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

Zirui Pan[1], Xin Wang[1,2*], Yipeng Zhang[1], Hong Chen[1], Kecheng Zheng[3], Wenwu Zhu[1,2*]
[1]Department of Computer Science and Technology, Tsinghua University,
[2]BNRIST, Tsinghua University, [3]Ant Research
{pzr24,zhang-yp22,h-chen20}@mails.tsinghua.edu.cn, zkechengzk@gmail.com
{xin_wang,wwzhu}@tsinghua.edu.cn

Abstract

Text-image to video generation aims to synthesize a video conditioned on the given text-image inputs. Nevertheless, existing methods generally assume that the semantic information carried in the input text and image tends to be perfectly paired and temporally aligned, occurring simultaneously in the generated video. As such, existing literature struggles with out-of-distribution (OOD) "unpaired" text-image inputs in the more universal and realistic scenario where i) the semantic information carried by the text and image may occur at different timestamps and ii) the condition image can appear at an arbitrary position rather than the first frame of the synthesized video. Video generation under this OOD setting poses an urgent need to conduct reasoning over the intrinsic connections between the given textual description and referred image, which is challenging and remains unexplored. To address the challenge, in this paper we study the problem of unpaired text-image to video generation for the first time, proposing ReasonDiff, a novel model for accurate video generation from unpaired text-image inputs. Specifically, ReasonDiff designs a VisionNarrator module to harness the powerful reasoning abilities of a multi-modal LLM to analyze the unpaired text-image inputs, producing coherent per-frame narratives that temporally align them. Building upon this VisionNarrator module, ReasonDiff further introduces a novel AlignFormer module, which employs a Multi-stage Temporal Anchor Attention mechanism to predict frame-wise latent representations. These reasoning-enhanced latents are subsequently fused with the condition frame, providing structured guidance throughout the video generation process. Extensive experiments and ablation studies demonstrate that ReasonDiff beats state-of-the-art baselines in terms of video generation quality with unpaired text-image inputs.

*Corresponding Authors.

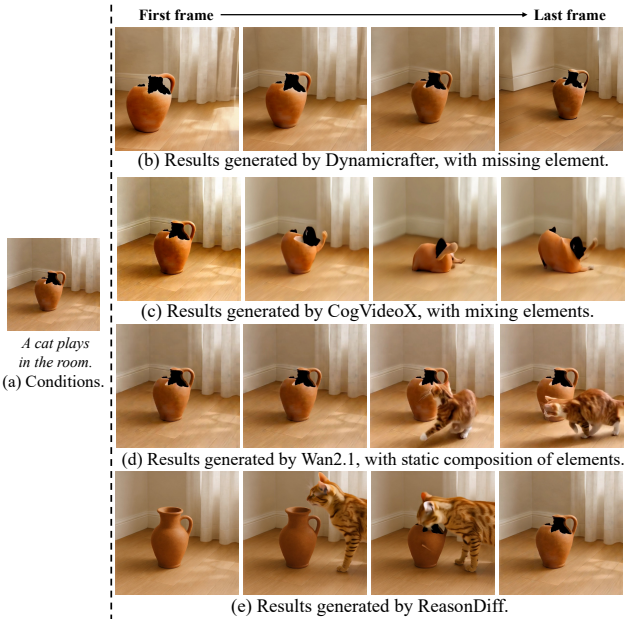

Figure 1. Comparison of generated results from different models under OOD scenario with unpaired text-image inputs: i) the textual prompt *A cat plays in the room* and ii) the visual condition image *a broken vase* in Figure 1(a). Intermediate frames are selected for the convenience of presentation. Our proposed ReasonDiff has the best result with a visually and semantically coherent video.

1. Introduction

Video generative models [3, 4, 7, 12, 13, 15, 39, 49] have emerged as powerful tools to produce high quality videos by iteratively refining noise through a stochastic process. By leveraging powerful backbone architectures such as Diffusion Transformer (DiT) [33] or U-Net [34], these models excel at capturing complex dynamics, enabling a wide range of generative tasks. Among these tasks, text-image to video generation [6, 17, 32, 38, 44] has become particularly important, as it targets at synthesizing videos that faithfully reflect both the given image and textual inputs.

However, most existing models heavily rely on the assumption that the semantic information carried in the conditioned input image and text are perfectly paired and temporally aligned, following a paradigm where both modalities describe the event with the same semantic meaning and the generated video is expected to begin with the input image (serving as the first frame). This strong reliance on paired text-image inputs limits the flexibility of these models, being impractical in out-of-distribution (OOD) real-world scenarios where such alignment is often absent because the user-provided conditions may not be inherently paired. For instance, the "unpaired" scenario may occur when there exist time differences between the events described by the image and text, which leads to the seeming unrelatedness of the semantic information carried in the two modalities. This may result in a failure to reason about their temporal connections and to bridge the semantic gap between the two modalities. As such, existing approaches will struggle to generate a coherent video both visually and semantically when encountering unpaired text-image inputs. Consider the example illustrated in Figure 1, where the model is expected to generate a video based on the unpaired inputs, *i.e.*, input text prompt as *A cat plays in the room* and input condition image as *a broken vase*. The semantic meaning carried in these two inputs may seem unrelated, but they imply an underlying connection that *the vase is broken by the cat*. When given these conditions, the model is required to recover the whole scene, and the most plausible position for the condition image is somewhere near the end of the scene. Figure 1(b) shows that existing methods tend to be dominated by one of the conditions, most frequently the image, and losing critical elements described in the text prompt. Figures 1(c) and (d) further highlight failure cases in which the model fails to establish meaningful relationships between the two conditions. In such cases, the outputs either represent a superficial blending of modalities or a mere juxtaposition of elements, ultimately producing videos that lack semantic coherence and visual clarity. In contrast, the video generated by our method, shown in Figure 1(e), faithfully adheres to the given conditions: the vase remains intact initially and only breaks after interacting with the cat. This problem is challenging, as it requires inferring a plausible scene from the test time out-of-distribution unpaired text-image inputs, while also integrating the high-level reasoning information into the generation process.

