ImagerySearch: Adaptive Test-Time Search for Video Generation Beyond Semantic Dependency Constraints

Meiqi Wu[1,3*], **Jiashu Zhu**[2], **Xiaokun Feng**[3], **Chubin Chen**[4], **Chen Zhu**[5], **Bingze Song**[2], **Fangyuan Mao**[6], **Jiahong Wu**[2†], **Xiangxiang Chu**[2], **Kaiqi Huang**[3‡]

[1]School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 100049, China
[2]AMAP, Alibaba Group, Beijing 100012, China
[3]The Key Laboratory of Cognition and Decision Intelligence for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[4] Tsinghua University, Beijing 100084, China
[5] Southeast University, Nanjing 210096, China
[6] Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
wumeiqi18@mails.ucas.ac.cn, kqhuang@nlpr.ia.ac.cn

Abstract

Video generation models have achieved remarkable progress, particularly excelling in realistic scenarios; however, their performance degrades notably in imaginative scenarios. These prompts often involve rarely co-occurring concepts with long-distance semantic relationships, falling outside training distributions. Existing methods typically apply test-time scaling for improving video quality, but their fixed search spaces and static reward designs limit adaptability to imaginative scenarios. To fill this gap, we propose **ImagerySearch**, a dynamic test-time scaling law strategy inspired by imagery that adaptively adjusts the inference search space and rewards guided by prompts, effectively enhancing generation quality in imaginative scenarios. Furthermore, we introduce **LDT-Bench**, the first benchmark targeting long-distance semantic prompts, designed to evaluate the creativity of video generation models. It comprises 2,839 challenging concept pairs from diverse recognition datasets and incorporates an automatic evaluation protocol to assess creative capacity. Extensive experiments on LDT-Bench demonstrate that our approach consistently outperforms general generation models and test-time scaling approaches. Additionally, ImagerySearch achieves strong performance on VBench, confirming its effectiveness in improving video generation quality under diverse conditions.

Code — https://github.com/AMAP-ML/ImagerySearch

1 Introduction

Imagine describing a surreal scene–"a panda playing violin on Mars during a sandstorm"–and instantly seeing it come to life as a video. Text-to-video generation turns language into dynamic, vivid worlds. Recent video generation models have made significant progress in generating realistic scenes (Yang et al. 2024; Peng et al. 2025; OpenAI 2025; Wan Team et al. 2025); however, their performance drops sharply when handling subjectively imaginative scenarios, hindering the advancement of truly creative video generation. *Why is imagination so hard to generate?*

This limitation arises from two primary factors. **(1) The model's semantic dependency**: Generative models exhibit strong semantic dependency constraints on long-distance semantic prompts, making it difficult to generalize to imaginative scenarios beyond the training distribution (Fig. 1). **(2) The scarcity of imaginative training data**: Mainstream video datasets (Huang et al. 2024b; Liao et al. 2025; Ling et al. 2025) predominantly contain realistic scenarios, offering limited imaginative combinations characterized by long-distance semantic relationships (Fig. 3(d)). Recent test-time scaling approaches (Liu et al. 2025a; He et al. 2025a) alleviate data scarcity by sampling multiple candidates and selecting the most promising one. However, their predefined sampling spaces and static reward functions constrain adaptability to the open-ended nature of creative generation.

The Imagery Construction theory (Pylyshyn 2002) posits that humans create mental scenes for imaginative scenarios by iteratively refining visual imagery in response to language. Motivated by this principle, we introduce **ImagerySearch**, a test-time search strategy that enhances prompt-based visual generation. ImagerySearch comprises two core components: (i) **S**emantic-distance-**a**ware **D**ynamic **S**earch **S**pace (**SaDSS**), which adaptively modulates sampling granularity according to the semantic span of the prompt; and (ii) **A**daptive **I**magery **R**eward (AIR), which incentivizes outputs that align more closely with the intended semantics.

To evaluate models' imaginative capability, we propose **LDT-Bench**, the first benchmark designed for long-distance semantic prompts. It comprises 2,839 challenging concept pairs, constructed by maximizing semantic distance across

*Work done during the internship at AMAP, Alibaba Group.
†Project leader.
‡Corresponding author.

