

Technical Appendix for Reasoning Diffusion for Unpaired Test Time Out-of-distribution Text-Image to Video Generation

Zirui Pan¹, Xin Wang^{1,2*}, Yipeng Zhang¹, Hong Chen¹, Kecheng Zheng³, Wenwu Zhu^{1,2*}

¹Department of Computer Science and Technology, Tsinghua University,

²BNRIST, Tsinghua University, ³Ant Research

{pzr24, zhang-yp22, h-chen20}@mails.tsinghua.edu.cn, zkechengzk@gmail.com

{xin.wang, wwzhu}@tsinghua.edu.cn

Abstract

In this Technical Appendix, we first provide detailed information on the implementation and evaluation of ReasonDiff to support reproducibility. We then present ablations on the use of VisionNarrator and include illustrative reasoning examples. Next, we report three additional experiments that further validate the effectiveness of our method: (1) a detailed analysis of the naive baseline to supplement the discussion in Section 4.3 of the main paper; (2) an examination of ReasonDiff’s reasoning capability through attention maps, and (3) a demonstration of one potential application—prompt-driven customization. Finally, we include additional qualitative results and the complete quantitative tables referenced in the main paper. Generated video samples can be found in the Multimedia Appendix.

1. Implementation Details

1.1. Model and Hyper-parameters

Our base generative model is built upon Wan2.1 [7]. Specifically, Wan2.1 adopts a Diffusion Transformer (DiT) architecture, and employs mask mechanism and decoupled cross attention. The model is designed to take a condition image and a text prompt as inputs, and generates a video starting with that image. The condition image is first concatenated with zero-filled frames along the temporal axis, and then compressed using Wan-VAE. More details can be found in their official report (<https://arxiv.org/abs/2503.20314>, page 26). The version we use is Wan2.1-I2V-14B-480P¹. For VisionNarrator,

we employ the state-of-the-art MLLM GPT-4o [5]. We posit that the multi-modal reasoning supplied by VisionNarrator is crucial for guiding the downstream base generator. In particular, stronger reasoning capabilities in the MLLM more effectively maintain that the generated video remains consistent with the unpaired inputs. Nevertheless, we also evaluate the performance and efficiency of using different MLLMs as the VisionNarrator and present a detailed analysis in Section 3.1. For AlignFormer—whose architecture is illustrated in Figure 2 of the main paper—we design a multi-stage temporal anchor attention mechanism. This module employs multiple cross-attention layers to inject high-level reasoning signals into the base video generator. Each multi-head attention layer is configured with 8 heads.

1.2. Training and Inference

In this section, we give the details on the implementation for training and inference. During training, the amount of data we use is approximately 10k, which is processed from WebVid dataset [1]. We set batch size to 8, sampling frames at 0.2 second intervals, and fix the number of frames and the frame resolution to 33 and 512×512 , respectively. We add LoRA layers to the base Wan2.1 model, while randomly initializing the AlignFormer module, and we choose the AdamW optimizer with a learning rate of $1e-5$. Furthermore, we conduct one epoch of training in the first stage, and an additional epoch in the second stage, with the balancing parameter β set to 0.2. All training is conducted on one A100 80G GPU.

During inference, in qualitative comparison experiment we fix the length of the generated video to 81 frames. Moreover, We set the number of timesteps in the denoising process to 50, and fix the scale of the classifier-free guidance to 5.0.

Additionally, we give the overall implementation of the

*Corresponding Authors.

