UNIVID: THE OPEN-SOURCE UNIFIED VIDEO MODEL

Anonymous authors
Paper under double-blind review

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, achieveing competitive results across various tasks.

ABSTRACT

Unified video modeling combining generation and understanding capabilities is increasingly important, yet faces two key challenges: maintaining semantic faithfulness during flow-based generation due to text-visual token imbalance and the suboptimality of uniform cross-modal attention across the flow trajectory, and efficiently extending image-centric MLLMs to video without costly retraining. We present **UniVid**, a unified architecture that couples an MLLM with a diffusion decoder through a lightweight adapter, enabling both video understanding and generation. We introduce *Temperature Modality Alignment* to improve prompt adherence and *Pyramid Reflection* for efficient temporal reasoning via dynamic keyframe selection. Extensive experiments on standard benchmarks demonstrate the state-of-the-art performance of our unified video model, achieving a **2.2%** improvement on VBench-Long total score compared to the previous SOTA method EasyAnimateV5.1, and **1.0%** and **3.3%** accuracy gains on MSVD-QA and ActivityNet-QA, respectively, compared with the best prior 7B baselines.

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 converged 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) and Chameleon (Lu et al., 2023). The second is a hybrid diffusion–AR route: a multimodal AR backbone governs understanding and

control signals, while a diffusion video decoder renders high-fidelity frames from high-level visual tokens; recent works such as Transfusion (Zhou et al., 2024a) and Show-O (Xie et al., 2025a) follow this pattern. In this work, we adopt the hybrid route to retain high-quality rendering while leveraging an MLLM for semantic control and interpretability.

However, even within this hybrid setting, unified video modeling faces two key challenges. First, maintaining semantically faithful conditioning in video diffusion across the flow trajectory is difficult. Text prompts convey high-level intent but under-specify pixel-aligned details; in MM-DiT-style Esser et al. (2024) models, the cross-modal signal can be diluted by the numerical imbalance between few text tokens and many visual tokens, and the role of guidance is inherently timestep-dependent—early steps benefit more from strong semantic constraints, whereas later steps benefit from visual detail refinement, yielding prompt–video drift that worsens with longer, higher-resolution clips. Second, extending image-centric MLLMs to video faces two key challenges: the computational cost of temporal modeling (dedicated encoders, long-context handling, large-scale training) that risks destabilizing existing capabilities, and the mismatch between video's vast temporal information and the typically small subset relevant to any question. Traditional approaches either process all frames uniformly, causing inefficiency and noise, or use fixed sampling that may miss critical evidence. Furthermore, different question types demand different strategies—static questions need distinctive keyframes while dynamic questions require understanding temporal transitions.

To address these challenges, our motivation is twofold. First, on the generation side, we leverage multimodal understanding to construct structure-aware tokens in the language space that encode both global semantics and localized cues; these tokens are used as faithful semantic conditioning for a diffusion video decoder, and we schedule cross-modal attention over flow steps so that early integration emphasizes textual intent while later steps emphasize visual refinement. Second, on the understanding side, we develop an adaptive evidence selection approach that extends image-centric MLLMs to video without substantial architectural changes. This requires a mechanism that can iteratively explore and refine the evidence set based on feedback, balance exploration of new frames with exploitation of current evidence, and learn from failure signals to improve future selections. This suggests a sequential decision-making framework, but rather than traditional parameter updates, we implement a form of verbal test-time reinforcement learning. We develop Pyramid Reflection, where policy improvement occurs through natural language refinement—the Reflector verbally adjusts search queries based on feedback, while SigLIP2 (Tschannen et al., 2025) enables query-driven keyframe selection that iteratively expands or prunes the evidence set.

Hence, we propose **UniVid**, a unified architecture that couples a multimodal LLM with a diffusion video decoder via a lightweight conditioning adapter: the LLM ingests text and salient visual evidence and outputs rich semantic understandable tokens that both support reasoning and condition the decoder for text/image-to-video generation. To stabilize guidance in MM-DiT (Esser et al., 2024), we introduce *Temperature Modality Alignment*, a timestep-aware, temperature-adjusted cross-modal attention schedule that emphasizes semantic intent early and visual refinement late, mitigating text suppression and improving prompt faithfulness. To enable efficient understanding with minimal change, we introduce *Pyramid Reflection*, which implements sequential decision-making through SigLIP2-based keyframe selection and an Actor–Evaluator–Reflector loop that verbally adjusts search strategies while progressively expanding or pruning context. Through extensive evaluation on standard benchmarks, we validate the superior capability of our unified approach, which consistently outperforms 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**, a unified paradigm that couples an MLLM with a diffusion video decoder via a lightweight conditioning adapter; the MLLM produces rich, understandable semantic tokens that both support reasoning and condition text/image-to-video generation.
- We propose *Temperature Modality Alignment*, a timestep-aware, temperature-adjusted cross-modal attention schedule in MM-DiT that strengthens early semantic guidance and later shifts emphasis to visual refinement; we further develop *Pyramid Reflection* with SigLIP2-based keyframe selection to enable efficient temporal reasoning with minimal architectural change and training.
- We conduct comprehensive experiments on MSVD-QA (Piergiovanni et al., 2022), MSRVTT-QA (Piergiovanni et al., 2022), TGIF-QA (Jang et al., 2017), and ActivityNet-QA (Yu et al.,

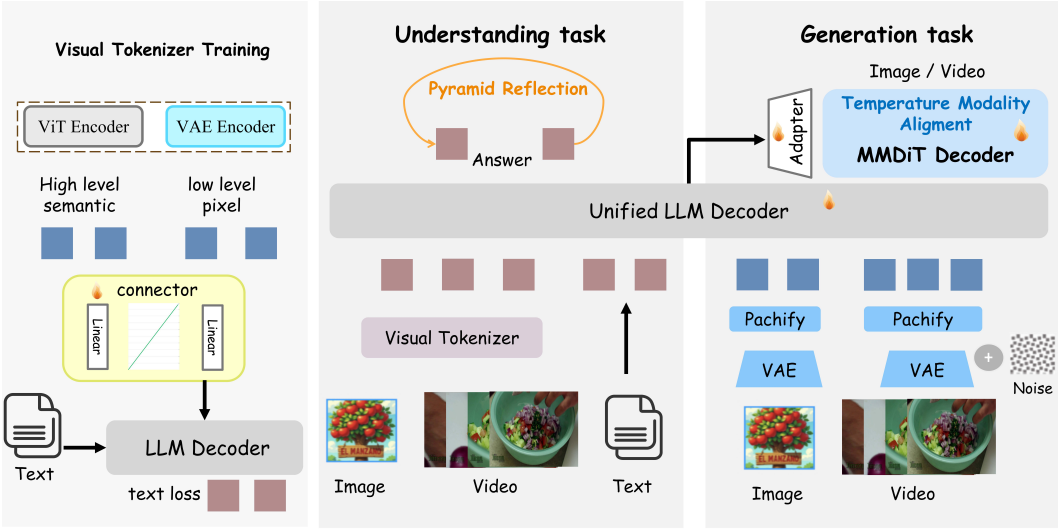


Figure 2: Overall architecture of our proposed UniVid for unified video understanding and generation. Notably, for the understanding task, we adopt only the ViT encoder to achieve a better efficiency–accuracy trade-off.

