

# UniVid: The Open-Source Unified Video Model

Anonymous ACL submission



Figure 1: We present **UniVid**, an open-source unified video model for both understanding and generation tasks. Our model requires only a small amount of high-quality data for fine-tuning, achieving competitive results across various tasks.

## Abstract

Recent advances in unified understanding and generation have significantly improved visual generation and understanding in the image domain. However, extending such unified modelling paradigms to video remains challenging, as training unified video models from scratch is prohibitively expensive, while a substantial gap between understanding and generation persists and becomes more pronounced in the video domain. Therefore we present **UniVid**, a unified architecture that couples a curated understanding encoder with a diffusion decoder through a lightweight connector, trained with a multi-stage strategy to align semantic and low-level information and achieve mutual benefits. Furthermore, we introduce Temperature Modality Alignment to improve prompt adherence and Pyramid Reflection for efficient temporal reasoning via dynamic keyframe selection. Extensive experiments on diverse benchmarks demonstrate the state-of-the-art performance of our unified video model, for example, 2.2%

improvement on VBench-Long total score compared to generative models and unified models, and competitive performance in state-of-the-art and conventional video understanding benchmarks compared to most understanding models and unified models.

## 1 Introduction

Video intelligence encompasses two core capabilities: generation and understanding. Generation enables content creation, simulation, and data augmentation through diffusion and flow models (Shi et al., 2020; Podell et al., 2024; Wang et al., 2025; Blattmann et al., 2023a). Understanding powers perception, retrieval, analytics, and human-computer interaction via multimodal LLMs (Wang et al., 2024a; Chen et al., 2024c; Lin et al., 2024; Bai et al., 2025). Real-world applications increasingly demand unified systems that combine both capabilities within a single framework. Recent efforts toward unified video modeling have con-

001  
002  
003  
004  
005  
006  
007  
008  
009  
010  
011  
012  
013  
014  
015  
016  
017  
018  
019  
020  
021  
022  
023

024  
025  
026  
027  
028  
029  
  
030  
031  
032  
033  
034  
035  
036  
037  
038  
039  
040  
041  
042  
043

verged on two paradigms. The first is an autoregressive (AR)–centric route: all modalities (text, images, video) are projected into a shared discrete token space and a single Transformer is trained with next-token prediction over multimodal sequences; representative examples include Emu3 (Wang et al., 2024b), Chameleon (Lu et al., 2023), VILA-U (Wu et al., 2025). The second is a hybrid diffusion–AR route: a multimodal AR backbone governs understanding and control signals, while a generation decoder renders high-fidelity frames from high-level visual tokens; recent works such as Transfusion (Zhou et al., 2024a), Show-O series (Xie et al., 2025a,b), Omni-Video (Tan et al., 2025) follow this pattern. In this work, we adopt the hybrid route to curate a unified framework building upon strong pre-trained vision and language encoders as well as video diffusion models.

However, even within this hybrid setting, building a unified video model remains non-trivial. We identify three key issues. (i) **Synergy**. Existing hybrid systems often couple understanding and generation in a coarse manner, where semantic reasoning provides predominantly high-level control signals for synthesis. Such representations lack fine-grained, spatially grounded details that are crucial for video diffusion, and feedback from generation rarely refines the understanding module. As a result, understanding and generation fail to mutually reinforce each other during training or inference, limiting robustness and generalization across tasks. (ii) **Faithful conditioning for generation**. Maintaining semantically faithful conditioning throughout the video diffusion trajectory remains challenging. First, in MM-DiT-style models (Esser et al., 2024), conditional signals are often subject to conditional suppression, as a small number of conditioning tokens can be overwhelmed by the large population of visual tokens, leading to weakened semantic control. Second, existing guidance mechanisms are typically timestep-agnostic, which is misaligned with the intrinsic learning dynamics of diffusion models. Diffusion models tend to rely more heavily on external conditioning at early timesteps to establish global structure and semantic layout, and gradually shift toward low-level, pixel-wise refinement at later timesteps. Applying uniform conditioning strength across all timesteps therefore fails to respect this progression, resulting in prompt–video drift that becomes increasingly pronounced for long and high-resolution video generation. (iii) **Efficient and evidence-aware video**

**understanding**. Extending image-centric encoders to video understanding is primarily constrained by computational cost and context limitations. Videos contain far more frames than can be accommodated within the limited context window of encoders, although only a small subset is typically informative for a given query. Processing all frames uniformly is therefore inefficient and unnecessary; instead, effective video understanding hinges on identifying and leveraging the most informative frames to maximize reasoning performance under constrained context budgets.

To address these challenges, we formulate UniVid as a three-stage pipeline. **Stage I: Understanding-guided generation**. In the first stage, we focus on enabling multimodal understanding to more effectively guide video generation. A curated image-centric understanding encoder is designed to produce structure-aware embeddings that fuse high-level semantic intent with low-level, spatially grounded cues, providing faithful conditioning for a diffusion-based video decoder. To align semantic conditioning with the intrinsic learning dynamics of diffusion models, we introduce *Temperature Modality Alignment* (TMA), which modulates cross-modal attention across diffusion timesteps—emphasizing semantic guidance at early stages and progressively shifting toward visual detail refinement at later stages—thereby mitigating conditional suppression and prompt–video drift. **Stage II: Temporal adaptation for video understanding**. In the second stage, we adapt the image-centric understanding encoder to multi-frame visual inputs without substantial architectural changes. Instead of uniform frame sampling, we mitigate the computational and context constraints of video understanding with *Pyramid Reflection*, implementing sequential decision-making through keyframe selection and an Actor–Evaluator–Reflector loop that verbally adjusts search strategies while progressively expanding or pruning context. **Stage III: Joint refinement**. Finally, we perform a lightweight joint fine-tuning stage that further refines both the understanding and generation components, allowing the two capabilities to better reinforce each other within a unified framework. This final stage consolidates the benefits of the previous stages, yielding a coherent and robust unified video model.

Through extensive evaluation on standard benchmarks, we validate the superior capability of our unified approach, which consistently outperforms

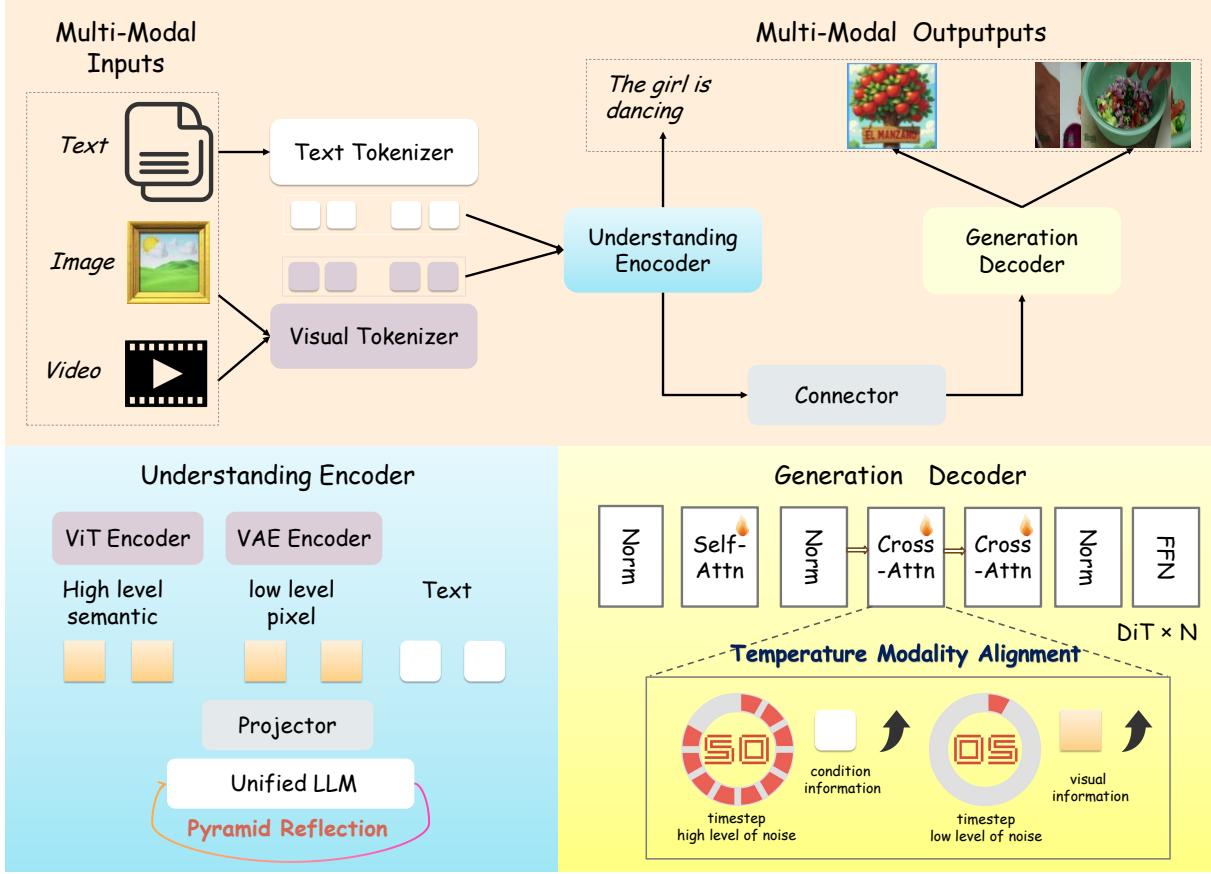


Figure 2: Overall architecture of our proposed UniVid for unified video understanding and generation. Notably, for the understanding task, we adopt only the ViT encoder to achieve a better efficiency–accuracy trade-off.

existing methods across multiple video-centric tasks, demonstrating the potential of unified modeling for comprehensive video intelligence.

Our contribution can be summarized below:

- We introduce **UniVid**, coupling a curated understanding encoder with a generation decoder via a lightweight conditioning adapter. We employ a three-stage training pipeline, enabling alignment and mutual reinforcement between video understanding and generation.
- On the generation side, we propose *Temperature Modality Alignment*: a timestep-aware, temperature-adjusted cross-modal attention schedule in MM-DiT that emphasizes semantic guidance early and shifts to visual refinement later. On the understanding side, we develop *Pyramid Reflection* to enable efficient temporal reasoning with minimal architectural and training overhead.
- We conduct comprehensive experiments on diverse benchmarks such as MMMU (Yue et al., 2024), MMBench (Fang et al., 2024), MSVD-QA (Piergiovanni et al., 2022), ActivityNet-

QA (Yu et al., 2018) for understanding, and on VBBench for generation, demonstrating competitive performance and efficiency. Ablations validate the contribution of each component.

## 2 The Proposed Method

### 2.1 Overview

Our goal is a unified multimodal video model that supports both generation and understanding within a single framework. To this end, we adopt a three-stage training recipe that first aligns the conditioning between the understanding encoder and the generation decoder, then finetunes understanding encoder and introduces Pyramid Reflection, which augments the understanding branch with temporal cues, and finally co-adapts both branches end-to-end. Fig. 2 presents the overall UniVid architecture.

### 2.2 Architecture

**Multimodal archiecture.** The multimodal large language model serves as the core reasoning engine. Text inputs are processed through a standard tokenizer, while visual inputs follow different encoding paths depending on the target branch. For

148  
149  
150  
151

152  
153  
154  
155  
156  
157

158  
159  
160  
161  
162  
163  
164  
165  
166

167  
168  
169  
170

171  
172  
173  
174

175

176

177  
178  
179

180  
181  
182  
183  
184  
185  
186

188  
189  
190  
191  
192

193 the generation branch, images are encoded using  
 194 both ViT (Dosovitskiy et al., 2021) for semantic  
 195 features and VAE (Kingma and Welling, 2019) for  
 196 pixel-level details. For the understanding branch,  
 197 only ViT encoding is employed, as video under-  
 198 standing tasks primarily rely on high-level semantic  
 199 understanding rather than fine-grained pixel details.  
 200 The encoded visual features are then projected into  
 201 the textual token space and concatenated with text  
 202 tokens, allowing the LLM to output unified multi-  
 203 modal representations.

204 **Generation branch.** The generation pathway em-  
 205 ploys a DiT-based model Wan 2.2 (Wang et al.,  
 206 2025) conditioned on rich semantic representa-  
 207 tions extracted from MLLM outputs through a  
 208 lightweight connector. The system processes video  
 209 generation in latent space using a 3D VAE (Zhao  
 210 et al., 2024), with conditioning signals integrated  
 211 via cross-attention mechanisms.

212 **Understanding branch.** For video understand-  
 213 ing, multi-frame evidence is encoded by the ViT  
 214 (Dosovitskiy et al., 2021) and fused with text; the  
 215 LLM produces an initial textual answer. We then  
 216 apply Pyramid Reflection, a query-driven, hier-  
 217 archical loop that iteratively expands or prunes  
 218 keyframe context via SigLIP2 (Tschanne et al.,  
 219 2025) selection and refines the frame space via an  
 220 Actor–Evaluator–Reflector process, yielding the  
 221 final answer without modifying the backbone.

222 Conclusively, our generation builds on the  
 223 MLLM’s strong comprehension, while video under-  
 224 standing uses Pyramid Reflection to leverage the  
 225 MLLM and collaborate with an LLM for efficient  
 226 and accurate answers.

### 227 228 2.3 Conditional Generation with Temperature Modality Alignment

229 Given fused tokens from the understanding path,  
 230 the MLLM output  $Z_u$  is mapped to time-indexed  
 231 conditions by a lightweight connector  $g_\phi$ :

$$232 C_t = g_\phi(Z_u, t) \in \mathbb{R}^{M_t \times d_c}, \quad (1)$$

233 where  $M_t$  is the number of conditioning tokens at  
 234 timestep  $t$  and  $d_c$  is the conditioning dimension.