¹<https://huggingface.co/Wan-AI/Wan2.1-I2V-14B-480P>

Algorithm 1 Training procedure of ReasonDiff.

```
1: Input: Data =  $\{(x_1 \in \mathbb{R}^{b \times c \times f \times h \times w}, h \in \mathbb{R}^{f \times l \times d})\}_{N_D}$ 
2: Parameter:  $\theta_R$ : parameters for ReasonDiff,  $\theta_V$ : parameters for base video generative model.
3: function MULTIFRAMEREASONING( $c_i, h, f$ )
4:    $c^* \leftarrow \text{repeat}(c_i, f)$ .
5:    $c^* \leftarrow \text{Self-Attn}(c^*), h \leftarrow \text{Self-Attn}(h)$ .
6:    $c^* \leftarrow \text{Cross-Attn}(\text{query} = c^*, \text{key} = h, \text{value} = h)$ .
7:    $c^* \leftarrow \text{MLP}(c^*) + c^*$ .
8:    $c_i^* \leftarrow c_i$ .  $\triangleright$  Use anchor  $c_i$  to replace  $i^{\text{th}}$  element of  $c^*$ .
9:   return  $c^*$ .
10: end function
11: BEGIN MAIN FUNCTION:
12: Initialize  $f, \beta$  as the number of frames and the weight for the auxiliary loss.
13: for  $\theta \in [\{\theta_R, \theta_V\}, \{\theta_R\}]$  do
14:   repeat
15:     Forward a batch of data  $\{(x_1, h)\}_{N_{\text{batch}}}$ .
16:     Sample  $x_0 \sim \mathcal{N}(0, 1)$ .
17:      $x_t \leftarrow tx_1 + (1 - t)x_0$ .
18:     Randomly select the condition index  $i \in \{1, \dots, f\}$ .
19:      $c \leftarrow \text{VIDEOFRAMEENCODER}(x_1)$ .
20:      $c^* \leftarrow \text{MULTIFRAMEREASONING}(c_i, h, f)$ .
21:      $\text{loss} \leftarrow \|u_{\theta_V}(x_t, h, c^*) - v(x_t)\|_2^2$ .
22:     if  $\theta = \{\theta_R\}$  then  $\triangleright$  Second stage training.
23:        $\text{loss} \leftarrow \text{loss} + \beta \cdot \text{MSE}(c, c^*)$ 
24:     end if
25:      $\theta \leftarrow \arg \min_{\theta} (\text{loss})$ .
26:   until One Epoch Ends
27: end for
```

training procedure in Algorithm 1, which consists of two stages, where we first train the base video generative model and the newly added AlignFormer module as a whole, and then we introduce an auxiliary reconstruction loss to fine-tune the AlignFormer individually. Note that we have simplified the *MultiFrameReasoning* function, *i.e.*, the role of AlignFormer module, in the algorithm. The notations are the same as in the main paper.

2. Evaluation Details

2.1. Baselines

We compare our method with the following baselines, which are the latest works for video generation that achieve good performances, specifically,

1. Dynamicrafter [8]²: Dynamicrafter studies the animation

²<https://huggingface.co/Doubiiu/DynamiCrafter/>

of open-domain images by using a query transformer to project the input image into a text-aligned rich context representation space and fuse the initial image in the diffusion process.

2. LTX-Video-2B [3]³: LTX-Video seamlessly integrate the video-vae and denoising transformer, jointly optimizing their interaction for improved efficiency and quality. Notably, it achieves faster-than-real-time generation.
3. CogVideoX1.5-5B [10]⁴: CogVideoX addresses the problem of the generation of long videos that are coherent with the conditions. They introduce a 3D-VAE to compress the videos, and further propose expert adaptive LayerNorm and progressive training to improve video quality.
4. Wan2.1-I2V-14B-480P [7]⁵: Wan2.1 is a comprehensive video foundation model comprising a novel spatial-temporal variational autoencoder and scalable pre-training strategies. It is built upon DiT architecture with parameters on the scale of billion.

2.2. Dataset

Since our work focuses on video generation under OOD scenarios with unpaired text-image conditions, we construct a custom evaluation dataset tailored to this setting. Specifically, we randomly sample 500 videos from ActivityNet [2] and extract a 16-frame clip from each. For every clip, we select either the first or the last frame as the condition image and use LLaMA-3.2-11B-Vision-Instruct to generate a caption for the opposite end (*i.e.*, the last or first frame, respectively) to serve as the text prompt. This design introduces temporal separation between the visual and textual conditions, effectively simulating an unpaired scenario. Additionally, we incorporate the public general-purpose dataset MSR-VTT [9], which contains paired conditions, by directly using its validation set, which contains approximately 500 video clips. For each clip, we use the first frame as the condition image.

2.3. Metrics

In this section, we elaborate more on the details of the metrics employed in the quantitative experiment, *i.e.*, *Imaging Quality*, *Motion Smooth*, *Dynamic Degree*, *CLIP Score (Image/Text)* and *User Rank*. The first three metrics, which are universal evaluation criteria for general videos, are supported by VBench [4], an open-sourced evaluation benchmark on video domain. As indicated by the name, *Imaging Quality* measures the aesthetic level of the generated

blob/main/model.ckpt

³<https://huggingface.co/Lightricks/LTX-Video/blob/main/ltxv-2b-0.9.6-distilled-04-25safetensors>

⁴<https://huggingface.co/zai-org/CogVideoX1.5-5B-I2V>

⁵<https://huggingface.co/Wan-AI/Wan2.1-I2V-14B-480P>

Table 1. Latency and quantitative comparisons between different MLLMs on ActivityNet under the same experimental settings. We measure the mean response latency and compute each model’s relative time cost with respect to the total inference time.