2018) for understanding, and on VBench for generation, demonstrating competitive performance and efficiency. Ablations verify the contribution of each component.

2 Related Work

Video generation. Video generation has seen remarkable advancements with the rise of diffusion models and generative adversarial networks tailored for temporal data. Recent diffusion or flow based frameworks, such as Video Diffusion Models (Ho et al., 2022b), Imagen Video (Ho et al., 2022a), VideoCrafter2 (Chen et al., 2024a) and Stable Video Diffusion (Blattmann et al., 2023b), have produced high-fidelity clips with improved temporal consistency, enabling applications in creative generation and simulation (Liu et al., 2025; Shi et al., 2025). Latent diffusion techniques (Blattmann et al., 2023c) further improve efficiency by operating in compressed latent spaces, enabling scalable video generation. In parallel, GAN methods like MoCoGAN (Tulyakov et al., 2018) and StyleGAN-V (Skorokhodov et al., 2022) explore alternative formulations. Despite these advances, maintaining long-term temporal consistency in extended sequences remains challenging, as summarized by recent surveys and analyses (Melnik et al., 2024; Yin et al., 2025).

Video understanding. Recent progress in video understanding has been driven by transformer-based architectures and self-supervised learning paradigms that effectively model spatio-temporal relationships. Methods like MViT (Fan et al., 2021), Video Swin Transformer (Liu et al., 2022), TimeSformer (Bertasius et al., 2021) and ViViT (Arnab et al., 2021) have advanced the field by capturing long-range dependencies across video frames, achieving strong performance on datasets such as Kinetics-700 (Carreira et al., 2019). Beyond supervised training, self-supervised approaches—including masked modeling (VideoMAE (Tong et al., 2022), MaskFeat (Wei et al., 2022), OmniMAE (Girdhar et al., 2023)) and early contrastive methods (VideoMoCo (Pan et al., 2021))—leverage unlabeled videos to learn robust, transferable representations, reducing dependence on costly annotations and benefiting action recognition and video segmentation.

Unified multimodal models. Unified multimodal modeling has progressed from joint vision–language pretraining to architectures that support both understanding and generation across modalities. Foundational systems like CLIP (Radford et al., 2021) establish large-scale alignment, while BEiT-3 (Wang et al., 2023) and UnifiedMLLM (Li et al., 2024) broaden task coverage. Pushing toward unified generation, Show-o (Xie et al., 2025a) integrates autoregression with discrete diffusion within a single Transformer to support VQA, text-to-image, and various editing tasks. In a complementary direction focused on robustness rather than general any-to-any generation, FLUID (Cuong et al., 2025) uses token-level distillation for cross-modal fusion. Open generalist systems

then aim to unify understanding and generation end-to-end: BAGEL (Deng et al., 2025) offers an open, decoder-only framework with parallel language and diffusion branches trained jointly, achieving competitive results across image-centric tasks, and BLIP3-o (Chen et al., 2025) releases a fully open family where a diffusion transformer is coupled to strong multimodal understanding, yielding unified image understanding and generation. Extending unification from images to video, Omni-Video (Tan et al., 2025) teaches an MLLM to emit continuous visual tokens that are adapted and consumed by a diffusion video decoder, enabling generation, editing, and understanding in one pipeline.

3 The Proposed Method

3.1 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 hierarchical training recipe that first aligns the conditioning between the MLLM and the generator, then finetunes the MLLM 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.

3.2 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 the generation branch, images are encoded using both ViT (Dosovitskiy et al., 2021) for semantic features and VAE (Kingma & Welling, 2019) for pixel-level details. For the understanding branch, only ViT encoding is employed, as video understanding tasks primarily rely on high-level semantic understanding rather than fine-grained pixel details. The encoded visual features are then projected into the textual token space and concatenated with text tokens, allowing the LLM to output unified multimodal representations.

Generation branch. The generation pathway employs a DiT-based model Wan 2.2 (Wang et al., 2025) conditioned on rich semantic representations extracted from MLLM outputs through a lightweight adapter. The system processes video generation in latent space using a 3D VAE (Zhao et al., 2024), with conditioning signals integrated via cross-attention mechanisms.

Understanding branch. For video understanding, multi-frame evidence is encoded by the ViT (Dosovitskiy et al., 2021) and fused with text; the LLM produces an initial textual answer. We then apply Pyramid Reflection, a query-driven, hierarchical loop that iteratively expands or prunes keyframe context via SigLIP2 (Tschannen et al., 2025) selection and refines the frame space via an Actor–Evaluator–Reflector process, yielding the final answer without modifying the backbone.

Conclusively, our generation builds on the MLLM's strong comprehension, while video understanding uses Pyramid Reflection to leverage the MLLM and collaborate with an LLM for efficient and accurate answers.

3.3 Conditional Generation with Temperature Modality Alignment

Given fused tokens from the understanding path, the MLLM output Z_u is mapped to time-indexed conditions by a lightweight adapter g_ϕ:

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

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

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

Inspired by TACA (Lv et al., 2025), we adapt its finding that text is suppressed in MM-DiT (Esser et al., 2024) because (i) the softmax over a much larger pool of visual tokens ($N_{\text{vis}} \gg N_{\text{txt}}$) dilutes

Algorithm 1 Pyramid Reflection as Test-time RL

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

attention mass on text keys, and (ii) conditioning plays different roles across timesteps (early semantics, late detail). We therefore strengthen the visual-to-text path in Wan 2.2 (Wang et al., 2025) with a simple schedule:

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

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

$$\alpha_{\text{txt}}(u) = \begin{cases} 1 + \dfrac{\lambda_{\text{txt}}}{2}\Big(1 + \cos\Big(\dfrac{\pi u}{0.4}\Big)\Big), & u \in [0,\, 0.4], \\ 1, & u \in (0.4,\, 1], \end{cases} \qquad \lambda_{\text{txt}} = 0.3. \tag{3}$$

Thus, text guidance is strongest early and decays to neutral ($\alpha_{\text{txt}} \to 1$) 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:

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

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

$$\alpha_{\text{img}}(u) = \begin{cases} 1, & u \in [0,\, 0.6], \\ 1 + \dfrac{\lambda_{\text{img}}}{2}\left(1 - \cos\left(\dfrac{\pi(u - 0.6)}{0.4}\right)\right), & u \in (0.6,\, 1], \end{cases} \qquad \lambda_{\text{img}} = 0.3. \tag{5}$$

3.4 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 π_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 π_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

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

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	**69.48**	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	**97.93**	99.49	98.99	**70.83**	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
UniVid (Ours)	**85.27**	**86.44**	**80.58**	**98.96**	97.76	**99.88**	**99.25**	61.83	64.21	**73.03**

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)	**97.83**	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	**97.20**	87.04	57.55	**56.57**	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	**24.89**	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
UniVid (Ours)	94.52	**77.45**	94.20	**92.10**	**80.70**	46.66	23.57	**25.91**	**27.60**

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 $[t_1 \to t_k]$ 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.5.