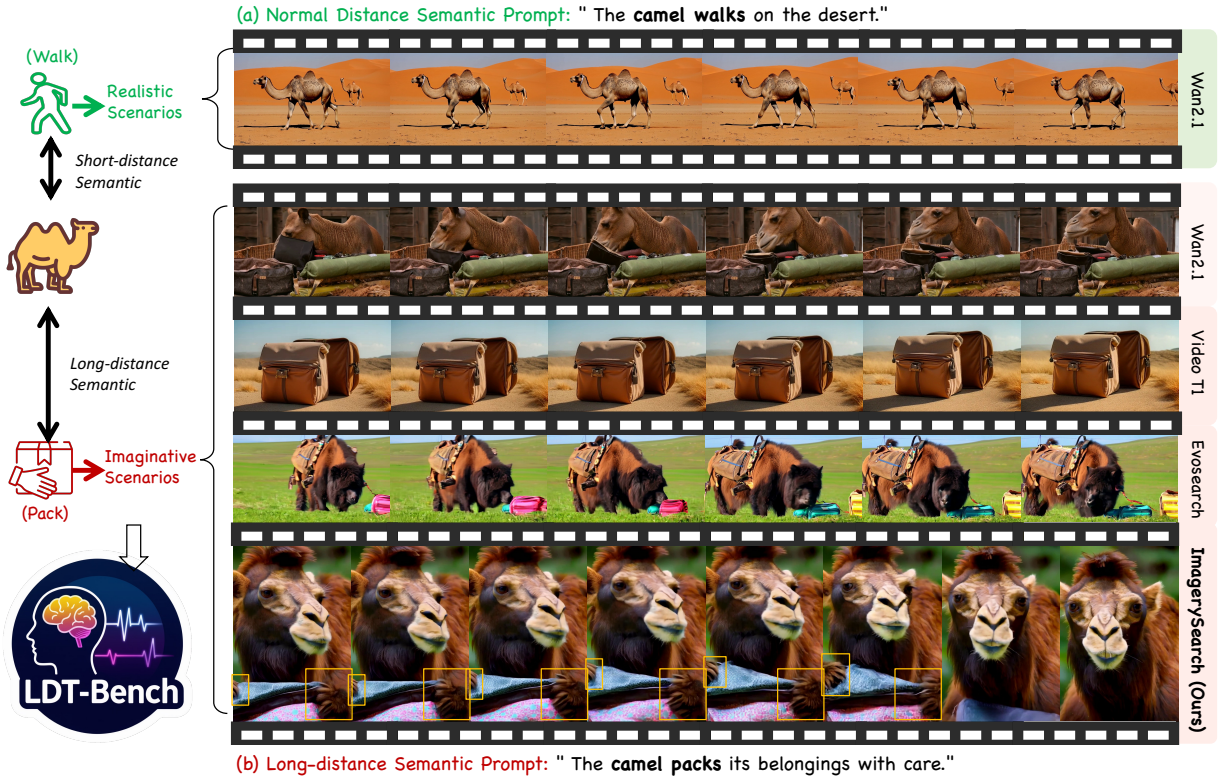


Figure 1: **The motivation of ImagerySearch. Left**: The distance depicts the corresponding strength of prompt tokens during the denoising process. *LDT-Bench* consists of imaginative scenarios with long-distance semantics. **Right**: Wan2.1 performs well on short-distance semantics but fails under long-distance. Test time scaling methods (*e.g.*, Video T1 (Liu et al. 2025a), Evosearch (He et al. 2025a)) also struggle. However, *ImagerySearch* generates coherent, context-aware motions (orange box).

object–action and action–action dimensions from diverse recognition datasets (*e.g.*, ImageNet-1K (Deng et al. 2009), Kinects-600 (Carreira et al. 2018)). In addition, LDT-Bench includes an automatic evaluation protocol, **ImageryQA**, which quantifies creative generation with respect to element coverage, semantic alignment, and anomaly detection.

Extensive experiments reveal that general models (*e.g.*, Wan14B (Wan Team et al. 2025), Hunyuan-13B (Kong et al. 2024), CogVideoX (Yang et al. 2024)) and TTS-based models (*e.g.*, VideoT1 (Liu et al. 2025a), EvoSearch (He et al. 2025a)) suffer from significant degradation in video quality and semantic alignment when conditioned on long-distance semantics. In contrast, our framework consistently improves generation fidelity and alignment, demonstrating superior capability in handling long-distance semantic prompts.

Our contributions can be summarized as follows:

- We propose ImagerySearch, a dynamic test-time scaling law strategy inspired by mental imagery that adaptively adjusts the inference search space and reward according to prompt semantics.
- We introduce LDT-Bench, a benchmark for video generation from long-distance semantic prompts, with 2,839 prompts and an automatic framework to evaluate creativity in imaginative scenarios.
- Extensive experiments reveal that our approach consistently improves imaging quality and semantic alignment under long-distance semantic prompts.

2 Related Work

Text-to-Video Generation Models. With increased training resources, large-scale T2V models (OpenAI 2025; Zheng et al. 2024a; Peng et al. 2025; Genmo Team 2024; Kong et al. 2024; Wan Team et al. 2025) have emerged, capable of generating coherent videos, understanding physics, and generalizing to complex scenarios. But they require massive data, and collecting enough long-range semantic prompts is impractical (Chu, Li, and Wang 2025). Although fine-tuning (Wallace et al. 2024) and post-training (Luo et al. 2023; Li et al. 2024a,b) methods mitigate data requirements to some extent, the extreme scarcity of long-distance semantic videos still hinders effective training. In contrast, the Test-Time Scaling (TTS) methods (Oshima et al. 2025; Xie et al. 2025; Yang et al. 2025; Liu et al. 2025a; He et al. 2025a) require no additional training and achieve strong performance through a highly general approach.

Test-Time Scaling in T2V Models. TTS improves performance by using rewards to select better outputs. In T2V generation, TTS are primarily explored in two aspects: selection strategies and reward strategies. Selection strategies mainly include Best-of-N, particle sampling, and beam search. The

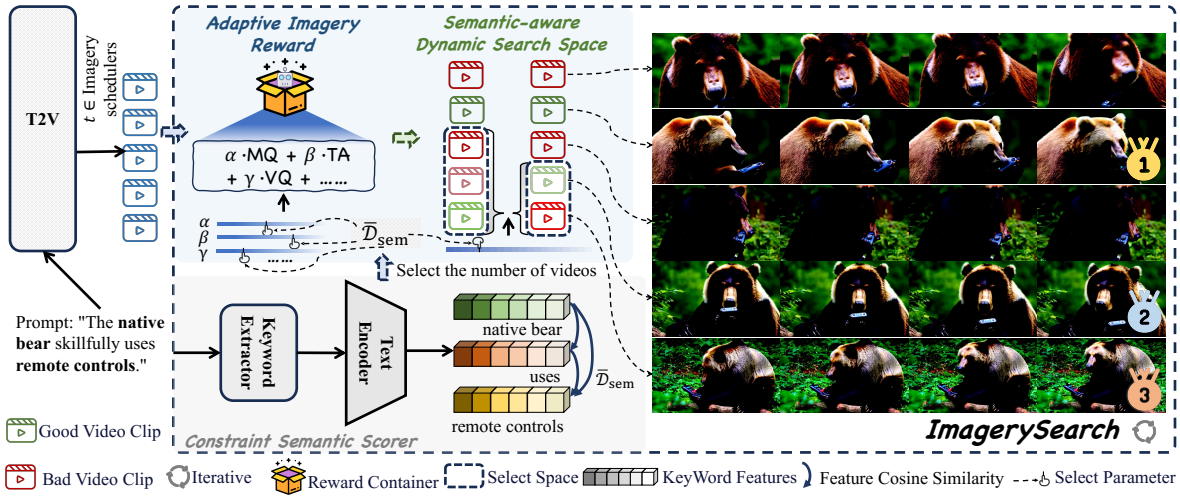


Figure 2: Overview of our ImagerySearch. The prompt is scored by the Constrained Semantic Scorer (producing $\bar{\mathcal{D}}_{\text{sem}}$) and simultaneously fed to the T2V backbone (Wan2.1). At every step t specified by the imagery scheduler, we sample a set of candidate clips, rank them with a reward function conditioned on $\bar{\mathcal{D}}_{\text{sem}}$, and retain only a $\bar{\mathcal{D}}_{\text{sem}}$-controlled subset. The loop repeats until generation completes. The figure shows the result after a single denoising step at $t = 5$.