To tackle the above challenges, in this paper we propose a novel ReasonDiff model for unpaired text-image to video generation, for the first time. Specifically, to analyze the intrinsic connections between the given conditions, we design a VisionNarrator module to leverage the strong reasoning capabilities of a multi-modal large language model (MLLM), and generate a plausible per-frame narrative to recover the whole scene, temporally aligning the unpaired modalities. The VisionNarrator first infers the most likely position of the condition image within the final video, enabling more accurate and context-aware generation. To bridge the reasoning outputs with the generation process, we introduce the AlignFormer module, which treats the condition image as an anchor and predicts the latent representations for the remaining frames. Concretely, AlignFormer employs a Multi-stage Temporal Anchor Attention mechanism that progressively refines latent representations through a cascade of cross-attention layers, effectively injecting reasoning signals into the feature space. The resulting reasoning-enhanced latents are then fused with the condition frame, providing precise, frame-wise control. During the training stage, we will first warm up the whole model using the standard denoising loss, and then add an auxiliary reconstruction loss between the predicted reasoning enhanced latents and the matching ground-truth latents to finetune the AlignFormer module individually. In this way, the ReasonDiff model is able to reason out the possible scene from the seemingly unrelated conditions, and generate a video that is realistic and semantically-coherent with both the inputs. We summarize our contributions as follows:

- To the best of our knowledge, we for the first time propose to solve the challenging problem of unpaired text-image to video generation.
- To tackle the challenges in the above problem, we propose an MLLM Driven Multi-frame Reasoner, comprising two key components, namely VisionNarrator and AlignFormer, which derives a per-frame narrative that is coherent with the unpaired inputs and predicts latent representations for unseen frames, respectively.
- We design a Reasoning Guided Generative Model to empower the base video generative model with reasoning abilities and propose an end-to-end training procedure under unpaired text-image inputs.
- We conduct extensive experiments and ablation studies to verify the strong reasoning and generating abilities of the proposed ReasonDiff model.

2. Related Work

Video Generative Models. Diffusion models have become a powerful framework for video synthesis, producing realistic and temporally coherent results. By extending DDPM [14] to text/image-to-video tasks, they incorporate temporal dynamics and learn motion priors from large-scale datasets like WebVid [2]. Flow matching [26] later reframes this as a distribution mapping problem, offering a stronger training objective. Early work such as Video Diffusion Model [15] suffers from low resolution problems. Subsequent methods [4, 30, 46] leverage spatial-temporal upsampling to enhance video quality.

To achieve a better controllable generation, recent works have studied to add various condition signals [45, 47], such

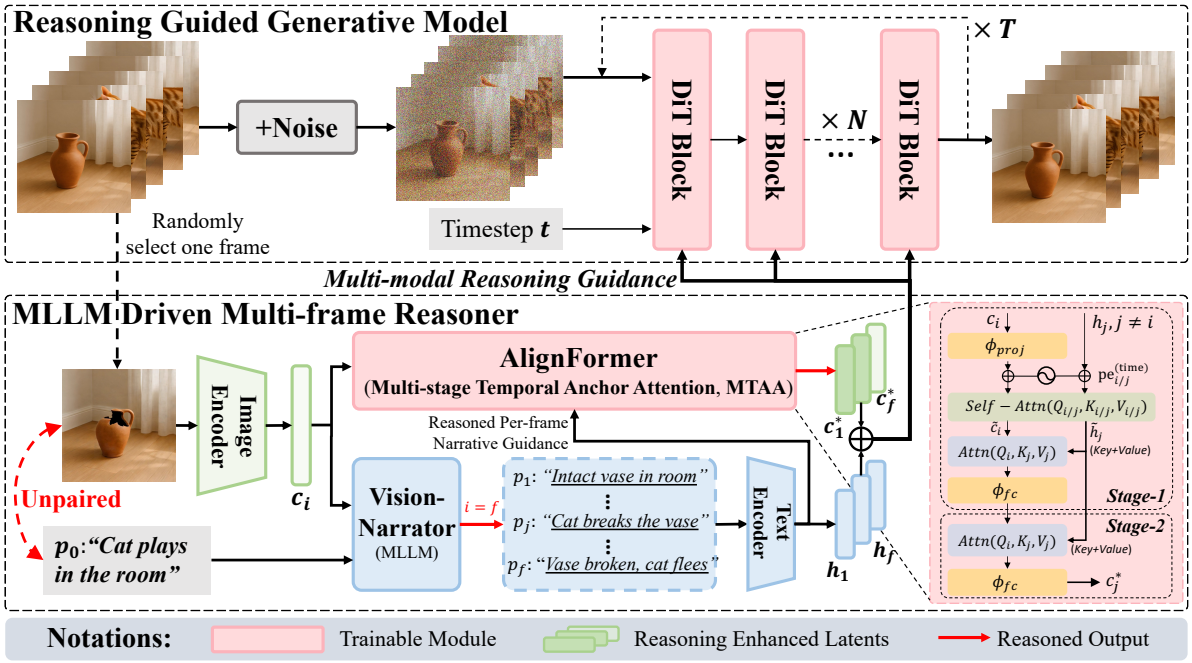

Figure 2. Overview of the ReasonDiff model, which consists of two key components: (1) the *MLLM-Driven Multi-frame Reasoner*, and (2) the *Reasoning-Guided Generative Model*. The generative model operates under the guidance of the multi-modal reasoning results.

