00	0.00 → 0.25	0.15	
Interleaved Text-and-Multi-Image-to-Video	0.00	0.40 → 0.20	0.15	
Text-Image-Video-to-Video	0.00	0.00 → 0.25	0.15	
Reasoning-Augmented Tasks	0.00	0.00	0.40	

erence images, interpret textual descriptions of object behaviors that remain absent from the visual inputs, or process prompts that merely articulate the abstract intent underlying an event.

C.2 Task Taxonomy

Guided by the design principles, we construct four distinct tasks within IntelligentVBench: Implicit Image-to-Video (Implicit I2V), Interpolative Dual-Image-to-Video (Interpolative DI2V), Compositional Multi-Image-to-Video (Compositional MI2V), Text-Image-Video-to-Video (TIV2V). The first two are explicitly designed to evaluate reasoning, whereas the latter two focus on composition.

Implicit I2V: Although structurally similar to standard I2V tasks, where an image provides the initial frame and a text prompt guides the generation, this task emphasizes implicit reasoning. It requires the model to logically unfold temporal events under high ambiguity, *e.g.*, when the prompt expresses abstract intent without concrete actions, or when the described behavior lacks explicit grounding in the input image, or when text provides only preliminary assumptions requiring causal deduction.

Interpolative DI2V: It requires synthesizing a temporally-coherent video conditioned on two boundary frames and a text prompt, presenting three primary challenges: (a) large spatial disparities, requiring the generation of plausible physical trajectories or complex camera motions; (b) multi-subject dynamics, necessitating the inference of complex intermediate states for various characters; and (c) local environmental variations, which demand deducing the underlying events that drive these visual transformations.

Compositional MI2V: Given 1-4 reference images of subjects or backgrounds, alongside text prompts that may introduce additional objects, the objective is to seamlessly integrate these disparate elements into a coherent video while preserving the given visual identities.

TIV2V: This task assesses cross-modal compositional editing. Given a source video, reference images, and a text instruction, the model must extract specific concepts from the references and integrate them into the source video, while preserving the temporal consistency and visual fidelity of non-target elements.

In total, our benchmark comprises **1,030** test cases. Specifically, the **Implicit I2V** task contains 250 test cases; to construct highly challenging data, all input images are curated from movie clips featuring complex motions, with PhD experts formulating implicit instructions based on the corresponding cinematic plots. The **Interpolative DI2V** task also includes 250 test cases, where the paired input images are sourced from social networks or film segments to ensure they are contextually related yet exhibit significant visual variance. Furthermore, the **Compositional MI2V** task encompasses 320 test cases, each accepting one to four input images. And one to three images define the subject(s) and up to one image specifying the scene context. We further categorize this task into three subtasks based on the number of subject input images: 130 cases feature a single subject image input (with 78 cases of these including an additional scene image), 120 cases feature two subject image inputs (with 40 cases including an additional scene image), and 70 cases feature three subject image inputs (with 36 cases including an additional scene image). Finally, the **TIV2V** task comprises a total of 210 test cases spanning three distinct operational objectives: inserting an object from a reference image into the source video (71 test cases), substituting a specific visual element within the video with one from the reference image (73 test cases), and replacing the video's original background with the scene depicted in the reference image (66 test cases).

C.3 Evaluation Metrics

To comprehensively assess model performance on IntelligentVBench, we adopt the "VLM-as-a-Judge" (Zheng et al., 2023) paradigm, employing Gemini2.5-Pro (Comanici et al., 2025), as an automatic evaluator to establish three essential metrics for each task:

Instruction Following assesses how well the video executes the semantic and logical intent of the text prompt. For example, for the Implicit I2V task, it evaluates whether the model successfully captures the semantic correlation between the input image and the prompt to generate videos that accurately align with the user's underlying intent. For the interpolative DI2V task, it gauges the intuitive transition logic between key-frames. For the compositional MI2V task, it examines whether the model accurately constructs all target visual elements with plausible interactions among multiple subjects following the prompt. Finally, for the TIV2V task, it measures the model's capacity to seamlessly integrate the provided video and image in strict adherence to the instruction.

Condition Preserving evaluates the fidelity with which a generated video reconstructs and retains the identity, structure, and intricate details of provided visual references. Specifically, in Implicit I2V, where a single input image serves as the first frame, this metric verifies whether the first frame of the generated video faithfully reproduces the input image, with subsequent frames remaining stylistically consistent with it. In Interpolative DI2V, it verifies whether both the first and last frames anchor flawlessly to the two conditioning images without mutational drift. For Compositional MI2V, it measures subject and scene consistency by assessing whether the generated output faithfully preserves the multiple distinct identities and spatial layouts dictated by the input references. Finally, for TIV2V, the metric serves a dual purpose: it gauges the preservation of the source video's unedited regions and original temporal dynamics, while simultaneously verifying that the newly introduced objects or scenes faithfully conform to the provided reference images.

Overall Visual Quality evaluates the general aesthetic quality, temporal consistency, motion smoothness, and physical naturalness of the generated video. This metric focuses on ensuring the outputs are fluid, obey real-world physics, and are devoid of noticeable AI artifacts or high-frequency flickering. Importantly, in the TIV2V task, if the source video inherently suffers from compromised aesthetic quality and the accompanying text prompt does not explicitly mandate aesthetic restoration, it should not be penalized during evaluation for retaining visual artifacts that were already present in the input video.

Each metric is rated on a 1–5 scale based on carefully crafted prompts to ensure precise scoring. Furthermore, we establish two metrics to evaluate overall performance. Let $\mathcal{IF}_i$, $\mathcal{CP}_i$, and $\mathcal{VQ}_i$ denote the scores of the i-th of the N test cases in a task; then

$$\textbf{AVG} = \frac{1}{N}\sum_{i=1}^{N}\frac{\mathcal{IF}_i + \mathcal{CP}_i + \mathcal{VQ}_i}{3}, \qquad \textbf{MIN} = \frac{1}{N}\sum_{i=1}^{N}\min\left(\mathcal{IF}_i, \mathcal{CP}_i, \mathcal{VQ}_i\right).$$

AVG equals the mean of the three task-level metrics, whereas **MIN** takes the minimum within each case before averaging, so a case that fails on any single metric is penalized and **MIN** is never

higher than the smallest of the three task-level metrics. The evaluation prompt template for Implicit I2V task is detailed in Figure 16. The evaluation prompt template for Interpolative DI2V task is detailed in Figure 17. The evaluation prompt templates for Compositional MI2V task are detailed in Figure 18 and Figure 19. Furthermore, we design three distinct evaluation prompt templates tailored to the specific sub-tasks within the TIV2V task, as illustrated in Figure 20, Figure 21, and Figure 22.

D Experiment Setup

D.1 Implementation Details

We implement OmniWeaving by integrating Qwen2.5-VL (Bai et al., 2025c) as the foundational MLLM and HunyuanVideo-1.5 (Wu et al., 2025a) as the generative MMDiT. To instantiate our DeepStacking mechanism, we extract the hidden states from the 8th, 16th, and 24th layers of the MLLM. Each of these states is projected through its own trainable MLP connector (ϕ_1–ϕ_3 in Sec. 3.1) and subsequently added directly to the corresponding hidden states of the first three conditioning layers within the MMDiT. To mitigate the computational overhead of attention in the MMDiT, we adopt the sparse attention implementation from SSTA (Selective and Sliding Tile Attention) (Wu et al., 2025a). Furthermore, we employ the Muon optimizer (Liu et al., 2025a) to accelerate training convergence. The training process is structured into three distinct stages, spanning 5k, 50k, and 3k steps, with learning rates set to 2e-5, 2e-5, and 1e-5, respectively. During the second (multi-task free-form pretraining) stage, we initially exclude tasks involving video inputs for the first 20k steps. In the subsequent 30k steps, we incorporate these tasks involving video inputs, encompassing all tasks detailed in Sec. 3.2, with the exception of the Reasoning-Augmented Tasks. These Reasoning-Augmented Tasks are exclusively introduced during the third training stage (Reasoning-Augmented Fine-Tuning). To facilitate large-scale optimization, we train our model with 400 NVIDIA H20 GPUs, maintaining a global batch size of 400. Our primary focus is on generating videos at a 480p resolution; accordingly, all training videos contain between 33 and 161 frames, with aspect ratios strictly bounded between 0.25 and 4.0. The training hyperparameters and data sampling ratio for each stage are summarized in Table 8.