Best-of-N (Ma et al. 2025; Liu et al. 2025a) selects the top N outputs from multiple generations. Particle sampling (Singh et al. 2025; Sunwoo Kim 2025) improves upon this by performing importance-based sampling across the denoising process. Beam search (Liu et al. 2025a; Yang et al. 2025; Liu et al. 2025a; He et al. 2025b) keeps multiple candidates at each step, expanding the sequence set over time. Reward strategies are based on various evaluation metrics, such as VisionReward (Xu et al. 2024), ImageReward (Xu et al. 2023), Aesthetic score (Schuhmann et al. 2022), which guide the selection process by quantifying the quality of generated output. Current TTS methods optimize search and reward strategies for general T2V generation. Here, we explore how TTS can improve performance on long-distance semantic prompts.

Evaluation of Video Generative Models. Early video-generation metrics are simplistic: some diverged from human judgment (Chen et al. 2025), while others reused real-video tests unsuited to synthetic clips (Soomro, Zamir, and Shah 2012; Xu et al. 2016). Later, studies (Liu et al. 2024; Huang et al. 2024c; Sun et al. 2025; Zheng et al. 2025) such as VBench (Huang et al. 2024b) evaluated AI-generated videos from a comprehensive, multi-dimensional perspective. Several studies (Yuan et al. 2025; Ling et al. 2025) refine evaluation along single dimensions such as frame realism or temporal coherence. Current benchmarks struggle to evaluate long-distance semantic prompts, which are crucial for advancing video generation.

3 ImagerySearch

We cast text-to-video generation as a search over noise inputs in diffusion sampling, structured by reward functions and search algorithms to improve video quality.

3.1 Preliminaries

In standard diffusion frameworks, sampling starts from Gaussian noise $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$, and the model iteratively denoises the latent through a learned network f_θ. As a widely used sampling paradigm, DDIM performs the following step-wise denoising update:

$$\mathbf{x}_{t-1} = \zeta_{t-1}\left(\frac{\mathbf{x}_t - \sigma_t f_\theta(\mathbf{x}_t, t, \mathbf{c})}{\zeta_t}\right) + \sigma_{t-1} f_\theta(\mathbf{x}_t, t, \mathbf{c}), \quad (1)$$

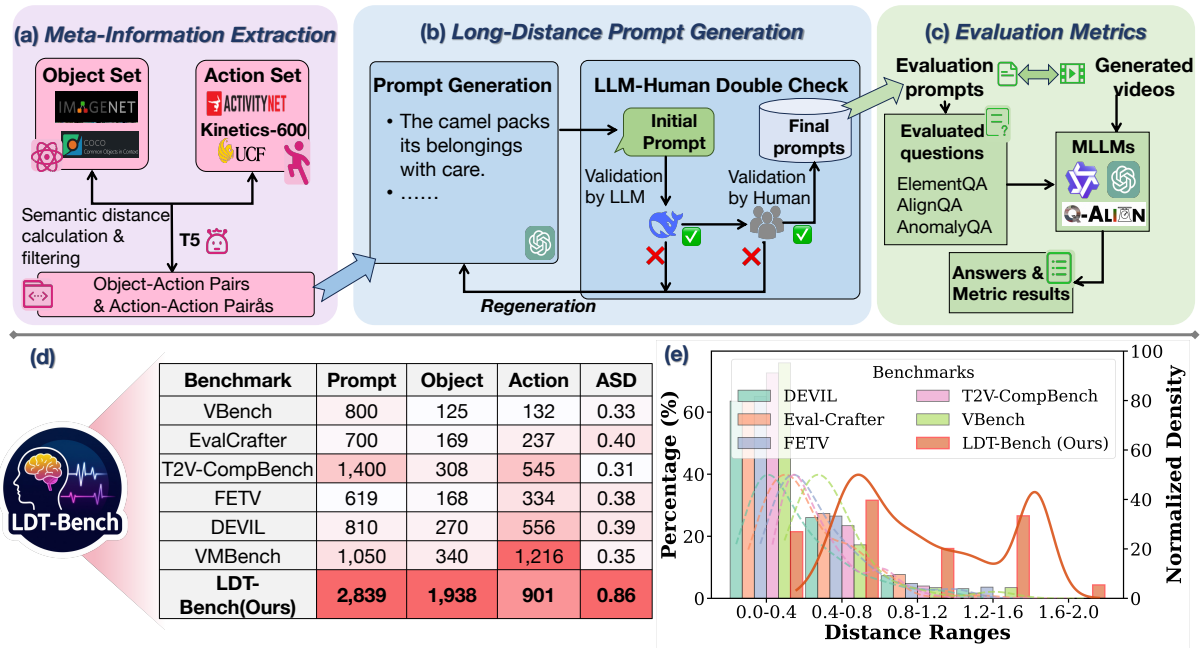
Where ζ_{t-1}, ζ_t, σ_{t-1} denote predefined schedules.

Prior test-time scaling approaches (Liu et al. 2025a; He et al. 2025a; Yang et al. 2025) operate within a fixed noise search space and use static reward functions to rank candidates. By contrast, our framework supports flexible reward design and adaptive noise selection, substantially improving both sample efficiency and generation quality.