235 Let the 3D VAE define the latent trajectory  $\{z_t\}$   
 236 along the flow, where  $z_t \in \mathbb{R}^{H \times W \times F \times C}$  repre-  
 237 sent the latent representation with spatial dimen-  
 238 sions  $H \times W$ , temporal frames  $F$ , and channels  
 239  $C$ . The Wan 2.2 DiT predicts the velocity field  
 240 under cross-attention to  $C_t$ , then we integrate the  
 241 probability–flow ODE to obtain  $\hat{z}_0$ , which the VAE  
 242 decoder converts to video frames.

---

### Algorithm 1 Pyramid Reflection as Test-time RL

---

Require: video  $V$ , question  $q$

- 1: Uniformly sample  $N=64$  frames; *encode once and cache* visual embeddings
  - 2: From 16 frames, summarize into a global caption  $C_g$
  - 3: Initialize state  $s_1 \leftarrow (q, C_g, W=\emptyset)$ , policy  $\pi$  with mode router **expand/shrink**
  - 4: **for**  $r = 1$  to  $R$  **do**  $\triangleright R \leq 3$
  - 5:     **Action:**  $a_r \sim \pi(s_r)$
  - 6:     **expand:** add frames most relevant to current search text
  - 7:     **shrink:** prune to diverse key frames using cached similarities
  - 8:     Update working set  $W$  accordingly using cached embeddings (index-only change)
  - 9:     **Actor:** answer using ordered  $W$  conditioned on  $C_g$
  - 10:     **Evaluator:** score  $\hat{r}_r \in [0, 1]$  as confidence signal
  - 11:     **if**  $\hat{r}_r \geq \tau$  **then**
  - 12:         **return** answer
  - 13:     **else**
  - 14:         **Reflector:** refine the search text  $q \leftarrow$  short declarative cue
  - 15:         Update state  $s_{r+1} \leftarrow (q, C_g, W)$  (verbal policy improvement)
  - 16:     **end if**
  - 17: **end for**
  - 18: **return** fallback answer from  $C_g$
- 

243 Inspired by TACA (Lv et al., 2025), we adapt its  
 244 finding that text is suppressed in MM-DiT (Esser  
 245 et al., 2024) because (i) the softmax over a much  
 246 larger pool of visual tokens ( $N_{\text{vis}} \gg N_{\text{txt}}$ ) dilutes  
 247 attention mass on text keys, and (ii) conditioning  
 248 plays different roles across timesteps (early seman-  
 249 tics, late detail). We therefore strengthen the visual-  
 250 to-text path in Wan 2.2 (Wang et al., 2025) with a  
 251 simple schedule:

$$\tilde{S}_{v \rightarrow t}(u) = \alpha_{\text{txt}}(u) S_{v \rightarrow t}, \quad u \in [0, 1], \quad (2)$$

253 where  $u$  is the normalized flow matching progress  
 254 (0 early, 1 late),  $S_{v \rightarrow t}$  denotes the visual-to-text  
 255 attention scores, and  $\tilde{S}_{v \rightarrow t}(u)$  represents the mod-  
 256 ulated attention scores. The modulation factor is  
 257 defined as:

$$\alpha_{\text{txt}}(u) = \begin{cases} 1 + \frac{\lambda_{\text{txt}}}{2} (1 + \cos(\frac{\pi u}{0.4})), & u \in [0, 0.4], \\ 1, & u \in (0.4, 1]. \end{cases} \quad (3)$$

Thus, text guidance is strongest early and decays to neutral ( $\alpha_{\text{txt}} \rightarrow 1$ ) late, improving prompt faithfulness without over-constraining details.

For reference-image that requires identity stability, we apply a small late-stage boost to visual cross-attention:

$$\tilde{S}_{v \rightarrow v}(u) = \alpha_{\text{img}}(u) S_{v \rightarrow v}, \quad (4)$$

where  $S_{v \rightarrow v}$  represents visual cross-attention scores and

$$\alpha_{\text{img}}(u) = \begin{cases} 1, & u \in [0, 0.6], \\ 1 + \frac{\lambda_{\text{img}}}{2} (1 - \cos(\frac{\pi(u-0.6)}{0.4})), & u \in (0.6, 1]. \end{cases} \quad (5)$$

Unless otherwise specified, we set  $\lambda_{\text{txt}} = \lambda_{\text{img}} = 0.3$ .

## 2.4 Pyramid Reflection for understanding

**Formulation.** We cast video question answering as test-time reinforcement learning over a small, ordered evidence set. The state at round  $r$  is  $(s_r, W_r, C_g)$ , where  $s_r$  is a short search text,  $W_r$  is an ordered subset of frames, and  $C_g$  is a global caption distilled once from uniformly sampled seeds. The action is to reconfigure  $W_r$  given  $s_r$ , either by adding frames (expand) or by pruning to a diverse core (shrink). The policy  $\pi_s$  is a retrieval rule driven by text-image similarity and a diversity term; it maps  $s$  to a distribution over frame indices. The environment returns an answer  $a$  produced by the Actor and a scalar reward  $r \in [0, 1]$  from the Evaluator. Policy improvement is carried out verbally: the Reflector emits a refined  $s_{r+1}$  that concentrates on disambiguating cues such as before/after, first/last, motion phase, color, or role. The loop stops early when  $r$  exceeds a confidence threshold.

**Policy class.** We instantiate  $\pi_s$  with a cached-embedding retriever. All  $N$  candidate frames are embedded once by a vision encoder; the text side uses  $\phi(s)$ . For expand we add the highest-scoring unseen frames by cosine similarity  $\langle \mathbf{v}_i, \phi(s) \rangle$ , which suits static questions whose evidence is sparse but distinctive. For shrink we start broad to preserve chronology, then apply a Maximal

Marginal Relevance objective that balances relevance to  $\phi(s)$  and pairwise dissimilarity within  $W$ , which suits dynamic questions where ordering, repetition, or transitions matter. In both regimes  $W$  is kept in temporal order so the Actor can compare events across  $[t_1 \rightarrow t_k]$  rather than hallucinate transitions.

**Value and critic signals.** The Evaluator provides a calibrated confidence that serves as a value proxy. Its scalar reward  $r$  both triggers early stopping and conditions the Reflector. When  $r$  is low, the Reflector returns a short declarative refinement of  $s$  that encodes the suspected failure mode: missing entity, wrong time span, ambiguous referent, or occluded phase. This verbal update reshapes the retrieval distribution without touching model weights, yielding a form of policy gradient in the space of prompts. Our Pyramid Reflection procedure is summarized in Algorithm 1, and the high-level understanding pipeline is shown in Fig. 8. The theoretical details of Pyramid Reflection as test-time RL are provided in Appendix A.6.

The design achieves efficiency by caching frame embeddings once and reducing exploration to lightweight index updates, while the Actor reasons over compact, temporally ordered evidence with fixed global context to maintain scene priors under tight token budgets. The adaptive routing between expansion and MMR-based shrinking aligns retrieval strategies with question structure, enabling effective temporal reasoning at low computational cost.

Nevertheless, this efficiency-oriented retrieval scheme inherently operates on a sparse temporal subset rather than the full dense sequence. As a result, its ability to infer subtle motion cues, fine-grained temporal continuity, or high-frequency dynamics may be limited compared to methods that process all frames end-to-end. These dense approaches often provide more precise motion understanding and object interaction modeling, particularly in tasks where small spatial shifts or rapid temporal transitions are critical for accurate reasoning.

## 3 Experiments

### 3.1 Dataset and Metrics

**Datasets.** We evaluate UniVid on established benchmarks for both video generation and understanding. For generation, we train on curated samples from OpenVid-1M, a large-scale text-to-video

Method	Overall Scores				Technical Quality				Aesthetic Quality		
	Total Score↑	Quality↑	Semantic↑	Subject↑	Background↑	Temporal↑	Motion↑	Dynamic↑	Aesthetic↑	Imaging↑	
EasyAnimateV5.1 (Fu et al., 2024b)	83.42	85.03	77.01	98.00	97.41	99.19	98.02	57.15	<b>69.48</b>	68.61	
MiniMax-Video-01 (MiniMax, 2024)	83.41	84.85	77.65	97.51	97.05	99.10	99.22	64.91	63.03	67.17	
Kling 1.6 (Technology, 2025)	83.40	85.20	76.99	97.40	96.84	99.64	99.13	62.22	64.81	69.70	
Wan2.1-T2V-1.3B (Wang et al., 2025)	83.31	85.23	76.95	97.56	97.93	99.55	98.52	65.19	65.46	67.01	
Wan2.2-TI2V-5B (Wang et al., 2025)	83.59	85.64	76.53	97.66	98.03	99.10	98.71	65.76	65.52	67.51	
HunyuanVideo (Kong et al., 2024)	83.24	85.86	75.82	97.32	<b>97.93</b>	99.49	98.99	<b>70.83</b>	60.36	67.56	
Gen-3 (Runway, 2024)	82.32	84.11	75.17	97.01	96.62	99.61	99.23	60.14	63.34	66.82	
Vchitect-2.0 (VEnhancer) (Fan et al., 2025)	82.24	83.54	77.06	96.83	96.66	98.97	98.98	63.89	60.41	65.35	
CogVideoX1.5-5B (Yuan et al., 2024)	82.17	82.78	79.76	96.87	97.35	98.88	98.31	50.93	62.79	65.02	
Omni-Video (Tan et al., 2025)	83.00	84.27	77.92	98.39	97.68	99.87	99.10	56.67	62.48	64.56	
<b>UniVid (Ours)</b>	<b>85.27</b>	<b>86.44</b>	<b>80.58</b>	<b>98.96</b>	97.76	<b>99.88</b>	<b>99.25</b>	61.83	64.21	<b>73.03</b>	

Method	Semantic Fidelity								
	Object↑	Multi-Obj↑	Action↑	Color↑	Spatial↑	Scene↑	Appearance↑	Temporal↑	Overall↑
EasyAnimateV5.1 (Fu et al., 2024b)	89.57	66.85	95.60	77.86	76.11	54.31	23.06	24.61	26.47
MiniMax-Video-01 (MiniMax, 2024)	<b>97.83</b>	76.04	92.40	90.36	75.50	50.68	20.06	25.63	27.10
Kling 1.6 (Technology, 2025)	93.34	73.99	96.20	81.26	79.08	55.57	20.75	24.51	26.04
Wan2.1-T2V-1.3B (Wang et al., 2025)	88.81	74.83	94.00	82.00	73.04	41.96	21.81	23.13	25.50
Wan2.2-TI2V-5B (Wang et al., 2025)	89.21	75.23	94.09	82.43	72.90	42.36	21.89	23.78	26.03
HunyuanVideo (Kong et al., 2024)	86.10	71.66	93.42	91.60	68.09	53.69	19.80	23.89	26.44
Gen-3 (Runway, 2024)	87.81	53.64	96.40	80.90	65.03	54.57	24.31	24.71	26.69
Vchitect-2.0 (VEnhancer) (Fan et al., 2025)	86.61	68.84	<b>97.20</b>	87.04	57.55	<b>56.57</b>	23.73	25.01	27.57
CogVideoX1.5-5B (Yuan et al., 2024)	87.47	69.65	97.20	87.55	80.25	52.91	<b>24.89</b>	25.19	27.30
Omni-Video (Tan et al., 2025)	93.54	71.06	93.60	88.89	73.15	44.33	23.45	25.81	26.99
<b>UniVid (Ours)</b>	94.52	<b>77.45</b>	94.20	<b>92.10</b>	<b>80.70</b>	46.66	23.57	<b>25.91</b>	<b>27.60</b>

Table 1: T2V performance on VBench-Long (Huang et al., 2024).

dataset, and evaluate on VBench, a comprehensive benchmark suite for video generative models that provides fine-grained evaluation metrics across multiple dimensions. For understanding, we train on ActivityNet-QA train dataset (Yu et al., 2018) and evaluate on four comprehensive video QA benchmarks: MSVD-QA (Piergiovanni et al., 2022) with 1,970 video clips and 50.5K QA pairs, MSRVTT-QA (Piergiovanni et al., 2022) with 10K videos, 243K QA pairs, TGIF-QA (Jang et al., 2017) containing 165K QA pairs for animated GIFs, and the ActivityNet-QA test dataset (Yu et al., 2018) with 58,000 QA pairs on 5,800 complex web videos. These datasets cover diverse temporal reasoning scenarios across short to medium-length video clips, ranging from brief animated sequences to multi-minute activity videos.

**Evaluation metrics.** For video generation, we evaluate on VBench across multiple fine-grained dimensions: Technical Quality metrics including Subject consistency, Background preservation, Temporal flickering, Motion smoothness, and Dynamic degree; Aesthetic Quality measures covering overall visual appeal and imaging quality; and Semantic Fidelity metrics assessing Object accuracy, Multi-object handling, Action fidelity, Color accuracy, Spatial relationships, Scene consistency, Appearance preservation, and Temporal coherence. For video understanding, we report average accuracy and scores on each benchmark dataset.

## 3.2 Implementation Details

We adopt a three-stage hierarchical training recipe. It initializes UniVid from strong public checkpoints to reduce compute. For curated understanding encoder, we reuse ViT (Dosovitskiy et al., 2021) and VAE (Kingma and Welling, 2019) as visual encoder and 7B qwen2 (Yang et al., 2024) as unified LLM. For generation decoder, we build upon Wan 2.2 5B TI2V model (Wang et al., 2025). For connector, we use convolutional layers to bridge the gap between understanding encoder and generation decoder. In the first stage, we optimize the diffusion decoder and train a lightweight connector so that the curated understanding encoder can more effectively guide generation. Specifically, we train the self attention and cross attention layers in the diffusion model. During this stage, Temperature Modality Alignment (TMA) is applied to modulate cross-modal attention across diffusion timesteps. In the second stage, we adapt the image-centric encoder to multi-frame visual inputs by fine-tuning the connector and the last two ViT blocks, and the VAE blocks. Pyramid Reflection is employed to support adaptive frame selection during training and inference. In the final stage, we jointly fine-tune the understanding and generation components by updating the connector and task-specific modules, while keeping the core backbones frozen. Additional details are provided in Appendix A.3.