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.

4 Experiments

4.1 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 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 20k samples from the 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

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

Video QA Performance									
Method	**LLM size**	**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	**65.8**	**3.6**	**78.7**	**4.2**	55.5	3.4
UniVid (Ours)	7B	**80.1**	**4.2**	61.4	3.4	75.0	4.1	**58.8**	**3.6**

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.

4.2 Implementation Details

We adopt a three-stage hierarchical training recipe. It initializes UniVid from strong public checkpoints to reduce compute. For generation, we couple the BAGEL-7B (Deng et al., 2025) with Wan 2.2 5B TI2V model (Wang et al., 2025) via a textual adapter and LoRA on DiT (Peebles & Xie, 2023), keeping other weights frozen. For understanding, we tune only the connector and the last two ViT blocks on ActivityNet QA (Yu et al., 2018) with dialog style supervision while the LLM remains frozen. Finally, we co-train both tasks to refine the connector and obtain additive gains. Sequence parallelism enables long high-resolution clips. For details, see Appendix A.2.

For generation, we use a flow-matching ODE sampler with classifier-free guidance and a universal negative prompt. Unless noted, videos are sampled at 1280×704 resolution, 121 frames at 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 $4 \to 8 \to 16$ keyframe schedule. Dynamic questions follow $64 \to 32 \to 16$ 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.4.

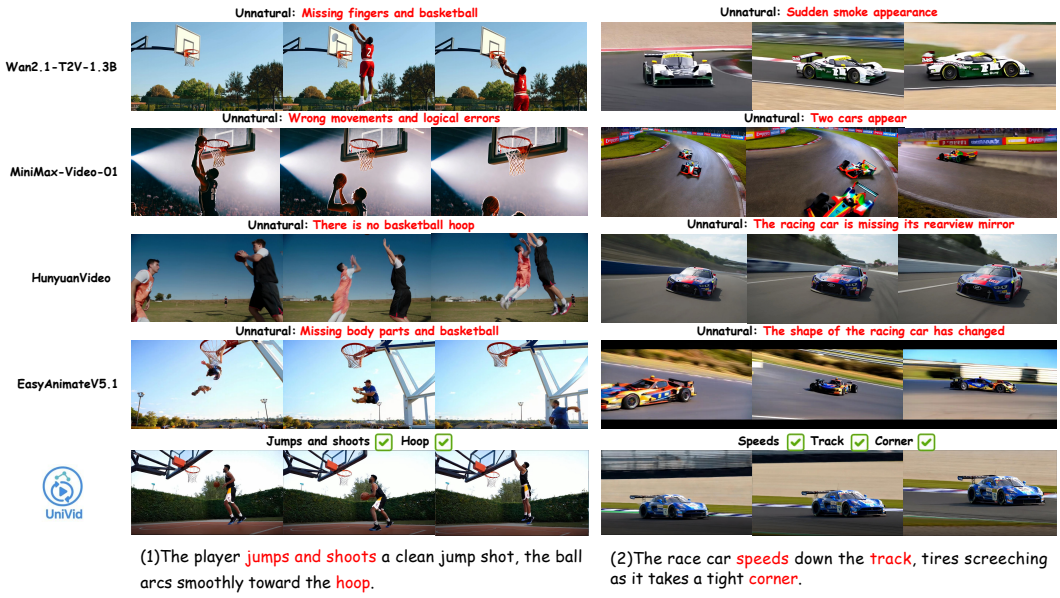


Figure 3: Comparisons with State-of-the-Art Video Generation Models (Wang et al., 2025; MiniMax, 2024; Kong et al., 2024; Fu et al., 2024b).

4.3 Main Results

Generation quantitative results. We evaluate UniVid on the challenging VBench-Long benchmark (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.3.

Generation qualitative results. Fig. 3 compares UniVid with Wan2.1-T2V-1.3B (Wang et al., 2025), MiniMax-Video-01 (MiniMax, 2024), HunyuanVideo (Kong et al., 2024), and EasyAnimateV5.1 (Fu et al., 2024b). Competing models often show missing basketballs or distorted cars, while UniVid generates coherent jump shots and realistic racing scenes with stable dynamics and faithful semantics.

Understanding quantitative evaluation. Across MSVD-QA (Piergiovanni et al., 2022), MSRVTT-QA (Piergiovanni et al., 2022), TGIF-QA (Jang et al., 2017), and ActivityNet-QA (Yu et al., 2018), UniVid sets the 7B-scale state of the art on MSVD-QA and ActivityNet-QA and remains competitive on the other two (Tab. 2), despite a smaller post-training set and no test-time ensembling. Joint finetuning of generation and understanding with Pyramid Reflection strengthens the abilities these datasets emphasize: better action–entity binding and object or attribute grounding

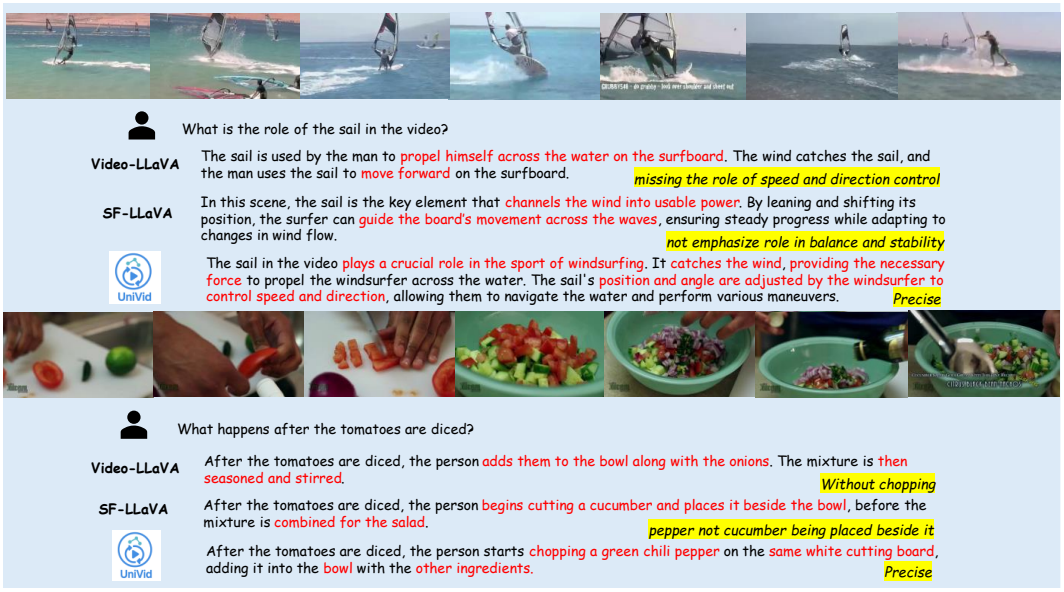


Figure 4: Comparisons of State-of-the-Art Video Understanding Models (Lin et al., 2024; Xu et al., 2024).