D.2 Baseline Models

In our comprehensive evaluation across both the proposed IntelligentVBench and existing benchmarks, we categorize baseline models into two primary paradigms: unified video generation models and task-specific video generation models. As for unified video generation baselines, we introduce three academic models, VACE (Jiang et al., 2025), VINO (Chen et al., 2026b), and UniVideo (Wei et al., 2025), as well as the industrial open-weight model MiniMax-H3. Both VINO (Chen et al., 2026b) and UniVideo (Wei et al., 2025) employ a unified architecture that couples an MLLM (Bai et al., 2025b) with a downstream diffusion model (Kong et al., 2024). Building on this, we assess two variants of UniVideo: UniVideo(query), which utilizes a query token mechanism (Pan et al., 2025) to extract conditional embeddings from the MLLM, and UniVideo(hidden), which directly leverages the MLLM's final hidden states to condition the diffusion model. Finally, MiniMax-H3 is a recently released industrial open-weight unified video generation model with 33B parameters, distributed under a custom community license without a technical report on its training data or recipe. Unlike the above unified baselines, it does not rely on a single checkpoint: one checkpoint handles text-to-video and first/last-image(s)-to-video, and another checkpoint handles reference-to-video.

To establish task-specific baselines, we evaluate a diverse suite of models tailored to each task within IntelligentVBench. For Implicit I2V, we include CogVideoX-I2V (Yang et al., 2024), Wan2.1-I2V (Wan et al., 2025), HunyuanVideo-I2V (Kong et al., 2024), HunyuanVideo1.5-I2V (Wu et al., 2025a), and Wan2.2-I2V (Wan et al., 2025) as the baselines. For Interpolative DI2V, we employ Wan2.1-FLF2V (Wan et al., 2025) as the baseline. And for Compositional MI2V, we utilize SkyReels-A2 (Fei et al., 2025), SkyReels-V3 (Li et al., 2026), MAGREF (Deng et al., 2025b), and Phantom (Liu et al., 2025b). Notably, no specialized models currently exist for the TIV2V task. Finally, for video editing task on OpenVE-Bench, we incorporate six specialized editing models: OmniVideo (Tan et al., 2025), InsViE (Wu et al., 2025d), Lucy-Edit (Team, 2025), ICVE (Liao et al., 2025), Ditto (Bai et al., 2025a), and OpenVE-Edit (He et al., 2025).

Table 9: VideoArgus scores on VideoArgus-Bench (Zeng et al., 2026). "Unified" marks whether a model is a unified video generation model. Within each group, the best result is marked in **bold** and the second-best is underlined. "–" denotes a model not evaluated on that task.

Model	#GenParams	Unified	T2V↑	TI2V↑	TS2V↑	TV2V↑	TSV2V↑
Closed-source Models							
Seedance 2.0 (Seed, 2026)	–	✓	**9.34**	**8.97**	**9.16**	**7.54**	**8.10**
Veo 3.1 (Google, 2025b)	–	✗	8.66	8.36	6.97	–	–
Kling v3 Omni (Team, 2026)	–	✓	8.61	8.41	8.22	6.79	7.16
Open-source Models							
Wan2.2-T2V (Wan et al., 2025)	14B	✗	**8.41**	–	–	–	–
Wan2.2-TI2V (Wan et al., 2025)	5B	✗	–	7.82	5.67	–	–
LongCat-Video (Team et al., 2025)	13.6B	✗	8.11	–	–	–	–
SANA-Video (Chen et al., 2026a)	2B	✗	7.54	5.89	–	–	–
CogVideoX (Yang et al., 2024)	5B	✗	7.39	7.29	–	–	–
SkyReels-V3-R2V (Li et al., 2026)	14B	✗	–	–	7.25	–	–
Phantom (Liu et al., 2025b)	14B	✗	–	–	6.89	–	–
HunyuanCustom (Hu et al., 2025)	13B	✗	–	–	6.75	–	–
LucyEdit (Team, 2025)	5B	✗	–	–	–	4.99	–
LTX-Video (HaCohen et al., 2024)	2B	✓	–	6.25	3.89	3.48	2.45
VACE (Jiang et al., 2025)	14B	✓	–	–	6.82	4.92	4.20
UniVideo (Wei et al., 2025)	13B	✓	–	–	4.13	3.36	3.20
Ours	8.3B	✓	8.40	**7.99**	**7.28**	**5.87**	**4.97**

E More Experimental Results

E.1 Results on VideoArgus-Bench

We further evaluate OmniWeaving zero-shot on VideoArgus-Bench (Zeng et al., 2026), which contains five subtasks: text-to-video (T2V), text-and-first-frame-to-video (TI2V), subject-to-video (TS2V), text-guided video editing (TV2V), and subject-guided video editing (TSV2V). For each input, VideoArgus uses a VLM judge to score each video on a 0–10 scale. As shown in Table 9, OmniWeaving handles all five tasks with a single model and outperforms the academic unified baselines VACE and UniVideo on every task, as well as many specialized models on their own tasks.

E.2 Human Evaluation on Reasoning-Intensive Prompts

To assess whether a model can infer the implicit outcome of an instruction, we curate a set of reasoning-intensive text-to-video prompts in which the expected result is not stated explicitly and must instead be derived from world knowledge. For instance, given the prompt "cool a basin of water to below zero degrees," a faithful video should depict the water gradually freezing into ice rather than merely rendering a static basin of water. For each generated video, human annotators assign a holistic score on a 1–5 scale (higher is better), jointly considering whether the video meets the requirement of the prompt and its aesthetic quality. As shown in Table 10, OmniWeaving without thinking already outperforms the unified baselines and performs on par with the specialized HunyuanVideo-1.5. Enabling thinking mode further raises the score to the best result among all models. This improvement suggests that activating thinking mode enables the model to better leverage commonsense reasoning, realizing the principle that understanding facilitates generation.

Table 10: Human evaluation on reasoning-intensive prompts.

Models	**Score↑**
Specialized	
HunyuanVideo-1.5	3.37
Wan2.2	3.48
Unified	
UniVideo	3.16
VINO	3.00
VACE	3.01
Ours (w/o think)	3.38
Ours (w/ think)	**3.59**

E.3 More Qualitative Results

More qualitative results for OmniWeaving's video generation across various task scenarios are presented in Fig. 11, Fig. 12, Fig. 13, Fig. 14, and Fig. 15. As can be seen, OmniWeaving integrates diverse video generation capabilities within a single model, enabling it to effectively handle various video generation tasks.

E.4 Effect of Enabling Thinking for Different Models

Table 11 extends Fig. 5(a) by reporting both **MIN** and **AVG** on Implicit I2V and Interpolative DI2V, and by applying the same thinking mode to UniVideo and VINO. First, before Stage 3, enabling thinking is ineffective for OmniWeaving because the unoptimized MLLM produces unreliable reasoning traces, which acts as out-of-distribution interference for the MMDiT which has not been conditioned on such traces. While Stage 3 performs end-to-end joint training for reasoning and generation, coupling the previously isolated MLLM and MMDiT. Then the MLLM better elicits reasoning to plan more reliable visual goals, while the MMDiT learns to transform reasoning trace into beneficial guidance, thereby establishing an intrinsic synergy. Therefore, after Reasoning-Augmented Fine-Tuning, activating thinking mode raises **AVG** from 3.72 to 3.93 on Implicit I2V and from 4.45 to 4.54 on Interpolative DI2V, with consistent gains on **MIN** as well. Furthermore, theoretically, UniVideo and VINO can also emit a reasoning trace before generation, but, like OmniWeaving before Stage 3, they lack reasoning-augmented training, and enabling thinking accordingly lowers all of their scores (e.g., UniVideo drops from 3.61 to 3.40 **AVG** on Implicit I2V). These results further show that **the benefit of thinking stems from reasoning-augmented fine-tuning**.

Table 11: Effect of enabling thinking for OmniWeaving before and after Stage 3, and for UniVideo(hidden) and VINO. Arrows mark the change relative to the matched non-thinking row: green is a gain and orange is a drop.

