Lumos-Nexus: Efficient Frequency Bridging with Homogeneous Latent Space for Video Unified Models

Jiazheng Xing[*1,4,2], Hangjie Yuan[*,‡2,3,1], Lingling Cai[1], Xinyu Liu[5], Yujie Wei[6], Fei Du[2,3], Tao Feng[7], Hai Ci[4], Jiasheng Tang[2,3], Weihua Chen[†2,3], Fan Wang[2], and Yong Liu[†1]

[1] Zhejiang University, Hangzhou, China,
[2] DAMO Academy, Alibaba Group, Hangzhou, China,
[3] Hupan Lab, Hangzhou, China,
[4] National University of Singapore, Singapore,
[5] Hong Kong University of Science and Technology, Hong Kong, China,
[6] Fudan University, Shanghai, China,
[7] Tsinghua University, Beijing, China
jiazhengxing@zju.edu.cn, kugang.cwh@alibaba-inc.com,
yongliu@iipc.zju.edu.cn
Project Page: https://jiazheng-xing.github.io/nexus-lumos-home/

[*] Equal contribution. [‡] Project lead. [†] Corresponding authors.

Abstract. Connector-based video unified models have demonstrated strong capability in instruction-grounded video synthesis, but integrating a large high-fidelity generator into the unified training loop is computationally prohibitive, limiting achievable visual quality. We therefore propose **Lumos-Nexus**, a training-efficient unified video generation framework that facilitates the development of strong reasoning-driven generation capabilities while significantly enhancing visual fidelity. Lumos-Nexus adopts a two-stage design: 1) During training, only a lightweight generator is aligned with the understanding block to learn to take in reasoning-driven semantic control. 2) During inference, we introduce Unified Progressive Frequency Bridging (UPFB) to progressively hand off generation to a high-capacity pretrained generator in the shared latent space, enabling coarse-to-fine refinement and producing high-fidelity videos without compromising reasoning quality. To fill the gap in reasoning-driven video generation benchmarks, we introduce VR-Bench, which assesses a model's capability to translate inferred intent into coherent and semantically aligned video content. Extensive experiments demonstrate that Lumos-Nexus achieves substantial gains in visual realism and temporal coherence on VBench, while exhibiting strong reasoning-based generative performance on VR-Bench.

Keywords: Video Unified Models · Video Generation · Video Generation Benchmark

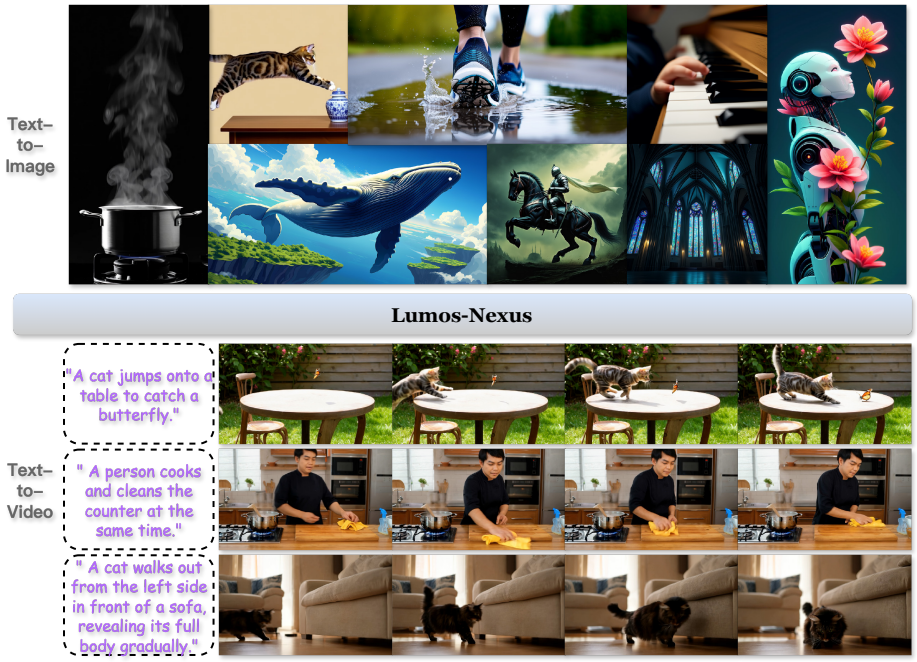

Fig. 1: Visualization of examples generated by **Lumos-Nexus**. Our Lumos-Nexus supports both text-to-image (T2I) and text-to-video (T2V) generation, demonstrating reasoning-aware generation with high-fidelity visual outputs.

1 Introduction

Unified models [1, 4, 7, 9, 10, 25, 27, 33, 41, 42, 52] have emerged as a promising paradigm that integrates multimodal understanding and generative modeling into a unified system, offering strong potential for the two processes to mutually reinforce one another. In particular, it has been shown that the understanding block supplies structured semantic priors to the visual generator, enabling it to interpret complex instructions and produce outputs aligned with coherent logical intent. Extending this paradigm to video modeling is particularly crucial, as video is not a static visual form but a sequence of events unfolding over time. Generating videos requires maintaining temporal consistency, causal progression, and coherent motion, which inherently demands stronger reasoning than image generation. As a result, video unified models [23,30,38,43] must effectively bridge high-level semantic understanding with temporally consistent generation, making the interaction between understanding and generation even more crucial.

From the architectural perspective, video unified models can be broadly divided into the joint-attention [23,38,43] and connector-based [30] models. Joint-attention models enable long-context interaction between multimodal understanding and generation through shared self-attention, offering stronger scala-

bility but requiring substantially more extensive training. In contrast, connector-based models introduce an explicit connector, which can be designed to align the understanding block's representation into the visual generator's condition injection space, thus avoiding joint optimization of both the understanding block and the generation block. However, despite this decoupled design, connector-based video unified models still require prohibitive fine-tuning overhead used for aligning understanding output with the generation input, due to the substantially large-scale diffusion generator (*e.g.*, Wan2.1-14B [34])—making it challenging to simultaneously achieve strong semantic alignment and high visual fidelity in practice.

Given practical computational constraints, we focus on the connector-based paradigm. However, directly fine-tuning a large diffusion generator within this framework remains computationally expensive. To address this, we explore whether a smaller diffusion generator—*homogeneous in latent space with the larger generator*—can be used during training instead. Since fine-tuning in unified models does not alter the latent representation space of the diffusion backbone, maintaining a shared latent space between the two generators establishes a foundation for bridging them during inference. Our key idea is to allow the small generator to learn how to absorb and encode high-level semantic priors from the understanding block within the unified training loop. Then, at inference time, this small generator acts as a semantic initiator, transforming the understanding-derived semantic representation into coherent structural priors that can be seamlessly inherited by the large generator. The large generator, pretrained on extensive video data, subsequently contributes its stronger high-fidelity synthesis ability and further reinforces the execution of reasoning-driven semantics. In this way, we achieve video generation that is both semantically accurate and visually high-quality, without requiring full-scale training of the large generator inside the unified model.