in short open-domain clips, stronger temporal reasoning over frame sequences, and more reliable long-range evidence retrieval in untrimmed videos.

As illustrated before, UniVid performs robust multi-frame reasoning with our Pyramid Reflection loop. Starting from a global caption and automatic type detection, the system first produces an initial answer, which is then scored by the evaluator; when evidence is insufficient, the reflector issues a refined, declarative query that re-ranks keyframes toward the true scene. This Pyramid Reflection steers attention from opening credits to the lane shots, yielding a consistent interpretation of roles (in the example of Fig. 8: bowler and nearby teammate/coach) grounded in the visual context rather than spurious cues. The dynamic keyframe schedule reduces the number of inspected frames while maintaining accuracy, demonstrating both evidence tracing and efficiency gains in short-clip understanding. More examples of video understanding can be seen in Appendix A.3.

Understanding qualitative results. We compare UniVid with Video-LLaVA (Lin et al., 2024) and SF-LLaVA (Xu et al., 2024) on video QA; as shown in Fig. 4, baselines often give plausible but incomplete statements. These examples highlight UniVid's stronger action–entity binding, temporal reasoning, and resistance to distractor frames, yielding precise and concise answers. Additionally, we conduct systematic ablation experiments to validate the contributions of UniVid. The results and analyses are provided in the Appendix A.6.

5 Conclusion

We introduced UniVid, a unified video model that couples an MLLM with a diffusion decoder via a lightweight conditioning adapter to both understand and generate videos. Two key mechanisms enable this: Temperature Modality Alignment schedules cross-modal attention across flow steps to preserve prompt faithfulness while refining details, and Pyramid Reflection performs query-driven keyframe selection for efficient temporal reasoning. With these components, UniVid achieves state-of-the-art or competitive results on VBench-Long and multiple video-QA benchmarks while avoiding costly retraining of image-centric backbones. We release UniVid to support research on practical, controllable, and truly unified video intelligence.

REFERENCES

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. GQA: training generalized multi-query transformer models from multi-head checkpoints. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 4895–4901. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.298. URL https://doi.org/10.18653/v1/2023.emnlp-main.298.

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucic, and Cordelia Schmid. Vivit: A video vision transformer. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 6816–6826, 2021.

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, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923, 2025.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pp. 813–824, 2021.

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. Stable video diffusion: Scaling latent video diffusion models to large datasets. *CoRR*, abs/2311.15127, 2023a.

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

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

Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 7310–7320, 2024a.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, Le Xue, Caiming Xiong, and Ran Xu. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *CoRR*, abs/2505.09568, 2025.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *CoRR*, abs/2412.05271, 2024b. doi: 10.48550/ARXIV.2412.05271. URL https://doi.org/10.48550/arXiv.2412.05271.

Zhe Chen, Jiannan Wu, Wenhai 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. Intern VL:

scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 24185–24198. IEEE, 2024c.

Van Duc Cuong, Ta Dinh Tam, Tran Duc Chinh, and Nguyen Thi Hanh. Fluid: Flow-latent unified integration via token distillation for expert specialization in multimodal learning, 2025.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Shi Guang, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *CoRR*, 2025.

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. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=FPnUhsQJ5B.

Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6824–6835, 2021.

Weichen Fan, Chenyang Si, Junhao Song, Zhenyu Yang, Yinan He, Long Zhuo, Ziqi Huang, Ziyue Dong, Jingwen He, Dongwei Pan, et al. Vchitect-2.0: Parallel transformer for scaling up video diffusion models. *arXiv preprint arXiv:2501.08453*, 2025.

Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *arXiv preprint arXiv:2406.14515*, 2024.

Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024a.

Jaskie Fu, Kun-Hao Yeh, Zhaofan Zha, Xinyu Wang, Chenghao Li, Han-Yi Shaw, Chao-Yi Li, and Pin-Yu Chen. Easyanimate: An easy-to-use framework for creating high-quality and controllable videos from a single image. *arXiv preprint arXiv:2403.04416*, 2024b.

Rohit Girdhar, Alaaeldin El-Nouby, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Omnimae: Single model masked pretraining on images and videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10406–10417, 2023.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021a.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021b.

Alex Henry, Prudhvi Raj Dachapally, Shubham Shantaram Pawar, and Yuxuan Chen. Query-key normalization for transformers. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pp. 4246–4253. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.FINDINGS-EMNLP.379. URL https://doi.org/10.18653/v1/2020.findings-emnlp.379.

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. Imagen video: High definition video generation with diffusion models. *CoRR*, 2022a.

Jonathan Ho, Tim Salimans, Alexey A. Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *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*, 2022b.

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

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. TGIF-QA: toward spatio-temporal reasoning in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 1359–1367. IEEE Computer Society, 2017.

Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 13700–13710. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01300.

Wonkyun Kim, Changin Choi, Wonseok Lee, and Wonjong Rhee. An image grid can be worth a video: Zero-shot video question answering using a vlm, 2024. URL https://arxiv.org/abs/2403.18406.

Diederik P. Kingma and Max Welling. An introduction to variational autoencoders. *Found. Trends Mach. Learn.*, 12(4):307–392, 2019. doi: 10.1561/2200000056. URL https://doi.org/10.1561/2200000056.

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

Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *CoRR*, abs/2305.06355, 2023.

Zhaowei Li, Wei Wang, YiQing Cai, Xu Qi, Pengyu Wang, Dong Zhang, Hang Song, Botian Jiang, Zhida Huang, and Tao Wang. Unifiedmllm: Enabling unified representation for multi-modal multi-tasks with large language model. *arXiv preprint arXiv:2408.02503*, 2024.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 5971–5984. Association for Computational Linguistics, 2024.

Akide Liu, Zeyu Zhang, Zhexin Li, Xuehai Bai, Yizeng Han, Jiasheng Tang, Yuanjie Xing, Jichao Wu, Mingyang Yang, Weihua Chen, et al. Fpsattention: Training-aware fp8 and sparsity co-design for fast video diffusion. *arXiv preprint arXiv:2506.04648*, 2025.