For generation, we use a flow-matching ODE sampler with classifier-free guidance and a universal negative prompt. Unless noted, videos are sampled at  $1280 \times 704$  resolution, 121 frames at

Method	LLM size	Video QA Performance							
		MSVD-QA		MSRVTT-QA		TGIF-QA		ActivityNet-QA	
		Acc↑	Score↑	Acc↑	Score↑	Acc↑	Score↑	Acc↑	Score↑
FrozenBiLM (Yang et al., 2022)	1B	32.2	–	16.8	–	41.0	–	24.7	–
VideoChat (Li et al., 2023)	7B	56.3	2.8	45.0	2.5	34.4	2.3	–	2.2
LLaMA-Adapter (Zhang et al., 2023b)	7B	54.9	3.1	43.8	2.7	–	–	34.2	2.7
Video-LLAMA (Zhang et al., 2023a)	7B	51.6	2.5	29.6	1.8	–	–	12.4	1.1
Video-ChatGPT (Maaz et al., 2024)	7B	64.9	3.3	49.3	2.8	51.4	3.0	35.2	2.7
Chat-UniVi (Jin et al., 2024)	7B	65.0	3.6	54.6	3.1	60.3	3.4	45.8	3.2
Video-LLaVA (Lin et al., 2024)	7B	70.7	3.9	59.2	3.5	70.0	4.0	45.3	3.3
BT-Adapter (Liu et al., 2024)	7B	67.5	3.7	57.0	3.2	–	–	45.7	3.2
Valley-v3 (Luo et al., 2023)	7B	60.5	3.3	51.1	2.9	–	–	45.1	3.2
FreeVA (Wu, 2024)	7B	73.8	4.1	60.0	3.5	–	–	51.2	3.5
DeepStack-L (Meng et al., 2024)	7B	76.0	4.0	–	–	–	–	49.3	3.1
IG-VLM (LLaVA-v1.6) (Kim et al., 2024)	7B	78.8	4.1	63.7	3.5	–	4.0	54.3	3.4
SF-LLaVA-7B (Xu et al., 2024)	7B	79.1	4.1	<b>65.8</b>	<b>3.6</b>	<b>78.7</b>	<b>4.2</b>	55.5	3.4
<b>UniVid (Ours)</b>	<b>7B</b>	<b>80.1</b>	<b>4.2</b>	61.4	3.4	75.0	4.1	<b>58.8</b>	<b>3.6</b>

Table 2: Comparison on four video QA benchmarks (Piergiovanni et al., 2022; Jang et al., 2017; Yu et al., 2018).

24 fps; the guidance scale is set to 5.0 for both T2V and I2V with 50 inference steps. At input time, the LLM receives the text prompt together with image ViT embeddings and VAE latents; it outputs conditional textual tokens. During generation, Wan 2.2 consumes these conditional textual tokens and image via cross-attention. Our Temperature Modality Alignment schedule applies a cosine-scheduled text gain that transitions from  $\alpha_{\text{txt}} = 1.3$  to 1.0 over the first 40% of denoising steps ( $u \in [0, 0.4]$ ), then maintains  $\alpha_{\text{txt}} = 1.0$  for the remaining steps. This enhances text guidance during early denoising when structural decisions are made, while allowing finer details to emerge in later stages.

For understanding, we uniformly sample a pool of  $N = 64$  frames per video and cache their SigLIP2 image embeddings; subsequent selection reuses cached features. Global context is a caption summarized from 16 uniformly spaced seed frames. Query-image ranking uses SigLIP2 cosine similarity with L2-normalized features and batch size 64. Static questions follow a  $4 \rightarrow 8 \rightarrow 16$  keyframe schedule. Dynamic questions follow  $64 \rightarrow 32 \rightarrow 16$  with MMR down-selection,  $\lambda = 0.5$ . Confidence is accepted when the Evaluator’s score is at least 0.7 or the verdict is accept, with at most  $R \leq 3$  rounds. The LLM determines routing between static and dynamic modes. For implementation, we use DeepSeek v3.1 to serve as the Evaluator and determine the type of questions and Qwen-plus to serve as the Reflector. Full prompt texts are listed in the Appendix A.5.

### 3.3 Main Results

**Generation quantitative results.** We evaluate UniVid on the challenging VBench-Long bench-

mark (Huang et al., 2024). As shown in Tab. 1, UniVid establishes a new state of the art with an overall score of 85.27, outperforming prior leading systems such as EasyAnimateV5.1 (Fu et al., 2024b), MiniMax-Video-01 (MiniMax, 2024), and Kling 1.6 (Technology, 2025). In particular, UniVid exhibits clear advantages in semantic alignment (80.58), highlighting its superior capability in faithfully rendering objects, actions, and multi-object interactions. On the technical side, it attains near-perfect temporal (99.88) and motion (99.25) consistency, validating the effectiveness of our long-context dynamics module. Moreover, UniVid delivers the best imaging score (73.03), reflecting sharper details and more stable visual quality compared with prior systems, as shown in Fig. 1, which demonstrates high-quality visual generation.

Beyond overall scores, UniVid demonstrates consistent gains in semantic fidelity. As summarized in the Semantic Fidelity block of Tab. 1, it achieves leading results on multi-object reasoning (77.45), color faithfulness (92.10), and spatial grounding (80.70), while remaining competitive in action depiction and appearance consistency. These improvements suggest that our design choices—particularly the integration of hierarchical scene representation with dynamic frame alignment—substantially enhance both controllability and alignment with textual prompts. Taken together, the results indicate that UniVid pushes forward the frontier of long-horizon text-to-video generation by simultaneously ensuring high-fidelity semantics and strong technical as well as aesthetic quality. More examples of video generation can be seen in Appendix A.4.

482 **Generation qualitative results.** Fig. 3 compares UniVid with Wan2.1-T2V-1.3B (Wang et al.,  
483 2025), MiniMax-Video-01 (MiniMax, 2024), Hun-  
484 yuanVideo (Kong et al., 2024), and EasyAni-  
485 mateV5.1 (Fu et al., 2024b). Competing models  
486 often show missing basketballs or distorted cars,  
487 while UniVid generates coherent jump shots and  
488 realistic racing scenes with stable dynamics and  
489 faithful semantics.  
490

491 **Understanding quantitative evaluation.** Across  
492 MSVD-QA (Piergiovanni et al., 2022), MSRVTT-  
493 QA (Piergiovanni et al., 2022), TGIF-QA (Jang  
494 et al., 2017), and ActivityNet-QA (Yu et al., 2018),  
495 UniVid sets the 7B-scale state of the art on MSVD-  
496 QA and ActivityNet-QA and remains competitive  
497 on the other two (Tab. 2), despite a smaller post-  
498 training set and no test-time ensembling. Joint  
499 finetuning of generation and understanding with  
500 Pyramid Reflection strengthens the abilities these  
501 datasets emphasize: better action–entity binding  
502 and object or attribute grounding in short open-  
503 domain clips, stronger temporal reasoning over  
504 frame sequences, and more reliable long-range evi-  
505 dence retrieval in untrimmed videos.  
506

507 As illustrated before, UniVid performs robust  
508 multi-frame reasoning with our Pyramid Reflection  
509 loop. Starting from a global caption and automatic  
510 type detection, the system first produces an initial  
511 answer, which is then scored by the evaluator;  
512 when evidence is insufficient, the reflector issues a  
513 refined, declarative query that re-ranks keyframes  
514 toward the true scene. This Pyramid Reflection  
515 steers attention from opening credits to the lane  
516 shots, yielding a consistent interpretation of roles  
517 (in the example of Fig. 8: bowler and nearby team-  
518 mate/coach) grounded in the visual context rather  
519 than spurious cues. The dynamic keyframe sched-  
520 ule reduces the number of inspected frames while  
521 maintaining accuracy, demonstrating both evidence  
522 tracing and efficiency gains in short-clip under-  
523 standing. More examples of video understanding  
524 can be seen in Appendix A.4.  
525

526 **Understanding qualitative results.** We compare  
527 UniVid with Video-LLaVA (Lin et al., 2024) and  
528 SF-LLaVA (Xu et al., 2024) on video QA; as  
529 shown in Fig. 4, baselines often give plausible but  
530 incomplete statements. These examples highlight  
531 UniVid’s stronger action–entity binding, tempo-  
532 ral reasoning, and resistance to distractor frames,  
533 yielding precise and concise answers. Additionally,  
534 we conduct systematic ablation experiments to val-  
535

536 idate the contributions of UniVid. The results and  
537 analyses are provided in the Appendix A.7.  
538

## 4 Conclusion

539 We introduced UniVid, a unified video model that  
540 couples an MLLM with a diffusion decoder via  
541 a lightweight conditioning adapter to both un-  
542 derstand and generate videos. Two key mech-  
543 anisms enable this: Temperature Modality Align-  
544 ment schedules cross-modal attention across flow  
545 steps to preserve prompt faithfulness while refining  
546 details, and Pyramid Reflection performs query-  
547 driven keyframe selection for efficient temporal rea-  
548 soning. With these components, UniVid achieves  
549 state-of-the-art or competitive results on VBench-  
550 Long and multiple video-QA benchmarks while  
551 avoiding costly retraining of image-centric back-  
552 bones. We release UniVid to support research on  
553 practical, controllable, and truly unified video in-  
554 telligence.  
555

## Limitations

556 While UniVid unifies an autoregressive MLLM  
557 with a DiT-based video diffusion decoder, the cur-  
558 rent interaction between the two modules remains  
559 relatively shallow. Most MLLM parameters are  
560 frozen, and the diffusion branch only receives lim-  
561 ited semantic guidance, restricting the potential  
562 mutual benefits between understanding and genera-  
563 tion. As a consequence, the MLLM gains little  
564 improvement in deeper reasoning, and the genera-  
565 tion branch relies primarily on data-driven priors  
566 rather than task-aware adaptive conditioning.  
567

568 These limitations manifest in characteristic fail-  
569 ure modes during generation, as illustrated in Fig. 9  
570 UniVid can exhibit temporal inconsistencies in  
571 long sequences (e.g., static lightning), loss of fine-  
572 grained details in distant regions, and occasional  
573 structural artifacts such as missing body parts in an-  
574 thropomorphized characters. These reflect inherent  
575 challenges of long-horizon diffusion sampling and  
576 the lack of stronger semantic–structural feedback  
577 between the two branches.  
578

<b>References</b>	
Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai.	632
2023. <b>GQA</b> : training generalized multi-query transformer models from multi-head checkpoints. In <i>Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023</i> , pages 4895–4901. Association for Computational Linguistics.	633
Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucic, and Cordelia Schmid.	634
2021. Vivit: A video vision transformer. In <i>2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021</i> , pages 6816–6826.	635
Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others.	636
2025. <b>Qwen2.5-vl</b> technical report. <i>CoRR</i> , abs/2502.13923.	637
Gedas Bertasius, Heng Wang, and Lorenzo Torresani.	638
2021. Is space-time attention all you need for video understanding? In <i>Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event</i> , pages 813–824.	639
Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach.	640
2023a. <b>Stable video diffusion: Scaling latent video diffusion models to large datasets</b> . <i>CoRR</i> , abs/2311.15127.	641
Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, and 1 others.	642
2023b. Stable video diffusion: Scaling latent video diffusion models to large datasets. <i>arXiv preprint arXiv:2311.15127</i> .	643
Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis.	644
2023c. Align your latents: High-resolution video synthesis with latent diffusion models. In <i>Proceedings of the IEEE/CVF conference on computer vision and pattern recognition</i> , pages 22563–22575.	645
Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman.	646
2019. A short note on the kinetics-700 human action dataset. <i>arXiv preprint arXiv:1907.06987</i> .	647
Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan.	648
2024a. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In <i>IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024</i> , pages 7310–7320.	649
Jiahao Wang, Tan Jiang, Bo Wang, and 21 others.	650
2024b. <b>Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling</b> . <i>CoRR</i> , abs/2412.05271.	651
Zhe Chen, Jiannan Wu, Wenhui Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai.	652
2024c. Intern VL: scaling up vision foundation models and aligning for generic visual-linguistic tasks. In <i>IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024</i> , pages 24185–24198. IEEE.	653
Van Duc Cuong, Ta Dinh Tam, Tran Duc Chinh, and Nguyen Thi Hanh.	654
2025. <b>Fluid: Flow-latent unified integration via token distillation for expert specialization in multimodal learning</b> . <i>Preprint, arXiv:2508.07264</i> .	655
Chaurui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Shi Guang, and Haoqi Fan.	656
2025. Emerging properties in unified multimodal pretraining. <i>CoRR</i> .	657
Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby.	658
2021. <b>An image is worth 16x16 words: Transformers for image recognition at scale</b> . In <i>9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021</i> . OpenReview.net.	659
Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach.	660
2024. <b>Scaling rectified flow transformers for high-resolution image synthesis</b> . In <i>Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024</i> . OpenReview.net.	661
Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer.	662
2021. Multiscale vision transformers. In <i>Proceedings of the IEEE/CVF international conference on computer vision</i> , pages 6824–6835.	663
	664
	665
	666
	667
	668
	669
	670
	671
	672
	673
	674
	675
	676
	677
	678
	679
	680
	681
	682
	683
	684
	685
	686
	687