3.2 Dynamic Search Space

Inspired by imagery cognitive theory–which posits that humans expend more effort to construct mental imagery for semantically distant concepts–we adapt the candidate-video search space to a prompt's semantic distance: shrinking it for short-distance prompts to boost test-time efficiency, and enlarging it for long-distance prompts to explore a broader range of possibilities. Therefore, we propose a **S**emantic-distance-**a**ware **D**ynamic **S**earch **S**pace (SaDSS). As shown in Fig. 2, this adaptive resizing is driven by a **Constrained Semantic Scorer**, which dynamically modulates the search space. Specifically, we define semantic distance as the average embedding distance between key entities (objects and actions) extracted from the prompt. Given a prompt $\mathbf{p}$, we extract its compositional units $\{p_i\}_{i=1}^n$ and compute:

$$\bar{\mathcal{D}}_{\text{sem}}(\mathbf{p}) = \frac{1}{|E|} \sum_{(i,j)\in E} \|\phi(p_i) - \phi(p_j)\|_2, \quad (2)$$



Benchmark	Prompt	Object	Action	ASD
VBench	800	125	132	0.33
EvalCrafter	700	169	237	0.40
T2V-CompBench	1,400	308	545	0.31
FETV	619	168	334	0.38
DEVIL	810	270	556	0.39
VMBench	1,050	340	1,216	0.35
LDT-Bench(Ours)	**2,839**	**1,938**	**901**	**0.86**

Figure 3: Overview of+ LDT-Bench. **Upper**: (a) LDT-Bench is built by first extracting meta-information from existing recognition datasets; (b) GPT-4o is then used to generate candidate prompts, which are filtered jointly by DeepSeek and humans to obtain the final prompt set; (c) We design a set of three MLLM-based QA tasks that serve as the creativity metric. **Lower**: (d) Compared with other benchmarks, LDT-Bench covers far richer categories. (e) its prompts also exhibit a semantic-distance distribution that is shifted toward substantially longer ranges. Note that "ASD" denotes the average semantic distance of prompts.

where $\phi(\cdot)$ denotes the embedding function ($e.g.$, T5 encoder), and E is the set of key entity pairs in the prompt.

At inference time, we adapt the sampling procedure based on $\bar{\mathcal{D}}_{\text{sem}}$. Specifically, the search space dynamically adapts based on semantic distance. Formally, the number of candidates N_t at timestep t is dynamically adjusted as:

$$N_t = N_{\text{base}} \cdot \left(1 + \lambda \cdot \bar{\mathcal{D}}_{\text{sem}}(\mathbf{p})\right), \tag{3}$$

where N_{base} is the base number of samples, and λ is a scaling factor that controls the sensitivity to semantic distance. In this work, we set $\lambda = 1$.

By tailoring the search scope to the inherent difficulty of the prompt, SaDSS encourages the model to explore more diverse visual hypotheses when needed, improving visual plausibility under challenging conditions, without incurring unnecessary computational costs for simple prompts.

3.3 Adaptive Imagery Reward

Based on our observations, adjacent denoising steps alter the latent video only marginally, so we invoke ImagerySearch at a few key noise levels $\mathcal{S} = \{5, 10, 20, 45\}$, termed the *Imagery Schedule* (see App. A). As shown in Fig. 2, given a partially denoised latent representation $\mathbf{x}_t$ conditioned on c, We perform a single ODE (Ordinary Differential Equation) denoising step:

$$\hat{\mathbf{x}}_0 = \frac{1}{\zeta_t}\left(\mathbf{x}_t - \sigma_t f_\theta(\mathbf{x}_t, t, c)\right), \tag{4}$$

The reward assessment is then conducted on this one-step denoised output, enabling us to analyze the impact of different denoising progress stages on the final video quality.

To enhance semantic alignment between generated videos and prompts with long-distance semantics, we introduce an **A**daptive **I**magery **R**eward (AIR) that modulates evaluation feedback based on the prompt's semantic difficulty. Specifically, we incorporate the semantic distance as a soft reweighting factor into the reward formulation. The reward $R_{\text{AIR}}(\hat{\mathbf{x}}_0)$ for each candidate video $\mathbf{x}_0$ is defined as:

$$R_{\text{AIR}}(\hat{\mathbf{x}}_0) = (\alpha \cdot \text{MQ} + \beta \cdot \text{TA} + \gamma \cdot \text{VQ} + \omega \cdot R_{any}) \cdot \bar{\mathcal{D}}_{\text{sem}}(\hat{\mathbf{x}}_0), \tag{5}$$

where α, β, γ, and ω are scaling factors that adjust the reward based on the prompt semantic distance $\bar{\mathcal{D}}$sem. MQ, TA, and VQ follow VideoAlign (Liu et al. 2025b), and Rany denotes an extensible reward ($e.g.$, VideoScore (He et al. 2024), VMBench (Ling et al. 2025)).

4 LDT-Bench

We introduce LDT-Bench from two perspectives: the construction of the prompt suite and the design of evaluation metrics. The core components are illustrated in Fig. 3.

4.1 Prompt Suite

Meta-information Extraction. Considering that objects and actions are the main entities in text prompts, we construct our prompts using the following two structural types. (1) **Object–Action**: An object combined with an uncommon

Figure 4: Visualization of examples. **Upper**: Results from general models. **Lower**: ImagerySearch versus other test-time scaling methods. Ours produces more vivid actions under long-distance semantic prompts.

or incompatible action. (2) **Action–Action**: Two semantically distant or even contradictory actions.

To cover a wide range of objects and actions, we build our object and action sets from representative large-scale datasets. Specifically, the object set is derived from ImageNet-1K (Deng et al. 2009) and COCO (Lin et al. 2014) (covering 1,938 objects), while the action set is collected from ActivityNet (Caba Heilbron et al. 2015), UCF101 (Soomro, Zamir, and Shah 2012), and Kinetics-600 (Carreira et al. 2018) (covering 901 actions). These collections serve as the foundation for subsequent prompt generation.