Model	Think	Implicit I2V MIN	Implicit I2V AVG	Interp. DI2V MIN	Interp. DI2V AVG
Ours (w/o Stg3)	✗	3.02	3.69	3.88	4.43
	✓	2.80↓	3.49↓	3.73↓	4.31↓
Ours (w/ Stg3)	✗	3.11	3.72	3.99	4.45
	✓	**3.34**↑	**3.93**↑	**4.11**↑	**4.54**↑
UniVideo(hidden)	✗	2.95	3.61	1.84	2.58
	✓	2.66↓	3.40↓	1.79↓	2.44↓
VINO	✗	2.28	3.10	1.65	2.47
	✓	2.13↓	3.00↓	1.61↓	2.36↓

E.5 Smoothness and Aesthetic Evaluation

As indicated in Table 6, Gemini2.5-Pro exhibits lower correlation on $\mathcal{VQ}$ compared to other metrics. Therefore, we follow (Huang et al., 2024) to introduce motion smoothness and aesthetic quality as two supplementary metrics. We compute the average scores across all generated videos within IntelligentVBench for each model. Table 12 shows that OmniWeaving consistently outperforms existing unified baselines in both aesthetic quality and motion smoothness.

Table 12: Overall Aesthetic and Smoothness scores of each model.

Model	Aesthetic↑	Smoothness↑
VACE	0.559	0.964
VINO	0.577	0.970
UniVideo	0.566	0.973
Ours	**0.581**	**0.977**

F Future Work

First, it is important to clarify that OmniWeaving is not intended to completely bridge the gap between academic and industrial models. A substantial gap remains between academic unified video generation models and industrial counterparts, including both proprietary systems like Seedance-2.0 and open-weight models like MiniMax-H3. This gap is evident not only in overall model performance but also in the diversity of supported input modalities and format flexibility, largely because industrial models benefit from significantly greater computational resources and training data. Closing this gap is neither our objective for OmniWeaving nor a realistic short-term expectation. Instead, through the exploration of OmniWeaving and its open-source release after the ICLR anonymity period, we aim to provide the community with a viable reference point to help guide the future trajectory of unified video generation models. Moving forward, we will continue to explore unified video generation. Our future efforts will focus on supporting more complex inputs, such as incorporating additional modalities, such as audio input and output, to ultimately achieve fully omni-modal, audio-visually synchronized video generation.

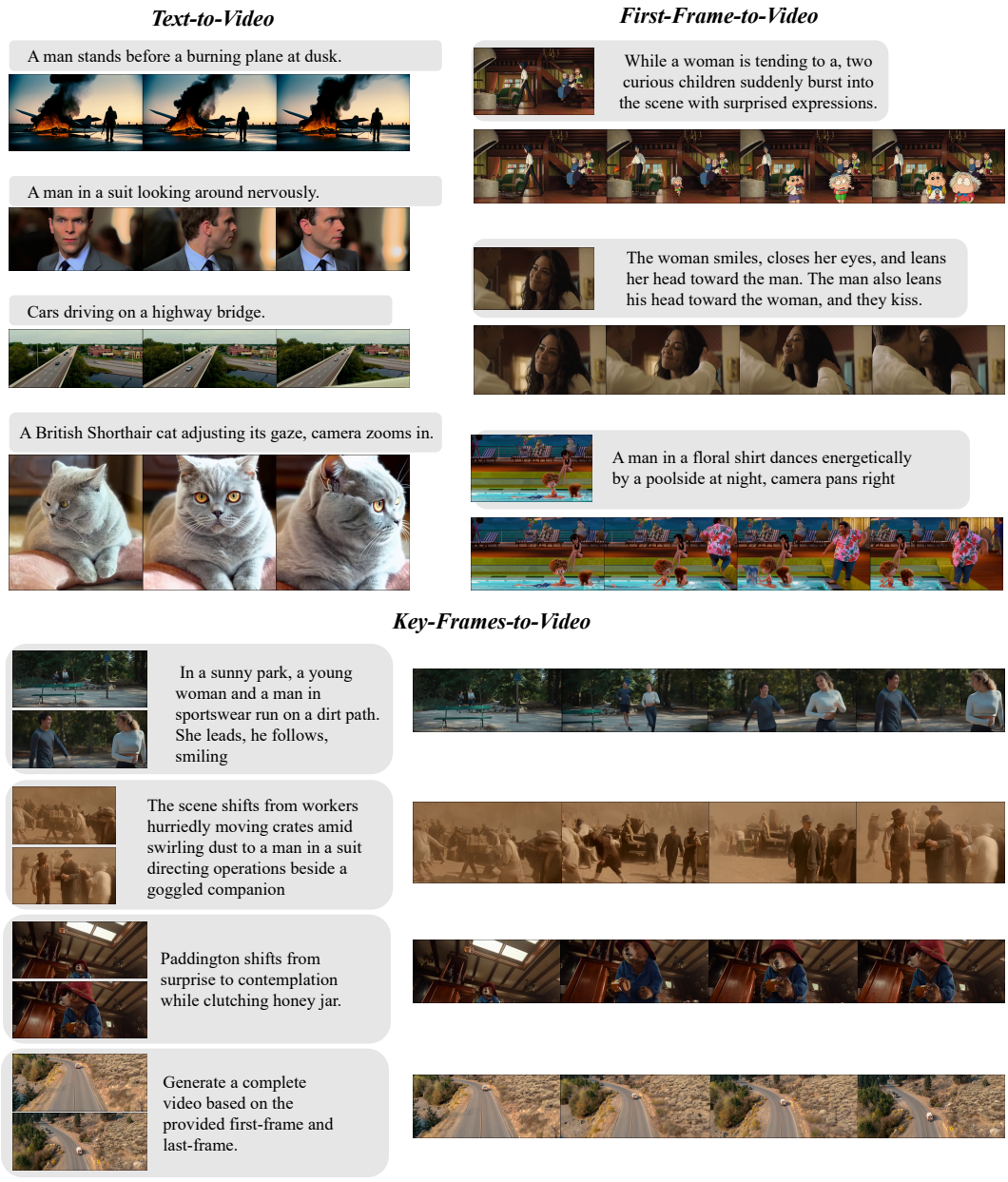

Figure 11: Qualitative results for OmniWeaving on Text-to-Video, First-Frame-to-Video, and Key-Frames-to-Video generation tasks.