688	Weichen Fan, Chenyang Si, Junhao Song, Zhenyu Yang, Yinan He, Long Zhuo, Ziqi Huang, Ziyue Dong, Jingwen He, Dongwei Pan, and 1 others. 2025. Vchitect-2.0: Parallel transformer for scaling up video diffusion models. <i>arXiv preprint arXiv:2501.08453</i> .	746
689		747
690		748
691		749
692		750
693	Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. 2024. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. <i>arXiv preprint arXiv:2406.14515</i> .	751
694		752
695		753
696		754
697		755
698	Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2024a. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. <i>arXiv preprint arXiv:2405.21075</i> .	756
699		757
700		758
701		759
702		760
703		761
704	Jaskie Fu, Kun-Hao Yeh, Zhaofan Zha, Xinyu Wang, Chenghao Li, Han-Yi Shaw, Chao-Yi Li, and Pin-Yu Chen. 2024b. Easyanimate: An easy-to-use framework for creating high-quality and controllable videos from a single image. <i>arXiv preprint arXiv:2403.04416</i> .	762
705		763
706		764
707		765
708		766
709		767
710	Rohit Girdhar, Alaaeldin El-Nouby, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Omnimae: Single model masked pre-training on images and videos. In <i>Proceedings of the IEEE/CVF conference on computer vision and pattern recognition</i> , pages 10406–10417.	768
711		769
712		770
713		771
714		772
715		773
716	Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021a. Aligning ai with shared human values. <i>Proceedings of the International Conference on Learning Representations (ICLR)</i> .	774
717		775
718		776
719		777
720		778
721	Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021b. Measuring massive multitask language understanding. <i>Proceedings of the International Conference on Learning Representations (ICLR)</i> .	779
722		780
723		781
724		782
725		783
726	Alex Henry, Prudhvi Raj Dachapally, Shubham Shantaram Pawar, and Yuxuan Chen. 2020. <b>Query-key normalization for transformers</b> . In <i>Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020</i> , volume EMNLP 2020 of <i>Findings of ACL</i> , pages 4246–4253. Association for Computational Linguistics.	784
727		785
728		786
729		787
730		788
731		789
732		790
733	Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey A. Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. 2022a. Imagen video: High definition video generation with diffusion models. <i>CoRR</i> .	791
734		792
735		793
736		794
737		795
738		796
739	Jonathan Ho, Tim Salimans, Alexey A. Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. 2022b. Video diffusion models. In <i>Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022</i> .	797
740		798
741		799
742		800
743		801
744		802
745		
688	Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, and 1 others. 2024. Vbench: Comprehensive benchmark suite for video generative models. In <i>Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition</i> , pages 21807–21818.	751
689		752
690		753
691		754
692		755
693		756
694		757
695		758
696		759
697		
698	Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. TGIF-QA: toward spatio-temporal reasoning in visual question answering. In <i>2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017</i> , pages 1359–1367. IEEE Computer Society.	760
699		761
700		762
701		763
702		764
703		765
704	Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. 2024. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In <i>IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024</i> , pages 13700–13710. IEEE.	766
705		767
706		768
707		769
708		770
709		771
710	Wonkyun Kim, Changin Choi, Wonseok Lee, and Won-jong Rhee. 2024. An image grid can be worth a video: Zero-shot video question answering using a vlm. <i>Preprint</i> , arXiv:2403.18406.	772
711		773
712		774
713		775
714		776
715		777
716	Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, and 1 others. 2024. Hunyuanvideo: A systematic framework for large video generative models. <i>arXiv preprint arXiv:2412.03603</i> .	778
717		779
718		780
719		781
720		782
721	Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wen-hai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023. <b>Videochat: Chat-centric video understanding</b> . <i>CoRR</i> , abs/2305.06355.	783
722		784
723		785
724		786
725		787
726	Zhaowei Li, Wei Wang, YiQing Cai, Xu Qi, Pengyu Wang, Dong Zhang, Hang Song, Botian Jiang, Zhida Huang, and Tao Wang. 2024. Unifiedmllm: Enabling unified representation for multi-modal multi-tasks with large language model. <i>arXiv preprint arXiv:2408.02503</i> .	788
727		789
728		790
729		791
730		792
731		793
732		794
733	Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024. Video-llava: Learning united visual representation by alignment before projection. In <i>Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024</i> , pages 5971–5984. Association for Computational Linguistics.	795
734		796
735		797
736		798
737		799
738		800
739	Akide Liu, Zeyu Zhang, Zhexin Li, Xuehai Bai, Yizeng Han, Jiasheng Tang, Yuanjie Xing, Jichao Wu, Mingyang Yang, Weihua Chen, and 1 others. 2025. Fpsattention: Training-aware fp8 and sparsity co-design for fast video diffusion. <i>arXiv preprint arXiv:2506.04648</i> .	801
740		802

803	Ruyang Liu, Chen Li, Yixiao Ge, Thomas H. Li, Ying Shan, and Ge Li. 2024. <b>Bt-adapter: Video conversation is feasible without video instruction tuning.</b> In <i>IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024</i> , pages 13658–13667. IEEE.	861
804		862
805		863
806		864
807		865
808		866
809	Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. 2022. Video swin transformer. In <i>Proceedings of the IEEE/CVF conference on computer vision and pattern recognition</i> , pages 3202–3211.	867
810		868
811		869
812		870
813		871
814	Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. 2023. Chameleon: Plug-and-play compositional reasoning with large language models. In <i>Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023</i> .	872
815		873
816		874
817		875
818		876
819		877
820		878
821		879
822	Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Minghui Qiu, Pengcheng Lu, Tao Wang, and Zhongyu Wei. 2023. <b>Valley: Video assistant with large language model enhanced ability.</b> <i>CoRR</i> , abs/2306.07207.	880
823		881
824		882
825		883
826		884
827	Zhengyao Lv, Tianlin Pan, Chenyang Si, Zhaoxi Chen, Wangmeng Zuo, Ziwei Liu, and Kwan-Yee K. Wong. 2025. <b>Rethinking cross-modal interaction in multimodal diffusion transformers.</b> <i>Preprint</i> , arXiv:2506.07986.	885
828		886
829		887
830		888
831		889
832	Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, and Fahad Khan. 2024. <b>Video-chatgpt: Towards detailed video understanding via large vision and language models.</b> In <i>Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024</i> , pages 12585–12602. Association for Computational Linguistics.	890
833		891
834		892
835		893
836		894
837		895
838		896
839		897
840		898
841	Andrew Melnik, Michal Ljubljancic, Cong Lu, Qi Yan, Weiming Ren, and Helge J. Ritter. 2024. Video diffusion models: A survey. <i>Trans. Mach. Learn. Res.</i>	899
842		900
843		901
844	Lingchen Meng, Jianwei Yang, Rui Tian, Xiyang Dai, Zuxuan Wu, Jianfeng Gao, and Yu-Gang Jiang. 2024. <b>Deepstack: Deeply stacking visual tokens is surprisingly simple and effective for lmms.</b> In <i>Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024</i> .	902
845		903
846		904
847		905
848		906
849		907
850		908
851		909
852	MiniMax. 2024. Minimax video generation api is now available. <a href="https://www.minimaxi.com/en/news/video-generation-api">https://www.minimaxi.com/en/news/video-generation-api</a> . Accessed: 2025-07-24.	910
853		911
854		912
855	Tian Pan, Yibing Song, Tianyu Yang, Wenhao Jiang, and Wei Liu. 2021. Videomoco: Contrastive video representation learning with temporally adversarial examples. In <i>Proceedings of the IEEE/CVF conference on computer vision and pattern recognition</i> , pages 11205–11214.	913
856		914
857		915
858		916
859		
860		
861	A. J. Piergiovanni, Kairo Morton, Weicheng Kuo, Michael S. Ryoo, and Anelia Angelova. 2022. Video question answering with iterative video-text co-tokenization. In <i>Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXVI</i> , volume 13696 of <i>Lecture Notes in Computer Science</i> , pages 76–94. Springer.	861
862		862
863		863
864		864
865		865
866		866
867		867
868		868
869	Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2024. SDXL: improving latent diffusion models for high-resolution image synthesis. In <i>The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024</i> . OpenReview.net.	869
870		870
871		871
872		872
873		873
874		874
875		875
876	Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In <i>International conference on machine learning</i> , pages 8748–8763. PMLR.	876
877		877
878		878
879		879
880		880
881		881
882		882
883	Runway. 2024. Gen-3 alpha: A new frontier for video generation. Technical report, Runway. Accessed: 2025-07-24.	883
884		884
885		885
886	Noam Shazeer. 2020. <b>GLU variants improve transformer.</b> <i>CoRR</i> , abs/2002.05202.	886
887		887
888	Jingwei Shi, Zeyu Zhang, Biao Wu, Yanjie Liang, Meng Fang, Ling Chen, and Yang Zhao. 2025. Presentagent: Multimodal agent for presentation video generation. <i>arXiv preprint arXiv:2507.04036</i> .	888
889		889
890		890
891		891
892	Zhan Shi, Xu Zhou, Xipeng Qiu, and Xiaodan Zhu. 2020. <b>Improving image captioning with better use of captions.</b> <i>Preprint</i> , arXiv:2006.11807.	892
893		893
894		894
895	Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. <b>Reflexion: language agents with verbal reinforcement learning.</b> In <i>Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023</i> .	895
896		896
897		897
898		898
899		899
900		900
901		901
902	Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. 2022. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In <i>IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022</i> , pages 3616–3626.	902
903		903
904		904
905		905
906		906
907		907
908		908
909	Jianlin Su, Murtadha H. M. Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. <b>Roformer: Enhanced transformer with rotary position embedding.</b> <i>Neurocomputing</i> , 568:127063.	909
910		910
911		911
912		912
913	Zhiyu Tan, Hao Yang, Luozheng Qin, Jia Gong, Mengping Yang, and Hao Li. 2025. <b>Omni-video: Democratizing unified video understanding and generation.</b> <i>CoRR</i> , abs/2507.06119.	913
914		914
915		915
916		916

- 917 Kuaishou Technology. 2025. Kling. <https://klingai.kuaishou.com/>. Accessed: 2025-07-24. 918
- 919 Zhan Tong, Yibing Song, Jue Wang, and Limin 920 Wang. 2022. Videomae: Masked autoencoders 921 are data-efficient learners for self-supervised 922 video pre-training. In *Advances in Neural 923 Information Processing Systems 35: Annual 924 Conference on Neural Information Processing 925 Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 926 2022*.
- 927 Michael Tschannen, Alexey A. Gritsenko, Xiao Wang, 928 Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, 929 Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, 930 Ye Xia, Basil Mustafa, Olivier J. Hénaff, Jeremiah 931 Harmsen, Andreas Steiner, and Xiaohua Zhai. 2025. 932 **Siglip 2: Multilingual vision-language encoders with** 933 **improved semantic understanding, localization, and** 934 **dense features.** *CoRR*, abs/2502.14786.
- 935 Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and 936 Jan Kautz. 2018. Mocogan: Decomposing motion 937 and content for video generation. In *2018 IEEE 938 Conference on Computer Vision and Pattern Recognition, 939 CVPR 2018, Salt Lake City, UT, USA, June 18-22, 940 2018*, pages 1526–1535.
- 941 Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei 942 Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao 943 Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, 944 Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, 945 Kang Zhao, Keyu Yan, Lianghua Huang, and 42 946 others. 2025. **Wan: Open and advanced large-scale** 947 **video generative models.** *CoRR*, abs/2503.20314.
- 948 Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhi- 949 hao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin 950 Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei 951 Du, Xuancheng Ren, Rui Men, Dayiheng Liu, 952 Chang Zhou, Jingren Zhou, and Junyang Lin. 2024a. 953 **Qwen2-vl: Enhancing vision-language model's per-** 954 **ception of the world at any resolution.** *CoRR*, 955 abs/2409.12191.
- 956 Wenhui Wang, Hangbo Bao, Li Dong, Johan 957 Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, 958 Owais Khan Mohammed, Saksham Singhal, Subho- 959 jit Som, and 1 others. 2023. Image as a foreign 960 language: Beit pretraining for vision and vision- 961 language tasks. In *Proceedings of the IEEE/CVF 962 Conference on Computer Vision and Pattern Recog- 963 nition*, pages 19175–19186.
- 964 Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, 965 Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, 966 Yueze Wang, Zhen Li, Qiying Yu, Yingli Zhao, Yu- 967 long Ao, Xuebin Min, Tao Li, Boya Wu, Bo Zhao, 968 Bowen Zhang, Liangdong Wang, Guang Liu, and 6 969 others. 2024b. **Emu3: Next-token prediction is all** 970 **you need.** *CoRR*, abs/2409.18869.
- 971 Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, 972 Alan L. Yuille, and Christoph Feichtenhofer. 2022. 973 Masked feature prediction for self-supervised visual
- 974 pre-training. In *IEEE/CVF Conference on Computer 975 Vision and Pattern Recognition, CVPR 2022, New 976 Orleans, LA, USA, June 18-24, 2022*, pages 14648– 977 14658.
- 978 Wenhao Wu. 2024. **Freeva: Offline MLLM as training- 979 free video assistant.** *CoRR*, abs/2405.07798.
- 980 Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian 981 Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze 982 Xie, Hongxu Yin, Li Yi, Song Han, and Yao Lu. 983 2025. **VILA-U: a unified foundation model integrat- 984 ing visual understanding and generation.** In *The Thir- 985 teenth International Conference on Learning Repre- 986 sentations, ICLR 2025, Singapore, April 24-28, 2025.* 987 OpenReview.net.
- 988 Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao 989 Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao 990 Gu, Zhiping Chen, Zhenheng Yang, and Mike Zheng 991 Shou. 2025a. Show-o: One single transformer to 992 unify multimodal understanding and generation. In 993 *The Thirteenth International Conference on Learning 994 Representations, ICLR 2025, Singapore, April 24-28, 995 2025.*
- 996 Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. 997 2025b. **Show-o2: Improved native unified multi- 998 modal models.** *CoRR*, abs/2506.15564.
- 999 Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. 1000 2025c. **Show-o2: Improved native unified multi- 1001 modal models.** *CoRR*, abs/2506.15564.
- 1002 Mingze Xu, Mingfei Gao, Zhe Gan, Hong-You Chen, 1003 Zhengfeng Lai, Haiming Gang, Kai Kang, and Af- 1004 shin Dehghan. 2024. **Slowfast-llava: A strong 1005 training-free baseline for video large language 1006 models.** *Preprint*, arXiv:2407.15841.
- 1007 An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, 1008 Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, 1009 Chengen Huang, Chenxu Lv, Chujie Zheng, Day- 1010 iheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao 1011 Ge, Haoran Wei, Huan Lin, Jialong Tang, and 40 1012 others. 2025. **Qwen3 technical report.** *CoRR*, 1013 abs/2505.09388.
- 1014 An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, 1015 Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan 1016 Li, Dayiheng Liu, Fei Huang, Guanting Dong, Hao- 1017 ran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian 1018 Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 1019 43 others. 2024. **Qwen2 technical report.** *CoRR*, 1020 abs/2407.10671.
- 1021 Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, 1022 and Cordelia Schmid. 2022. Zero-shot video ques- 1023 tion answering via frozen bidirectional language mod- 1024 els. In *Advances in Neural Information Processing 1025 Systems 35: Annual Conference on Neural Infor- 1026 mation Processing Systems 2022, NeurIPS 2022, New 1027 Orleans, LA, USA, November 28 - December 9, 2022.*
- 1028 Zhiyu Yin, Kehai Chen, Xuefeng Bai, Ruili Jiang, Jun- 1029 tao Li, Hongdong Li, Jin Liu, Yang Xiang, Jun Yu,