We first encode each object and action element text_i using a pretrained T5 text encoder (Raffel et al. 2020), obtaining a high-dimensional textual feature $\mathbf{h}_i \in \mathbb{R}^d$. These embeddings are then projected into a shared 2D semantic space via Principal Component Analysis (PCA):

$$\mathbf{z}_i = \text{PCA}(\mathbf{h}_i) = \text{PCA}(\text{T5}(\text{text}_i)), \quad \mathbf{z}_i \in \mathbb{R}^2, \qquad (6)$$

where $\mathbf{z}_i$ represents the semantic position of the i-th element in the 2D space. T5 can be replaced with other encoders, such as CLIP (Radford et al. 2021) (App. B.1).

To measure semantic divergence, we compute the Euclidean distance between each pair of elements as a criterion for selecting long-distance semantic prompts. We then form two candidate sets: pairing each object with its most distant action (1,938 object–action pairs) and matching each action with its most distant counterpart (901 action–action pairs). From each set, we select the 160 most distant pairs, resulting in 320 high-distance prompts that challenge the model with long-distance semantic combinations. For more analysis of the prompt suite, please refer to App. B.2.

Long-distance Prompt Generation. Based on the obtained text element pairs, we employ a large language model, *i.e.*, GPT-4o (Hurst et al. 2024), to generate fluent and complete text prompts by filling in necessary sentence components. Subsequently, each prompt is double-checked by both DeepSeekR1 (Guo et al. 2025) and human annotators to ensure quality, resulting in our final prompt suite. The generation process and examples are shown in Fig. 3(b).

4.2 Imagery Evaluation Metrics

We design three core metric dimensions to quantitatively evaluate video generation under long-distance semantics. Questions are generated from text prompts, and semantically strong MLLMs are used to analyze the generated videos and produce quantitative scores (Chu et al. 2025; Cho et al. 2023; Feng et al. 2025).

ElementQA. Because our prompts focus on objects and actions, ElementQA primarily consists of targeted questions revolving around these elements. For example, given the prompt "The traffic light is dancing.", we can generate two questions: "Does the traffic light appear in the video?" and "Is the traffic light performing a dancing action?"

AlignQA. In addition to the basic semantic information cov-

ered by ElementQA, we also evaluate the generated videos in terms of visual quality and aesthetics (Murray, Marchesotti, and Perronnin 2012). Given the challenging and inherently subjective nature of this assessment, we employ recently developed MLLMs that have been specifically optimized for alignment with human perception to perform the evaluation (Huang et al. 2024a; Wu et al. 2023).

AnomalyQA. We have observed that video generation models frequently produce anomalous outputs. Consequently, we leverage MLLMs to analyze the generated frames and answer targeted questions aimed at detecting anomalies.

Implementation Details. For ElementQA, we employ Qwen2.5-VL-72B-Instruct (Bai et al. 2025) as the underlying MLLM, whereas for AlignQA we adopt Q-Align (Wu et al. 2023), a model specifically optimized for rating visual quality and aesthetics. Given the broader generalization required by AnomalyQA, we utilize the more powerful GPT-4o (OpenAI 2024) for evaluation. We refer to these three components as ImageryQA. Details are in App. B.3.

5 Experiments

5.1 Experimental Setup

Datasets & Metrics. To assess the imaginative capacity of video-generation models, we evaluate them on both LDT-Bench and VBench (Huang et al. 2024b), using each benchmark's full prompt suite and associated metrics.

Compared Models. We compare two categories of models: (1) *General models*: Hunyuan (Kong et al. 2024), Wan2.1 (Wan Team et al. 2025), Open-Sora (Zheng et al. 2024b), CogVideoX (Yang et al. 2024); (2) *TTS methods*: Video-T1 (Liu et al. 2025a) and EvoSearch (He et al. 2025a). We use Wan2.1 as the base model and generate 33-frame clips with the default settings (see App. C for details).

Experimental Environment. All experiments are run on a server equipped with 8 × NVIDIA H20 GPUs (96 GB each), an Intel Xeon Gold 6348 CPU (32 cores, 2.6 GHz), and 512 GB of RAM, under Ubuntu 20.04 LTS (kernel 5.15). We used Python 3.9 with PyTorch 2.5.1 (CUDA 12.4, cuDNN 9.1), torchvision 0.20.1, and Transformers 4.50.3.

5.2 Comparison with Other Generation Models

Performance on LDT-Bench. As shown in Tab. 1, we adopt Wan2.1 as the base model. Our method achieves a significant improvement of 8.83%. Furthermore, compared to other test-time scaling approaches, ImagerySearch also delivers consistently superior performance. These results highlight the effectiveness of our method in long-distance semantic prompts and its robustness in imagination scenarios.

Performance on VBench. For balanced evaluation on VBench, we compare general generators (upper rows of Tab. 2) with test-time scaling methods (lower rows). All models are evaluated on prompts from LDT-Bench using the VBench metrics. ImagerySearch achieves the best overall score, indicating its strong ability to preserve prompt fidelity under wide semantic gaps. Fig. 4 illustrates this strength: ImagerySearch accurately reproduces both the specified subjects (*e.g.*, *bear*, *controls*) and their associated actions (*e.g.*, *uses*). App. D examples further show its robustness.

Model	LDT-Bench (%) ↑			
	ElementQA	AlignQA	AnomalyQA	ImageryQA (All)
Wan2.1	1.66	31.62	15.00	48.28
Video-T1	1.91	**38.16**	14.68	54.75
Evosearch	1.92	36.10	16.46	54.48
Ours	**2.01**	36.82	**18.28**	**57.11**

Table 1: Quantitative comparison on LDT-Bench. ImagerySearch achieves the best average performance.