Model	Latency(s)(↓)	Imaging Quality(↑)	Motion Smooth(↑)	Dynamic Degree(↑)	CLIP Score (Text)(↑)	CLIP Score (Image)(↑)
GPT-4o (ReasonDiff)	14.96 (+1.44%)	0.528	0.986	0.936	0.261	0.528
Qwen	13.20 (+1.27%)	0.509	0.985	0.908	0.247	0.515
LLaVA	9.38 (+0.90%)	0.434	0.947	0.868	0.209	0.501
Baseline (Wan2.1)	-	0.512	0.980	0.810	0.224	0.518

frames, and *Motion Smooth* judges whether the movements and transformations in the video are photorealistic, while *Dynamic Degree* favors those with more motion dynamics. The rest of the metrics are as follows:

CLIP Score (Image/Text) We employ a CLIP encoder⁶ [6] to extract features from the generated video frames, the condition image and the text prompt, and compute the average cosine similarity as the final evaluation metric. Specifically, for *CLIP Score (Text)*, we utilize LLaMA-3.2-11B-Vision-Instruct to generate captions for the ground-truth video frames, then calculate the similarity between each caption and its corresponding generated frame. This approach ensures that higher scores reflect stronger semantic alignment between the generated video and the ground-truth sequence.

User Rank In the user study, each participants are given multiple randomly-chosen sets of questions, each containing unpaired input image and text, with five corresponding and randomly-arranged video samples generated from ReasonDiff and the baseline models, which should be ranked based on their coherence with the input conditions, intuitive feeling and overall quality, etc. Specifically, we ask 20 participants to each answer ten questions, and average the scores for each model respectively as the final *User Rank* for the two datasets.

3. VisionNarrator

3.1. Choice of MLLM

In this section, we ablate the choice of MLLM used in VisionNarrator and provide a detailed analysis. In the full ReasonDiff model, we employ GPT-4o as the VisionNarrator due to its strong multi-modal reasoning capabilities, which are beneficial for guiding the downstream video generator. Here, we compare GPT-4o with two widely used

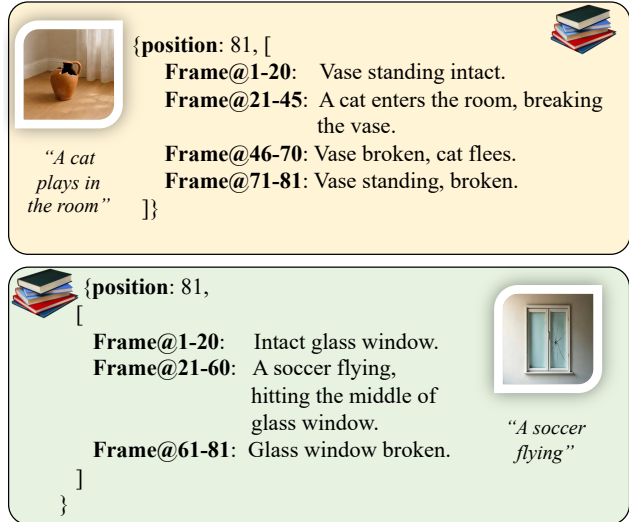


Figure 1. Reasoning examples of the proposed VisionNarrator, where we give the deduced position of the condition image and the detailed per-frame narrative following the specific format.

Table 2. Additional ablation study on the robustness of VisionNarrator, using ActivityNet dataset.

Model	Imaging Quality	Motion Smooth	Dynamic Degree	CLIP Text	CLIP Image
Narrator-err.	0.416	0.976	0.933	0.248	0.521
Anchor-miss.	0.450	0.973	0.916	0.237	0.497
ReasonDiff	0.528	0.986	0.936	0.261	0.528

MLLMs—Qwen3-VL-8B-Instruct⁷ and LLaVA-v1.5-7B⁸. Our goal is not to rigorously benchmark the standalone reasoning abilities of these models, but rather to examine how different levels of multi-modal reasoning impact our proposed method.