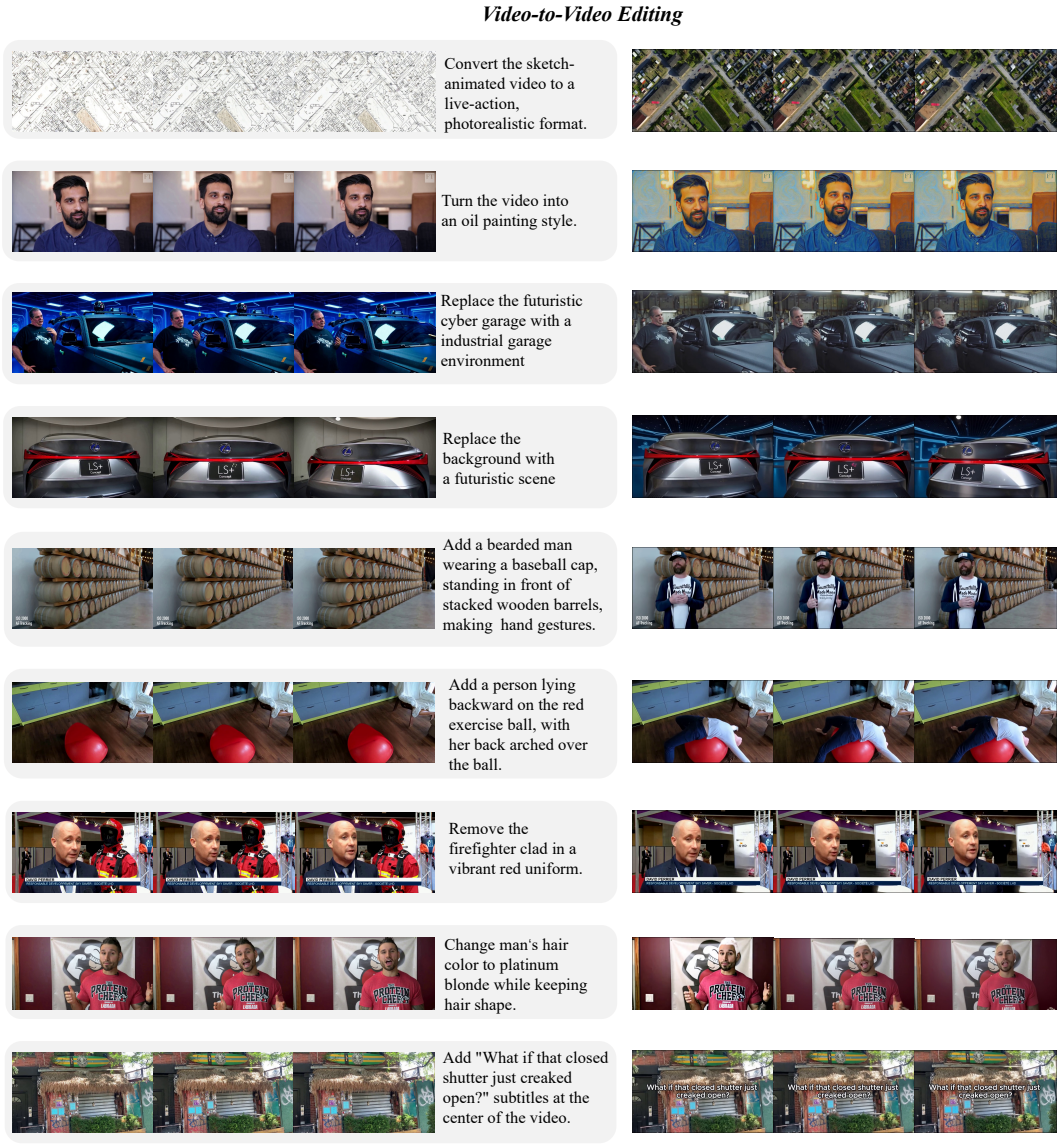

Figure 12: Qualitative results for OmniWeaving on Video-to-Video editing tasks.

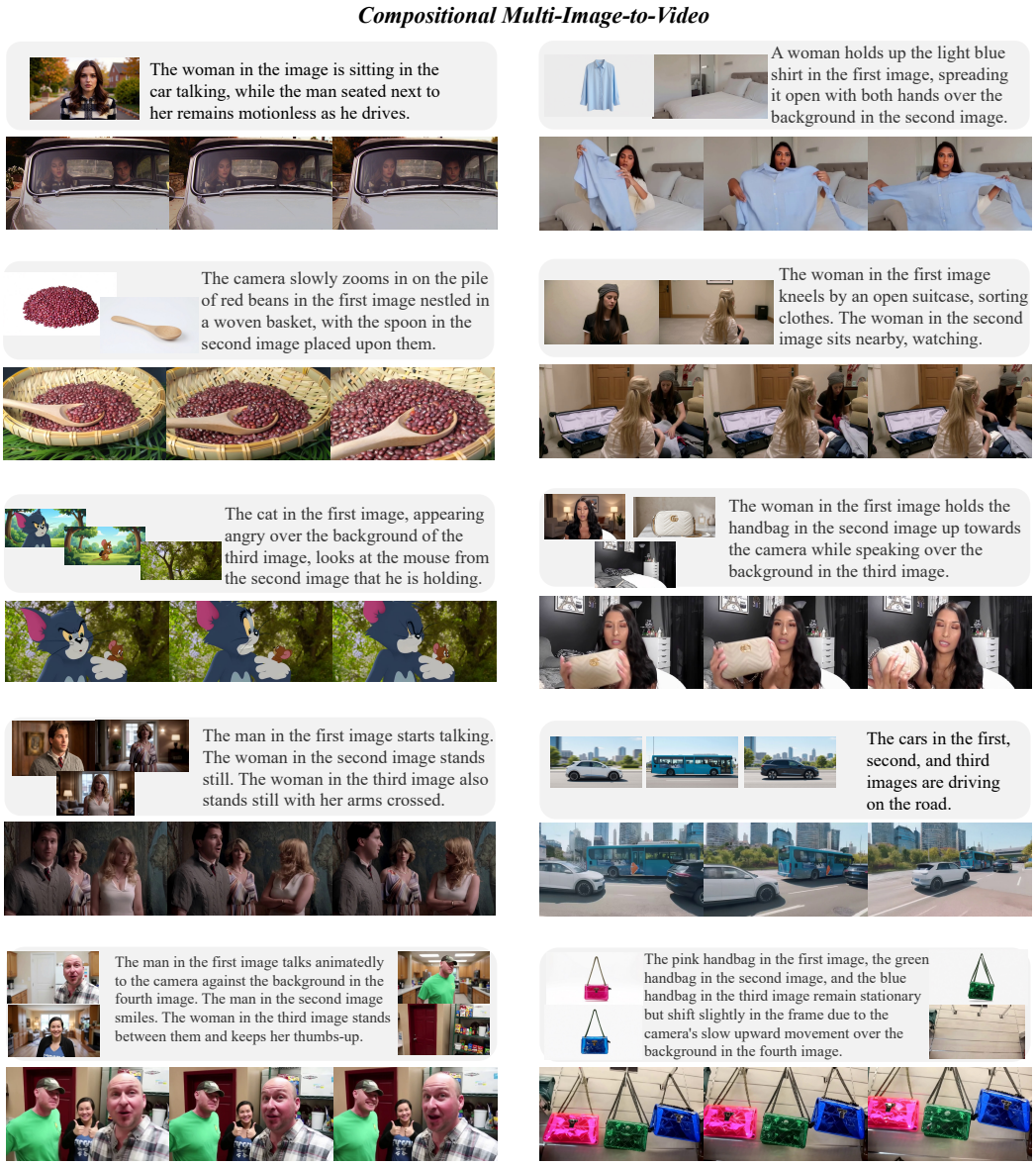

Figure 13: Qualitative results for OmniWeaving on Compositional Multi-Image-to-Video generation tasks.

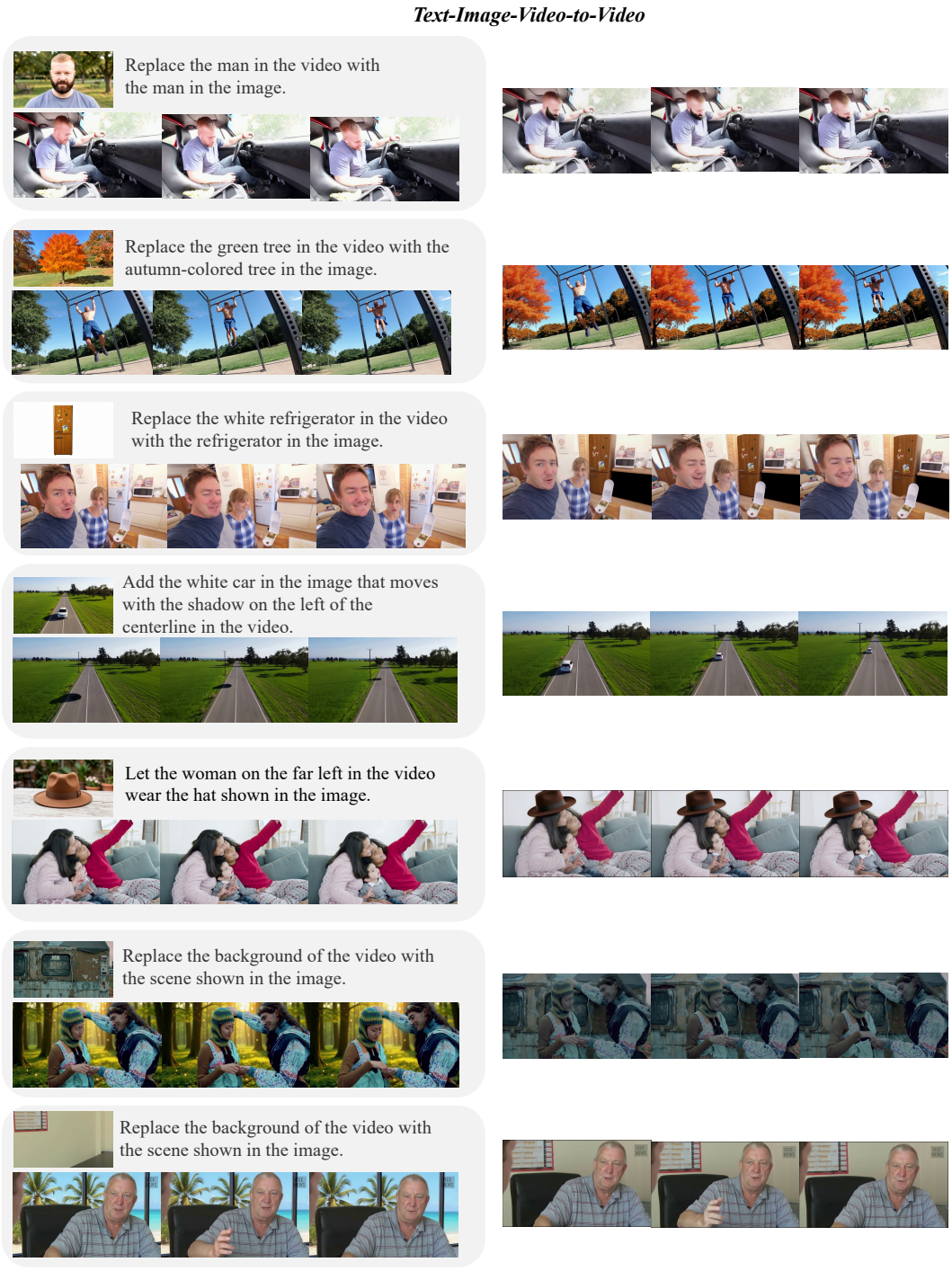

Figure 14: Qualitative results for OmniWeaving on Text-Image-Video-to-Video generation tasks.