- 1030 and Min Zhang. 2025. Asurvey: Spatiotemporal  
1031 consistency in video generation.
- 1032 Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang,  
1033 Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen,  
1034 Yuxiang Huang, Yuanqian Zhao, Bokai Xu, Junbo  
1035 Cui, Yingjing Xu, Liqing Ruan, Luoyuan Zhang,  
1036 Hanyu Liu, Jingkun Tang, Hongyuan Liu, Qining  
1037 Guo, and 15 others. 2025. [Minicpm-v 4.5: Cooking](#)  
1038 [efficient mllms via architecture, data, and training](#)  
1039 [recipe](#). *CoRR*, abs/2509.18154.
- 1040 Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yuet-  
1041 ing Zhuang, and Dacheng Tao. 2018. Activitynet-qa:  
1042 A dataset for understanding complex web videos via  
1043 question answering. In *The Thirty-Third AAAI Con-*  
1044 *ference on Artificial Intelligence, AAAI 2019, The*  
1045 *Thirty-First Innovative Applications of Artificial In-*  
1046 *telligence Conference, IAAI 2019, The Ninth AAAI*  
1047 *Symposium on Educational Advances in Artificial*  
1048 *Intelligence, EAAI 2019, Honolulu, Hawaii, USA,*  
1049 *January 27 - February 1, 2019*, pages 9127–9134.  
1050 AAAI Press.
- 1051 Zhen Yuan, Yifei Chen, Shuo Zhao, Wen yi Wang, Ming-  
1052 Hao Zhang, Zhiping Wang, Le Zhang, Boxi Zhao,  
1053 Jian Li, Zhi-Yuan Wu, Ming Ding, and Jie Tang.  
1054 2024. Cogvideox: A general-purpose video genera-  
1055 tion model. *arXiv preprint arXiv:2406.06511*.
- 1056 Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang,  
1057 Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu  
1058 Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao  
1059 Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan  
1060 Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and  
1061 3 others. 2024. [MMMU: A massive multi-discipline](#)  
1062 [multimodal understanding and reasoning benchmark](#)  
1063 [for expert AGI](#). In *IEEE/CVF Conference on Com-*  
1064 *puter Vision and Pattern Recognition, CVPR 2024,*  
1065 *Seattle, WA, USA, June 16-22, 2024*, pages 9556–  
1066 9567. IEEE.
- 1067 Biao Zhang and Rico Sennrich. 2019. [Root mean](#)  
1068 [square layer normalization](#). In *Advances in Neural*  
1069 *Information Processing Systems 32: Annual Confer-*  
1070 *ence on Neural Information Processing Systems 2019,*  
1071 *NeurIPS 2019, December 8-14, 2019, Vancouver, BC,*  
1072 *Canada*, pages 12360–12371.
- 1073 Hang Zhang, Xin Li, and Lidong Bing. 2023a. Video-  
1074 llama: An instruction-tuned audio-visual language  
1075 model for video understanding. In *Proceedings of*  
1076 *the 2023 Conference on Empirical Methods in Nat-*  
1077 *ural Language Processing, EMNLP 2023 - System*  
1078 *Demonstrations, Singapore, December 6-10, 2023*,  
1079 *pages 543–553*. Association for Computational Lin-  
1080 guistics.
- 1081 Renrui Zhang, Jiaming Han, Chris Liu, Aojun Zhou,  
1082 Pan Lu, Yu Qiao, Hongsheng Li, and Peng Gao.  
1083 2023b. Llama-adapter: Efficient fine-tuning of large  
1084 language models with zero-initialized attention. In  
1085 *The Twelfth International Conference on Learning*  
1086 *Representations, ICLR 2024, Vienna, Austria, May*  
1087 *7-11, 2024*. OpenReview.net.
- 1088 Sijie Zhao, Yong Zhang, Xiaodong Cun, Shaoshu Yang,  
1089 Muyao Niu, Xiaoyu Li, Wenbo Hu, and Ying Shan.  
1090 2024. [CV-VAE: A compatible video VAE for la-](#)  
1091 [tent generative video models](#). In *Advances in Neural*  
1092 *Information Processing Systems 38: Annual Confer-*  
1093 *ence on Neural Information Processing Systems 2024,*  
1094 *NeurIPS 2024, Vancouver, BC, Canada, December*  
1095 *10 - 15, 2024*.
- 1096 Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala,  
1097 Michihiro Yasunaga, Leonid Shamis, Jacob Kahn,  
1098 Xuezhe Ma, Luke Zettlemoyer, and Omer Levy.  
1099 2024a. [Transfusion: Predict the next token and dif-](#)  
1100 [fuse images with one multi-modal model](#). *Preprint*,  
1101 *arXiv:2408.11039*.
- 1102 Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao,  
1103 Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang,  
1104 and Zheng Liu. 2024b. Mlvu: A comprehensive  
1105 benchmark for multi-task long video understanding.  
1106 *arXiv preprint arXiv:2406.04264*.
- 1107 Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu,  
1108 Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan,  
1109 Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xue-  
1110 hui Wang, Yue Cao, Yangzhou Liu, Xinguang Wei,  
1111 Hongjie Zhang, Haomin Wang, Weiye Xu, and 32  
1112 others. 2025. [Internvl3: Exploring advanced train-](#)  
1113 [ing and test-time recipes for open-source multimodal](#)  
1114 [models](#). *CoRR*, abs/2504.10479.

## 1115 A Example Appendix

### 1116 A.1 LLM Use Declaration

1117 Large Language Models (ChatGPT) were used ex-  
1118clusively to improve the clarity and fluency of En-  
1119glish writing. They were not involved in research  
1120 ideation, experimental design, data analysis, or in-  
1121 terpretation. The authors take full responsibility  
1122 for all content.

### 1123 A.2 Related Work

1124 **Video generation.** Video generation has seen  
1125 remarkable advancements with the rise of diffu-  
1126 sion models and generative adversarial networks  
1127 tailored for temporal data. Recent diffusion or  
1128 flow based frameworks, such as Video Diffusion  
1129 Models (Ho et al., 2022b), Imagen Video (Ho  
1130 et al., 2022a), VideoCrafter2 (Chen et al., 2024a)  
1131 and Stable Video Diffusion (Blattmann et al.,  
1132 2023b), have produced high-fidelity clips with im-  
1133 proved temporal consistency, enabling applications  
1134 in creative generation and simulation (Liu et al.,  
1135 2025; Shi et al., 2025). Latent diffusion techniques  
1136 (Blattmann et al., 2023c) further improve efficiency  
1137 by operating in compressed latent spaces, enabling  
1138 scalable video generation. In parallel, GAN meth-  
1139 ods like MoCoGAN (Tulyakov et al., 2018) and  
1140 StyleGAN-V (Skorokhodov et al., 2022) explore  
1141 alternative formulations. Despite these advances,  
1142 maintaining long-term temporal consistency in ex-  
1143 tended sequences remains challenging, as summa-  
1144 rized by recent surveys and analyses (Melnik et al.,  
1145 2024; Yin et al., 2025).

1146 **Video understanding.** Recent progress in video  
1147 understanding has been driven by transformer-  
1148 based architectures and self-supervised learning  
1149 paradigms that effectively model spatio-temporal  
1150 relationships. Methods like MViT (Fan et al.,  
1151 2021), Video Swin Transformer (Liu et al., 2022),  
1152 TimeSformer (Bertasius et al., 2021) and ViViT  
1153 (Arnab et al., 2021) have advanced the field by  
1154 capturing long-range dependencies across video  
1155 frames, achieving strong performance on datasets  
1156 such as Kinetics-700 (Carreira et al., 2019).  
1157 Beyond supervised training, self-supervised ap-  
1158 proaches—including masked modeling (Video-  
1159 MAE (Tong et al., 2022), MaskFeat (Wei et al.,  
1160 2022), OmniMAE (Girdhar et al., 2023)) and  
1161 early contrastive methods (VideoMoCo (Pan et al.,  
1162 2021))—leverage unlabeled videos to learn robust,  
1163 transferable representations, reducing dependence

1164 on costly annotations and benefiting action recog-  
1165 nition and video segmentation.

1166 **Unified multimodal models.** Unified multi-  
1167 modal modeling has progressed from joint vi-  
1168 sion–language pretraining to architectures that sup-  
1169 port both understanding and generation across  
1170 modalities. Foundational systems like CLIP (Rad-  
1171 ford et al., 2021) establish large-scale alignment,  
1172 while BEiT-3 (Wang et al., 2023) and UnifiedML-  
1173 LM (Li et al., 2024) broaden task coverage. Push-  
1174 ing toward unified generation, Show-o (Xie et al.,  
1175 2025a) integrates autoregression with discrete dif-  
1176 fusion within a single Transformer to support VQA,  
1177 text-to-image, and various editing tasks. In a  
1178 complementary direction focused on robustness  
1179 rather than general any-to-any generation, FLUID  
1180 (Cuong et al., 2025) uses token-level distillation for  
1181 cross-modal fusion. Open generalist systems then  
1182 aim to unify understanding and generation end-to-  
1183 end: BAGEL (Deng et al., 2025) offers an open,  
1184 decoder-only framework with parallel language  
1185 and diffusion branches trained jointly, achieving  
1186 competitive results across image-centric tasks, and  
1187 BLIP3-o (Chen et al., 2025) releases a fully open  
1188 family where a diffusion transformer is coupled  
1189 to strong multimodal understanding, yielding uni-  
1190 fied image understanding and generation. Extending  
1191 unification from images to video, Omni-Video  
1192 (Tan et al., 2025) teaches an MLLM to emit contin-  
1193 uous visual tokens that are adapted and consumed  
1194 by a diffusion video decoder, enabling generation,  
1195 editing, and understanding in one pipeline.

### 1196 A.3 Hierarchical Post Training

1197 **Initialization.** To avoid the prohibitive cost of  
1198 training a unified video model from scratch, we  
1199 bootstrap UniVid from strong, publicly avail-  
1200 able checkpoints and finetune only small sub-  
1201 sets of parameters. Our architecture follows the  
1202 BAGEL (Deng et al., 2025) design framework,  
1203 adopting its multimodal integration approach with  
1204 three key components: Qwen2 (Yang et al., 2024)  
1205 as the LLM backbone with standard architec-  
1206 tural choices such as RMSNorm (Zhang and Sen-  
1207 nrich, 2019), SwiGLU (Shazeer, 2020), RoPE (Su  
1208 et al., 2024), GQA (Ainslie et al., 2023), and QK-  
1209 Norm (Henry et al., 2020) for training stability,  
1210 SigLIP2-so400m/14 (Tschanne et al., 2025) as the  
1211 ViT (Dosovitskiy et al., 2021) encoder for vi-  
1212 sual understanding with NaViT support for native  
1213 aspect ratios, and a pre-trained FLUX VAE with

1214  
1215  
1216  
1217  
1218  
1219

8x downsampling and frozen weights. The framework interleaves text, ViT, and VAE tokens within the LLM using generalized causal attention, where tokens attend to all preceding modality splits while maintaining appropriate attention patterns within each modality.

1220  
1221  
1222  
1223  
1224  
1225  
1226  
1227  
1228  
1229  
1230  
1231  
1232

**Data curation and formatting.** For understanding, we align our data format with the dialog style used by Video-ChatGPT (Maaz et al., 2024). ActivityNet-QA annotations ( $\text{video\_id}, q, a$ ) are converted into structured conversations. Specifically, each sample is represented as a JSON object containing three fields: (1) an identifier, (2) a video reference, and (3) a conversations array consisting of two turns, a user query and the corresponding model response. For generation, we curate a subset of OpenVid-1M to form text/image to video pairs. Videos are uniformly sub-sampled and pre-processed identically to inference.