The results are presented in Table 1. As shown, incorporating VisionNarrator introduces an additional latency of roughly 10 seconds, which is negligible relative to the total inference time (less than 1.5%). The base Wan2.1 model takes 1040s (± 17.22 s), whereas ReasonDiff takes

⁶<https://huggingface.co/openai/clip-vit-large-patch14>

⁷<https://huggingface.co/Qwen/Qwen3-VL-8B-Instruct>

⁸<https://huggingface.co/liuhaotian/llava-v1.5-7b>

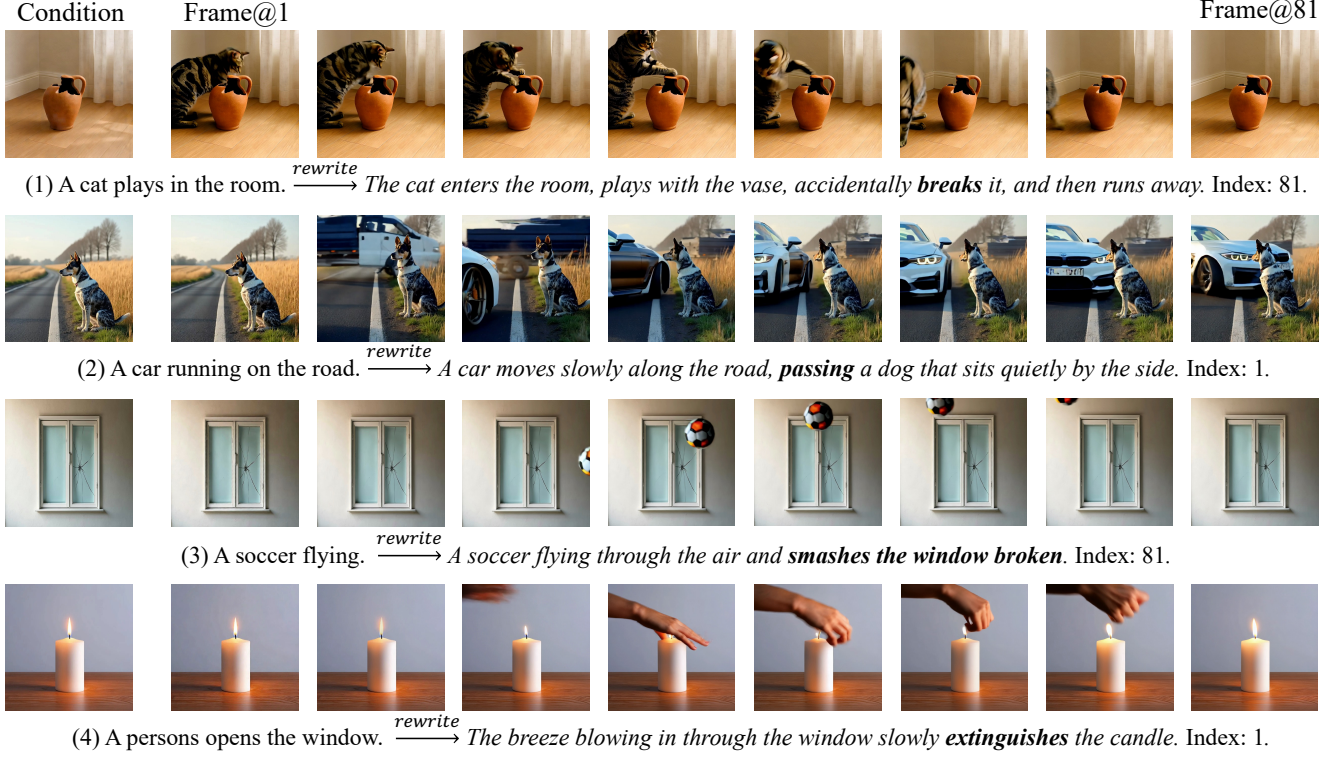


Figure 2. Qualitative examples of the naive approach based on vanilla Wan2.1. We rewrite the prompt using GPT-4o and manually select the conditioning frame index.

1083s (± 16.3 s), amounting to a 4.1% increase. Since VisionNarrator and the video generator operate sequentially, the MLLM can be offloaded once video generation begins, making the additional memory overhead minimal. Using Qwen3-VL as the VisionNarrator results in a slight performance drop, though ReasonDiff still outperforms the base Wan2.1 model. However, replacing GPT-4o with LLaVA-1.5-7B leads to a significant degradation. Due to LLaVA’s limited reasoning ability, it often fails to infer meaningful narrative information from the unpaired inputs. This behavior aligns with our expectations: ReasonDiff relies on VisionNarrator to infer a plausible scene narrative and the semantic relationship between the provided image and text. When VisionNarrator cannot uncover these intrinsic connections, the method may fail to generate a video that is semantically coherent with both inputs. Failure cases are analyzed in Section 6.