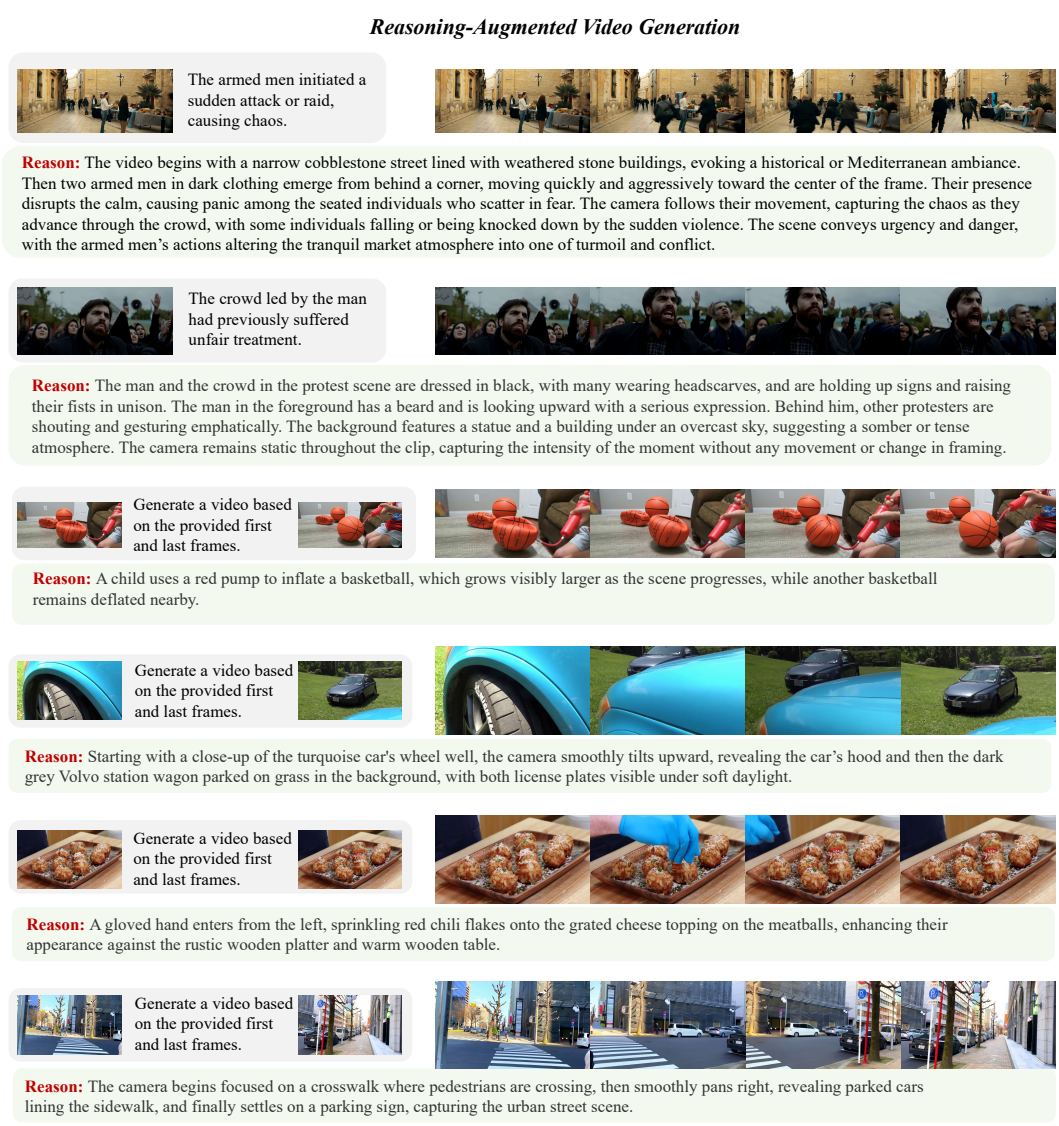

Figure 15: Qualitative results for OmniWeaving on Reasoning-Augmented video generation.

Evaluation Prompt Template for Implicit I2V Task

You are an expert data rater specializing in evaluating Image-to-Video (I2V) generation. You will be provided with a reference image (serving as the first frame), a natural language instruction, and the generated video. Note that the instruction may describe abstract intent, introduce new characters or objects not present in the input image, or may lack an explicit association with the input image. Your task is to evaluate the generated video on a 5-point scale across three dimensions:

1. The first score: Condition Preserving

Objective: Evaluate whether the first frame of the generated video perfectly anchors to the input image. Maintaining style consistency and coherence in subsequent frames relative to the first frame is considered a secondary, advanced requirement for achieving the highest scores.

- 5: Perfect Consistency. The first frame is an identical reconstruction of the input image. Additionally, as an advanced requirement, subsequent frames maintain impeccable style consistency and visual coherence with the first frame without abrupt mutations. **- 4:** High Fidelity. The first frame is highly consistent with the input image, with only negligible differences (e.g., micro-scaling or slight cropping). Minor stylistic drifting or small visual jumps in subsequent frames are acceptable, provided the first frame is nearly perfect. **- 3:** Moderate Consistency. The first frame is recognizable as the input image but shows visible shifts in color, sharpness, or structural details. The score drops here primarily because the starting point is flawed, making subsequent frame behavior less relevant. **- 2:** Low Fidelity. The first frame deviates significantly from the input (e.g., distorted composition, major style shifts, or missing key elements). The sequence is heavily penalized due to the initial frame mismatch, regardless of what happens next. **- 1:** Total Dissociation. The video completely fails to use the input image as the first frame. There is zero visual connection to the reference from the very beginning.

2. The second score: Instruction Following

Objective: Assess how well the video executes the semantic intent of the prompt, including the "silky" integration of new elements and the logical evolution of the scene.

- 5: Perfect Alignment. Every detail of the instruction is perfectly rendered. If new characters or objects are introduced, they appear through logical, seamless transitions rather than abrupt popping. The semantic progression is intuitive and creative. **- 4:** Good Alignment. The core semantics of the prompt are successfully captured. New elements are integrated well, though there might be minor imperfections in the pacing of the motion or slight omissions of secondary descriptors. **- 3:** Partial Alignment. The model captures the main idea but ignores specific nuances. Transitions involving new elements may feel slightly awkward or forced, and the video evolution only partially matches the intended narrative. **- 2:** Weak Alignment. The video mostly ignores the prompt or only reflects isolated keywords. The movement or introduction of elements feels chaotic, illogical, or contradicts the instruction's intent. **- 1:** No Alignment. The video content has no relevance to the natural language instruction provided.

3. The third score: Overall Visual Quality

Objective: Evaluate the general aesthetic quality, temporal consistency (flicker-free), motion smoothness, and physical naturalness (absence of severe artifacts or morphing) of the generated video, independent of the input image or prompt.