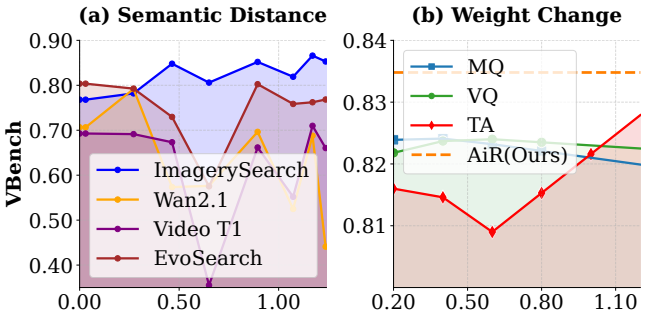


Figure 5: (a) Effect of semantic distance across different models. (b) Effect of reward weight.

Robustness Analysis Across Semantic Distances. As shown in Fig. 5(a), our method keeps VBench scores stable, whereas others fluctuate markedly. This stability highlights the superior robustness of our model across a wide range of semantic distances. Error analysis is in App E.

5.3 Test-time Scaling Law Analysis

We measure the inference-time computation by the number of function evaluations (NFEs). As shown in Fig. 6(a–d), where performance is assessed with the MQ, TA, and VQ metrics from VideoAlign (Liu et al. 2025b), ImagerySearch exhibits monotonic performance improvements as inference-time computation increases. Notably, on Wan2.1 (Wan Team et al. 2025), ImagerySearch continues to gain as NFEs grow, whereas baseline methods plateau at roughly 1×10^3 NFEs (corresponding to the 30th timestep). Moreover, our method achieves a clear advantage in the overall VideoAlign score, as illustrated in Fig. 6(d).

5.4 Ablation Study

Effect of SaDSS and AIR. As shown in Tab. 3, adding either the SaDSS or the AIR module individually already surpasses the baseline, while combining SaDSS with AIR achieves the best performance, confirming the complementary nature of semantic guidance and adaptive selection.

Effect of Search Space Size. The *SaDSS–static weight* rows in Tab. 3 compare fixed and dynamic search-space configurations. With static weights of 0.5, and 0.9, performance improves gradually, reaching a VBench score of 81.22%. In contrast, the dynamic approach attains a markedly higher score of 83.48%, demonstrating its superior ability to optimize the search space and thus boost model performance.

	Model	VBench (%) ↑						
		Aesthetic Quality	Background Consistency	Dynamic Degree	Imaging Quality	Motion Smoothness	Subject Consistency	Average
General	Wan2.1 (Wan Team et al. 2025)	50.50	91.80	82.85	58.25	97.50	90.25	78.53
	Opensora (Peng et al. 2025)	48.80	95.25	73.15	61.35	99.05	92.95	78.43
	CogvideoX (Yang et al. 2024)	48.80	95.30	47.20	65.05	98.55	94.65	74.93
	Hunyuan (Kong et al. 2024)	50.45	92.65	85.00	59.55	95.75	90.55	78.99
TTS	Video-T1 (Liu et al. 2025a)	57.20	95.65	54.05	60.25	**99.30**	94.80	76.88
	Evosearch (He et al. 2025a)	55.55	94.80	80.95	68.90	97.70	94.55	82.08
	ImagerySearch (Ours)	**57.70**	**96.00**	**84.05**	**69.20**	98.00	**95.90**	**83.48**

Table 2: Quantitative comparison of video generation models on VBench. **ImagerySearch** achieves the best average performance across multiple metrics, indicating better alignment and generation quality.

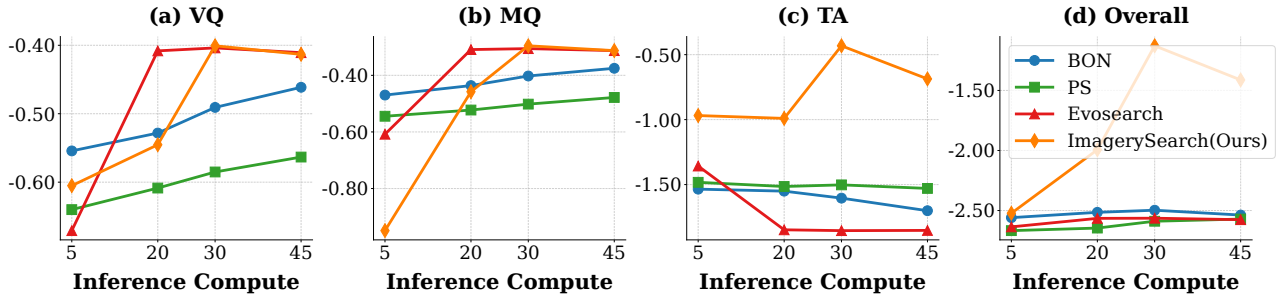


Figure 6: (a-d) Our AIR consistently delivers superior performance. Scaling behavior of ImagerySeach and baselines as inference-time computation increases. From left to right, the y-axes represent the score changes for MQ, TA, VQ, and Overall.

	Model	VBench (%)						
		Aesthetic Quality	Background Consistency	Dynamic Degree	Imaging Quality	Motion Smoothness	Subject consistency	Average
Baseline	Wan2.1 (Wan Team et al. 2025)	50.50	91.80	82.85	58.25	97.50	90.25	78.53
Modules	w/o AIR	56.25	94.60	81.85	68.05	97.50	94.40	82.11
	w/o SaDSS	55.35	95.10	77.20	68.00	97.60	94.55	81.30
SaDSS-static weight	0.5	57.25	96.15	70.00	70.75	97.45	95.45	81.18
	0.9	57.40	**96.05**	70.00	70.80	97.55	**95.50**	81.22
Search	BON (Ma et al. 2025)	57.40	95.00	83.01	68.10	**97.70**	94.63	82.64
	Particle Sampling (Ma et al. 2025)	56.51	93.52	81.72	67.04	96.18	93.38	81.39
	ImagerySearch (Ours)	**57.70**	96.00	**84.05**	**68.50**	97.65	94.70	**83.10**

Table 3: Ablation Study. "Baseline" is the plain backbone; "Modules" successively add our two novel modules; "SaDSS-static weight" denotes the performance when the selection space is kept at a fixed size; "Search" swaps in alternative search strategies.