Ruyang Liu, Chen Li, Yixiao Ge, Thomas H. Li, Ying Shan, and Ge Li. Bt-adapter: Video conversation is feasible without video instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 13658–13667. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01296.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3202–3211, 2022.

Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. Chameleon: Plug-and-play compositional reasoning with large language models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *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*, 2023.

Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Minghui Qiu, Pengcheng Lu, Tao Wang, and Zhongyu Wei. Valley: Video assistant with large language model enhanced ability. *CoRR*, abs/2306.07207, 2023.

Zhengyao Lv, Tianlin Pan, Chenyang Si, Zhaoxi Chen, Wangmeng Zuo, Ziwei Liu, and Kwan-Yee K. Wong. Rethinking cross-modal interaction in multimodal diffusion transformers, 2025. URL https://arxiv.org/abs/2506.07986.

Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 12585–12602. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.679.

Andrew Melnik, Michal Ljubljanac, Cong Lu, Qi Yan, Weiming Ren, and Helge J. Ritter. Video diffusion models: A survey. *Trans. Mach. Learn. Res.*, 2024.

Lingchen Meng, Jianwei Yang, Rui Tian, Xiyang Dai, Zuxuan Wu, Jianfeng Gao, and Yu-Gang Jiang. Deepstack: Deeply stacking visual tokens is surprisingly simple and effective for lmms. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/29cd7f8331d13ede6dc6d6ef3dfacb70-Abstract-Conference.html.

MiniMax. Minimax video generation api is now available. https://www.minimaxi.com/en/news/video-generation-api, October 2024. Accessed: 2025-07-24.

Tian Pan, Yibing Song, Tianyu Yang, Wenhao Jiang, and Wei Liu. Videomoco: Contrastive video representation learning with temporally adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11205–11214, 2021.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 4172–4182. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00387. URL https://doi.org/10.1109/ICCV51070.2023.00387.

A. J. Piergiovanni, Kairo Morton, Weicheng Kuo, Michael S. Ryoo, and Anelia Angelova. Video question answering with iterative video-text co-tokenization. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXVI*, volume 13696 of *Lecture Notes in Computer Science*, pp. 76–94. Springer, 2022.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Runway. Gen-3 alpha: A new frontier for video generation. Technical report, Runway, July 2024. Accessed: 2025-07-24.

Noam Shazeer. GLU variants improve transformer. *CoRR*, abs/2002.05202, 2020. URL https://arxiv.org/abs/2002.05202.

Jingwei Shi, Zeyu Zhang, Biao Wu, Yanjie Liang, Meng Fang, Ling Chen, and Yang Zhao. Presentagent: Multimodal agent for presentation video generation. *arXiv preprint arXiv:2507.04036*, 2025.

Zhan Shi, Xu Zhou, Xipeng Qiu, and Xiaodan Zhu. Improving image captioning with better use of captions, 2020. URL https://arxiv.org/abs/2006.11807.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *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*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.

Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 3616–3626, 2022.

Jianlin Su, Murtadha H. M. Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/J.NEUCOM.2023.127063. URL https://doi.org/10.1016/j.neucom.2023.127063.

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

Kuaishou Technology. Kling. https://klingai.kuaishou.com/, 2025. Accessed: 2025-07-24.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *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*, 2022.

Michael Tschannen, Alexey A. Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier J. Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *CoRR*, abs/2502.14786, 2025.

Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 1526–1535, 2018.

Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Xiaofeng Meng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *CoRR*, abs/2503.20314, 2025.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *CoRR*, abs/2409.12191, 2024a.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19175–19186, 2023.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Tao Li, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Tiejun Huang, and Zhongyuan Wang. Emu3: Next-token prediction is all you need. *CoRR*, abs/2409.18869, 2024b.

Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan L. Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 14648–14658, 2022.

Wenhao Wu. Freeva: Offline MLLM as training-free video assistant. *CoRR*, abs/2405.07798, 2024. doi: 10.48550/ARXIV.2405.07798. URL https://doi.org/10.48550/arXiv.2405.07798.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*, 2025a.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *CoRR*, abs/2506.15564, 2025b. doi: 10.48550/ARXIV.2506.15564. URL https://doi.org/10.48550/arXiv.2506.15564.

Mingze Xu, Mingfei Gao, Zhe Gan, Hong-You Chen, Zhengfeng Lai, Haiming Gang, Kai Kang, and Afshin Dehghan. Slowfast-llava: A strong training-free baseline for video large language models, 2024. URL https://arxiv.org/abs/2407.15841.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report. *CoRR*, abs/2407.10671, 2024. doi: 10.48550/ARXIV.2407.10671. URL https://doi.org/10.48550/arXiv.2407.10671.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jian Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *CoRR*, abs/2505.09388, 2025. doi: 10.48550/ARXIV.2505.09388. URL https://doi.org/10.48550/arXiv.2505.09388.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *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*, 2022.

Zhiyu Yin, Kehai Chen, Xuefeng Bai, Ruili Jiang, Juntao Li, Hongdong Li, Jin Liu, Yang Xiang, Jun Yu, and Min Zhang. Asurvey: Spatiotemporal consistency in video generation, 2025.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, Bokai Xu, Junbo Cui, Yingjing Xu, Liqing Ruan, Luoyuan Zhang, Hanyu Liu, Jingkun Tang, Hongyuan Liu, Qining Guo, Wenhao Hu, Bingxiang He, Jie Zhou, Jie Cai, Ji Qi, Zonghao Guo, Chi Chen, Guoyang Zeng, Yuxuan Li, Ganqu Cui, Ning Ding, Xu Han, Yuan Yao, Zhiyuan Liu, and Maosong Sun. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. *CoRR*, abs/2509.18154, 2025. doi: 10.48550/ARXIV.2509.18154. URL https://doi.org/10.48550/arXiv.2509.18154.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pp. 9127–9134. AAAI Press, 2018.

Zhen Yuan, Yifei Chen, Shuo Zhao, Wen yi Wang, Ming-Hao Zhang, Zhiping Wang, Le Zhang, Boxi Zhao, Jian Li, Zhi-Yuan Wu, Ming Ding, and Jie Tang. Cogvideox: A general-purpose video generation model. *arXiv preprint arXiv:2406.06511*, 2024.

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 9556–9567. IEEE, 2024. doi: 10.1109/CVPR52733.2024.00913. URL https://doi.org/10.1109/CVPR52733.2024.00913.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 12360–12371, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/1e8a19426224ca89e83cef47f1e7f53b-Abstract.html.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In Yansong Feng and Els Lefever (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023 - System Demonstrations, Singapore, December 6-10, 2023*, pp. 543–553. Association for Computational Linguistics, 2023a.