1233  
1234  
1235  
1236  
1237  
1238  
1239  
1240  
1241  
1242  
1243  
1244  
1245  
1246

**Stage I generation branch alignment.** We couple the MLLM with Wan 2.2 and adapt the conditioning path so that MLLM-produced tokens can reliably steer synthesis. Concretely, we (i) insert a textual adapter between the LLM tokens, with dynamic sequence length adaptation, and (ii) apply LoRA to the DiT cross-attention layers; all other DiT/MLLM weights remain frozen. Training uses a standard flow-matching objective with classifier-free guidance dropout on text, optimizing only the context projector and LoRA parameters. This stage preserves MLLM’s native understanding while aligning Wan’s generation to the rich semantics emitted by MLLM.

1247  
1248  
1249  
1250  
1251  
1252  
1253  
1254  
1255  
1256  
1257  
1258  
1259  
1260

**Stage II understanding adaptation.** We finetune for video QA on ActivityNet-QA using 20k samples from the dataset. Each sample concatenates the question with a <video> placeholder, and we feed a multi-frame clip obtained by uniform sampling. Frames are encoded by the ViT into visual tokens and projected to the LLM space via the connector. We adopt instruction SFT for video: compute autoregressive cross-entropy only on the assistant turns; user tokens are fully masked to prevent label leakage. To keep compute moderate while injecting temporal cues, we finetune only the last two ViT blocks and the connector for 4 epochs, keeping the LLM frozen.

1261  
1262  
1263

**Stage III joint training.** Finally, we co-train generation and understanding to let the two branches benefit from each other. During joint training, we

1264  
1265  
1266  
1267  
1268  
1269  
1270  
1271  
1272  
1273  
1274  
1275  
1276  
1277  
1278  
1279  
1280  
1281  
1282  
1283  
1284  
1285  
1286

gradually activate the VAE feature branch and alternate which modules are trainable. In the initial phase, only the last few ViT layers and connector are trainable and the adapter is frozen, so the MLLM is forced to encode features in ViT that are beneficial for both understanding and generation. Next, we freeze the ViT and progressively introduce VAE features into the adapter input, enabling the DiT LoRA to learn how to generate guided by different mixing ratios of ViT and VAE features. Finally, we co-train both the ViT layers, connector, adapter and DiT to fully exploit the complementary strengths of semantic ViT features and detail-rich VAE features. On the understanding side, Fig. ?? shows that joint training leads to faster convergence and consistently higher MSVD-QA (Piergiovanni et al., 2022) accuracy than single-task training. On the generation side, Stage III further improves the VBench-Long overall score from 79.28 to 85.27 and boosts most technical and semantic dimensions, as summarized in Tab. 6, confirming that better video understanding feedback translates into higher-quality video generation.

#### A.4 More examples of Video generation and understanding.

We provide more examples of video understanding and generation in Fig. 5 and Fig. 6

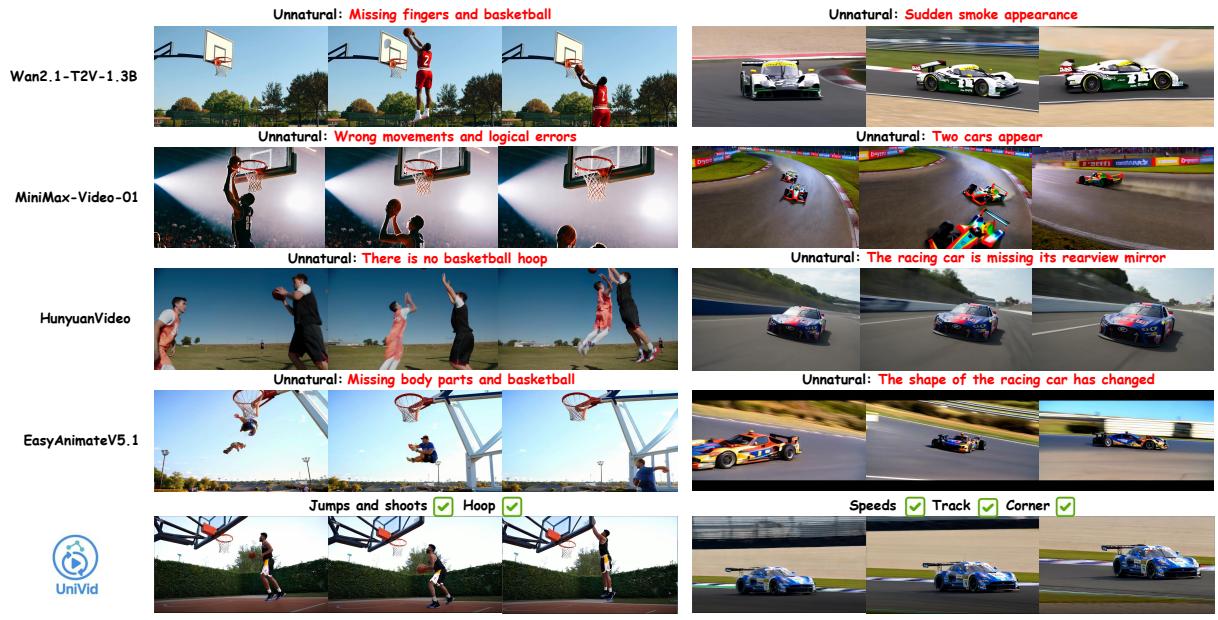


Figure 3: Comparisons with State-of-the-Art Video Generation Models (Wang et al., 2025; MiniMax, 2024; Kong et al., 2024; Fu et al., 2024b).

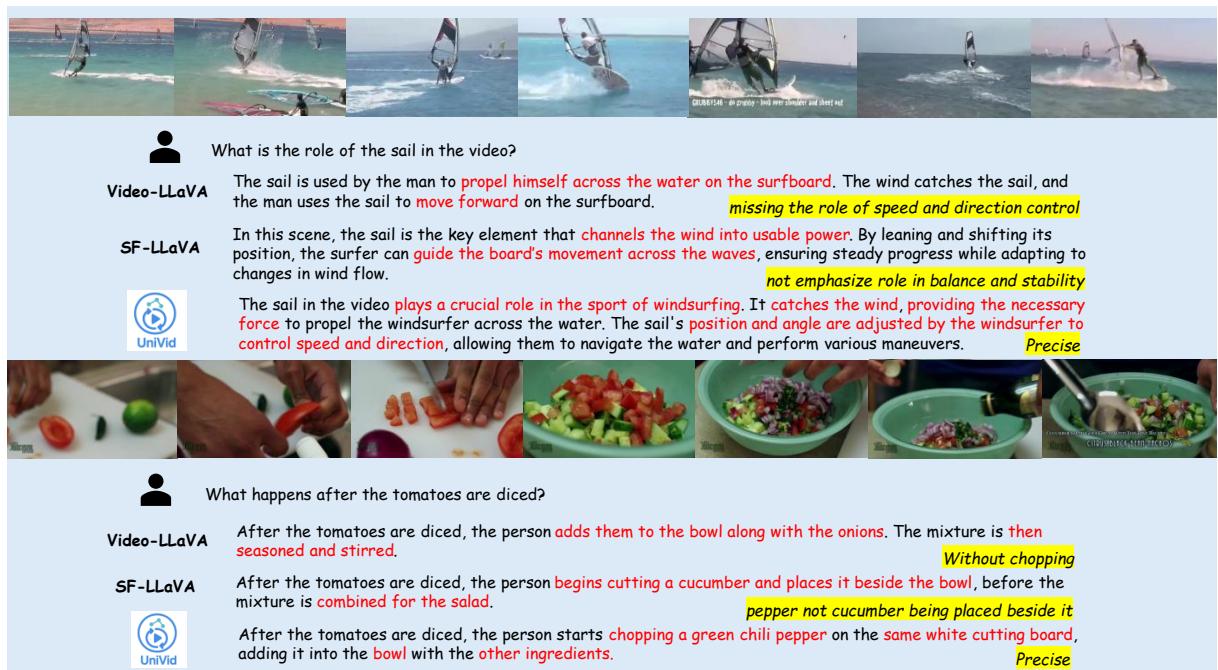


Figure 4: Comparisons of State-of-the-Art Video Understanding Models (Lin et al., 2024; Xu et al., 2024).



-  who is laughing?
-  The person **on the right** is laughing.
-  What letterman interviewing comedic actor bill murray?
-  The video shows a scene from **"The Tonight Show Starring David Letterman,"** where the host **David Letterman** is interviewing the comedic actor Bill Murray.



-  What is a person playing on a television?
-  Maybe the person is playing the **guitar**.
-  What is a woman doing?
-  The woman in the image appears to be **playing a rhythm game** on a screen, possibly a music **video game or a dance game**. She is positioned in front of a virtual stage with a guitar-like interface, indicating that she is likely performing or following a rhythm pattern.



-  What is the role of the sail in the video?
-  The sail in the video plays a crucial role in the sport of windsurfing. It catches the wind, providing the necessary force to propel the windsurfer across the water. The sail's **position and angle are adjusted by the windsurfer to control speed and direction**, allowing them to navigate the water and perform various maneuvers.
-  How safe is this game in the video?
-  The game shown in the video is windsurfing, which can be considered relatively safe for experienced individuals. However, it **does carry some risks**, such as falling into the water, collisions with other windsurfers or objects, and potential injuries from the sail or board. It is important for participants to wear appropriate safety gear, such as life jackets and helmets, and to be aware of their surroundings and wind conditions. Additionally, taking lessons from a qualified instructor can help ensure a safer experience.

Figure 5: The qualitative results of the video understanding. Blue for static questions, green for dynamic questions.

## Text to Video



A dolphin leaps out of the ocean, splashing water as it dives back in.



Two anthropomorphic cats in comfy boxing gear and bright gloves fight intensely on a spotlit stage.



A futuristic drone weaves quickly between skyscrapers, lights glowing in the night sky.

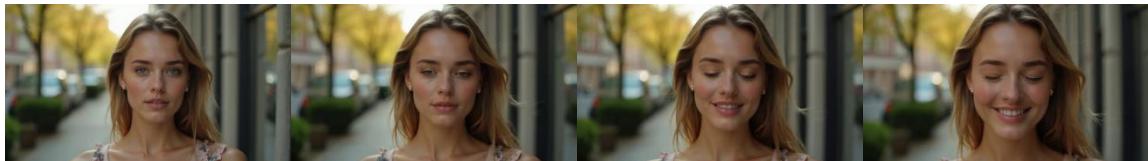


A high-speed train rushes past the station, its motion blurring in the background.

## Text and Image to Video



A hawk soars above the mountains, wings spread wide against the sunset.  
(from image)



A cinematic video of a young woman with natural makeup and long blonde hair, standing on a sunlit street with blurred trees and cars in the background. The camera slowly moves closer as her hair gently flows with the breeze. She softly smiles and blinks, creating a natural and elegant moment. Warm golden hour lighting, realistic style, high detail, 4K.  
(from image)

Figure 6: The qualitative results of T2V and TI2V generation.

1291

## A.5 Text Prompts used in the understanding

**Role.** Classify a video question as `static` or `dynamic`. Output JSON only.

**Definitions.**

- `dynamic`: requires temporal reasoning such as counting, repetition, order, or changes over time (e.g., “how many times”, “before/after”, “first/last”).
- `static`: can be answered from a small set of unordered frames (identity, attribute, location, scene, one-shot action).

**Question.** *{question}*

**Return.** Single-line JSON with fields: `qtype` (“`static`” or “`dynamic`”), `rationale` (1–2 short phrases; no extra text).

1292

**Role.** Summarize chronologically ordered frame notes into a compact global caption. Do not invent facts.

**Input.** Frame-wise notes (earlier → later):

- *{note\_1}*

- *{note\_2}*

...

**Write.** One global caption (2–4 sentences) that connects multiple frames, focusing on: (1) moving entities with consistent appearance and actions across time; (2) static scene objects and their states; (3) temporal hints only if explicitly evidenced (e.g., “then”, “later”, “repeatedly”). Style: terse and factual; no bullet lists, storytelling, or frame-by-frame recitation.

1294

## 2: Frame Summarization Prompt

1295

**Role.** Precise evaluator for video-QA. Return a *single-line* JSON only (no Markdown/code).

**Keys.** `score` (float 0..1), `verdict` (“`accept`” if `score`  $\geq 0.7$  else “`reject`”), `brief_reason` (1–2 short bullets).

**Example user.** *{one\_shot\_user}*

**Example assistant.** *{one\_shot\_assistant}*

**Your task.** Given the current case, output the JSON only.

1296

1293

## 1: Question Type Classification Prompt

## 3: Answer Evaluation Prompt

1297

**Role.** Reflector in a video-understanding pipeline. You receive the question, a global caption (from 16 uniformly sampled frames), the last answer (low confidence/rejected), and its evaluation JSON.

**Objective.** Analyze why the answer likely fails (missing object, wrong span, ambiguity, etc.) and produce a single short *declarative* retrieval text for the next round of keyframe selection.

**Strict rules.** (1) Output JSON only with key `refined_query`. (2)  $\text{refined\_query} \leq 25$  tokens, declarative statement (not a question), capturing disambiguating cues (entities, attributes, actions, temporal hints, viewpoint). (3) If confidence is already good (score  $\geq 0.7$  or verdict="accept"), return an empty string. (4) Prefer concrete visual cues (colors, clothing, object names, motion phase, timestamps, left/right, first/last). (5) No speculation or unseen entities.

**Inputs.** Question: `{question}` Global caption: `{global_caption}` Last answer: `{last_answer}` Evaluation JSON: `{eval_json}`

**Return.** `{"refined_query": "..."}`

1298

1299

## 4: Reflection Prompt

**Role.** Assist video understanding via per-frame analysis. Describe the main objects and actions in *this single frame* concisely.

**Focus.** (1) Living entities: distinct entities (appearance, clothing, color, species), likely roles, and what each is doing (verb phrases). (2) Static objects & scene: salient items and states (color, shape, on/off, open/closed, broken/intact), plus scene context (indoor/outdoor, location hints).