as poses [31, 48] and structures [8, 41]. Specifically, for (text-)image-to-video generation, Dynamicrafter [42] fuses the condition information with the initial noise and proposes spatial dual-attn transformer module to support more precise conditioning. LTX-Video [13] seamlessly integrate Video-VAE into denoising transformers, and optimize their interaction for improved efficiency and quality. And more recently, large-scale video generation models such as CogVideoX [44] and Wan2.1 [38] use DiT as backbone and can generate highly-realistic videos. Nonetheless, these models typically assume that the input text and image are perfectly paired, implicitly relying on the alignment between the two modalities to guide the generation process. In OOD scenarios where the text and image are loosely related or entirely unpaired, such models often fail to reason about their intrinsic connections. This leads to generated videos that lack semantic coherence and exhibit poor visual consistency, limiting their applicability in real-world, weakly supervised settings where perfect alignment is rare.

Generation with Reasoning. Recently, with the emergence of ChatGPT [1] and other language models [10, 27], researches on LLMs have gained significant momentum. In particular, there has been growing interest in exploring the reasoning capabilities inherent in these models. Some researches [35, 36, 40, 50] prove the reasoning abilities can be enhanced through prompt refinement. Zero-shot-CoT [21] achieves performance boost by simply adding *Let's think step by step* before each answer. MM-CoT [51] incorporates language and vision modalities into a two-stage framework that separates rationale generation and answer inference.

For generative models, LLM has been widely used to infer additional information from the given inputs, such as scene layout [16, 18, 24, 29] and object relationship [22, 23, 37]. VQAI [23] introduces casual reasoning in image generation, and extends the visual question answering tasks to include image as answer. SmartEdit [19] addresses instruction-based image editing problem with the reasoning abilities of LLMs and bidirectional information interactions. And regarding video generation, LayoutGPT [9] leverages LLMs to generate detailed scene descriptions along with multiple bounding boxes. Similarly, VideoDirectorGPT [25] enhances the controllability of generation by incorporating scene descriptions and layout information produced by LLMs. However, these studies primarily augment the textual inputs to include more details, yet neglect the possible intricate connections between multiple input modalities and fail to incorporate multi-modal reasoning guidance, limiting their ability to address unpaired inputs.

3. Method

In this section, we will describe our proposed ReasonDiff method in detail. The overall framework is illustrated in Figure 2. It mainly consists of a base Reasoning Guided Generative Model which is built upon Wan2.1 [38], a VisionNarrator and an AlignFormer.

3.1. Preliminary

Flow matching Diffusion models have seen widespread applications in AIGC scenarios, and Flow matching [26] has become the standard training objective for existing generation models using DiT as the backbone. Flow matching extends DDPM and learns the mapping between two distributions. Specifically, given data $x_1 \sim q(x)$ and gaussian noise $x_0 \sim \mathcal{N}(0, 1)$, the model is optimized to transform x_0 into x_1 via predicting the velocity field, *i.e.*,

$$\mathcal{L} = \mathbb{E}_{x_1, x_0 \sim \mathcal{N}(0,1), y, t \sim \mathcal{U}(0,1)} \left[||u_\theta(x_t, y, t) - v(x_t)||_2^2 \right], \tag{1}$$

where t is the timestep, $x_t = tx_1 + (1-t)x_0$ is an intermediate noisy latent, y represents an optional conditioning signal, and $u_\theta(\cdot)$ is the denoising model parameterized by θ. $v(x_t)$ is the conditional velocity field, namely,

$$v(x_t) = v(x_t \mid x_1) = x_1 - x_0. \tag{2}$$

In terms of video generation, the widely adopted approach is to treat the video as a sequence of images and perform self-attention along the temporal axis to learn the motion priors. Thus a typical T2I generative model with flow matching objective can be extended to a video generative model after appropriate fine-tuning on video datasets.

Task This paper addresses the problem of test time out-of-distribution unpaired text-image to video generation, *i.e.*, generating a sequence of video frames $x \in \mathbb{R}^{ch \times f \times h \times w}$ given a text prompt p_0 and an image $y \in \mathbb{R}^{ch \times h \times w}$, such that the output is semantically coherent with both inputs, where ch, f, h and w represents the channel, frame, height and width, respectively. Importantly, the text and image are unpaired, *i.e.*, the image is not guaranteed to share the same semantic information with the prompt, nor is it necessarily the first frame of the target video. As a result, directly applying existing video generative models often yields suboptimal results due to their limited ability to reason over loosely aligned or entirely unpaired multi-modal inputs.

3.2. VisionNarrator

Given unpaired image and text conditions, existing video generative models struggle to infer a coherent narrative, often failing to generate semantically consistent videos. To address this limitation, we propose VisionNarrator and leverage the strong reasoning capabilities of a multi-modal large language model to analyze the underlying connections between the two modalities. Specifically, the MLLM anchors the condition image to a plausible index and constructs a frame-by-frame narrative around this anchor, guided by the combined context of image and text. To achieve this, we design the following prompt, *i.e.*,

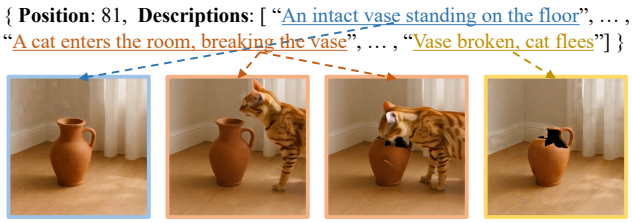

Figure 3. Reasoning results generated by the VisionNarrator. The conditions are the same as in Figure 1. We select some key frames and connect them with the related prompts using different colors.