Renrui Zhang, Jiaming Han, Chris Liu, Aojun Zhou, Pan Lu, Yu Qiao, Hongsheng Li, and Peng Gao. Llama-adapter: Efficient fine-tuning of large language models with zero-initialized attention. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2023b.

Sijie Zhao, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Muyao Niu, Xiaoyu Li, Wenbo Hu, and Ying Shan. CV-VAE: A compatible video VAE for latent generative video models. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15,*

2024, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/1787533e171dcc8549cc2eb5a4840eec-Abstract-Conference.html.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model, 2024a. URL https://arxiv.org/abs/2408.11039.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024b.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *CoRR*, abs/2504.10479, 2025. doi: 10.48550/ARXIV.2504.10479. URL https://doi.org/10.48550/arXiv.2504.10479.

A Appendix

A.1 LLM Use Declaration

Large Language Models (ChatGPT) were used exclusively to improve the clarity and fluency of English writing. They were not involved in research ideation, experimental design, data analysis, or interpretation. The authors take full responsibility for all content.

A.2 Hierarchical Post Training

Initialization. To avoid the prohibitive cost of training a unified video model from scratch, we bootstrap UniVid from strong, publicly available checkpoints and finetune only small subsets of parameters. Our architecture follows the BAGEL (Deng et al., 2025) design framework, adopting its multimodal integration approach with three key components: Qwen2 (Yang et al., 2024) as the LLM backbone with standard architectural choices such as RMSNorm (Zhang & Sennrich, 2019), SwiGLU (Shazeer, 2020), RoPE (Su et al., 2024), GQA (Ainslie et al., 2023), and QK-Norm (Henry et al., 2020) for training stability, SigLIP2-so400m/14 (Tschannen et al., 2025) as the ViT (Dosovitskiy et al., 2021) encoder for visual understanding with NaViT support for native aspect ratios, and a pre-trained FLUX VAE with 8× 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.

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 $(\texttt{video\_id}, q, a)$ 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 preprocessed identically to inference.

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.

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.

Stage III joint training. Finally, we co-train generation and understanding to let the two branches benefit from each other. During joint training, we 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. 10 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

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.

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.3 More examples of Video generation and understanding.

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

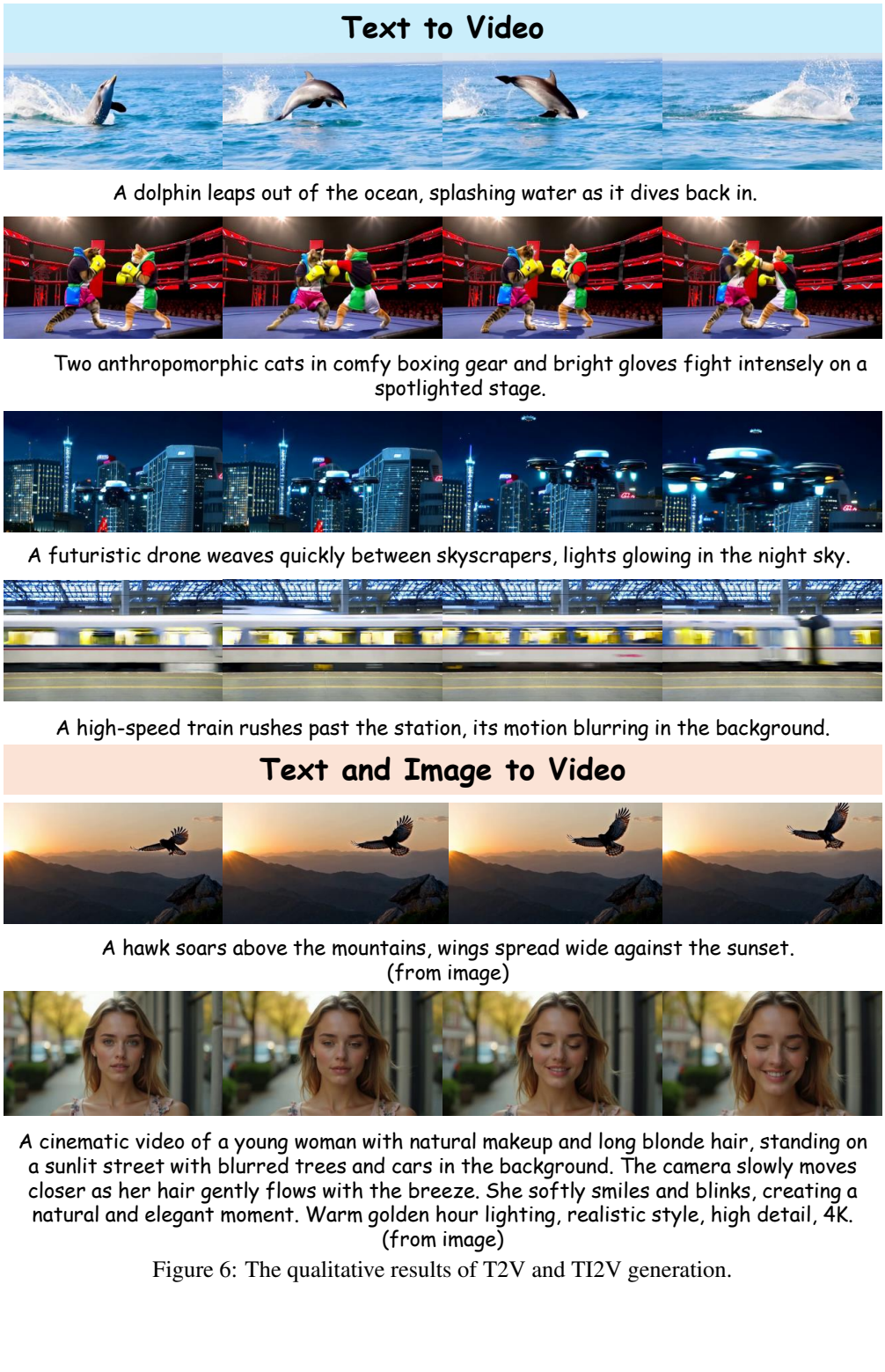


Figure 6: The qualitative results of T2V and TI2V generation.

A.4 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).

1: Question Type Classification Prompt

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.

2: Frame Summarization Prompt

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.