To build a training-efficient and high-quality unified video generation framework, we propose **Lumos-Nexus**, which significantly enhances visual fidelity while strengthening reasoning-driven generative capability, as shown in Fig 1. Lumos-Nexus is structured in two stages, disentangling the acquisition of semantic alignment from the process of high-fidelity video synthesis. (1) *Training stage*. We align only a lightweight diffusion generator with the understanding block so that it learns to transform semantic and reasoning cues into structured generative signals. (2) *Inference stage*. We introduce Unified Progressive Frequency Bridging (UPFB), which progressively transfers the generative responsibility from the lightweight generator to a pretrained high-capacity generator operating in the shared homogeneous latent space. This controlled handoff produces a natural coarse-to-fine refinement process, enabling high-fidelity, temporally coherent videos while further reinforcing reasoning behavior learned during training. Meanwhile, to address the lack of benchmarks for reasoning-driven video generation and to verify that our method preserves the unified model's reasoning ability, we introduce **VR-Bench**, which systematically evaluates the alignment between inferred intent and generated video content across

eight dimensions spanning physical-world reasoning, commonsense reasoning and embodied interactions. Extensive experiments demonstrate that Lumos-Nexus achieves substantial improvements in visual realism and temporal coherence on VBench, while maintaining strong reasoning-based generative performance on VR-Bench. The contribution of our Lumos-Nexus can be summarized as follows:

- We propose Lumos-Nexus, a training-efficient unified video generation framework that aligns reasoning-guided semantics using a lightweight generator during training, and employs Unified Progressive Frequency Bridging (UPFB) at inference to progressively hand off generation to a high-capacity generator, achieving high-fidelity video synthesis while further enhancing reasoning capability.
- We introduce VR-Bench, which evaluates the alignment between inferred intent and generated video content across multiple dimensions in video generation models.
- Extensive experiments show that Lumos-Nexus significantly improves visual realism and temporal coherence on VBench, while maintaining strong reasoning-guided generation on VR-Bench.

2 Related Works

Video Generation Models. Recent progress in video generation has been driven primarily by advances in both autoregressive [8, 13, 15, 17, 20, 21, 32, 37, 45, 50] and diffusion-based [2, 3, 16, 18, 22, 24, 28, 29, 31, 34, 35, 40, 44, 48, 51, 53] modeling paradigms. Autoregressive approaches transform videos into discrete token sequences and generate them step-by-step through large transformers, enabling explicit temporal modeling but typically suffering from high inference latency and accumulated error over long sequences. Diffusion-based video models instead synthesize video by denoising latent representations, and have demonstrated strong temporal smoothness and visual quality. Early diffusion-based methods [3, 16, 35, 40] largely extended 2D U-Net architectures used for text-to-image generation to the video domain, which limited temporal expressiveness. More recent architectures [18, 22, 24, 31, 34, 48, 51, 53] employ diffusion transformers (DiTs) with spatiotemporal attention, improving motion consistency and scalability to high resolutions.

Video Unified Models. Unified models [1, 4, 6, 7, 27, 33, 36, 39, 42, 47, 54] integrate multimodal understanding and visual generation within a single framework, where the two components mutually reinforce one another, and existing architectures are commonly categorized into autoregressive-based [36, 39], diffusion-based [47], and hybrid formulations [1, 4, 7, 27, 33, 42, 54]. Video unified models [5, 23, 30, 38, 43] extend the unified modeling paradigm to the video domain, enabling semantic grounding and logical reasoning within video synthesis. In terms of fusion strategy between understanding and generation, video unified models can be grouped into joint-attention [5, 23, 38, 43] designs, which perform long-context interaction via shared self-attention, and connector-based [30] designs, which inject understanding features into the generator through a dedicated

connector. The former offers stronger scalability but demands heavy training cost, while the latter is more computation-efficient yet still faces difficulty when scaling to large diffusion generators for high-fidelity video synthesis. Considering practical computational constraints, we focus on the connector-based paradigm in this work.

3 Methods

3.1 Preliminary

In connector-based unified models, prior studies [4,26,27,30,33] have shown that extracting world knowledge from pretrained vision–language models (VLMs) used for understanding and injecting it into diffusion transformers can effectively enhance conditional generation. Especially in text-to-video (T2V) generation, when confronted with complex or indirectly specified textual instructions, such unified models can interpret semantic intent, perform logical reasoning, and subsequently synthesize visual content grounded in the inferred reasoning process. For connector-based image or video unified models, a crucial step is injecting the output of the understanding block as a conditioning signal into the generation block, which can be formally expressed as:

$$\mathbf{c}_{\mathcal{U}} = f_{\mathcal{C}}\left(\mathcal{U}\left(x;\theta_{\mathcal{U}}\right);\theta_{\mathcal{C}}\right), \qquad \hat{y}_{\mathcal{G}} \sim p_{\theta_{\mathcal{G}}}(y \mid \mathbf{c}_{\mathcal{U}}), \tag{1}$$

where x denotes the textual instruction input. $\mathcal{U}$ serves as the understanding block parameterized by $\theta_{\mathcal{U}}$ that processes the textual input into a world-knowledge-grounded semantic representation, which is further transformed by the connector $f_{\mathcal{C}}(\,\cdot\,;\theta_{\mathcal{C}})$ into a generator-compatible conditioning signal $\mathbf{c}_{\mathcal{U}}$. Finally, the generator parameterized by $\theta_{\mathcal{G}}$ predicts $\hat{y}_{\mathcal{G}}$ conditioned on $\mathbf{c}_{\mathcal{U}}$. In this conditioning process, the generator must be fine-tuned to effectively utilize the semantic output from the understanding block. The training cost is relatively high in connector-based video unified models compared with image unified models, as the introduction of temporal sequences greatly increases input length and computational complexity, especially with large diffusion generators.

3.2 Overview

While integrating a large generator into the end-to-end understanding-to-generation pipeline is computationally expensive due to the need for large-scale fine-tuning, its capacity to produce high-fidelity visual details remains highly appealing. To balance computational cost and generation quality, we propose Lumos-Nexus, which leverages small and large generators that share a homogeneous latent space within connector-based unified video models, as shown in Fig. 2. During training, only the small generator is used and fine-tuned within the unified model framework, enabling it to effectively learn to incorporate the semantic knowledge from the understanding block at an acceptable cost. Building on the shared homogeneous latent space, we design the Unified Progressive Frequency Bridging

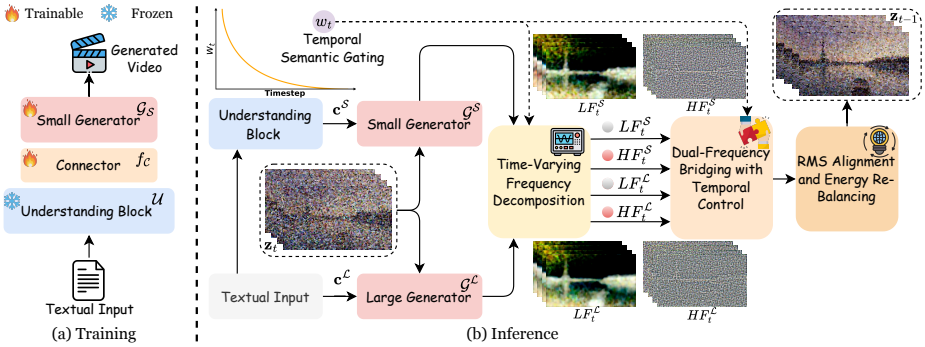

Fig. 2: Overview of Lumos-Nexus. (a): The connector and small generator are fine-tuned within the connector-based video unified model during training. (b): inference performs Unified Progressive Frequency Bridging (UPFB) to combine the small generator's semantic guidance with high-fidelity details from the large generator for high-quality video generation.