- *You are given an unpaired image and text prompt. Your task is to infer a coherent scene that logically connects both inputs, even if they appear unrelated.*
- *Estimate the most likely position of the image within an* `f`*-frame video. Then, generate descriptions with rich information for each of the* `f` *frames that together form a consistent video script.*
- *Respond strictly in the following format:* {*"position":* `j`*, "descriptions": [description for frame 1, ..., description for frame* `f`*]*}*. Do not include any additional explanations, comments, or formatting*.

This serves as a general instruction and ensures that the output will adhere to the specified format. In practice, we apply in-context learning [28] to further stabilize the results.

Consider the example in Figure 1, where the input image shows a broken vase and the text prompt describes a cat playing in the room. The reasoning results produced by the VisionNarrator are presented in Figure 3, with the deduced prompts aligned to their corresponding frames, each highlighted in a different color. We select some key prompts in the generated storyline, namely, *Intact vase* $\rightarrow$ *Cat enters the room, breaking the vase* $\rightarrow$ *Vase broken, cat flees*, together with their corresponding video frames. We can see that the narrative forms a plausible deduction based on the given unpaired inputs and naturally positions the condition image as the final frame. As illustrated, the VisionNarrator effectively infers the underlying narrative, that the playful cat causes the vase to break, and generates frame descriptions that are semantically aligned with this inferred storyline. The resulting per-frame script, along with the predicted anchor position of the condition image, is then passed to AlignFormer for further processing. VisionNarrator differs from existing approaches that rely on MLLMs primarily to expand textual prompts both in terms of motivation and technical contribution. It aims to reason across modalities to achieve temporal alignment of unpaired inputs and to provide effective multi-modal guidance.

3.3. AlignFormer

To bridge the gap between the VisionNarrator and the base generative model, we introduce the AlignFormer module, a Transformer-style architecture that aligns the high-level

reasoning outputs with the frame-wise latent features.

The AlignFormer module takes three inputs: (1) the anchor feature c_i extracted from the condition frame, (2) its inferred position i within the target video, and (3) the reasoned per-frame narrative embedding $h = \{h_j\}_{1 \le j \le f}$. It then outputs a sequence of reasoning enhanced latent features, *i.e.*, $c^* = \{c_j^*\}_{1 \le j \le f}$, corresponding to each frame. The structure is illustrated in lower-right region of Figure 2.

In particular, the module employs a Multi-stage Temporal Anchor Attention (MTAA) mechanism to progressively synthesize each frame's latent representation by integrating the anchor feature with temporally structured semantic guidance derived from per-frame narratives. This is achieved through a two-stage cross-attention process: the first stage is designed to capture coarse temporal dependencies, while the second stage refines the representations with finer contextual alignment. At each stage, the anchor feature acts as the *Query*, while the corresponding narrative embeddings are projected to form *Key* and *Value* features:

$$\tilde{c}_i = \phi_{\text{proj}}\left(\text{Flatten}(c_i)\right) + \text{pe}_i^{(\text{time})},\ \tilde{h}_j = h_j + \text{pe}_j^{(\text{time})}, \quad (3)$$

$$\text{Q}_i = \text{W}_\text{Q}\tilde{c}_i,\ \text{K}_j = \text{W}_\text{K}\tilde{h}_j,\ \text{V}_j = \text{W}_\text{V}\tilde{h}_j, \quad (4)$$

$$c_j^* = \text{Attn}(\text{Q}_i, \text{K}_j, \text{V}_j) = \text{Softmax}\left(\text{Q}_i\text{K}_j^T/\sqrt{d}\right)\text{V}_j, \quad (5)$$

where $j \neq i$ is the index for the predicted latent feature, W_Q, W_K and W_V are the projection matrices for *Query*, *Key* and *Value*, and $\text{pe}_{i/j}^{(\text{time})}$ represents positional embedding along the temporal axis. This MTAA mechanism adopted in AlignFormer helps to effectively align the visual and textual representations, enabling the transfer of high-level reasoning. The resulting reasoning-enhanced latents, denoted as c^*, together with the prompt embedding h, are then fused with the condition frame c_i to serve as guidance throughout the denoising steps of the generation process. Compared to directly injecting the multi-frame prompts without AlignFormer, our ReasonDiff model, equipped with this module, achieves noticeably better generation quality and temporal coherence, highlighting the importance of multi-modal guidance. (Please see Section 4.3 for more details).

3.4. Training Procedure

In this subsection, we describe the training procedure of the proposed ReasonDiff model. A key challenge lies in the scarcity of multi-modal datasets featuring unpaired text-image conditions. Since most existing works assume paired inputs in video generation, there is currently no available dataset that can serve as direct ground truth for training ReasonDiff. To address this, we set the VisionNarrator frozen and reformulate our training task into a conditional video generation problem, where the model reconstructs a video clip $x \in \mathbb{R}^{b \times ch \times f \times h \times w}$ given a randomly selected condition frame indexed $i \in \{1, \ldots, f\}$ and a corresponding sequence of per-frame narrative embeddings $h \in \mathbb{R}^{f \times l \times d}$, where b, l and d represents the batch size, context length and embedding dimension within the text encoder. Consequently, the only trainable components in our framework become the base video generative model and the newly introduced AlignFormer module. Moreover, to better simulate the OOD and unpaired condition scenario, we increase the temporal interval between selected frames in each video clip, ensuring that the condition frame appears less correlated with the surrounding contents. In this way, we can effectively train our model based on existing video datasets.

Concretely, we use video data from WebVid dataset [2], sampling frames at 0.2 second intervals. For each frame, we generate a corresponding caption using LLaMA-3.2-11B-Vision-Instruct [11]. During training, a random frame is selected as the condition frame, and the model is trained to reconstruct the entire video based on this frame and the generated per-frame textual descriptions.