3: Answer Evaluation Prompt

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) `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": "..."`}

4: Reflection Prompt

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 (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
UniVid (Full)	**85.27**	**86.44**	**80.58**	**98.96**	**97.76**	**99.88**	**99.25**	**61.83**	**64.21**	**73.03**

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
UniVid (Full)	**94.52**	**77.45**	**94.20**	**92.10**	**80.70**	**46.66**	**23.57**	**25.91**	**27.60**

Table 4: Ablation study on TMA schedules 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
UniVid (Consine)	**85.27**	**86.44**	**80.58**	**98.96**	**97.76**	**99.88**	**99.25**	**61.83**	**64.21**	**73.03**

Model	Semantic Fidelity								
	Object↑	Multi-Obj↑	Action↑	Color↑	Spatial↑	Scene↑	Appearance↑	Temporal↑	Overall↑
UniVid (w/o TMA)	91.51	75.42	91.53	89.33	77.58	44.61	21.03	23.62	24.13
UniVid (Constant)	92.52	76.81	92.40	90.81	79.13	45.29	22.01	24.19	25.41
UniVid (Step)	91.78	75.81	91.41	89.88	78.13	44.89	21.78	23.54	24.31
UniVid (Linear)	92.80	76.32	92.11	90.98	79.61	45.25	22.56	24.21	26.91
UniVid (Consine)	**94.52**	**77.45**	**94.20**	**92.10**	**80.70**	**46.66**	**23.57**	**25.91**	**27.60**

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.

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."*

6: Global Answer Prompt

A.5 Pyramid Reflection as Test-time RL

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.

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 (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
UniVid (Ours, VAE & ViT Encoder)	**85.27**	**86.44**	**80.58**	**98.96**	**97.76**	**99.88**	**99.25**	**61.83**	**64.21**	**73.03**

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
UniVid (Ours, VAE & ViT Encoder)	**94.52**	**77.45**	**94.20**	**92.10**	**80.70**	**46.66**	**23.57**	**25.91**	**27.60**

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**.

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

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
UniVid (Joint, Stage III)	**94.52**	**77.45**	**94.20**	**92.10**	**80.70**	**46.66**	**23.57**	**25.91**	**27.60**

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:

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

We define a soft retrieval policy over the pool P:

$$\pi(i \mid s) = \frac{\exp\big(\mathrm{sim}(i, s)/\tau\big)}{\sum_{j \in P} \exp\big(\mathrm{sim}(j, s)/\tau\big)}. \tag{7}$$

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

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

$$\Delta_t = \underset{i \in P \setminus S_{\mathrm{sel}}}{\arg\max}{}^{m}\ \mathrm{sim}(i, s_{t-1}), \qquad S_{\mathrm{sel}} \leftarrow S_{\mathrm{sel}} \cup \Delta_t, \quad m = K_t - |S_{\mathrm{sel}}|. \tag{8}$$

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

$$S_{\mathrm{sel}} = \underset{S \subseteq S_{\mathrm{sel}},\, |S| = K_t}{\arg\max} \sum_{i \in S} \Big[\lambda\, \mathrm{sim}(i, s_{t-1}) - (1 - \lambda) \max_{j \in S \setminus \{i\}} \mathrm{sim}(i, j) \Big]. \tag{9}$$

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

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

with

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

Using the likelihood–ratio identity with a baseline b yields

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

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

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

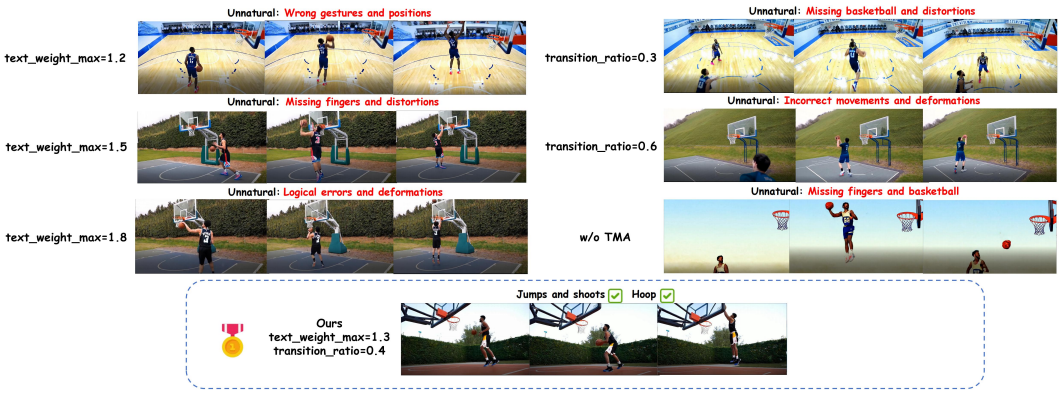
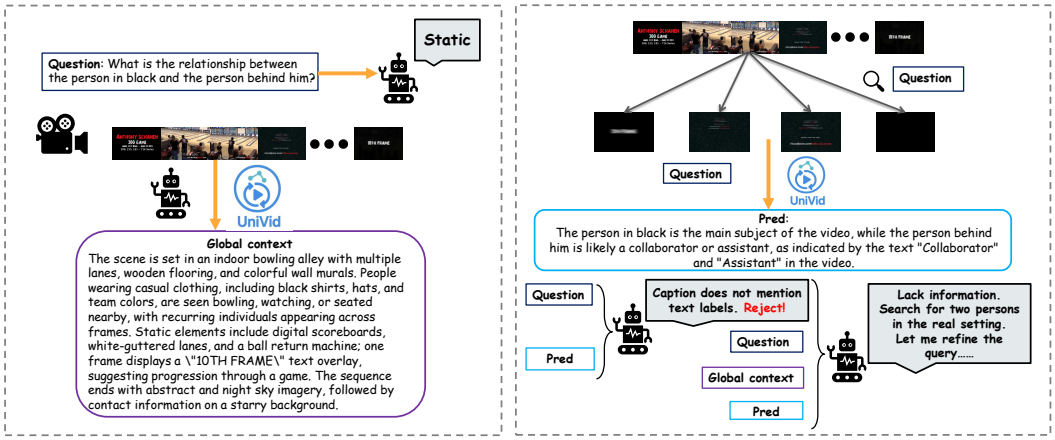
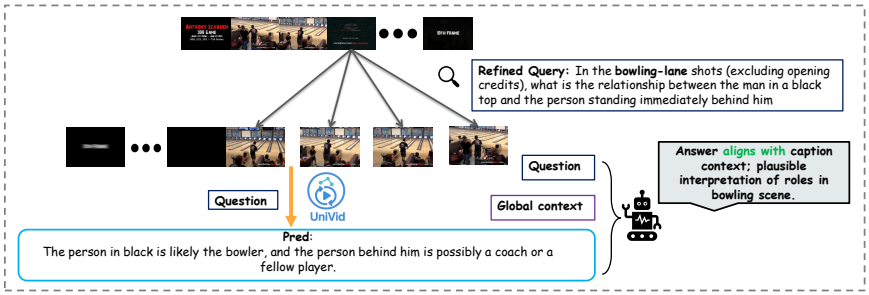


Figure 7: Ablation Study on Temperature Modality Alignment.