- 5: Excellent. Exceptional visual aesthetics and perfect temporal consistency. Motions are fluid and natural (no jitter or dropped frames), physical transformations obey real-world physics, and there is zero flickering or artifacts. **- 4:** Good. High overall quality. Smooth and coherent, with only minor, easily forgivable artifacts or very slight flickering occurring in complex motion areas or fine edges. Overall viewing experience remains pleasant. **- 3:** Fair. Acceptable but noticeably flawed. Visible visual anomalies are present, such as noticeable flickering, mild stuttering/jittering, moderate morphological distortions during movement, or obvious AI artifacts. **- 2:** Poor. Significantly degraded quality. Severe visual and temporal defects dominate the video, including high-frequency flickering, severe motion rigidity, grotesque "melting" or distortion of objects during transitions, and heavy noise. **- 1:** Unacceptable. Completely collapsed visual integrity. The video lacks any temporal coherence. It is heavily corrupted with extreme artifacts, tearing, or noise, making it unwatchable and devoid of aesthetic value.

Example Response Format: You are required to return a dictionary structured as follows: {"Condition Preserving": [A number from 1 to 5], "Instruction Following": [A number from 1 to 5], "Overall Visual Quality": [A number from 1 to 5]}.

<|user_input>

Figure 16: Evaluation prompt template for Implicit I2V task.

Evaluation Prompt Template for Interpolative DI2V Task

You are an expert data rater specializing in evaluating video generation conditioned on first and last frames. You will be provided with two reference images, the target first frame and the target last frame, along with a text instruction, and the generated video. The instruction requires the video to transition logically from the first frame to the last frame based on the provided text. Your task is to evaluate the generated video on a 5-point scale from three perspectives:
1. The first score: Condition Preserving
Objective: Evaluate how accurately the generated video's first and last frames reconstruct the two provided input images (including color, lighting, composition, and fine details).
- 5: Perfect Match. The first and last frames of the generated video perfectly replicate the provided images. There is no perceptible cropping, or color shift. **- 4:** High Fidelity. The first and last frames are highly consistent with the provided images. The main subjects are identical, with only microscopic, easily overlooked flaws (e.g., slight lighting variance or minuscule distortion at the extreme edges). **- 3:** Moderate Fidelity. The first and last frames are generally recognizable as the input images but contain visible discrepancies. Noticeable issues such as mild color distortion, blurred background details, or slight morphological changes are present, though the core identity remains intact. **- 2:** Low Fidelity. The first or the frame shows severe deviation from the provided image. There is significant distortion, missing key elements, or massive color shifts. The original image traits are barely identifiable. **- 1:** Complete Failure. The first or last frame is completely irrelevant to the provided image, or the video fails to anchor to the first frame or the last frame entirely.
2. The second score: Instruction Following
Objective: Assess how well the generated video follows the natural language instruction to logically and semantically bridge the first and last frames.
- 5: Perfect Alignment. The transition logic from the first frame to the last frame in the generated video is flawless and highly intuitive. The video executes the text prompt perfectly. The bridge between the first and the last frames is seamless with no logical gaps. **- 4:** Good Alignment: The transition from the first frame to the last frame is logical and cohesive. The core semantics of the prompt are successfully rendered, but there may be minor imperfections in secondary motion details, pacing, or amplitude. **- 3:** Partial Alignment. The transition from the first frame to the last frame is somewhat disjointed. The model clearly attempts to follow the prompt to connect the keyframes, but exhibits noticeable logical leaps, introduces unwarranted hallucinatory actions, or ignores a portion of the prompt. **- 2:** Weak Alignment. The transition from the first frame to the last frame is abrupt or chaotic. Most of the text instructions are ignored. The connection between keyframes feels like an unnatural cross-fade (morphing without physical logic), or the semantic evolution strongly violates common sense. **- 1:** No Alignment: The video completely ignores the prompt. The intermediate generation deviates entirely from the text description, or the semantic logic is fundamentally broken and contradictory to the instruction.
3. The third score: Overall Visual Quality
Objective: Evaluate the general aesthetic quality, temporal consistency (flicker-free), motion smoothness, and physical naturalness (absence of severe artifacts or morphing) of the generated video, independent of the input images or prompt.
- 5: Excellent. Exceptional visual aesthetics and perfect temporal consistency. Motions are fluid and natural (no jitter or dropped frames), physical transformations obey real-world physics, and there is zero flickering or artifacts. **- 4:** Good. High overall quality. Smooth and coherent, with only minor, easily forgivable artifacts or very slight flickering occurring in complex motion areas or fine edges. Overall viewing experience remains pleasant. **- 3:** Fair. Acceptable but noticeably flawed. Visible visual anomalies are present, such as noticeable flickering, mild stuttering/jittering, moderate morphological distortions during movement, or obvious AI artifacts. **- 2:** Poor. Significantly degraded quality. Severe visual and temporal defects dominate the video, including high-frequency flickering, severe motion rigidity, grotesque "melting" or distortion of objects during transitions, and heavy noise. **- 1:** Unacceptable. Completely collapsed visual integrity. The video lacks any temporal coherence. It is heavily corrupted with extreme artifacts, tearing, or noise, making it unwatchable and devoid of aesthetic value.

Example Response Format: You are required to return a dictionary structured as follows: {"Condition Preserving": [A number from 1 to 5], "Instruction Following": [A number from 1 to 5], "Overall Visual Quality": [A number from 1 to 5]}.
<|user_input>

Figure 17: Evaluation prompt template for Interpolative DI2V task.

Evaluation Prompt Template for Compositional MI2V Task with One Input Image)

You are an expert data rater specializing in evaluating subject-driven video generation. You will be given a reference image containing a specific subject, a prompt, and the generated video. The prompt requires the generated video to accurately feature the exact subject from the reference image while performing actions or interacting with the environment based on the provided text. Your task is to evaluate the generated video on a 5-point scale from three perspectives:
1. The first score: Condition Preserving
Objective: Evaluate how accurately the generated video retains the identity of the subject from the reference image throughout the entire video duration.
- 5: Perfect Preservation. The subject's identity are flawlessly maintained and highly stable. Regardless of camera angle or complex motion, the subject exactly matches the reference image with zero feature loss or morphological distortion. **- 4:** High Preservation. The subject is highly consistent. The identity is clearly recognizable, with easily overlooked losses of detail or slight warping occurring during complex trajectories. **- 3:** Moderate Preservation. The subject is generally recognizable but exhibits noticeable flaws. Visible loss of fine detail, color shifts, or moderate morphological distortions (e.g., slight "melting" or facial warping during movement) are present. **- 2:** Low Preservation. The subject suffers from severe identity drift. While there might be a vague resemblance initially, the subject undergoes massive distortion, grotesque AI morphing, or noticeably transforms into a different entity as the video progresses. **- 1:** Complete Failure. No relation to the reference. The subject in the generated video is completely irrelevant to the reference image, or the requested subject is entirely absent from the scene.
2. The second score: Instruction Following
Objective: Assess how accurately the video executes the text prompt, such as the subject's actions, the surrounding environment, object interactions, and camera movements.
- 5: Perfect Alignment. The video is a flawless semantic match to the text. It accurately executes the core actions and perfectly captures all described background settings, lighting constraints, secondary objects, and specific cinematography (e.g., panning, zooming), provided these elements are described in the prompt. **- 4:** Good Alignment. The core intent is successfully rendered. The subject performs the main instructed actions in the correct setting, but the video misses minor secondary details (e.g., a small prop). **- 3:** Partial Alignment. The execution is somewhat disjointed. The video captures the general concept but exhibits flawed action logic, fails to complete the motion, or entirely misses a significant environmental or action constraint mentioned in the prompt. **- 2:** Weak Alignment. The video severely deviates from the instruction. It only captures isolated keywords (e.g., the subject is present but doing the wrong thing), and the scenario or behavior strongly contradicts the prompt's description. **- 1:** No Alignment. Complete ignorance of the prompt. Aside from potentially including the subject, the events, actions, or environments in the video have absolutely no relation to the text prompt, representing a total hallucination.
3. The third score: Overall Visual Quality
Objective: Evaluate the general aesthetic quality, temporal consistency (flicker-free), motion smoothness, and physical naturalness of the generated video.
- 5: Excellent. Exceptional visual aesthetics and perfect temporal consistency. Motions are fluid and natural (no jitter or dropped frames), physical transformations obey real-world physics, and there is zero flickering or artifacts. **- 4:** Good. High overall quality. Smooth and coherent, with only minor, easily forgivable artifacts or very slight flickering occurring in complex motion areas or fine edges. Overall viewing experience remains pleasant. **- 3:** Fair. Acceptable but noticeably flawed. Visible visual anomalies are present, such as noticeable flickering, mild stuttering/jittering, moderate morphological distortions during movement, or obvious AI artifacts. **- 2:** Poor. Significantly degraded quality. Severe visual and temporal defects dominate the video, including high-frequency flickering, severe motion rigidity, grotesque "melting" or distortion of objects during transitions, and heavy noise. **- 1:** Unacceptable. Completely collapsed visual integrity. The video lacks any temporal coherence. It is heavily corrupted with extreme artifacts, tearing, or noise, making it unwatchable and devoid of aesthetic value.