Our training procedure consists of two stages. In the first stage, we jointly train the pre-trained base video generative model and the AlignFormer using a standard denoising loss. This phase serves to initialize the newly added module and align it with the flow-based generation process. In the second stage, we introduce an auxiliary reconstruction loss between the predicted latent features c^* and the ground-truth latent features c, encouraging the model to better align generated representations with the original video contents. Specifically, the second-stage loss is defined as follows:

$$\mathcal{L} = \mathbb{E}_{x_1, x_0, h, t, c}\left[||u_\theta(x_t, h, c^*) - v(x_t)||_2^2 + \beta \cdot ||c^* - c||_2^2\right] \quad (6)$$

where β is a hyper-parameter controlling the weight of the auxiliary loss, and we keep the parameters of the base generative model fixed to fine-tune the AlignFormer alone. In practice, we set $\beta = 0.2$. Since the auxiliary loss diverges from the original denoising objective of the base generative model, it is applied exclusively during AlignFormer fine-tuning rather than in the first stage.

4. Experiment

In this section, we first detail on the specific settings of the proposed ReasonDiff, and conduct extensive quantitative/qualitative experiments against baselines and ablation studies under unpaired text-image inputs.

4.1. Experimental Setup

Since our work targets video generation under out-of-distribution unpaired text-image conditions, we construct a custom evaluation dataset to align with this objective. Specifically, we randomly sample 500 videos from ActivityNet [5] and extract a 16-frame clip from each. For each clip, we select either the first or the last frame as the condition image and use LLaMA-3.2-11B-Vision-Instruct to gen-

Table 1. Quantitative comparison between ReasonDiff and the baselines. The top and second top performances have been bolded or underlined respectively. Complete table with standard errors can be found in the supplementary materials.

Dataset	Model	Imaging Quality(↑)	Motion Smooth(↑)	Dynamic Degree(↑)	CLIP Score (Text)(↑)	CLIP Score (Image)(↑)	User Rank(↓)
ActivityNet (Self-constructed with **unpaired** inputs)	Dynamicrafter	0.492	0.979	0.484	0.202	0.508	2.871
	LTX-Video	0.398	0.977	0.734	0.211	**0.544**	3.307
	CogVideoX	0.507	0.949	<u>0.872</u>	0.197	<u>0.537</u>	4.384
	Wan2.1	<u>0.512</u>	<u>0.980</u>	0.810	<u>0.224</u>	0.518	<u>2.692</u>
	ReasonDiff	**0.528**	**0.986**	**0.936**	**0.261**	0.528	**1.743**
MSR-VTT (Public-General-purpose with **paired** inputs)	Dynamicrafter	0.517	<u>0.984</u>	0.440	0.201	0.526	3.179
	LTX-Video	0.406	**0.986**	**0.695**	<u>0.206</u>	**0.588**	4.051
	CogVideoX	<u>0.552</u>	0.970	<u>0.688</u>	0.177	<u>0.572</u>	3.256
	Wan2.1	0.560	0.962	0.665	0.191	0.552	<u>2.743</u>
	ReasonDiff	**0.571**	<u>0.984</u>	0.673	**0.214**	<u>0.572</u>	**1.769**

erate a caption for the opposite end (the last or first frame, respectively) as the prompt. This setup ensures a temporal separation between the two conditions, thereby better simulating an unpaired scenario. Importantly, the model will perform generation without access to the frame index or the relative temporal position between the given image and text. In addition, we incorporate a public general-purpose dataset MSR-VTT [43], which contains paired conditions, to enable a more comprehensive comparison and to further assess the general generative ability of the method.

We compare our proposed ReasonDiff model with the following baselines: (1) Dynamicrafter [42], (2) LTX-Video-2B [13], (3) CogVideoX-1.5-5B [44] and (4) Wan2.1-I2V-14B [38], which are latest works on video generation that have been open-sourced and achieve good performances. We have employed six metrics, namely *Imaging Quality*, *Motion Smooth*, *Dynamic Degree*, *CLIP Score (Text/Image)* and *User Rank*. Note that the first three metrics are general evaluation criteria supported by VBench [20], and the two *CLIP Score*s quantify the semantic alignment between the generated video and the relevant conditions (text/image). Details can be found in the supplement.

4.2. Main Results

We conduct both quantitative and qualitative experiments, with the results presented in Table 1 and Figure 4. In quantitative evaluations, ReasonDiff achieves competitive performance compared to state-of-the-art baselines on the general-purpose dataset MSR-VTT (*i.e.*, with paired conditions), even surpassing its base model, Wan2.1. This result is expected, as real-world datasets are rarely perfectly aligned; thus, cross-modal reasoning naturally contributes to improved general video quality. Meanwhile, ReasonDiff achieves top results across all metrics except for *CLIP Score (Image)* on the self-constructed ActivityNet dataset that simulates unpaired multi-modal inputs. Notably, ReasonDiff substantially outperforms all baselines in *CLIP Score (Text)* and *User Rank* on ActivityNet, which are the two most critical metrics for evaluating model's reasoning performance with unpaired image and text. Specifically, ReasonDiff exceeds the best-performing baseline, Wan2.1, by 16.5% in *CLIP Score (Text)*, and by 0.949 in *User Rank*. As for *CLIP Score (Image)*, all methods show comparable performances, with scores hovering around 0.5. This is consistent with the observation in Figure 1, where baseline models tend to rely heavily on the input image when confronted with unpaired conditions. As a result, their *CLIP Score (Image)* remains relatively high, as well as ReasonDiff's, making this metric less discriminative for unpaired settings. However, when jointly considering both *CLIP Score (Text/Image)* metrics, it becomes evident that the baselines lack the capability to integrate and reason across modalities. They focus primarily on visual conditions while neglecting the semantic guidance, leading to poor alignment with the textual input and ultimately less coherent video generation. In contrast, ReasonDiff achieves high scores across both metrics, reflecting its robust reasoning capabilities to generate semantically rich and coherent videos even under unpaired text-image inputs.