Effect of Search Strategy. The *Search* rows in Tab. 3 compare different search strategies (*e.g.*, BON, Particle Sampling (Ma et al. 2025)). The experimental results demonstrate that our search strategy delivers the best performance.

Effect of Reward Dynamic Adjustment Mechanism. Fig. 5(b) demonstrates the impact of varying reward weights on VBench scores across different models (MQ, TA, VQ). As weights change from 0.2 to 1.2, TA shows notable improvement while MQ and VQ maintain relatively stable performance. The dashed line indicates the consistent superiority of our method, validating the effectiveness of dynamic reward adjustment under varying weights.

6 Conclusion

In this study, we propose ImagerySearch, an adaptive test-time search method that improves video-generation quality for long-distance semantic prompts drawn from imaginative scenarios. Additionally, we present LDT-Bench, the first benchmark designed to evaluate such challenging prompts. ImagerySearch achieves state-of-the-art performance on LDT-Bench and VBench, highlighting its effectiveness for long-range semantic text-to-video generation. In future work, we will explore more flexible reward mechanisms and search strategies to further improve video generation performance.

Acknowledgments

This work was supported in part by the National Science and Technology Major Project, Grant No. 2022ZD0116403.

References

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Caba Heilbron, F.; Escorcia, V.; Ghanem, B.; and Carlos Niebles, J. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, 961–970.

Carreira, J.; Noland, E.; Banki-Horvath, A.; Hillier, C.; and Zisserman, A. 2018. A Short Note about Kinetics-600. *CoRR*, abs/1808.01340.

Chen, R.; Sun, L.; Tang, J.; Li, G.; and Chu, X. 2025. Finger: Content aware fine-grained evaluation with reasoning for ai-generated videos. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 3517–3526.

Cho, J.; Hu, Y.; Garg, R.; Anderson, P.; Krishna, R.; Baldridge, J.; Bansal, M.; Pont-Tuset, J.; and Wang, S. 2023. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. *arXiv preprint arXiv:2310.18235*.

Chu, X.; Huang, H.; Zhang, X.; Wei, F.; and Wang, Y. 2025. Gpg: A simple and strong reinforcement learning baseline for model reasoning. *arXiv preprint arXiv:2504.02546*.

Chu, X.; Li, R.; and Wang, Y. 2025. Usp: Unified self-supervised pretraining for image generation and understanding. *arXiv preprint arXiv:2503.06132*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Feng, X.; Yu, H.; Wu, M.; Hu, S.; Chen, J.; Zhu, C.; Wu, J.; Chu, X.; and Huang, K. 2025. NarrLV: Towards a Comprehensive Narrative-Centric Evaluation for Long Video Generation Models. *arXiv preprint arXiv:2507.11245*.

Genmo Team. 2024. Mochi 1. https://github.com/genmoai/models. Accessed: Oct 22, 2024.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

He, H.; Liang, J.; Wang, X.; Wan, P.; Zhang, D.; Gai, K.; and Pan, L. 2025a. Scaling Image and Video Generation via Test-Time Evolutionary Search. arXiv:2505.17618.

He, H.; Liang, J.; Wang, X.; Wan, P.; Zhang, D.; Gai, K.; and Pan, L. 2025b. Scaling Image and Video Generation via Test-Time Evolutionary Search. arXiv:2505.17618.

He, X.; Jiang, D.; Zhang, G.; Ku, M.; Soni, A.; Siu, S.; Chen, H.; Chandra, A.; Jiang, Z.; Arulraj, A.; et al. 2024. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. *arXiv preprint arXiv:2406.15252*.

Huang, Y.; Sheng, X.; Yang, Z.; Yuan, Q.; Duan, Z.; Chen, P.; Li, L.; Lin, W.; and Shi, G. 2024a. Aesexpert: Towards multi-modality foundation model for image aesthetics perception. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 5911–5920.

Huang, Z.; He, Y.; Yu, J.; Zhang, F.; Si, C.; Jiang, Y.; Zhang, Y.; Wu, T.; Jin, Q.; Chanpaisit, N.; et al. 2024b. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21807–21818.

Huang, Z.; Zhang, F.; Xu, X.; He, Y.; Yu, J.; Dong, Z.; Ma, Q.; Chanpaisit, N.; Si, C.; Jiang, Y.; et al. 2024c. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *arXiv preprint arXiv:2411.13503*.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Kong, W.; Tian, Q.; Zhang, Z.; Min, R.; Dai, Z.; Zhou, J.; Xiong, J.; Li, X.; Wu, B.; Zhang, J.; et al. 2024. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*.

Li, J.; Feng, W.; Fu, T.-J.; Wang, X.; Basu, S.; Chen, W.; and Wang, W. Y. 2024a. T2v-turbo: Breaking the quality bottleneck of video consistency model with mixed reward feedback. *Advances in neural information processing systems*, 37: 75692–75726.

Li, J.; Long, Q.; Zheng, J.; Gao, X.; Piramuthu, R.; Chen, W.; and Wang, W. Y. 2024b. T2v-turbo-v2: Enhancing video generation model post-training through data, reward, and conditional guidance design. *arXiv preprint arXiv:2410.05677*.

Liao, M.; Ye, Q.; Zuo, W.; Wan, F.; Wang, T.; Zhao, Y.; Wang, J.; Zhang, X.; et al. 2025. Evaluation of text-to-video generation models: A dynamics perspective. *Advances in Neural Information Processing Systems*, 37: 109790–109816.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Ling, X.; Zhu, C.; Wu, M.; Li, H.; Feng, X.; Yang, C.; Hao, A.; Zhu, J.; Wu, J.; and Chu, X. 2025. VMBench: A Benchmark for Perception-Aligned Video Motion Generation. *arXiv preprint arXiv:2503.10076*.