Stage 1: Extract the global caption and determine the question type. **Stage 2**: Initial response with subsequent evaluation and reflection.



Stage 3: Further information augmentation with a focus on understanding and evaluation.

Figure 8: The pipeline of the video understanding.

where we use the softmax score function with $g_i(s) := \nabla_s \mathrm{sim}(i,s)$ and $\bar{g}(s) := \mathbb{E}_{j\sim\pi(\cdot|s)} g_j(s)$: $\nabla_s \log \pi(i \mid s) = \tau^{-1}\big(g_i(s) - \bar{g}(s)\big)$, 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 $\mathrm{sim}(i,s)$ 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):

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

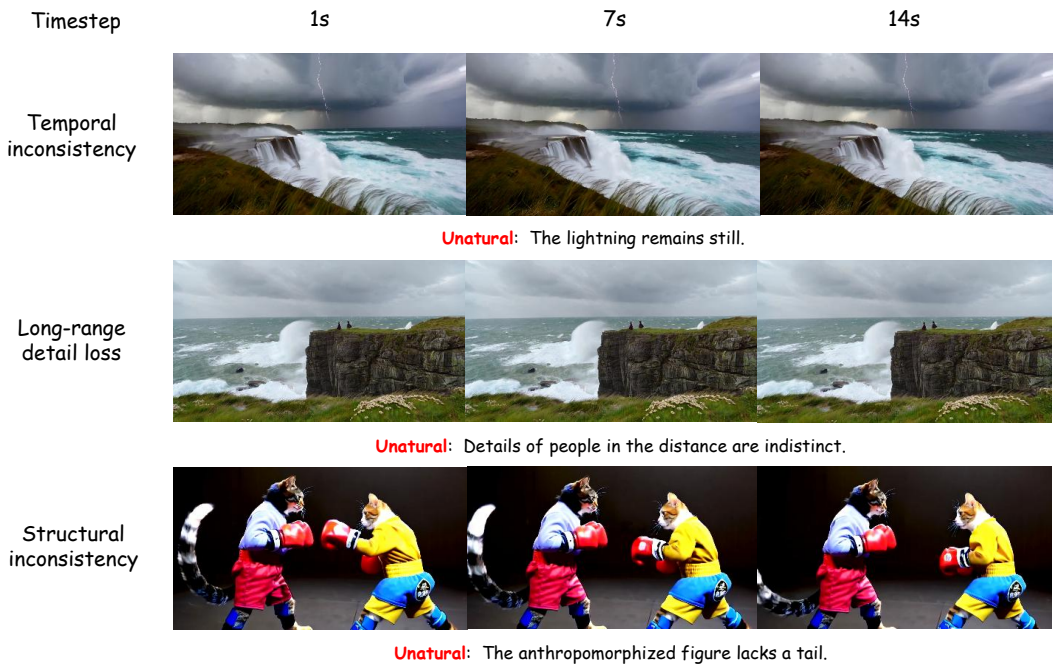


Figure 9: Categorized Failure Modes in Video Generation.

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

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
UniVid (Full)	**80.1**	**4.2**	**61.4**	**3.4**	**75.0**	**4.1**	**58.8**	**3.6**

Since $\partial\,\mathrm{sim}(i,s)/\partial s$ 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 η the expected first-order improvement satisfies

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

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

$$\text{stop at round } r \ \text{ if } \ R_r \geq \tau, \qquad \tau = 0.7. \tag{16}$$

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.6 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 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.

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).

Methods	MSVD-QA		MSRVTT-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
UniVid (Ours)	80.1	4.2	61.4	3.4	75.0	4.1	58.8	3.6

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**.

Methods	MSVD-QA		MSRVTT-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
UniVid (Ours, ViT only)	**80.1**	**4.2**	**61.4**	**3.4**	**75.0**	**4.1**	**58.8**	**3.6**

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.

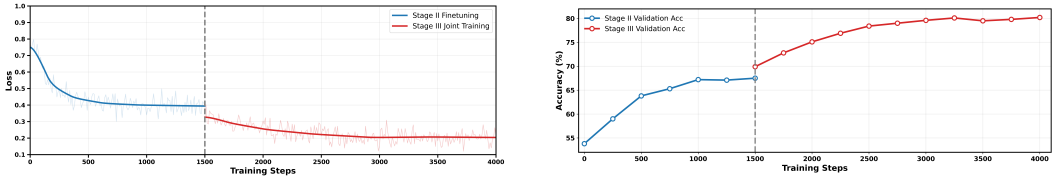
Ablation on video understanding. Tab. 7 compares four variants: a lightweight base model without our training or reasoning additions, a version w/o finetune that removes Stage-II video-QA finetuning, a version w/o Reflection that keeps finetuning but disables the Pyramid Reflection loop, and the Full 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, achieving 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 com-

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.

Model	MMLU↑	MMMU↑	MME↑	MME(S&M)↑	MMBench↑	MLVU↑
Frame Num	**32**	**32**	**64**	**64**	**64**	**64**
Und.Only Models						
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)	**71.6**	**69.9**	**71.4**	**89.7**	**2.55**	**78.1**
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)	57.71	47.97	66	79.5	1.69	67.964
Unified Models						
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-o2-7B (Xie et al., 2025b)	45.77	53.99	66.87	76.62	1.67	68.92
Ours-7B	49.88	59.41	62.68	78.4	1.85	70.77



(a) Training loss curve across dual stages. (b) Validation accuracy during training.

Figure 10: Training loss (left) and validation accuracy (right) curves for UniVid's understanding branch. Notably, red line refers to co-training period in Stage III.

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

A.7 Limitation and Future Work

While UniVid unifies an autoregressive MLLM with a DiT-based video diffusion decoder, the current interaction between the two modules remains relatively shallow. Most MLLM parameters are frozen, and the diffusion branch only receives limited semantic guidance, restricting the potential mutual benefits between understanding and generation. As a consequence, the MLLM gains little improvement in deeper reasoning, and the generation branch relies primarily on data-driven priors rather than task-aware adaptive conditioning.

These limitations manifest in characteristic failure modes during generation, as illustrated in Fig. 9 UniVid can exhibit temporal inconsistencies in long sequences (e.g., static lightning), loss of fine-grained details in distant regions, and occasional structural artifacts such as missing body parts in anthropomorphized characters. These reflect inherent challenges of long-horizon diffusion sampling and the lack of stronger semantic–structural feedback between the two branches.

In future work, we plan to develop deeper bidirectional coupling mechanisms that allow MLLM reasoning signals to shape the diffusion trajectory dynamically, while generated visual feedback reinforces semantic learning. Another promising direction is integrating native dense video encoders to support substantially longer videos with richer motion dynamics. Although these extensions require greater training resources, they offer the potential for more stable long-range generation and more emergent capabilities from cross-modal co-training.