3.2. Robustness

In Table 2, we report two additional ablation studies evaluating robustness to narrator errors and anchor misplacement. For *Narrator-err.*, we randomly drop 30% of the per-frame prompts with a probability of 50%. For *Anchor-miss.*, we randomly assign misplaced anchors to AlignFormer with the same probability. As expected, ReasonDiff ex-

hibits a slight performance degradation, primarily in Imaging Quality, while other metrics show marginal changes. These results indicate that our method maintains relatively strong robustness under possibly imperfect narration and anchor alignment.

3.3. Reasoning Examples

In this section, we give some detailed reasoning examples of VisionNarrator, and present the results in Figure 1. For instance, given the image of a *broken vase* and the text prompt of *A cat plays in the room*, the VisionNarrator divides the whole scene into four clips: (1) Frame@1-20, with the vase intact at first; (2) Frame@21-45, with the cat entering the room and breaking the vase; (3) Frame@46-70, with the cat fleeing and lastly, (4) Frame@71-81, with the vase already broken, corresponding to the given condition image. From the results we can observe that the VisionNarrator is capable of reasoning out complicated connections and generate a semantically coherent per-frame narrative to guide the generation process.

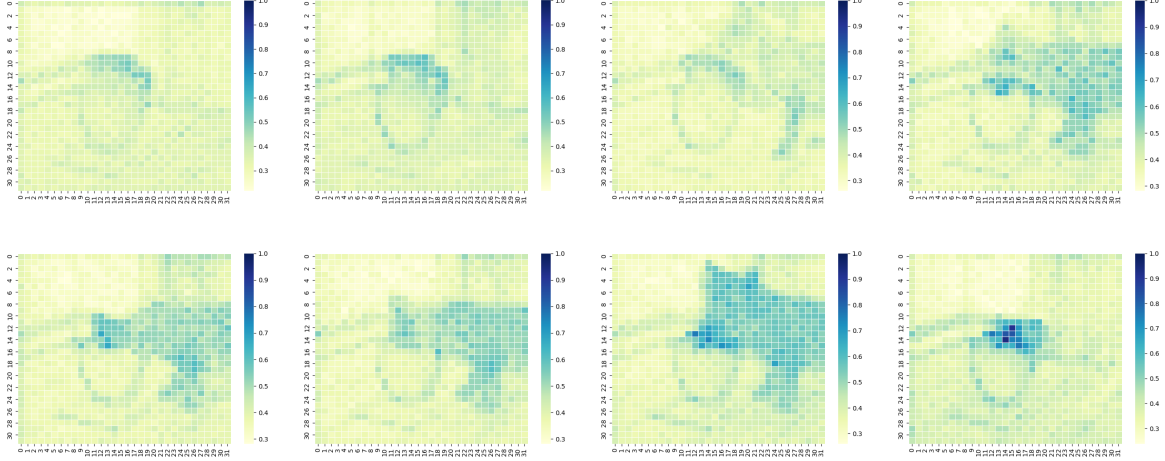


Figure 3. Attention map of the *vase-cat* example in Figure 7(c) (last line, the generated sample from ReasonDiff). Specifically, we compute the attention scores between the CLIP-encoded embedding of the word *breaking* and the hidden states of the video frames in the second transformer block, and illustrate them in time order. The deeper the color, the higher the attention score.

4. Additional Experiments

4.1. Naive approach

In this section, we present more detailed qualitative examples—including additional cases and more frames per case—of the naive baseline, i.e., rewriting the prompt and manually selecting a conditioning frame index for existing video generation models. These examples further support the discussion in Section 4.2 of the main paper. Because Wan 2.1 does not natively support flexible frame conditioning, we apply minimal modifications to enable this functionality for comparison.