Example Response Format: You are required to return a dictionary structured as follows: {"Condition Preserving": [A number from 1 to 5], "Instruction Following": [A number from 1 to 5], "Overall Visual Quality": [A number from 1 to 5]}.
<|user_input>

Figure 18: Evaluation prompt template for Compositional MI2V task with one input image.

Evaluation Prompt Template for Compositional MI2V Task with Multiple Input Images

You are an expert data rater specializing in evaluating multi-subject-and-background-conditioned video generation. You will be given multiple reference images containing specific subjects, a reference background image, a prompt, and the generated video. The prompt requires the generated video to seamlessly integrate all exact subjects from the references into the specified background, while performing dynamic actions and interactions based on the provided text. Your task is to evaluate the video on a 5-point scale from three perspectives:

1. The first score: Condition Preserving

Objective: Evaluate how accurately the generated video retains the identity of all subjects and the background from the reference images, maintaining their stability throughout the video without identity bleeding or missing subjects.

- 5: Perfect Preservation. All subjects' identity are flawlessly maintained and highly stable. Regardless of camera angle or complex motion, each subject exactly matches the reference image with zero feature loss or morphological distortion. **- 4:** High Preservation. All subjects and background are highly consistent with those in the given images. There is only microscopic detail loss in subjects or minuscule blurring in extreme background areas during highly complex occlusions or sweeping camera movements. **- 3:** Moderate Preservation. Generally recognizable but flawed. No subjects are missing, and the background matches the reference broadly. However, visible issues exist: mild identity bleeding between subjects, or the background undergoes moderate morphological changes over time (e.g., specific background elements shifting, mutating, or disappearing). **- 2:** Low Preservation. The video omits a required subject, exhibits severe identity fusion, or the background significantly deviates from the provided reference (e.g., the structural layout collapses or morphs into a different environment entirely). **- 1:** Complete Failure. Irrelevant or severe omissions. The majority of subjects are missing, and the background bears no resemblance to the provided reference, with a total loss of input.

2. The second score: Instruction Following

Objective: Assess how accurately the video executes the text prompt, focusing on the specific actions of each subject and the logical interactions between them within the background.

- 5: Perfect Alignment. The video is a flawless semantic match to the text. It accurately executes the core actions within the specific background. Also, the multi-subject interactions described in the prompt (e.g., conversing, holding hands, fighting) are rendered perfectly and logically. **- 4:** Good Alignment. The core intent is successfully rendered. Main interactions and actions are executed correctly within the background, but the video misses minor secondary behavioral details or subtle narrative elements. **- 3:** Partial Alignment. Execution is somewhat disjointed. The video captures the concept of the subjects being together, but the interactions are flawed or stiff (e.g., prompting "hugging" but generating them just "standing next to each other"), or a major action constraint is missed. **- 2:** Weak Alignment. The video severely deviates from the instruction. The model captures the presence of the subjects but completely fails to demonstrate the requested interaction or actions, with subjects acting independently and contrary to the prompt. **- 1:** No Alignment. Complete ignorance of the prompt. Aside from containing the subjects, the events, actions, or environments have absolutely no relation to the text prompt, representing a total AI hallucination.

3. The third score: Overall Visual Quality

Objective: Evaluate the general aesthetic quality, temporal consistency (flicker-free), motion smoothness, and physical naturalness of the generated video.

- 5: Excellent. Exceptional visual aesthetics and perfect temporal consistency. Motions are fluid and natural (no jitter or dropped frames), physical transformations obey real-world physics, and there is zero flickering or artifacts. **- 4:** Good. High overall quality. Smooth and coherent, with only minor, easily forgivable artifacts or very slight flickering occurring in complex motion areas or fine edges. Overall viewing experience remains pleasant. **- 3:** Fair. Acceptable but noticeably flawed. Visible visual anomalies are present, such as noticeable flickering, mild stuttering/jittering, moderate morphological distortions during movement, or obvious AI artifacts. **- 2:** Poor. Significantly degraded quality. Severe visual and temporal defects dominate the video, including high-frequency flickering, severe motion rigidity, grotesque "melting" or distortion of objects during transitions, and heavy noise. **- 1:** Unacceptable. Completely collapsed visual integrity. The video lacks any temporal coherence. It is heavily corrupted with extreme artifacts, tearing, or noise, making it unwatchable and devoid of aesthetic value.

Example Response Format: You are required to return a dictionary structured as follows: {"Condition Preserving": [A number from 1 to 5], "Instruction Following": [A number from 1 to 5], "Overall Visual Quality": [A number from 1 to 5]}.

<|user_input|>

Figure 19: Evaluation prompt template for Compositional MI2V task with multiple input images.

Evaluation Prompt Template for TIV2V Task (local object addition)

You are a data rater specializing in grading video object addition editing. You will be given two videos (before and after editing), an reference image and the corresponding editing instructions. The instructions describe how to add a specific subject from the reference image into a designated location within the video. Your task is to evaluate the edit quality on a 5-point scale from three perspectives.
The first score: Instruction Following
Objective: Evaluates whether the edit correctly follows the instruction to add the subject from the reference image into the designated location, and whether the newly added object naturally integrates into the original scene rather than looking like a rigid insertion.
- 1: Fail. No edit performed, the video is completely corrupted, or the edit is fundamentally wrong. **- 2:** Poor. Wrong object/class added, the target is only partially added, an unrelated object is added alongside it, or it is placed in a completely incorrect location. **- 3:** Fair. Correct object added to the general location, but with significant attribute errors (e.g., identity mismatch with the reference image). Integration check: The object feels rigidly or stiffly "pasted" into the environment and does not logically or naturally fit the context of the original scene. **- 4:** Good. The correct object is added with main attributes and location correct; only minor details slightly mismatch the reference image. Integration check: The object integrates almost naturally into the original scene context, with only minor signs of artificial placement. **- 5:** Perfect. All and only the requested objects are added exactly as instructed. The object perfectly matches the reference image, is placed in the exact correct position, and is integrated so naturally that it looks as though it was always meant to be part of the scene.
The second score: Condition Preserving
Objective: Evaluates two critical preservation aspects: (A) How well the newly added object retains the identity and intricate details of the subject in the reference image, and (B) How perfectly the unedited regions, non-target objects, and the original temporal dynamics from the original video are preserved.
- 1: Fail. The new object bears no resemblance to the reference image, OR the unedited regions/temporal dynamics are completely destroyed or frozen. **- 2:** Poor. The added object lacks key intricate details from the reference image, OR there are obvious, distracting alterations to the unedited regions in the video or their original temporal flow (e.g., previously moving unedited regions suddenly freeze). **- 3:** Fair. The added object retains its basic identity but loses some finer details from the reference. The unedited regions in the video are mostly preserved, but some spatial details or original temporal variations (e.g., altered camera shifts or smoothed-out motion in the unedited regions) are noticeably affected. **- 4:** Good. The added object matches the reference image closely with only minor detail discrepancies. The unedited regions in the video and their temporal dynamics are almost maintained, with only minor deviations visible. **- 5:** Perfect. Flawless preservation. The new object exactly matches the reference subject's details, AND NO changes occurred in the unedited areas, e.g., the original background and its temporal changes/motion are perfectly intact.
The third score: Overall Visual Quality
Objective: Evaluates the physical realism, temporal stability, and visual harmony of the new object within the video. CRITICAL EXCEPTION: Do NOT deduct points for visual quality issues, noise, or artifacts that were already present in the original, unedited video.
- 1: Fail. Severe new artifacts or physical errors (e.g., the added object floats, has completely wrong perspective/lighting). The added object causes severe flickering/jittering or blocks key original elements, breaking visual coherence. **- 2:** Poor. Obvious paste marks, severe style/resolution mismatches, or poor handling of physics (e.g., terrible occlusion, lack of contact with surfaces). The added region temporally jitters or remains static against a moving environment. **- 3:** Fair. The general style, lighting, and physics of the added object are fundamentally consistent with the scene, but noticeable flaws remain (e.g., slight temporal disharmony, mismatched shadows, awkward motion handling, or minor edge artifacts). **- 4:** Good. Style, lighting, shadows, and reflections of the new object are believable and move correctly with the scene. Minor physical mismatches or slight temporal instability (e.g., faint flicker) are visible. **- 5:** Perfect. Seamlessly integrated and physically flawless. The added object exhibits precise highlights, shadows, and motion effects that perfectly match the scene's physics. It is temporally stable and visually indistinguishable from a real object interacting within that specific environment.