Liu, F.; Wang, H.; Cai, Y.; Zhang, K.; Zhan, X.; and Duan, Y. 2025a. Video-t1: Test-time scaling for video generation. *arXiv preprint arXiv:2503.18942*.

Liu, J.; Liu, G.; Liang, J.; Yuan, Z.; Liu, X.; Zheng, M.; Wu, X.; Wang, Q.; Qin, W.; Xia, M.; et al. 2025b. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*.

Liu, Y.; Cun, X.; Liu, X.; Wang, X.; Zhang, Y.; Chen, H.; Liu, Y.; Zeng, T.; Chan, R.; and Shan, Y. 2024. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22139–22149.

Luo, S.; Tan, Y.; Huang, L.; Li, J.; and Zhao, H. 2023. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*.

Ma, N.; Tong, S.; Jia, H.; Hu, H.; Su, Y.-C.; Zhang, M.; Yang, X.; Li, Y.; Jaakkola, T.; Jia, X.; et al. 2025. Inference-time scaling for diffusion models beyond scaling denoising steps. *arXiv preprint arXiv:2501.09732*.

Murray, N.; Marchesotti, L.; and Perronnin, F. 2012. AVA: A large-scale database for aesthetic visual analysis. In *2012 IEEE conference on computer vision and pattern recognition*, 2408–2415. IEEE.

OpenAI. 2024. GPT-4o: OpenAI's new flagship model. https://openai.com/index/gpt-4o-and-gpt-4-api-updates/. Accessed: June 5, 2024.

OpenAI. 2025. Sora. https://openai.com/index/sora/. Accessed: Feb 25, 2025.

Oshima, Y.; Suzuki, M.; Matsuo, Y.; and Furuta, H. 2025. Inference-Time Text-to-Video Alignment with Diffusion Latent Beam Search. *arXiv preprint arXiv:2501.19252*.

Peng, X.; Zheng, Z.; Shen, C.; Young, T.; Guo, X.; Wang, B.; Xu, H.; Liu, H.; Jiang, M.; Li, W.; and et al. 2025. Open-Sora 2.0: Training a Commercial-Level Video Generation Model in $200k. *arXiv preprint arXiv:2503.09642*.

Pylyshyn, Z. W. 2002. Mental imagery: In search of a theory. *Behavioral and brain sciences*, 25(2): 157–182.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140): 1–67.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35: 25278–25294.

Singh, A.; Mukherjee, S.; Beirami, A.; and Rad, H. J. 2025. CoDe: Blockwise Control for Denoising Diffusion Models. *ArXiv*, abs/2502.00968.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Sun, K.; Huang, K.; Liu, X.; Wu, Y.; Xu, Z.; Li, Z.; and Liu, X. 2025. T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 8406–8416.

Sunwoo Kim, D. P., Minkyu Kim. 2025. Test-time Alignment of Diffusion Models without Reward Over-optimization. In *The Thirteenth International Conference on Learning Representations*.

Wallace, B.; Dang, M.; Rafailov, R.; Zhou, L.; Lou, A.; Purushwalkam, S.; Ermon, S.; Xiong, C.; Joty, S.; and Naik, N. 2024. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8228–8238.

Wan Team; Wang, A.; Ai, B.; Wen, B.; Mao, C.; Xie, C.-W.; Chen, D.; Yu, F.; Zhao, H.; Yang, J.; et al. 2025. Wan: Open and Advanced Large-Scale Video Generative Models. *arXiv preprint arXiv:2503.20314*.

Wu, H.; Zhang, Z.; Zhang, W.; Chen, C.; Liao, L.; Li, C.; Gao, Y.; Wang, A.; Zhang, E.; Sun, W.; et al. 2023. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*.

Xie, E.; Chen, J.; Zhao, Y.; Yu, J.; Zhu, L.; Wu, C.; Lin, Y.; Zhang, Z.; Li, M.; Chen, J.; et al. 2025. Sana 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer. *arXiv preprint arXiv:2501.18427*.

Xu, J.; Huang, Y.; Cheng, J.; Yang, Y.; Xu, J.; Wang, Y.; Duan, W.; Yang, S.; Jin, Q.; Li, S.; et al. 2024. Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation. *arXiv preprint arXiv:2412.21059*.

Xu, J.; Liu, X.; Wu, Y.; Tong, Y.; Li, Q.; Ding, M.; Tang, J.; and Dong, Y. 2023. ImageReward: learning and evaluating human preferences for text-to-image generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 15903–15935.

Xu, J.; Mei, T.; Yao, T.; and Rui, Y. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5288–5296.

Yang, H.; Tang, F.; Hu, M.; Li, Y.; Liu, Y.; Peng, Z.; He, J.; Ge, Z.; and Razzak, I. 2025. ScalingNoise: Scaling Inference-Time Search for Generating Infinite Videos. *arXiv preprint arXiv:2503.16400*.

Yang, Z.; Teng, J.; Zheng, W.; Ding, M.; Huang, S.; Xu, J.; Yang, Y.; Hong, W.; Zhang, X.; Feng, G.; et al. 2024. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. *arXiv preprint arXiv:2408.06072*.

Yuan, S.; He, X.; Deng, Y.; Ye, Y.; Huang, J.; Lin, B.; Luo, J.; and Yuan, L. 2025. Opens2v-nexus: A detailed benchmark and million-scale dataset for subject-to-video generation. *arXiv preprint arXiv:2505.20292*.

Zheng, D.; Huang, Z.; Liu, H.; Zou, K.; He, Y.; Zhang, F.; Zhang, Y.; He, J.; Zheng, W.-S.; Qiao, Y.; et al. 2025. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*.

Zheng, Z.; Peng, X.; Yang, T.; Shen, C.; Li, S.; Liu, H.; Zhou, Y.; Li, T.; and You, Y. 2024a. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*.

Zheng, Z.; Peng, X.; Yang, T.; Shen, C.; Li, S.; Liu, H.; Zhou, Y.; Li, T.; and You, Y. 2024b. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*.