Figure 2 shows four examples in which we use GPT-4o to rewrite the initial prompts, allowing it to infer the key connection between the unpaired inputs (*e.g.*, break, extinguish, *etc.*). As the results show, although the rewritten prompts partially capture the overall scene, the base generative model fails to incorporate this high-level reasoning into the actual video synthesis. Consequently, the generated videos often fail to depict the correct interactions and sometimes even produce visually confusing contents (*e.g.*, the deformed car in Example (2) and the unnatural interaction between the hand and the candle in Example (4)). These observations further demonstrate the necessity and effectiveness of our approach for generating videos under OOD scenarios with unpaired multi-modal inputs.

4.2. Attention-map Illustration

In this section, we conduct an additional experiment to show that ReasonDiff can successfully reason out the temporal connections and generate a video coherently, rather than just combining the two modalities. Specifically, considering the *vase-cat* example, we compute the attention scores

between the CLIP-encoded embedding of the word *breaking* and the hidden states of the video frames in the second transformer block. These attention maps are then averaged across channels to obtain a frame-wise attention distribution. The results are illustrated in Figure 3. From the results we can observe that, the areas with high attention scores follow the interaction of the vase and the cat, with little attention scores at first (when the cat hasn’t entered the room). This highlights the model’s ability to infer dynamic causal relationships from unpaired text-image inputs.

4.3. Application: Prompt-driven Customization

In this section, we show one possible application of ReasonDiff as *Prompt-driven Customization*. In video generation, there is often no single ground-truth answer due to the inherent uncertainty of the task. For instance, recovering a broken scene from an image of a shattered vase alone is ambiguous. However, when jointly conditioned on a corresponding prompt, such as *cat plays in the room*, the model can reason that the cat likely causes the vase to break. Since different prompts can imply different plausible causes, the model must be capable of customizing the generated video accordingly. As shown in Figure 4(b), we use the same condition image while varying the subject in the prompt (*e.g.*, cat, dog, or pig), and compare the outputs of ReasonDiff and Wan2.1. We can observe that Wan2.1 fail to uncover the connections between the two modalities, even generating confusing scenes under the subject *pig*. In contrast, ReasonDiff successfully infers the correct relationship between the unpaired input image and text, producing videos that align with the varying prompts and keep narrative coherence.

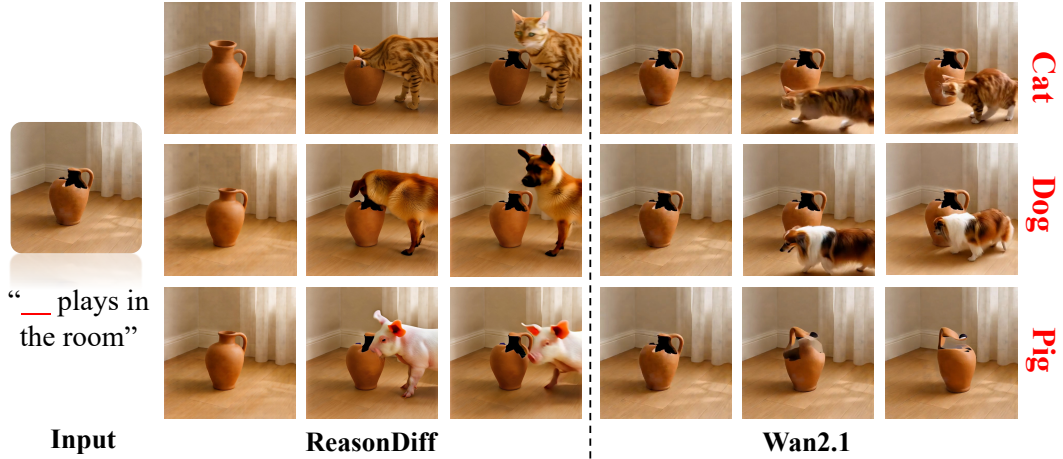


Figure 4. Qualitative comparison between ReasonDiff and Wan2.1 given the same condition image and text prompt with varying subjects, i.e., *Cat/Dog/Pig plays in the room* (causing the vase to be broken).