(UPFB) strategy: the small generator, trained within the unified model, primarily contributes to coherent semantics and global layout, while the large pretrained generator tends to refine details, enhance visual fidelity, and strengthens the execution of reasoning-driven semantics. This design allows the large generator to inherit unified reasoning ability learned during training while maintaining its high visual quality, achieving superior video generation performance at a fraction of the training cost.

3.3 Unified Progressive Frequency Bridging

Simply mixing or directly bridging the outputs of the two generators often results in unstable semantics, duplicated structures, and texture conflicts, caused by their heterogeneous architectures and frequency biases. To overcome this limitation, we propose Unified Progressive Frequency Bridging (UPFB), which fully exploits the small generator's access to semantic priors from the understanding block and the large generator's capacity for high-fidelity visual synthesis. UPFB dynamically bridges the two generators across temporal and frequency domains, treating the small generator $\mathcal{G}^{\mathcal{S}}$ as a semantic initiator responsible for early-stage layout and global structure, while the large generator $\mathcal{G}^{\mathcal{L}}$ serves as a detail refiner focusing on late-stage texture enhancement. This progressive bridging enables coherent semantic-to-detail transitions without additional training and can be seamlessly integrated into existing connector-based video unified model inference pipelines. Specifically, at sampling step t, both generators perform classifier-free guidance (CFG) on the same intermediate latent $\mathbf{z}_t$, but with distinct conditioning modalities:

$$\mathbf{v}_t^{\mathcal{S}} = \mathbf{v}^{\mathcal{S}}(\mathbf{z}_t, \mathbf{c}^{\mathcal{S}}, t) + s\Big[\mathbf{v}^{\mathcal{S}}(\mathbf{z}_t, \mathbf{c}^{\mathcal{S}}, t) - \mathbf{v}^{\mathcal{S}}(\mathbf{z}_t, \varnothing, t)\Big], \tag{2}$$

$$\mathbf{v}_t^{\mathcal{L}} = \mathbf{v}^{\mathcal{L}}(\mathbf{z}_t, \mathbf{c}^{\mathcal{L}}, t) + s\Big[\mathbf{v}^{\mathcal{L}}(\mathbf{z}_t, \mathbf{c}^{\mathcal{L}}, t) - \mathbf{v}^{\mathcal{L}}(\mathbf{z}_t, \varnothing, t)\Big], \tag{3}$$

where $\mathbf{v}^{\mathcal{S}}(\cdot)$ and $\mathbf{v}^{\mathcal{L}}(\cdot)$ denote the velocity fields predicted by $\mathcal{G}^{\mathcal{S}}$ and $\mathcal{G}^{\mathcal{L}}$, respectively. $\mathbf{c}^{\mathcal{S}}$ represents understanding-enhanced (VLM-aligned embeddings) conditioning provided to the small generator $\mathcal{G}^{\mathcal{S}}$, while $\mathbf{c}^{\mathcal{L}}$ denotes the direct conditioning (textual embeddings) used by the large generator $\mathcal{G}^{\mathcal{L}}$. $\varnothing$ is the unconditional input and s is the classifier-free guidance scale.

Temporal Semantic Gating. Early flow-matching updates require strong global semantic control, while later steps emphasize fine-detail refinement. To facilitate a smooth coarse-to-fine transition between the two generators, we introduce a monotonic temporal weighting strategy.

$$w_t = \frac{1}{2}\Big(1 + \cos\big(\pi(1-\tau_t)^{\gamma_w}\big)\Big), \tag{4}$$

where $\tau_t = \frac{T-1-t}{T-1}$, T denotes the total sampling steps, and γ_w controls the handoff sharpness. We explicitly design the temporal weight w_t to govern the dominance transition between the two generators: a larger w_t biases the update toward the small generator for semantic construction, whereas a smaller w_t gradually shifts the contribution to the large generator for detail refinement.

Time-Varying Frequency Decomposition. To mitigate the frequency bias between the two generators and achieve stable semantic-texture fusion, we explicitly separate low- and high-frequency components in the velocity domain. Specifically, a Gaussian low-pass operator $G_{\sigma_t}(\cdot)$ is applied to the predicted velocity fields $\mathbf{v}$ to separate global layout information from fine-grained details:

$$LF(\mathbf{v}) = G_{\sigma_t}(\mathbf{v}), \quad HF(\mathbf{v}) = \mathbf{v} - LF(\mathbf{v}). \tag{5}$$

where $LF(\mathbf{v})$ represents the low-frequency structure of the velocity field, encoding global semantics and coarse spatial layout, and $HF(\mathbf{v})$ retains residual high-frequency components such as edges and fine textures. The bandwidth parameter σ_t decays over time to encourage a coarse-to-fine transition:

$$\sigma_t = \sigma_{\min} + (\sigma_{\max} - \sigma_{\min})\, w_t, \tag{6}$$

where both $\sigma_{\max}$ and $\sigma_{\min}$ are inference hyperparameters that balance semantic stability and detail fidelity. A large σ_t suppresses high-frequency noise in the early denoising stage, while a small σ_t gradually restores fine structures for photorealistic refinement in later stages.

Dual-Frequency Bridging with Temporal Control. After frequency decoupling, the two generators exhibit complementary strengths across different frequency bands. To effectively bridge their outputs, we perform an asymmetric fusion in the velocity domain. The large generator $\mathcal{G}^{\mathcal{L}}$ provides reliable high-frequency textures, while the small generator $\mathcal{G}^{\mathcal{S}}$ focuses on maintaining semantic coherence. Therefore, we combine their decomposed components:

$$LF_t = w_t\, LF_t^{\mathcal{S}} + (1 - w_t)\, LF_t^{\mathcal{L}}, \tag{7}$$

$$HF_t = w_t\,\gamma_{hf} HF_t^{\mathcal{S}} + (1 - w_t)\,HF_t^{\mathcal{L}}, \tag{8}$$

$$\mathbf{v}_t = LF_t + HF_t, \tag{9}$$

where $LF_t^{\mathcal{S}}$ and $HF_t^{\mathcal{S}}$ denote the low- and high-frequency components of the small generator's guided velocity prediction $\mathbf{v}_t^{\mathcal{S}}$, obtained through the frequency decomposition in Eq. 5. Similarly, $LF_t^{\mathcal{L}}$ and $HF_t^{\mathcal{L}}$ are derived from the large generator's velocity field $\mathbf{v}_t^{\mathcal{L}}$ in the same manner. The coefficient $\gamma_{hf} \in [0.5, 0.8]$ is used to suppress noisy or inconsistent high-frequency components from the small generator. A larger w_t emphasizes semantic accuracy by relying more on the small generator, while a smaller w_t gradually shifts the dominance toward the large generator for fine-detail refinement.