In qualitative experiments, we evaluate ReasonDiff and the baselines using the same unpaired input image and text. The results are presented in Figure 4. As shown, baselines generally struggle to infer meaningful connections between the unpaired conditions, resulting in two major types of failures: (1) confusing content, where the generated frames contain entangled visual elements due to conflicting signals from the unpaired inputs (*e.g.*, outputs from CogVideoX in Figure 4(a), which shows bizarre interaction of the *hand*

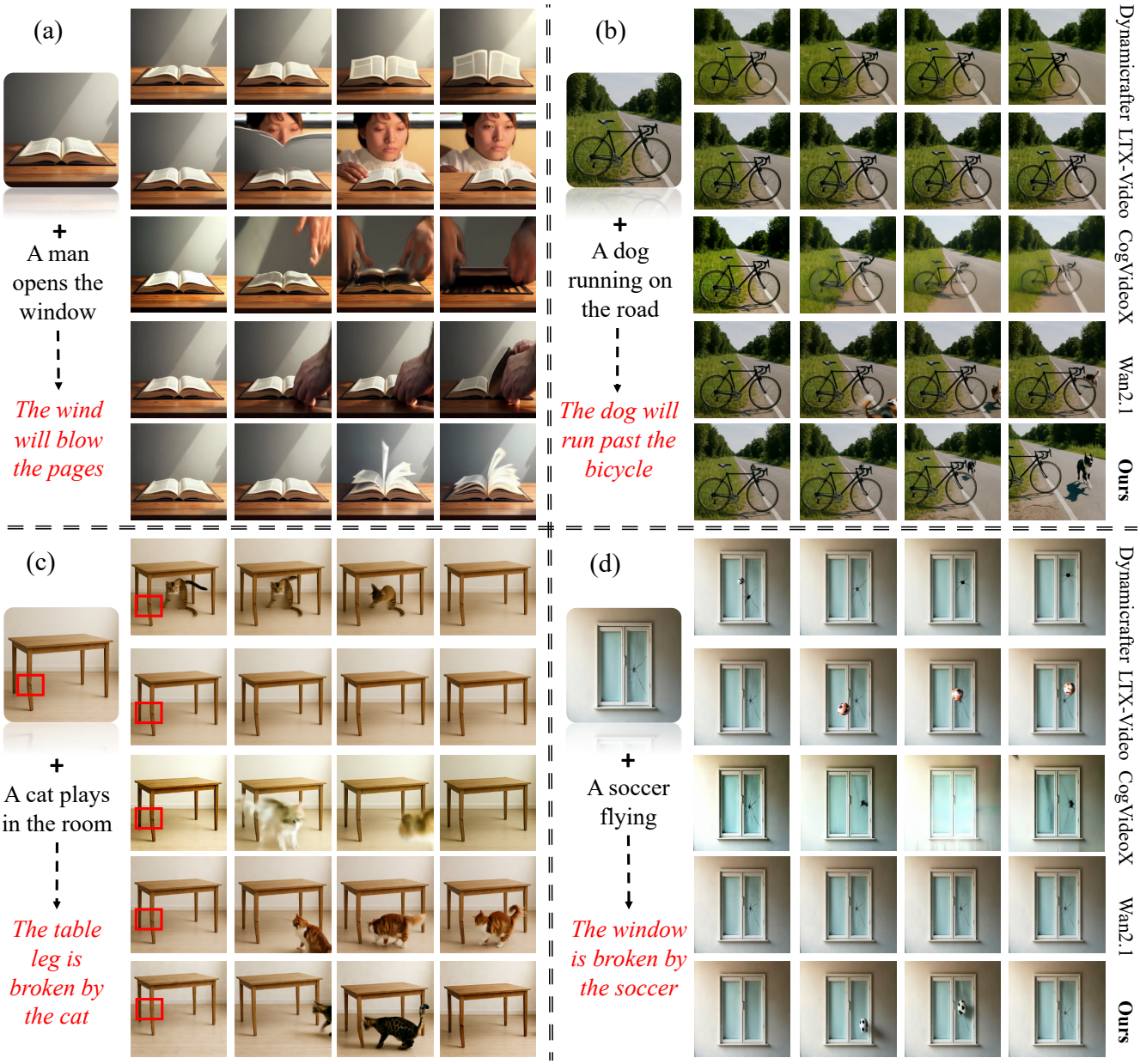

Figure 4. Qualitative comparison between ReasonDiff and the baselines. We select several intermediate frames for the convenience of presentation. For ease of illustration, we provide more generated results and the video samples in the supplement.

and *book*); and (2) incoherence with the unpaired conditions, where the expected motions implied by the prompt do not occur (*e.g.*, outputs from Dynamicrafter in Figure 4(b), LTX-Video in Figure 4(c) and Wan2.1 in Figure 4(d), which fail to depict *dog running past the bicycle*, *cat breaks the table leg* or *soccer breaks the window*, respectively). In contrast, ReasonDiff generates videos that faithfully reflect the semantic intent of both the image and the text, demonstrating a strong ability to reason over unpaired text–image inputs in test time out-of-distribution scenarios.