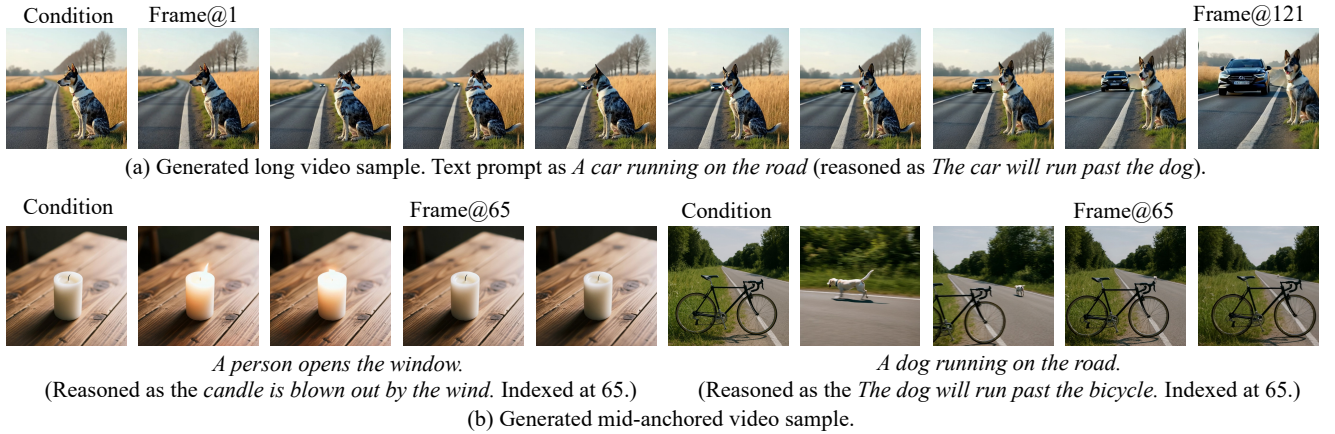


Figure 5. Qualitative examples of ReasonDiff generating two types of videos: (a) longer videos consisting of 121 frames, and (b) videos where the conditioning image is anchored at the middle of the sequence.

5. Extended Results and Visual Comparisons

5.1. Examples for Longer Video and Mid-anchored Video

In the qualitative experiments, we have already shown that our method generalizes to long video generation (81 frames) despite being trained only on 33-frame sequences, and that it works when the conditioning frame is placed at either the beginning or the end of the video. In this subsection, we further evaluate its ability to handle two additional scenarios: (a) generating even longer videos of 121 frames, and (b) generating videos where the conditioning image is anchored in the middle of the sequence. We note that scenario (b) is inherently incorporated in our problem setting, and conditioning at the beginning or end of a video is arguably more challenging than conditioning in the middle, as it creates a larger semantic gap between the multi-modal

inputs. Nevertheless, the results in Figure 5 show that our model continues to perform well when scaled to longer sequences, and it can consistently infer plausible connections between the unpaired inputs, regardless of where the conditioning image is placed, and produce coherent videos.

5.2. Complete Tables and Additional Qualitative Comparisons

In this section, we provide complete tables with standard deviations in Table 3 and Table 4, corresponding to the quantitative comparisons on the self-constructed ActivityNet dataset and the public general-purpose MSR-VTT dataset, respectively. And we give additional qualitative comparisons in Figure 7. As illustrated, we can see that ReasonDiff consistently outperforms baseline models, which often exhibit visual inconsistency, such as object mixing, semantic misalignment, where one of the input con-

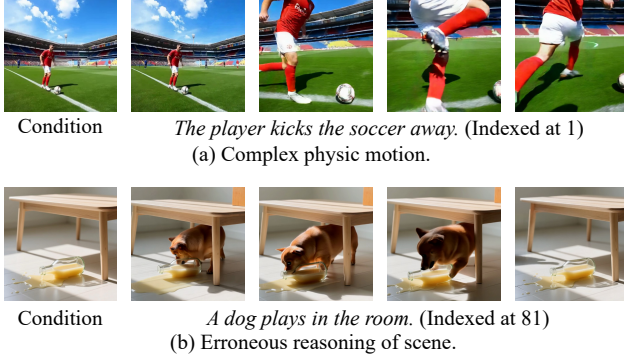


Figure 6. Failure cases. Our method is less effective in (a) generating complex physical motions and (b) scenarios where Vision-Narrator fails to infer the complete narrative.

ditions is ignored, or failing to identify the possible connections between the multi-modal conditions. In contrast, ReasonDiff generates more coherent and semantically aligned video content across diverse scenarios.

6. Discussions

Although video generation under test time out-of-distribution unpaired text-image conditions has not yet been explored in the literature, it holds significant real-world relevance. On the one hand, user-provided conditions cannot be assumed to be perfectly aligned, leading existing approaches to perform suboptimally as they generally fail to reason across modalities. On the other hand, in tasks that inherently require reasoning, users may deliberately provide unpaired conditions—for example, to recover past scenes or to predict future ones. Consequently, the reasoning capability introduced by our proposed ReasonDiff is essential, enabling more robust and contextually coherent video generation. Moreover, although our method is designed to take a single image as the condition for one frame, it can be seamlessly extended to multi-image conditioning without requiring changes to the core framework.