RMS Alignment and Energy Re-Balancing. To further strengthen the bridge between the two generators and ensure stable integration, we normalize their velocity magnitudes across timesteps. This alignment mitigates potential magnitude mismatch and stabilizes the sampling process by applying a pre-fusion adjustment followed by post-fusion re-balancing:

$$\mathbf{v}_t^{\mathcal{L}} \leftarrow \mathbf{v}_t^{\mathcal{L}} \cdot \frac{\mathrm{RMS}(\mathbf{v}_t^{\mathcal{S}})}{\mathrm{RMS}(\mathbf{v}_t^{\mathcal{L}})}, \tag{10}$$

$$\mathbf{v}_t \leftarrow \mathbf{v}_t \cdot \frac{\frac{1}{2}\big(\mathrm{RMS}(\mathbf{v}_t^{\mathcal{S}}) + \mathrm{RMS}(\mathbf{v}_t^{\mathcal{L}})\big)}{\mathrm{RMS}(\mathbf{v}_t)}. \tag{11}$$

Here, $\mathrm{RMS}(\cdot)$ denotes the root-mean-square magnitude, which quantifies the overall signal energy and ensures consistent normalization across timesteps. This energy normalization effectively prevents over-exposure and unstable activations, keeping the fused velocity prediction numerically stable throughout the denoising trajectory.

In summary, UPFB can be seamlessly applied to most connector-based unified video model paradigms, where a lightweight generator learns to inherit prior knowledge from the understanding block, while a high-capacity generator operates outside the training loop to refine visual details and further strengthen the execution of reasoning-driven semantics. By simply inserting our training-free UPFB into the inference process, unified models can retain strong multimodal reasoning capabilities while effectively leveraging the high-fidelity generation power of large generators with no additional training cost. This makes UPFB a practical and scalable solution for deploying high-quality video generation in real-world unified model systems.

3.4 VR-Bench: Benchmark for Reasoning-Driven Video Generation

While recent video generation benchmarks primarily focus on visual fidelity and temporal coherence, they overlook the reasoning dimension—the ability of a model to infer, plan, and act based on semantic intent. To fill this gap, we

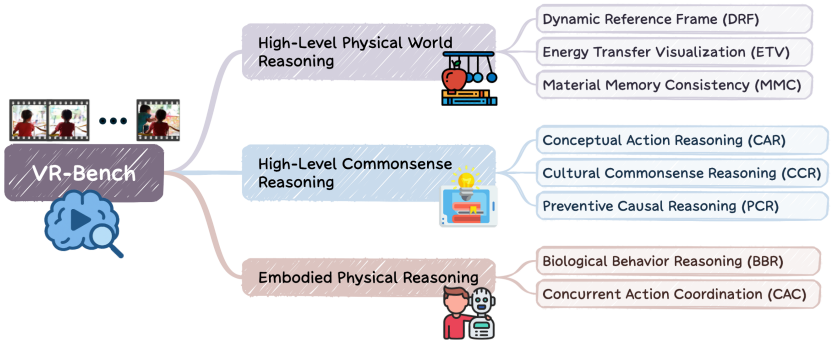

Fig. 3: Overview of VR-Bench.

introduce **VR-Bench**, an eight-dimensional benchmark suite designed to systematically evaluate reasoning-oriented video generation. As illustrated in Fig. 3, The detailed dimensions are described below.

Dynamic Reference Frame (DRF). Evaluates the model's ability to represent *relative motion* and *spatial relations* under changing viewpoints. Higher scores indicate improved *motion coordination*, *spatial consistency*, and *viewpoint invariance*.

Energy Transfer Visualization (ETV). Assesses whether generated motions reflect *momentum propagation* and *energy conservation*. Higher scores reflect stronger adherence to *Newtonian dynamics* and *temporal continuity*.

Material Memory Consistency (MMC). Measures the ability to reproduce realistic *material deformation and recovery*, reflecting physical memory. Higher scores indicate more natural *deformation* and *relaxation dynamics*.

Conceptual Action Reasoning (CAR). Tests understanding of *abstract relational actions* and *coherent state transitions*. Higher scores indicate stronger *intent abstraction* and *action-sequence coherence*.

Cultural Commonsense Reasoning (CCR). Assesses the ability to generate behaviors aligned with social and cultural context. Higher scores indicate stronger *symbol understanding* and *social appropriateness*.

Preventive Causal Reasoning (PCR). Evaluates *anticipatory causal reasoning* by determining whether the model can infer *preventive relations* instead of simple event sequences. Higher scores reflect clearer *causal intent* and *outcome consistency*.

Biological Behavior Reasoning (BBR). Measures *biomechanical plausibility* and *ecological coherence* in living entities. Higher scores indicate better *anatomical feasibility* and *environmental adaptation*.

Concurrent Action Coordination (CAC). Evaluates the representation of *simultaneous multi-action dynamics*. Higher scores indicate stronger *temporal synchronization* and *semantic coherence*.

4 Experiments

4.1 Experimental Settings

Model. We adopt the connector-based video unified model Omni-Video [30] as our baseline. The diffusion generator in Omni-Video is the Wan2.1-T2V-1.3B [34], which has already been fine-tuned within the unified framework. In our Lumos-Nexus, this generation model also serves as the small generator $\mathcal{G}^{\mathcal{S}}$. The large generator $\mathcal{G}^{\mathcal{L}}$ in Lumos-Nexus is Wan2.1-T2V-14B, which operates in the same homogeneous latent space as the small generator $\mathcal{G}^{\mathcal{S}}$.

Evaluations. For the text-to-video (T2V) task, we conduct a reasoning-driven evaluation using VBench [14] and our proposed VR-Bench. While VBench measures overall perceptual quality, VR-Bench targets reasoning alignment—assessing the consistency between inferred intent and generated content. For VR-Bench, we design 216 evaluation cases across eight dimensions, covering three high-level reasoning categories: High-Level Physical World Reasoning, High-Level Commonsense Reasoning, and Embodied Physical Reasoning. Each dimension is scored on a 0–1 scale (equivalently 0–100%), with higher scores indicating stronger reasoning performance. All evaluations are conducted on the Qwen3-VL-30B-A3B-Instruct [46] model.