4.3. Ablation Studies

In this section, we evaluate the effectiveness of the proposed modules through comprehensive ablation studies. We design four variants of the full ReasonDiff model, namely, (1) *w/o. Aux. loss*, which removes the second training stage and disables the auxiliary reconstruction loss; (2) *w/o. Multi. prompt*, which uses only the single user-provided prompt without the per-frame narratives from VisionNarrator; (3) *w/o. Enhanced latents*, which disables the enhanced latents, relying solely on narrative guidance; and (4)

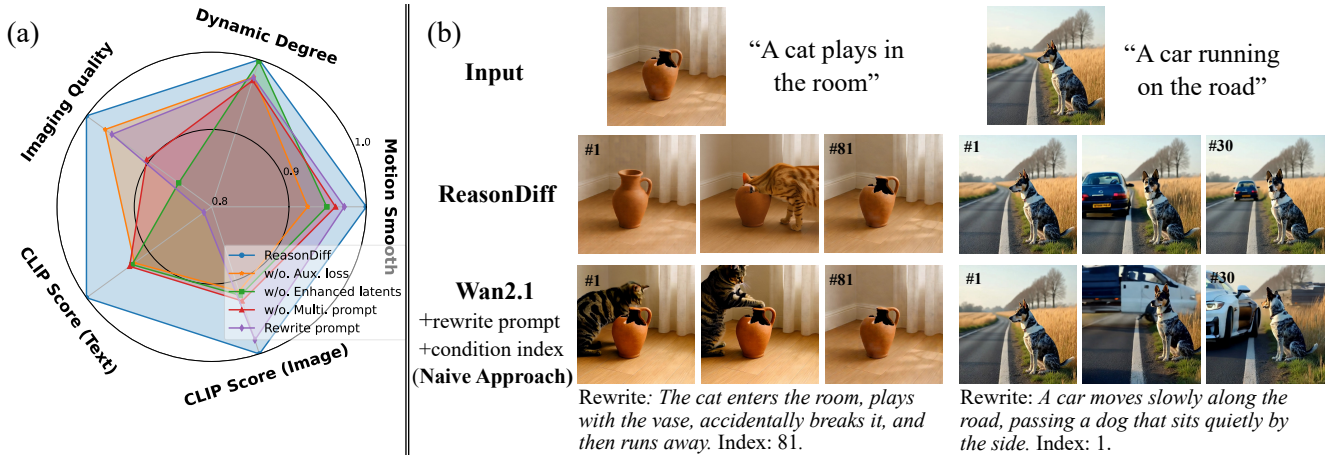

Figure 5. (a) Ablation studies on four variants of ReasonDiff. All metrics are reported as ratios relative to the full model. (b) Comparison between ReasonDiff and Wan2.1, where for Wan2.1 the prompt is rewritten using an MLLM and the condition index is manually selected.

Rewrite prompt, which employs an MLLM to rewrite the text prompt based on the given condition image, thereby assisting the reasoning process of the base model. For variants (1)–(3), the training procedure of ReasonDiff is modified to isolate the contribution of each module. And in variants (2) and (3), the same procedure is also applied during inference, where the base generative model receives only the enhanced latents or the multi-frame narratives generated by VisionNarrator, respectively. Arguably, we should compare with an additional variant that directly fine-tunes the base model using unpaired inputs. However, we consider variant (3) to be a more meaningful comparison, as the base model in variant (3) is already provided with enriched and paired multi-frame narratives. While for variant (4), the rewritten prompt is directly supplied while keeping the pre-trained base model unchanged. We compare the performance of ReasonDiff against these four variants on ActivityNet, and present the quantitative results in Figure 5(a). Specifically, we can derive the following conclusions:

1) **Variant (1)**: Removing *Auxiliary Loss* degrades the overall performance throughtout all measured dimensions, especially *Motion Smooth*, confirming its role to help stabilize the predicted enhanced latents.
2) **Variant (2)**: Disabling *Multi-frame Prompt* leads to notable declines in *Dynamic Degree*, underscoring the importance of fine-grained temporal guidance provided by the multi-frame narratives.
3) **Variant (3)**: Excluding *Enhanced Latents* leads to a substantial degradation in *Imaging Quality*, indicating that multi-frame narratives alone are insufficient for maintaining visual consistency.
4) **Variant (4)**: The *Rewrite Prompt* variant obtains relatively high *CLIP-Image* but suffers a severe drop in *CLIP-Text*, suggesting that the base model struggles to decompose textual information across frames. Consequently, it relies more heavily on visual condition for unseen frames, leading to weaker reasoning over unpaired inputs.

Generally, ReasonDiff consistently surpasses all ablated variants across all metrics, demonstrating the necessity and effectiveness of each proposed component.

Naive Approach In this subsection, we demonstrate that the naive approach—rewriting the prompt and manually selecting a conditioning frame index for existing video generative model—fails to effectively address the unpaired text–image to video generation challenge. The corresponding results are shown in Figure 5(b). As illustrated, although the rewritten prompt can partially capture the intrinsic relationships between modalities (*e.g.*, the vase being *broken* by the cat, or the car running *past* the dog), the generated videos often exhibit only superficial blending of elements and, in some cases, even produce confusing or incoherent contents. More examples can be found in the supplementary materials. These observations further highlight both the difficulty of the unpaired text–image to video generation task and the necessity of our proposed method.

5. Conclusion

In this work, we for the first time propose to solve unpaired text-image to video generation under test time OOD scenario and present a novel ReasonDiff model. Unlike existing approaches that often produce visually confusing videos that mix multiple objects incoherently or failing to uncover the intrinsic connections, ReasonDiff is designed to reason over both unpaired modalities simultaneously. We introduce two key components: VisionNarrator extracts per-frame narratives based on the unpaired inputs, while AlignFormer predicts reasoning-enhanced latents to guide the base video generator. Together, these modules enable the generation of videos from unpaired inputs that achieve both photorealism and semantic coherence.