However, our method still has several limitations. First, because it is built on Wan2.1 as the base video generative model, it inherits common shortcomings of current generative approaches. For instance, it has difficulty producing intricate structures such as detailed human poses or complex object interactions, *e.g.*, shaking hands or kicking a football, which are challenging to maintain consistently across frames (see Figure 6 (a)). Second, because our method relies on VisionNarrator to construct a plausible scene, it may degenerate into superficially mixing the multi-modal inputs when VisionNarrator fails to uncover their intrinsic connections (see Figure 6 (b)). A potential solution may be to jointly fine-tune VisionNarrator to improve its reasoning under unpaired scenarios, which we leave for future explo-

ration.

References

- [1] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1728–1738, 2021. 1
- [2] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961–970, 2015. 2
- [3] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024. 2
- [4] 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*, pages 21807–21818, 2024. 2
- [5] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 1
- [6] 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*, pages 8748–8763. PmLR, 2021. 3
- [7] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 1, 2
- [8] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pages 399–417. Springer, 2024. 2
- [9] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 2
- [10] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 2

Table 3. Complete table with standard deviation for the quantitative comparison between ReasonDiff and the baselines on the self-constructed ActivityNet dataset that simulates unpaired settings. The top and second top performances have been bolded or underlined respectively.

Model	Imaging Quality(\uparrow)	Motion Smooth(\uparrow)	Dynamic Degree(\uparrow)	CLIP Score (Text)(\uparrow)	CLIP Score (Image)(\uparrow)	User Rank(\downarrow)
Dynamicrafter	0.492 ± 0.111	0.979 ± 0.019	0.484 ± 0.499	0.202 ± 0.057	0.508 ± 0.087	2.871 ± 1.239
LTX-Video	0.398 ± 0.081	0.977 ± 0.008	0.734 ± 0.442	0.211 ± 0.051	0.544 ± 0.084	3.307 ± 1.259
CogVideoX	0.507 ± 0.086	0.949 ± 0.023	<u>0.872</u> ± 0.089	0.197 ± 0.039	<u>0.537</u> ± 0.078	4.384 ± 1.041
Wan2.1	<u>0.512</u> ± 0.103	<u>0.980</u> ± 0.023	0.810 ± 0.280	<u>0.224</u> ± 0.056	0.518 ± 0.079	<u>2.692</u> ± 1.079
ReasonDiff	0.528 ± 0.106	0.986 ± 0.048	0.936 ± 0.244	0.261 ± 0.061	0.528 ± 0.082	1.743 ± 1.044

Table 4. Complete table with standard deviation for the quantitative comparison between ReasonDiff and the baselines on the public and general-purpose MSR-VTT dataset. The top and second top performances have been bolded or underlined respectively.

Model	Imaging Quality(\uparrow)	Motion Smooth(\uparrow)	Dynamic Degree(\uparrow)	CLIP Score (Text)(\uparrow)	CLIP Score (Image)(\uparrow)	User Rank(\downarrow)
Dynamicrafter	0.517 ± 0.123	<u>0.984</u> ± 0.017	0.440 ± 0.496	0.201 ± 0.043	0.526 ± 0.091	3.179 ± 1.189
LTX-Video	0.406 ± 0.087	0.986 ± 0.010	0.695 ± 0.460	<u>0.206</u> ± 0.037	0.588 ± 0.078	4.051 ± 0.971
CogVideo	<u>0.552</u> ± 0.082	0.970 ± 0.015	<u>0.688</u> ± 0.323	0.177 ± 0.025	<u>0.572</u> ± 0.059	3.256 ± 1.481
Wan2.1	0.560 ± 0.111	0.962 ± 0.023	0.665 ± 0.183	0.191 ± 0.036	0.552 ± 0.075	<u>2.743</u> ± 1.140
ReasonDiff	0.571 ± 0.109	<u>0.984</u> ± 0.028	0.673 ± 0.470	0.214 ± 0.044	<u>0.572</u> ± 0.092	1.769 ± 1.245



Figure 7. Additional qualitative comparison between ReasonDiff and the baselines. We select several intermediate frames for the convenience of presentation.