Implementation Details. In UPFB, we adopt a coarse-to-fine transition by jointly configuring the temporal gating sharpness γ_w to 0.3, the bandwidth schedule with $\sigma_{\min} = 0.35$ and $\sigma_{\max} = 0.70$, and the frequency-decay coefficient γ_{hf} to 0.7, which together provide a stable balance between semantic consistency and detail refinement. Video generation is performed at 480p resolution, with each training clip containing 81 frames (5 seconds at 16 FPS). During inference, we use 50 sampling steps and set the classifier-free guidance (CFG) scale to 5 for both T2I and T2V tasks.

5 Main Results

5.1 Quantitative T2V Results

We evaluate our approach on the VBench-T2V [14] benchmark to assess comprehensive video generation quality. As shown in Tab. 1, our method, Lumos-Nexus, achieves the highest overall score of 84.12, surpassing both conventional video generation models [12,18,34,48–50] and recent video unified models [30,36,38,43]. Our Lumos-Nexus effectively integrates the strengths of both baselines, combining Omni-Video's [30] strong semantic understanding with Wan2.1-14B's [34] high-fidelity detail synthesis. This design alleviates the short-board effect, where Omni-Video's accurate semantic guidance from the understanding block is undermined by the limited generative capacity of its smaller diffusion generator. Through unified frequency bridging, Lumos-Nexus enhances the generator's expressive power, achieving higher semantic alignment (79.10 $\rightarrow$ 80.52) and superior overall video generation quality.

Table 1: Performance comparison on VBench-T2V benchmark. We list partial metrics due to space limits. The **boldfacen** and underline font indicate the highest and the second highest results.

Model	Total↑	Quality	Semantic	Temp. Flick.	Aes. Qua.	Obj. Class	Color	Spat. Rel.	Scene	App. Style	Overall Cons.
Video generation models											
Lumos-1 [50]	78.52	79.52	73.51	98.04	55.77	90.05	82.00	59.10	45.64	22.78	24.10
LTX-Video [12]	80.00	82.30	70.79	99.34	59.81	83.45	81.45	65.43	51.07	21.47	25.19
CogVideoX1.5-5B [48]	82.17	82.78	79.76	98.88	62.79	87.47	87.55	80.25	52.91	**24.89**	27.30
HunyuanVideo [18]	83.43	85.07	76.88	99.39	60.28	83.48	89.79	72.13	54.46	22.21	26.95
CausVid [49]	83.88	85.21	78.57	96.89	64.70	92.80	80.34	64.77	55.74	24.18	**27.53**
Wan2.1-1.3B [34]	83.31	85.23	75.65	99.55	65.46	88.81	89.20	73.04	41.96	21.81	25.50
Wan2.1-14B [34]	83.69	**85.59**	76.11	99.46	66.07	86.28	88.59	75.89	45.75	22.64	25.91
Video unified models											
Emu3 [36]	80.96	-	-	-	59.64	86.17	-	68.73	37.11	20.92	-
Show-o2 [43]	81.34	82.10	78.31	97.68	65.15	94.81	80.89	62.61	**57.67**	23.29	27.00
UniVideo [38]	82.58	-	-	-	-	-	-	-	-	-	-
Omni-Video [30]	83.82	85.00	79.10	99.67	66.37	94.56	90.61	73.82	46.80	23.20	26.95
Lumos-Nexus	**84.12**	85.03	**80.52**	**99.70**	**67.29**	**96.20**	**91.46**	**76.15**	52.47	23.49	27.23

Table 2: Performance comparison on VR-Bench. The eight metrics are grouped into three reasoning categories: High-Level Physical World Reasoning (HL-Phys.), High-Level Commonsense Reasoning (HL-Comm.), and Embodied Physical Reasoning (Emb.-Phys.). Lumos-Nexus* indicates the variant constructed by replacing the large generator with Wan2.2-T2V-A14B. **Bold** and underline indicate the best and the second-best results, respectively.

Model	Total↑	HL-Phys.	HL-Comm.	Emb.-Phys.	DRF	ETV	MMC	CAR	CCR	PCR	BBR	CAC
Closed-source models												
Veo 3.1 [11]	93.95	94.25	95.37	91.37	95.80	93.60	93.33	97.38	93.52	95.20	83.85	98.89
Kling 2.6 [19]	91.13	92.96	88.71	92.01	99.29	89.60	90.00	86.26	91.48	88.40	88.46	95.56
Open-source models												
CogVideoX1.5-5B [48]	66.27	71.92	67.03	56.67	93.70	60.40	61.67	66.82	76.67	57.60	69.63	43.70
HunyuanVideo [18]	75.38	76.79	69.73	81.39	94.81	60.00	75.56	73.47	78.64	57.07	**79.26**	83.52
Wan2.1-1.3B [34]	77.00	79.31	73.60	78.64	95.56	66.13	76.25	75.10	78.89	66.80	73.83	83.46
Wan2.1-14B [34]	78.23	**80.34**	75.96	78.46	96.27	66.73	**78.03**	**77.58**	82.96	67.33	76.79	80.12
Omni-Video [30]	72.78	72.39	72.79	73.33	94.94	52.93	69.31	72.09	83.21	63.07	74.07	72.59
Lumos-Nexus	**79.28**	79.49	**77.57**	**81.54**	**96.54**	**66.80**	75.14	76.28	**84.81**	**71.60**	78.89	**84.20**
Wan2.2-5B [34]	75.35	79.65	70.66	75.93	94.81	60.80	**83.33**	76.54	81.85	53.60	**80.74**	71.11
Wan2.2-A14B [34]	80.98	**83.14**	78.86	80.93	97.78	70.40	81.25	**88.46**	83.70	64.40	75.56	86.30
Lumos-Nexus*	**81.90**	83.02	**79.43**	**83.93**	**97.92**	**70.43**	80.71	86.23	**84.87**	**67.19**	78.87	**88.99**

We assess our method on the VR-Bench to measure reasoning-oriented video generation performance. In addition to our approach, we benchmark some strong closed-source models (including Veo 3.1 [11] and Kling 2.6 [19]) as well as a broad range of open-source models under the same evaluation protocol. As shown in Tab. 2, within the Wan2.1-level generation models (Rows 5–10), Lumos-Nexus achieves the highest overall score of 79.28. It attains the best results in HL-Comm. (77.57) and Emb.-Phys. (81.54), demonstrating superior spatial coherence and biomechanical realism. Moreover, Lumos-Nexus effectively alleviates the short-board effect observed in Omni-Video [30], where accurate semantic guidance from the understanding block does not always translate into correct execution in generated videos, due to the inherent limitations of the diffusion generator's performance. This limitation also accounts for why Omni-Video un-