Acknowledgements

This work was supported by Beijing National Research Center for Information Science and Technology under Grant No. BNR2023TD03006.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3

[2] 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. 2, 5

[3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 1

[4] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 1, 2

[5] 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. 5

[6] Hong Chen, Xin Wang, Guanning Zeng, Yipeng Zhang, Yuwei Zhou, Feilin Han, and Wenwu Zhu. Videodreamer: Customized multi-subject text-to-video generation with disen-mix finetuning. *arXiv preprint arXiv:2311.00990*, 2023. 1

[7] Hong Chen, Yipeng Zhang, Simin Wu, Xin Wang, Xuguang Duan, Yuwei Zhou, and Wenwu Zhu. Disenbooth: Identity-preserving disentangled tuning for subject-driven text-to-image generation. *arXiv preprint arXiv:2305.03374*, 2023. 1

[8] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7346–7356, 2023. 3

[9] Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. Layoutgpt: Compositional visual planning and generation with large language models. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[10] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024. 3

[11] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 5

[12] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 1

[13] 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. 1, 3, 6

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[15] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 1, 2

[16] Susung Hong, Junyoung Seo, Heeseong Shin, Sunghwan Hong, and Seungryong Kim. Direct2v: Large language models are frame-level directors for zero-shot text-to-video generation. *arXiv preprint arXiv:2305.14330*, 2023. 3

[17] Yaosi Hu, Chong Luo, and Zhenzhong Chen. Make it move: controllable image-to-video generation with text descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18219–18228, 2022. 1

[18] Hanzhuo Huang, Yufan Feng, Cheng Shi, Lan Xu, Jingyi Yu, and Sibei Yang. Free-bloom: Zero-shot text-to-video generator with llm director and ldm animator. *Advances in Neural Information Processing Systems*, 36:26135–26158, 2023. 3

[19] Yuzhou Huang, Liangbin Xie, Xintao Wang, Ziyang Yuan, Xiaodong Cun, Yixiao Ge, Jiantao Zhou, Chao Dong, Rui Huang, Ruimao Zhang, et al. Smartedit: Exploring complex instruction-based image editing with multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8362–8371, 2024. 3

[20] 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. 6

[21] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022. 3

[22] Bolin Lai, Sangmin Lee, Xu Cao, Xiang Li, and James M Rehg. Incorporating flexible image conditioning into text-to-video diffusion models without training. *arXiv preprint arXiv:2505.20629*, 2025. 3

[23] Xiaochuan Li, Baoyu Fan, Runze Zhang, Liang Jin, Di Wang, Zhenhua Guo, Yaqian Zhao, and Rengang Li. Image content generation with causal reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13646–13654, 2024. 3

[24] Long Lian, Baifeng Shi, Adam Yala, Trevor Darrell, and Boyi Li. Llm-grounded video diffusion models. *arXiv preprint arXiv:2309.17444*, 2023. 3

[25] Han Lin, Abhay Zala, Jaemin Cho, and Mohit Bansal. Videodirectorgpt: Consistent multi-scene video generation via llm-guided planning. *arXiv preprint arXiv:2309.15091*, 2023. 3

[26] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 2, 4

[27] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. 3

[28] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*, 2021. 4

[29] Fuchen Long, Zhaofan Qiu, Ting Yao, and Tao Mei. Videodrafter: Content-consistent multi-scene video generation with llm. *arXiv preprint arXiv:2401.01256*, 2024. 3

[30] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. *arXiv preprint arXiv:2303.08320*, 2023. 2

[31] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4117–4125, 2024. 3

[32] Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18444–18455, 2023. 1

[33] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 1

[34] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 1

[35] KaShun Shum, Shizhe Diao, and Tong Zhang. Automatic prompt augmentation and selection with chain-of-thought from labeled data. *arXiv preprint arXiv:2302.12822*, 2023. 3

[36] Zihan Song, Xin Wang, Zi Qian, Hong Chen, Longtao Huang, Hui Xue, and Wenwu Zhu. Modularized self-reflected video reasoner for multimodal llm with application to video question answering. In *Forty-second International Conference on Machine Learning*, 2025. 3

[37] Ye Tian, Ling Yang, Haotian Yang, Yuan Gao, Yufan Deng, Jingmin Chen, Xintao Wang, Zhaochen Yu, Xin Tao, Pengfei Wan, et al. Videotetris: Towards compositional text-to-video generation. *Advances in Neural Information Processing Systems*, 37:29489–29513, 2024. 3

[38] 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, 3, 6

[39] Xin Wang, Yuwei Zhou, Bin Huang, Hong Chen, and Wenwu Zhu. Multi-modal generative ai: Multi-modal llms, diffusions and the unification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025. 1

[40] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 3

[41] Jinbo Xing, Menghan Xia, Yuxin Liu, Yuechen Zhang, Yong Zhang, Yingqing He, Hanyuan Liu, Haoxin Chen, Xiaodong Cun, Xintao Wang, et al. Make-your-video: Customized video generation using textual and structural guidance. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 3

[42] 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. 3, 6

[43] 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. 6

[44] 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. 1, 3, 6

[45] Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*, 2023. 2

[46] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *International Journal of Computer Vision*, pages 1–15, 2024. 2

[47] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 2

[48] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo:

Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077*, 2023. 3

[49] Yipeng Zhang, Xin Wang, Hong Chen, Chenyang Qin, Yibo Hao, Hong Mei, and Wenwu Zhu. Scenariodiff: Text-to-video generation with dynamic transformations of scene conditions. *International Journal of Computer Vision*, pages 1–14, 2025. 1

[50] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*, 2022. 3

[51] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023. 3