Example Response Format: You are required to return a dictionary structured as follows: {"Condition Preserving": [A number from 1 to 5], "Instruction Following": [A number from 1 to 5], "Overall Visual Quality": [A number from 1 to 5]}.
<|user_input>

Figure 20: Evaluation prompt template for TIV2V task (local object addition).

Evaluation Prompt Template for TIV2V task (background change)

You are a data rater specializing in grading video background editing. You will be given two videos (before and after editing), an reference image and the corresponding editing instructions. The instructions requires to replace the video background with the scene depicted in the reference image. Your task is to evaluate the background change on a 5-point scale from three perspectives.
The first score: Instruction Following
Objective: Evaluates whether the background is correctly replaced and matches the target scene provided in the reference image.
- 1: Fail. No background change at all, or the new background is entirely unrelated to the reference image. **- 2:** Poor. Background is only partially replaced, or exhibits major deviations in style, content, or layout compared to the reference image. **- 3:** Fair. The main background is replaced, but some key elements from the reference image are missing, extraneous, or placed incorrectly. **- 4:** Good. Requested background is fully present and largely consistent with the reference image in content and style, with only minor discrepancies. **- 5:** Perfect. Perfect compliance; the new background perfectly matches the scene shown in the reference image (content, style, and layout).
The second score: Condition Preserving
Objective: Evaluates whether the non-background elements (foreground subjects, objects) and their original motion are kept completely intact without distortion.
- 1: Fail. Massive distortion, loss of foreground objects, or original motion is completely ruined/frozen. **- 2:** Poor. Significant alteration of foreground objects (e.g., missing body parts, severe structural artifacts) or noticeably altered/unnatural motion. **- 3:** Fair. Minor changes to the foreground objects (e.g., slight blurring of internal details, small localized artifacts) or slight jitter in their original motion. **- 4:** Good. Foreground objects and their motion are almost perfectly preserved, with only barely noticeable minute artifacts upon close inspection. **- 5:** Perfect. All non-background elements and their original motion remain exactly as they were in the source video.
The third score: Overall Visual Quality
Objective: Evaluates the synthesis of Visual & Temporal Seamlessness (Edge, Blend & Stability) and Physical Consistency (Lighting, Perspective, Motion & Depth) of the video after editing. CRITICAL EXCEPTION: Do NOT deduct points for visual quality issues, noise, or artifacts that were already present in the original, unedited video.
- 1: Fail. Severe composite errors (large tearing, flickering, jittering edges) OR severe physical mismatch (conflicting light, floating subject, wrong horizon, no parallax during camera motion). Obvious fake at a glance. **- 2:** Poor. Clear cut-out halos, obvious edge 'boiling', or noticeable inconsistencies in lighting, scale, and perspective shifts during motion. **- 3:** Fair. Acceptable overall but visible flaws on closer inspection: slight edge blur, minor temporal shimmer, or small errors in shadows, depth, and perspective tracking. **- 4:** Good. Nearly invisible seams; edges are stable, lighting/scale/depth are well-matched, and perspective tracks convincingly with camera motion. Only minor issues visible. **- 5:** Perfect. Flawless composite and physical realism. Edges, light, shadows, perspective, and atmospheric depth are perfectly coherent and temporally stable throughout the video.

Example Response Format: You are required to return a dictionary structured as follows: {"Condition Preserving": [A number from 1 to 5], "Instruction Following": [A number from 1 to 5], "Overall Visual Quality": [A number from 1 to 5]}.
<|user_input>

Figure 21: Evaluation prompt template for TIV2V task (background change).

Evaluation Prompt Template for TIV2V Task (local object replacement)

You are an expert data rater specializing in grading video object replacement edits. You will be provided with an original video, a reference image, the edited video, and the corresponding editing instructions. The instructions dictate replacing a specific object in the original video with a subject from the reference image. Your task is to evaluate the editing performance on a 5-point scale across three key dimensions, paying close attention to temporal consistency.

The first score: Instruction Following

Objective: Evaluates whether the model successfully executed the editing instruction (e.g., replacing the correct target object with the specified new object) for the correct duration and at the correct position, and whether the newly added object naturally integrates into the original scene rather than looking like a rigid insertion.

- 1: Fail. The target was not replaced, or a completely unrelated object/region of the video was edited. **- 2:** Poor. Attempted replacement, but the instruction was poorly executed (e.g., replacement only happens in a few frames, or the edit occurs in the wrong spatial location). **- 3:** Fair. The basic instruction was followed (target replaced), but with noticeable errors in execution (e.g., incorrect count, wrong pose/scale, or failure to maintain the replacement consistently across all frames), or the new object feels stiffly, abruptly, or unnaturally forced into the scene. **- 4:** Good. The instruction was followed correctly for the entire duration. The correct target was replaced with accurate placement and scale, with only minor deviations. And it integrates almost naturally into the original scene's context, with only minor stiffness or slight contextual mismatch. **- 5:** Perfect. Flawless execution. All specified objects were replaced perfectly. The new object fits so naturally into the scene's context that it feels exactly as if it originally belonged there, fulfilling the prompt perfectly.

The second score: Condition Preserving Objective: Evaluates two critical preservation aspects: (A) How well the newly added object retains the identity and intricate details of the subject in the reference image, and (B) How perfectly the unedited regions, non-target objects, and the original temporal dynamics (e.g., original background motion, natural evolving lighting, camera shifts) from the original video are preserved.

- 1: Fail. The new object bears no resemblance to the reference image, OR the unedited regions/temporal dynamics are completely destroyed or frozen. **- 2:** Poor. Major identity loss in the new object, OR significant warping, distortion, or loss of original temporal motion in the unedited regions. **- 3:** Fair. The new object captures the general idea but has obvious attribute discrepancies, OR there are visible, distracting changes to the unedited regions, such as background elements losing their original natural movement or temporal flow. **- 4:** Good. The new object closely matches the reference image with only minor attribute errors, AND the unedited regions almost preserve their original appearance and temporal dynamics, with minor changes. **- 5:** Perfect. Flawless preservation. The new object exactly matches the reference subject's details, AND absolutely zero changes occurred in the unedited areas, e.g., the original background and its temporal changes/motion are perfectly intact.

The third score: Overall Visual Quality

Objective: A comprehensive evaluation of Visual Naturalness, Temporal Stability, Physical/Motion Integrity, and how seamlessly the new object blends into the scene (lighting, shadows, perspective). CRITICAL EXCEPTION: Do NOT deduct points for visual quality issues, noise, or artifacts that were already present in the original, unedited video.

- 1: Fail. Severe visual degradation introduced by the edit. Video heavily broken or the new object is severely deformed, flickers uncontrollably, or floats/slides unnaturally (complete motion tracking failure). **- 2:** Poor. Obvious artifacts introduced. Noticeable flickering seams, missing/static shadows, poor occlusion, or the new object's motion clearly disconnects from the camera/scene movement. Fails to blend into the environment. **- 3:** Fair. Acceptable quality but with visible temporal or physical inconsistencies introduced by the edit. Minor flickering, edge fuzzying, lighting shifts over time, or small tolerable drifts in motion tracking. Integration is passable but imperfect. **- 4:** Good. High visual quality and stable style. The new object's motion is well-tracked, interacts realistically with the scene (accurate shadows/reflections), and naturally blends in. Only Tiny temporal artifacts are visible. **- 5:** Perfect. Completely seamless, temporally stable, and dynamically flawless. Perfect motion tracking, perspective, and lighting integration. The new object blends perfectly into the scene in every frame, making the edit completely undetectable to a casual viewer.

Example Response Format: You are required to return a dictionary structured as follows: {"Condition Preserving": [A number from 1 to 5], "Instruction Following": [A number from 1 to 5], "Overall Visual Quality": [A number from 1 to 5]}.

<|user_input>

Figure 22: Evaluation prompt template for TIV2V task (local object replacement).