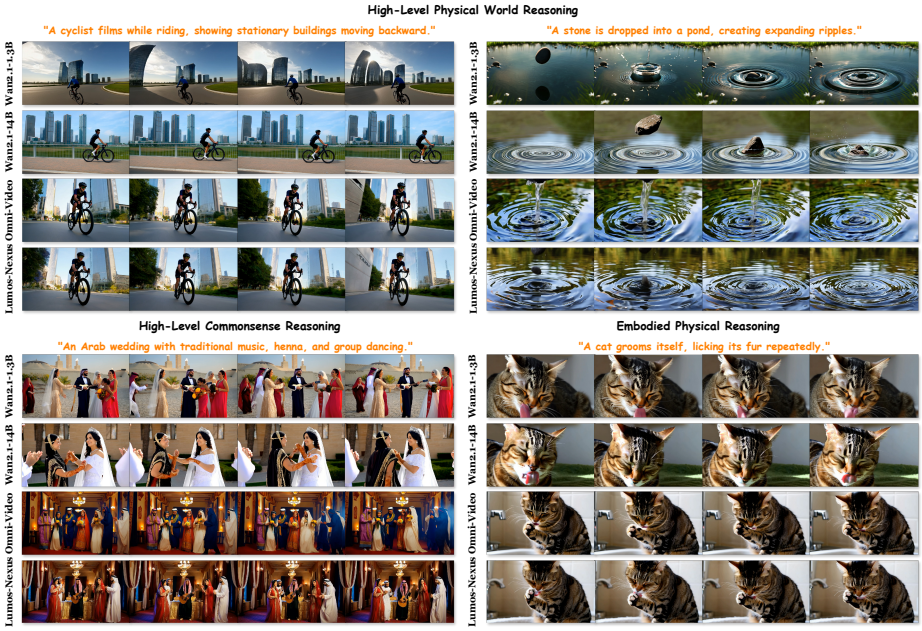

Fig. 4: VR-Bench T2V qualitative comparison across three reasoning dimensions: High-Level Physical World Reasoning, High-Level Commonsense Reasoning, and Embodied Physical Reasoning.

derperforms compared with conventional video generators like Wan2.1-T2V-1.3B [34] across multiple dimensions (*e.g.*, DRF, ETV, and CAC). In contrast, our method consistently delivers superior performance on these metrics. To further demonstrate the extensibility of our framework and evaluate its effectiveness on stronger open-source video generation models in VR-Bench, we conduct additional experiments on the Wan2.2 series (Rows 11–13). Since Wan2.2-T2V-A14B [34] shares a homogeneous latent space with the smaller generator used in Lumos-Nexus, we construct Lumos-Nexus* by replacing the large generator with Wan2.2-T2V-A14B. As shown in Tab. 2, Lumos-Nexus* achieves an overall score of 81.90, outperforming Wan2.2-TI2V-5B (75.35) and Wan2.2-T2V-A14B (80.98). It further improves HL-Comm. to 79.43 and Emb.-Phys. to 83.93, while consistently achieving superior scores across multiple metrics (*e.g.*, DRF 97.92, ETV 70.43, CCR 84.87, CAC 88.99). These results demonstrate that our framework improves reasoning performance and generalizes well across different video generation architectures.

5.2 Qualitative T2V Comparison

We conduct qualitative comparisons on the three reasoning dimensions of VR-Bench, as illustrated in Fig.4. For High-Level Physical World Reasoning, Wan-T2V [34] models exhibit limited ability to capture fine-grained physical regu-

larities, such as the relative motion of stationary buildings when filming from a moving bicycle (top-left case) or the ripple propagation after a stone drops into water (top-right case). Although Omni-Video [30] demonstrates partial awareness of such physical dynamics, its visual fidelity is limited by the capacity of its diffusion generator, even though its understanding block may provide accurate semantic priors. In contrast, our Lumos-Nexus produces significantly more realistic and physically consistent results, adopting the same semantic priors from the understanding block as Omni-Video. Similar observations are found in High-Level Commonsense Reasoning and Embodied Physical Reasoning. Omni-Video struggles to express nuanced semantics like musical rhythm in the wedding scene (bottom-left case) and accurate tongue–paw contact in the cat grooming example (bottom-right case). Overall, Lumos-Nexus effectively bridges the gap between Wan-2.1-14B's limited reasoning capability and Omni-Video's constrained generative capacity, achieving balanced improvements in high-quality video synthesis.

6 Ablation Study

6.1 Impact of varying γ_w on w_t

We ablate the effect of the temporal transition sharpness γ_w in UPFB, as shown in Tab. 3. As illustrated in Fig. 5, γ_w critically controls how quickly the small generator hands off semantic dominance to the large generator: when γ_w is larger (*e.g.*, $\gamma_w = 0.5$), the small generator remains influential for a longer portion of the sampling trajectory, causing the generated video to inherit more of Omni-Video's coarse layout and camera motion patterns; conversely, smaller γ_w values (*e.g.*, $\gamma_w = 0.2$) accelerate the transition toward the large generator, leading to richer textures, stronger high-frequency details, and more vivid object appearance. Quantitatively, a moderate value ($\gamma_w = 0.3$) yields the best overall performance across both VBench and VR-Bench: on VBench, $\gamma_w = 0.3$ achieves the highest total score (84.12) and the strongest semantic alignment (80.52), indicating that an appropriately smooth handoff between generators benefits coarse-to-fine semantic grounding; on VR-Bench, $\gamma_w = 0.3$ also delivers the best overall reasoning performance (79.28), particularly improving High-Level Commonsense reasoning (77.57) and Embodied Physical Reasoning (81.54). Overall, this progression shows that γ_w effectively modulates the semantic-to-detail transition—larger values preserve baseline structural behavior, whereas smaller values enable stronger high-fidelity refinement—while extremely small or large γ_w values weaken the balance between semantic guidance and detail refinement, confirming the necessity of a controlled, progressive transition.

6.2 Effect of $\sigma_{\min}$ and $\sigma_{\max}$ in UPFB

We investigate how different bandwidth schedules $(\sigma_{\min}, \sigma_{\max})$ influence the coarse-to-fine transition of UPFB. As shown in Tab. 4, a moderate setting $(\sigma_{\min} = 0.35, \sigma_{\max} = 0.70)$ achieves the best overall performance on VBench,

Table 3: Performance comparison with different γ_w on w_t.

Method	VBench			VR-Bench			
	Total↑	Quality	Semantic	Total↑	HL-Phys.	HL-Comm.	Emb-Phys.
$\gamma_w = 0.2$	84.08	85.09	80.04	79.09	**80.11**	76.57	79.63
$\gamma_w = 0.3$	**84.12**	85.03	**80.52**	**79.28**	79.49	**77.57**	**81.54**
$\gamma_w = 0.4$	84.02	85.12	79.63	76.49	75.77	75.70	78.77
$\gamma_w = 0.5$	84.05	**85.24**	79.29	75.10	73.07	75.89	76.98

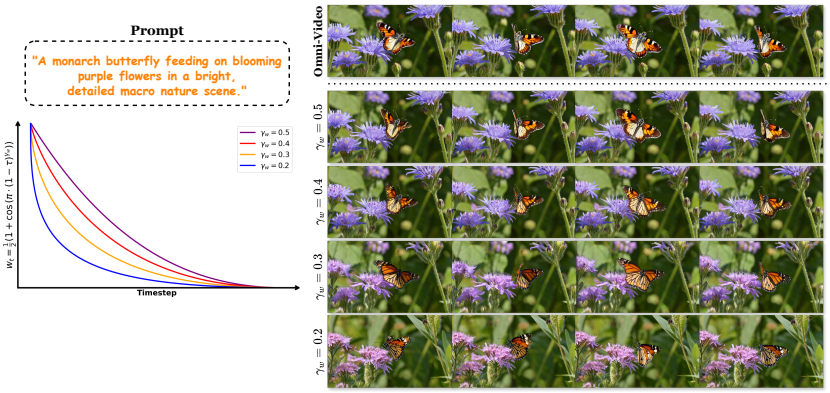

Fig. 5: Visualized qualitative comparison under varying γ_w on w_t.