**Style.** Specific but brief; no speculation; 2–4 short sentences.

1300

1301

## 5: Single-Frame Analysis Prompt

**Role.** Answer concisely using only the question and the global video caption.

**Inputs.** Question: `{question}` Global caption (may miss fine details): `{global_caption}`

**Instruction.** Produce one short answer (1–2 sentences). If information is insufficient, reply: "Not enough evidence from global caption."

1302

## 6: Global Answer Prompt

1303

### A.6 Pyramid Reflection as Test-time RL

1304

We cast Pyramid Reflection as a test-time reinforcement learning procedure operating on an ordered evidence set. At round  $r$ , the state is  $x_r = (s_r, W_r, C_g)$ , where  $s_r$  is a short search text,  $W_r$  is the ordered working set of frames, and  $C_g$  is a global caption distilled once from uniformly sampled seeds. The action reconfigures  $W_r$  given  $s_r$  via an expand or shrink policy. The Actor answers from  $(W_r, C_g)$ , and the Evaluator returns a score  $R_r \in [0, 1]$  and a verdict that controls early stopping. All frame embeddings are computed once and cached; later rounds update indices and similarity or diversity scores only.

1305

Frame selection uses a vision–language retriever with cosine similarity. Let  $\phi(s)$  be the text embedding and  $\{\mathbf{v}_i\}_{i=1}^N$  the cached frame embeddings:

1306

$$\text{sim}(i, s) = \langle \widehat{\mathbf{v}}_i, \widehat{\phi(s)} \rangle. \quad (6)$$

1307

We define a soft retrieval policy over the pool  $P$ :

1308

$$\pi(i | s) = \frac{\exp(\text{sim}(i, s)/\tau)}{\sum_{j \in P} \exp(\text{sim}(j, s)/\tau)}. \quad (7)$$

1309

Sampling sequentially without replacement with joint probability  $\prod_{\ell=1}^K \pi(i_\ell | s, i_{<\ell})$  and respecting chronology yields  $W_s$ .

1310

In the expand mode, at target size  $K_t$  we add the top  $m$  unseen frames by similarity (no duplicates):

1311

$$\Delta_t = \arg \max_{i \in P \setminus S_{\text{sel}}}^m \text{sim}(i, s_{t-1}), \quad (8)$$

1312

$$S_{\text{sel}} \leftarrow S_{\text{sel}} \cup \Delta_t, \quad m = K_t - |S_{\text{sel}}|.$$

1313

In the shrink mode, with current  $S_{\text{sel}}$  and target  $K_t \in \{32, 16\}$ , we apply Maximal Marginal Relevance:

1314

$$S_{\text{sel}} = \arg \max_{\substack{S \subseteq S_{\text{sel}} \\ |S|=K_t}} \sum_{i \in S} \left[ \lambda \text{sim}(i, s_{t-1}) - (1 - \lambda) \max_{j \in S \setminus \{i\}} \text{sim}(i, j) \right]. \quad (9)$$

1315

1316

1317

We adopt a verbal policy-improvement view (Shinn et al., 2023). Let the objective be the expected Evaluator value under the retrieval policy:

1318

1319

1320

$$J(s) = \mathbb{E}_{i_{1:K} \sim \pi(\cdot | s)} [V(W_s)], \quad (10)$$

1321

1322

1323

Model	Overall Scores				Technical Quality				Aesthetic Quality	
	Total Score↑	Quality↑	Semantic↑	Subject↑	Background↑	Temporal↑	Motion↑	Dynamic↑	Aesthetic↑	Imaging↑
UniVid (base)	76.25	77.11	72.82	93.82	93.43	94.15	94.04	57.16	58.47	65.65
UniVid (w/o MLLM)	77.82	78.69	74.32	94.55	94.78	95.19	94.79	58.08	59.88	66.01
UniVid (w/o TMA)	80.42	81.51	76.04	96.55	95.91	97.12	96.25	59.98	62.08	67.10
<b>UniVid (Full)</b>	<b>85.27</b>	<b>86.44</b>	<b>80.58</b>	<b>98.96</b>	<b>97.76</b>	<b>99.88</b>	<b>99.25</b>	<b>61.83</b>	<b>64.21</b>	<b>73.03</b>

Model	Semantic Fidelity								
	Object↑	Multi-Obj↑	Action↑	Color↑	Spatial↑	Scene↑	Appearance↑	Temporal↑	Overall↑
UniVid (base)	89.53	73.32	89.41	87.86	76.13	42.32	19.03	21.60	22.48
UniVid (w/o MLLM)	90.80	74.37	90.12	87.99	76.63	43.32	20.57	22.26	22.98
UniVid (w/o TMA)	91.51	75.42	91.53	89.33	77.58	44.61	21.03	23.62	24.13
<b>UniVid (Full)</b>	<b>94.52</b>	<b>77.45</b>	<b>94.20</b>	<b>92.10</b>	<b>80.70</b>	<b>46.66</b>	<b>23.57</b>	<b>25.91</b>	<b>27.60</b>

Table 3: Ablation study of UniVid on VBench-Long. w/o means “without”. Best results are **bold**.

Model	Overall Scores				Technical Quality				Aesthetic Quality	
	Total Score↑	Quality↑	Semantic↑	Subject↑	Background↑	Temporal↑	Motion↑	Dynamic↑	Aesthetic↑	Imaging↑
UniVid (w/o TMA)	80.42	81.51	76.04	96.55	95.91	97.12	96.25	59.98	62.08	67.10
UniVid (Constant)	82.72	83.96	77.78	97.81	96.41	98.12	98.01	60.11	63.47	70.65
UniVid (Step)	82.80	84.35	76.59	97.32	96.74	98.15	98.54	59.71	63.91	71.19
UniVid (Linear)	83.30	84.51	78.47	97.45	96.78	98.20	98.76	60.01	63.88	71.01
<b>UniVid (Cosine)</b>	<b>85.27</b>	<b>86.44</b>	<b>80.58</b>	<b>98.96</b>	<b>97.76</b>	<b>99.88</b>	<b>99.25</b>	<b>61.83</b>	<b>64.21</b>	<b>73.03</b>

Table 4: Ablation study on TMA schedules on VBench-Long. w/o means “without”. Best results are **bold**.

with

$$V(W_s) = \mathbb{E}[R | W_s, C_g]. \quad (11)$$

Using the likelihood-ratio identity with a baseline  $b$  yields

$$\nabla_s J(s) = \mathbb{E} \left[ \left( \sum_{t=1}^K \nabla_s \log \pi(i_t | s, i_{<t}) \right) (R - b) \right]. \quad (12)$$

A single ascent step motivates a verbal update to the search text:

$$s_{r+1} = s_r + \eta \left( \sum_{t=1}^K \nabla_s \log \pi(i_t | s_r, i_{<t}) \right) (R_r - b), \quad (13)$$

where we use the softmax score function with  $g_i(s) := \nabla_s \text{sim}(i, s)$  and  $\bar{g}(s) := \mathbb{E}_{j \sim \pi(\cdot|s)} g_j(s)$ :  $\nabla_s \log \pi(i | s) = \tau^{-1}(g_i(s) - \bar{g}(s))$ , so the edit in  $s$  aligns with frames that explain higher return through the text encoder  $\phi(\cdot)$ . Practically, the reflector inserts temporally and semantically discriminative cues (entities, colors, viewpoints, before/after, first/last, motion phase), which increases  $\text{sim}(i, s)$  for diagnostic frames and decreases it for distractors, implementing Eq. 13 in language space without parameter updates.

To connect the update with both expand and shrink, we use a piecewise-smooth set surrogate that trades relevance against redundancy (subgradients at ties):

$$\tilde{V}(W_s) = \frac{1}{K} \sum_{i \in W_s} \text{sim}(i, s) - \gamma \max_{i \neq j \in W_s} \text{sim}(i, j). \quad (14)$$

Since  $\partial \text{sim}(i, s) / \partial s$  points toward  $\mathbf{v}_i$  via  $\phi(s)$ , the gradient  $\nabla_s \tilde{V}(W_s)$  is aligned with the direction in Eq. 12. If the reflector’s edit correlates with the advantage  $A_r = R_r - b$ , then for a sufficiently small step size  $\eta$  the expected first-order improvement satisfies

$$\mathbb{E}[J(s_{r+1}) - J(s_r)] \approx \eta \mathbb{E} \left[ \left\langle \sum_t \nabla_s \log \pi(i_t | s_r, i_{<t}), s_{r+1} - s_r \right\rangle A_r \right] \geq 0. \quad (15)$$

Early stopping is triggered when the Evaluator score exceeds a fixed threshold:

$$\text{stop at round } r \text{ if } R_r \geq \tau, \quad \tau = 0.7. \quad (16)$$

With cached features, each round requires only similarity and diversity scoring together with reasoning over a compact, temporally ordered  $W_r$ , which concentrates the Actor on temporal relations under a tight token budget and improves video understanding with low computational cost.

## A.7 Ablation Study

**Ablation on video generation.** Tab. 3 presents an ablation on VBench-Long disentangling the roles of our two main components. Removing the multi-level language modeling module (w/o MLLM) mainly hurts the semantic-fidelity metrics that require precise spatial layout and appearance preservation, while the low-level technical quality remains relatively stable. In contrast, disabling Temperature Modality Alignment (w/o TMA) leads to a clear drop in temporal and motion-related

Model	Overall Scores				Technical Quality				Aesthetic Quality	
	Total Score↑	Quality↑	Semantic↑	Subject↑	Background↑	Temporal↑	Motion↑	Dynamic↑	Aesthetic↑	Imaging↑
UniVid (w/o ViT)	48.53	57.16	46.37	74.51	72.91	74.02	74.23	46.91	47.01	55.10
UniVid (w/o VAE)	71.78	71.90	71.75	89.43	88.75	90.19	89.80	57.23	58.86	67.12
<b>UniVid (Ours, VAE &amp; ViT Encoder)</b>	<b>85.27</b>	<b>86.44</b>	<b>80.58</b>	<b>98.96</b>	<b>97.76</b>	<b>99.88</b>	<b>99.25</b>	<b>61.83</b>	<b>64.21</b>	<b>73.03</b>

Model	Semantic Fidelity								
	Object↑	Multi-Obj↑	Action↑	Color↑	Spatial↑	Scene↑	Appearance↑	Temporal↑	Overall↑
UniVid (w/o ViT)	72.41	54.41	75.51	74.31	58.68	32.69	14.12	15.63	17.15
UniVid (w/o VAE)	87.23	69.54	87.34	88.92	74.32	39.27	20.54	21.61	22.12
<b>UniVid (Ours, VAE &amp; ViT Encoder)</b>	<b>94.52</b>	<b>77.45</b>	<b>94.20</b>	<b>92.10</b>	<b>80.70</b>	<b>46.66</b>	<b>23.57</b>	<b>25.91</b>	<b>27.60</b>

Table 5: Ablation study of the generation branch of UniVid to verify the effectiveness of encoder setting. w/o means “without”. Best results are **bold**.

Model	Overall Scores				Technical Quality				Aesthetic Quality	
	Total Score↑	Quality↑	Semantic↑	Subject↑	Background↑	Temporal↑	Motion↑	Dynamic↑	Aesthetic↑	Imaging↑
UniVid (Stage I)	79.28	80.38	74.90	94.23	94.19	95.31	96.32	58.98	61.91	70.11
<b>UniVid (Joint, Stage III)</b>	<b>85.27</b>	<b>86.44</b>	<b>80.58</b>	<b>98.96</b>	<b>97.76</b>	<b>99.88</b>	<b>99.25</b>	<b>61.83</b>	<b>64.21</b>	<b>73.03</b>

Model	Semantic Fidelity								
	Object↑	Multi-Obj↑	Action↑	Color↑	Spatial↑	Scene↑	Appearance↑	Temporal↑	Overall↑
UniVid (Stage I)	90.12	75.59	90.98	89.91	77.52	44.57	20.51	21.12	24.01
<b>UniVid (Joint, Stage III)</b>	<b>94.52</b>	<b>77.45</b>	<b>94.20</b>	<b>92.10</b>	<b>80.70</b>	<b>46.66</b>	<b>23.57</b>	<b>25.91</b>	<b>27.60</b>

Table 6: Stage I vs Stage III performance on VBench-Long to verify the effect of hierarchical joint training on video generation. w/o means “without”. Best results are **bold**.

scores, indicating that the denoising process becomes less stable over long horizons even though per-frame quality is still high. The full UniVid model consistently achieves the best performance across technical, aesthetic, and semantic dimensions, suggesting that multi-level language modeling and TMA are complementary: the former strengthens multi-object, spatial, and appearance grounding, whereas the latter enforces temporally coherent, prompt-faithful dynamics during generation.

Tab. 4 shows that removing TMA causes a noticeable drop in temporal stability, motion smoothness, and imaging quality, confirming its necessity for coherent long-horizon generation. Among different scheduling strategies, the cosine scheme consistently performs best. Its smooth transition from stronger early text guidance to later visual refinement yields better semantic fidelity and more stable dynamics than constant, step, or linear variants, highlighting the importance of a well-shaped modulation schedule.

Fig. 7 visualizes these issues: without TMA, generated players exhibit unnatural fingers, distorted poses, and implausible ball trajectories, whereas the full UniVid produces coherent jump shots with realistic ball arcs. Qualitative comparisons in Fig. 3 confirm that UniVid consistently avoids missing objects and deformations that plague prior models, achieving both semantic plausibility and temporal stability.