Table 4: Performance comparison with different $(\sigma_{\min}, \sigma_{\max})$ settings in the bandwidth schedule.

Method	VBench		
	Total↑	Quality	Semantic
$\sigma_{\min} = 0.05$, $\sigma_{\max} = 0.10$	83.99	84.90	80.34
$\sigma_{\min} = 0.35$, $\sigma_{\max} = 0.70$	**84.12**	**85.03**	**80.52**
$\sigma_{\min} = 1.00$, $\sigma_{\max} = 2.00$	83.56	84.86	78.36

Table 5: Performance comparison with and without RMS alignment and energy re-balancing in UPFB.

Method	VBench		
	Total↑	Quality	Semantic
w/o RMS	84.07	84.98	80.43
w/ RMS	**84.12**	**85.03**	**80.52**

yielding the highest total score (84.12). This indicates that a balanced frequency range helps maintain stable semantic grounding while progressively introducing high-frequency details. When the bandwidth is too narrow (0.05, 0.10) or too wide (1.00, 2.00), performance drops in both semantic and perceptual metrics. Overly small values restrict the semantic foundation of the generation process, while excessively large values introduce unstable high-frequency components, leading to weaker coherence.

6.3 Ablation of RMS Alignment and Energy Re-Balancing in UPFB

To evaluate the contribution of RMS alignment and energy re-balancing in UPFB, we compare performance with and without this component, as shown

in Tab. 5. Removing RMS normalization slightly destabilizes the fusion of velocity fields from the two generators, leading to degraded semantic grounding and visual coherence. With RMS alignment enabled, UPFB achieves consistent improvements across all VBench metrics, increasing the total score from 84.07 to **84.12**, the quality score from 84.98 to 85.03, and the semantic score from 80.43 to 80.52. These gains confirm that RMS alignment provides a more stable integration of multi-frequency components by preventing magnitude mismatch and ensuring smoother denoising dynamics.

7 Conclusion

In this work, we introduce Lumos-Nexus, a training-efficient unified video generation framework that preserves strong reasoning-driven generative capability while substantially enhancing visual fidelity. By aligning only a lightweight generator with the understanding block during training and applying Unified Progressive Frequency Bridging (UPFB) to progressively transition generation to a high-capacity pretrained generator during inference, Lumos-Nexus enables coherent coarse-to-fine video synthesis at low training cost. To evaluate reasoning-oriented video generation, we further propose VR-Bench, an eight-dimensional benchmark that measures the alignment between inferred intent and generated content. Extensive experiments demonstrate that Lumos-Nexus achieves significant gains in video generation.

Acknowledgments

Hangjie Yuan was supported in part by the Postdoctoral Science Preferential Funding of Zhejiang Province, China (ZJ2025005).

References

1. AI, I., Gong, B., Zou, C., Zheng, C., Zhou, C., Yan, C., Jin, C., Shen, C., Zheng, D., Wang, F., et al.: Ming-omni: A unified multimodal model for perception and generation. arXiv preprint arXiv:2506.09344 (2025)
2. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
3. Chen, H., Zhang, Y., Cun, X., Xia, M., Wang, X., Weng, C., Shan, Y.: Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7310–7320 (2024)
4. Chen, J., Xu, Z., Pan, X., Hu, Y., Qin, C., Goldstein, T., Huang, L., Zhou, T., Xie, S., Savarese, S., et al.: Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. arXiv preprint arXiv:2505.09568 (2025)
5. Chen, L., Gu, Y., Mao, Q.: Univid: Unifying vision tasks with pre-trained video generation models. arXiv preprint arXiv:2509.21760 (2025)

6. Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C.: Janus-pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811 (2025)
7. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., et al.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025)
8. Deng, H., Pan, T., Diao, H., Luo, Z., Cui, Y., Lu, H., Shan, S., Qi, Y., Wang, X.: Autoregressive video generation without vector quantization. arXiv preprint arXiv:2412.14169 (2024)
9. Dong, R., Han, C., Peng, Y., Qi, Z., Ge, Z., Yang, J., Zhao, L., Sun, J., Zhou, H., Wei, H., et al.: Dreamllm: Synergistic multimodal comprehension and creation. arXiv preprint arXiv:2309.11499 (2023)
10. Ge, Y., Zhao, S., Zhu, J., Ge, Y., Yi, K., Song, L., Li, C., Ding, X., Shan, Y.: Seed-x: Multimodal models with unified multi-granularity comprehension and generation. arXiv preprint arXiv:2404.14396 (2024)
11. Google DeepMind: Veo 3.1. https://deepmind.google/models/veo/ (10 2025)
12. HaCohen, Y., Chiprut, N., Brazowski, B., Shalem, D., Moshe, D., Richardson, E., Levin, E., Shiran, G., Zabari, N., Gordon, O., et al.: Ltx-video: Realtime video latent diffusion. arXiv preprint arXiv:2501.00103 (2024)
13. Hong, W., Ding, M., Zheng, W., Liu, X., Tang, J.: Cogvideo: Large-scale pretraining for text-to-video generation via transformers. arXiv preprint arXiv:2205.15868 (2022)
14. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., 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)
15. Jin, Y., Sun, Z., Li, N., Xu, K., Jiang, H., Zhuang, N., Huang, Q., Song, Y., Mu, Y., Lin, Z.: Pyramidal flow matching for efficient video generative modeling. arXiv preprint arXiv:2410.05954 (2024)
16. Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., Shi, H.: Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15954–15964 (2023)
17. Kondratyuk, D., Yu, L., Gu, X., Lezama, J., Huang, J., Schindler, G., Hornung, R., Birodkar, V., Yan, J., Chiu, M.C., et al.: Videopoet: A large language model for zero-shot video generation. arXiv preprint arXiv:2312.14125 (2023)
18. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
19. Kuaishou: Kling 2.6. https://www.kling26.com/ (12 2025)
20. Li, Z., Hu, S., Liu, S., Zhou, L., Choi, J., Meng, L., Guo, X., Li, J., Ling, H., Wei, F.: Arlon: Boosting diffusion transformers with autoregressive models for long video generation. arXiv preprint arXiv:2410.20502 (2024)
21. Liu, H., Liu, S., Zhou, Z., Xu, M., Xie, Y., Han, X., Pérez, J.C., Liu, D., Kahatapitiya, K., Jia, M., et al.: Mardini: Masked autoregressive diffusion for video generation at scale. arXiv preprint arXiv:2410.20280 (2024)
22. Lu, H., Yang, G., Fei, N., Huo, Y., Lu, Z., Luo, P., Ding, M.: Vdt: General-purpose video diffusion transformers via mask modeling. arXiv preprint arXiv:2305.13311 (2023)
23. Luo, J., Lin, J., Zhang, Z., Wu, B., Fang, M., Chen, L., Tang, H.: Univid: The open-source unified video model. arXiv preprint arXiv:2509.24200 (2025)

24. Ma, X., Wang, Y., Jia, G., Chen, X., Liu, Z., Li, Y.F., Chen, C., Qiao, Y.: Latte: Latent diffusion transformer for video generation. arXiv preprint arXiv:2401.03048 (2024)
25. Mai, Z., Zhang, K., Wang, F.E., Wang, Z.K., Chen, A.Y., Xia, L., Sun, M., Chao, W.L., Kuo, C.H.: Revisiting model stitching in the foundation model era. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 41342–41351 (2026)
26. Mai, Z., Zhang, P., Tu, C.H., Chen, H.Y., Nguyen, Q.H., Zhang, L., Chao, W.L.: Lessons and insights from a unifying study of parameter-efficient fine-tuning (peft) in visual recognition. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 14845–14857 (2025)
27. Pan, X., Shukla, S.N., Singh, A., Zhao, Z., Mishra, S.K., Wang, J., Xu, Z., Chen, J., Li, K., Juefei-Xu, F., et al.: Transfer between modalities with metaqueries. arXiv preprint arXiv:2504.06256 (2025)
28. Qiu, D., Fei, Z., Wang, R., Bai, J., Yu, C., Fan, M., Chen, G., Wen, X.: Skyreels-a1: Expressive portrait animation in video diffusion transformers. arXiv preprint arXiv:2502.10841 (2025)
29. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792 (2022)
30. Tan, Z., Yang, H., Qin, L., Gong, J., Yang, M., Li, H.: Omni-video: Democratizing unified video understanding and generation. arXiv preprint arXiv:2507.06119 (2025)
31. Team, M.L., Cai, X., Huang, Q., Kang, Z., Li, H., Liang, S., Ma, L., Ren, S., Wei, X., Xie, R., et al.: Longcat-video technical report. arXiv preprint arXiv:2510.22200 (2025)
32. Teng, H., Jia, H., Sun, L., Li, L., Li, M., Tang, M., Han, S., Zhang, T., Zhang, W., Luo, W., et al.: Magi-1: Autoregressive video generation at scale. arXiv preprint arXiv:2505.13211 (2025)
33. Tong, S., Fan, D., Li, J., Xiong, Y., Chen, X., Sinha, K., Rabbat, M., LeCun, Y., Xie, S., Liu, Z.: Metamorph: Multimodal understanding and generation via instruction tuning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17001–17012 (2025)
34. Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
35. Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., Zhang, S.: Modelscope text-to-video technical report. arXiv preprint arXiv:2308.06571 (2023)
36. Wang, X., Zhang, X., Luo, Z., Sun, Q., Cui, Y., Wang, J., Zhang, F., Wang, Y., Li, Z., Yu, Q., et al.: Emu3: Next-token prediction is all you need. arXiv preprint arXiv:2409.18869 (2024)
37. Wang, Y., Xiong, T., Zhou, D., Lin, Z., Zhao, Y., Kang, B., Feng, J., Liu, X.: Loong: Generating minute-level long videos with autoregressive language models. arXiv preprint arXiv:2410.02757 (2024)
38. Wei, C., Liu, Q., Ye, Z., Wang, Q., Wang, X., Wan, P., Gai, K., Chen, W.: Univideo: Unified understanding, generation, and editing for videos. arXiv preprint arXiv:2510.08377 (2025)
39. Wu, C., Chen, X., Wu, Z., Ma, Y., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C., et al.: Janus: Decoupling visual encoding for unified multimodal understanding and generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12966–12977 (2025)

40. Wu, J.Z., Ge, Y., Wang, X., Lei, S.W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7623–7633 (2023)
41. Wu, S., Fei, H., Qu, L., Ji, W., Chua, T.S.: Next-gpt: Any-to-any multimodal llm. In: Forty-first International Conference on Machine Learning (2024)
42. Xie, J., Mao, W., Bai, Z., Zhang, D.J., Wang, W., Lin, K.Q., Gu, Y., Chen, Z., Yang, Z., Shou, M.Z.: Show-o: One single transformer to unify multimodal understanding and generation. arXiv preprint arXiv:2408.12528 (2024)
43. Xie, J., Yang, Z., Shou, M.Z.: Show-o2: Improved native unified multimodal models. arXiv preprint arXiv:2506.15564 (2025)
44. Xing, J., Du, F., Yuan, H., Liu, P., Xu, H., Ci, H., Niu, R., Chen, W., Wang, F., Liu, Y.: Lumosx: Relate any identities with their attributes for personalized video generation. In: The Fourteenth International Conference on Learning Representations (2026)
45. Yan, W., Zhang, Y., Abbeel, P., Srinivas, A.: Videogpt: Video generation using vq-vae and transformers. arXiv preprint arXiv:2104.10157 (2021)
46. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
47. Yang, L., Tian, Y., Li, B., Zhang, X., Shen, K., Tong, Y., Wang, M.: Mmada: Multimodal large diffusion language models. arXiv preprint arXiv:2505.15809 (2025)
48. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)
49. Yin, T., Zhang, Q., Zhang, R., Freeman, W.T., Durand, F., Shechtman, E., Huang, X.: From slow bidirectional to fast autoregressive video diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22963–22974 (2025)
50. Yuan, H., Chen, W., Cen, J., Yu, H., Liang, J., Chang, S., Lin, Z., Feng, T., Liu, P., Xing, J., et al.: Lumos-1: On autoregressive video generation from a unified model perspective. arXiv preprint arXiv:2507.08801 (2025)
51. Zhang, D.J., Wu, J.Z., Liu, J.W., Zhao, R., Ran, L., Gu, Y., Gao, D., Shou, M.Z.: Show-1: Marrying pixel and latent diffusion models for text-to-video generation. International Journal of Computer Vision **133**(4), 1879–1893 (2025)
52. Zhang, P., Mai, Z., Nguyen, Q.H., Chao, W.L.: Revisiting semi-supervised learning in the era of foundation models. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025), https://openreview.net/forum?id=yDh9s1qPzB
53. Zheng, Z., Peng, X., Yang, T., Shen, C., Li, S., Liu, H., Zhou, Y., Li, T., You, Y.: Open-sora: Democratizing efficient video production for all. arXiv preprint arXiv:2412.20404 (2024)
54. Zhou, C., Yu, L., Babu, A., Tirumala, K., Yasunaga, M., Shamis, L., Kahn, J., Ma, X., Zettlemoyer, L., Levy, O.: Transfusion: Predict the next token and diffuse images with one multi-modal model. arXiv preprint arXiv:2408.11039 (2024)