**Ablation on video understanding.** Tab. 7 compares four variants: a lightweight base model without our training or reasoning additions, a version w/o finetune that removes Stage-II video-QA finetuning, a version w/o Reflection that keeps finetuning but disables the Pyramid Reflection loop, and the Full UniVid. Finetuning the understanding branch on ActivityNet-QA style instruction data already yields clear gains over the base, indicating that modest, task-aligned supervision substantially improves cross-modal grounding. Adding Pyramid Reflection further boosts accuracy, with similar trends in the QA scores, confirming that query-driven keyframe selection plus the Actor–Evaluator–Reflector loop improves temporal coherence and evidence retrieval. Overall, the full system combines data-efficient tuning with iterative reasoning to deliver competitive results across all four benchmarks.

Furthermore, we investigate the impact of scaling down the Evaluator and Reflector. Specifically, we replace the originally used large-scale language model (LLM) with a more lightweight 7B LLM. As shown in Tab. 8, the results demonstrate only a marginal performance drop. This is because the primary reasoning and semantic alignment are handled by the MLLM, while the Evaluator and Reflector mainly serve to refine information selection, a process that does not heavily rely on strong reasoning capability or extensive prior knowledge. This indicates that Pyramid Reflection can be efficiently executed using smaller models, achiev-

1390  
1391  
1392  
1393  
1394  
1395  
1396  
1397  
1398  
1399  
1400  
1401  
1402  
1403  
1404  
1405  
1406  
1407  
1408  
1409  
1410  
1411  
1412  
1413  
1414  
1415  
1416  
1417  
1418  
1419  
1420  
1421  
1422  
1423  
1424  
1425  
1426  
1427  
1428  
1429  
1430  
1431  
1432  
1433  
1434  
1435  
1436  
1437  
1438  
1439  
1440  
1441  
1442  
1443  
1444  
1445  
1446  
1447  
1448  
1449  
1450  
1451  
1452



Figure 7: Ablation Study on Temperature Modality Alignment.

Methods	MSVD-QA		MSRVTT-QA		TGIF-QA		ActivityNet-QA	
	Acc↑	Score↑	Acc↑	Score↑	Acc↑	Score↑	Acc↑	Score↑
UniVid (Base)	64.1	3.3	48.9	2.8	54.2	3.0	39.8	3.0
UniVid (w/o finetune)	71.1	3.9	52.2	3.0	63.5	3.6	46.5	3.2
UniVid (w/o Reflection)	73.1	4.0	55.0	3.1	64.6	3.6	52.0	3.4
<b>UniVid (Full)</b>	<b>80.1</b>	<b>4.2</b>	<b>61.4</b>	<b>3.4</b>	<b>75.0</b>	<b>4.1</b>	<b>58.8</b>	<b>3.6</b>

Table 7: Ablation study of UniVid on four video QA benchmarks. Acc. denotes accuracy (%), Score denotes average rating (0–5). Best results are **bold**.

ing a favorable trade-off between efficiency and accuracy. Notably, when we only substitute the Evaluator and Reflector with smaller LLMs while keeping the MLLM unchanged, performance degradation remains minimal, which further supports the above conclusion. Additionally, to mitigate potential understanding-evaluation(reflection) bias caused by using the same model family, we adopt different model types for Evaluator and Reflector, leading to moderate but consistent performance improvements.

Moreover, we evaluate our model on several recent benchmarks designed for unified video understanding, including MMLU (Hendrycks et al., 2021a,b), MMMU (Yue et al., 2024), MME (Fu et al., 2024a), MMBench (Fang et al., 2024), and MLVU (Zhou et al., 2024b). These datasets cover diverse multimodal reasoning tasks and reflect models’ comprehensive understanding capabilities. We compare our unified model with its understanding-only models and latest Open-Source Unified Video Model to highlight our model’s performance. As shown in Tab. 10, our method achieves competitive results on most benchmarks, particularly outperforming existing unified models. It is also worth noting that Video-MME includes longer videos ( $>10$  min), for which we further report results under short-video (S) and mid-length (M) subsets.

Our unified model shows more significant advantages on short-video scenarios, consistent with its design characteristics, while still maintaining strong overall comprehension capabilities.

**Ablation on encoding mechanism.** We study the internal encoding mechanism of UniVid. During training, we employ both a ViT and a VAE to encode visual information, where the ViT excels at capturing high-level semantics and the VAE is more effective in representing pixel-level details. We conduct ablation studies for both generation and understanding tasks to examine the role of each encoder.

For video generation, Tab. 5 shows that using only the ViT or only the VAE leads to significant degradation across almost all VBV-Bench-Long dimensions. In contrast, combining both encoders yields large improvements in overall score and boosts technical, aesthetic, and semantic fidelity metrics. This confirms that high-level semantic encoding and low-level detail encoding are complementary for long-horizon video synthesis.

For video understanding, Tab. 9 indicates that ViT alone is sufficient to achieve strong performance, while adding the VAE brings marginal or no further improvement. This aligns with the intuition that understanding tasks rely more on semantic abstraction than pixel-level reconstruction. Together,

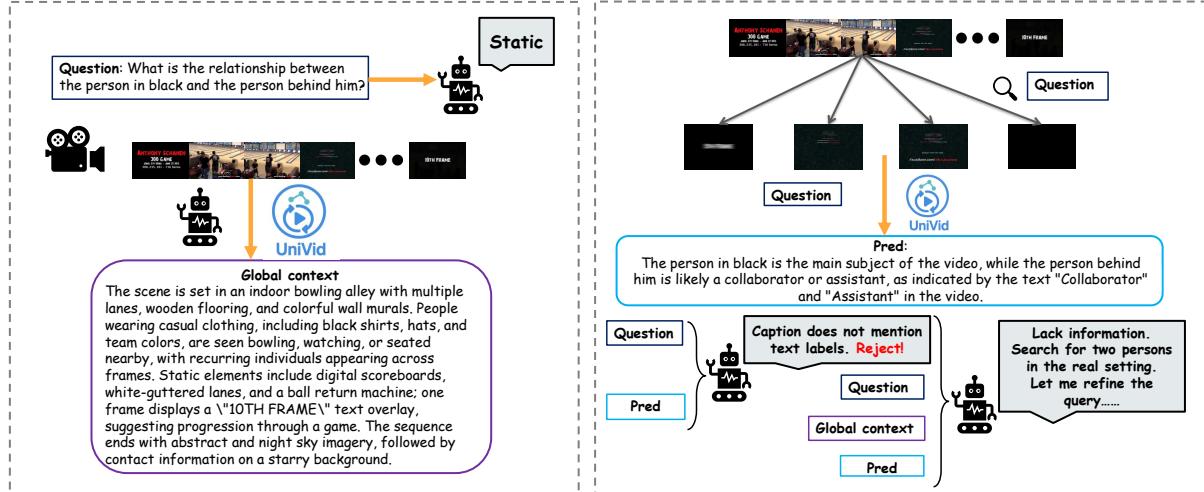


Figure 8: The pipeline of the video understanding.

Methods	MSVD-QA		MSRVT-QA		TGIF-QA		ActivityNet-QA	
	Acc↑	Score↑	Acc↑	Score↑	Acc↑	Score↑	Acc↑	Score↑
UniVid (Qwen2-7B E&R)	76.9	3.9	57.4	3.2	71.8	3.9	56.7	3.5
UniVid (LLaMA-3 8B E and LLaVA-1.6 7B R)	78.2	4.0	59.1	3.3	72.4	4.0	56.8	3.5
UniVid (Qwen2-7B R)	78.5	4.0	59.0	3.3	71.8	3.9	57.6	3.5
UniVid (Qwen2-7B E)	77.4	3.9	58.4	3.2	72.2	3.9	57.3	3.5
<b>UniVid (Ours)</b>	<b>80.1</b>	<b>4.2</b>	<b>61.4</b>	<b>3.4</b>	<b>75.0</b>	<b>4.1</b>	<b>58.8</b>	<b>3.6</b>

Table 8: Ablation on Evaluator/Reflector Model Size (Hereafter, we use E to denote the Evaluator and R to denote the Reflector). Acc. denotes accuracy (%), Score denotes average rating (0–5).

these results demonstrate that UniVid benefits from a hybrid encoding design for generation, while semantic encoders dominate in understanding.

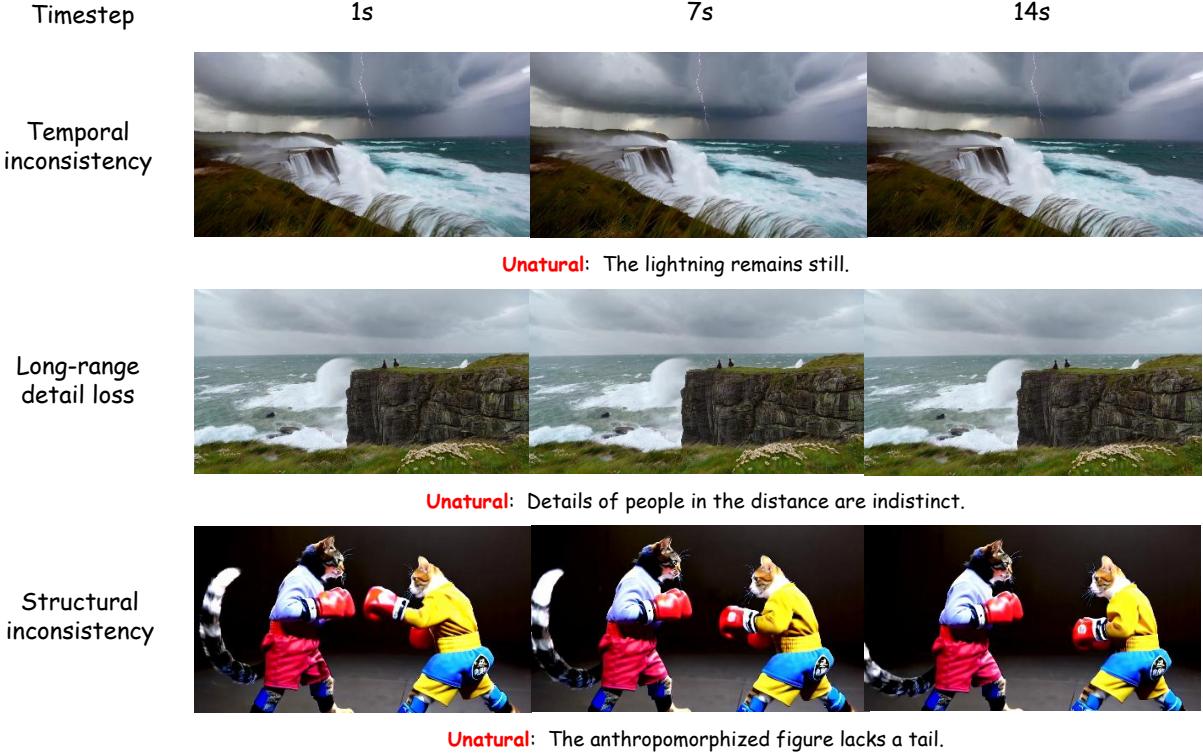


Figure 9: Categorized Failure Modes in Video Generation.

Methods	MSVD-QA		MSRVT-QA		TGIF-QA		ActivityNet-QA	
	Acc↑	Score↑	Acc↑	Score↑	Acc↑	Score↑	Acc↑	Score↑
UniVid (VAE Encoder)	49.1	3.2	44.7	2.7	52.9	2.9	38.5	3.0
UniVid (VAE & ViT Encoder)	78.6	4.1	56.9	3.2	72.8	3.9	57.1	3.5
<b>UniVid (Ours, ViT only)</b>	<b>80.1</b>	<b>4.2</b>	<b>61.4</b>	<b>3.4</b>	<b>75.0</b>	<b>4.1</b>	<b>58.8</b>	<b>3.6</b>

Table 9: Ablation study of the understanding branch of UniVid to verify the effectiveness of encoder setting. Acc. denotes accuracy (%), Score denotes average rating (0–5). w/o means “without”. Best results are **bold**.

Model	MMLU↑	MMMU↑	MME↑	MME(S&M)↑	MMBench↑	MLVU↑
Frame Num	32	32	64	64	64	64
<b>Und.Only Models</b>						
Qwen2-VL-7B (Wang et al., 2024a)	21.02	41.26	59.7	72.1	1.45	62.34
Qwen2.5-VL-7B (Bai et al., 2025)	24.17	47.44	62.8	75.9	1.49	62.052
Qwen3-VL-8B (Yang et al., 2025)	<b>71.6</b>	<b>69.9</b>	<b>71.4</b>	<b>89.7</b>	<b>2.55</b>	<b>78.1</b>
LLaVA-Video-7B (Lin et al., 2024)	15.89	36.11	63.7	78.1	1.6	67.66
MiniCPM-V-2.6-7B (Yu et al., 2025)	–	–	59.7	74.7	1.7	52.82
InternVL2_5-8B (Chen et al., 2024b)	52.47	43	63.7	77	1.68	63.94
InternVL3-8B (Zhu et al., 2025)	<u>57.71</u>	47.97	66	<u>79.5</u>	1.69	67.964
<b>Unified Models</b>						
Omni-Video-7B (Tan et al., 2025)	41.28	51.62	59.43	71.43	1.59	67.24
Emu3-8B (Wang et al., 2024b)	40.33	49.73	60.98	68.76	1.54	66.77
Show-02-7B (Xie et al., 2025c)	45.77	<u>53.99</u>	<u>66.87</u>	76.62	1.67	68.92
<b>Ours-7B</b>	49.88	<u>59.41</u>	62.68	78.4	<u>1.85</u>	<u>70.77</u>

Table 10: Comparison of Und.Only and Unified Models across major video benchmarks (Hendrycks et al., 2021a,b; Fu et al., 2024a; Fang et al., 2024; Zhou et al., 2024b; Yue et al., 2024). The best results are highlighted in **bold**, and the second-best are underlined. Notably, all methods are evaluated under a unified frame-setting for fair comparison and our method can utilize **at most** unified setting frames.