FancyVideo: Towards Dynamic and Consistent Video Generation via Cross-frame Textual Guidance

Jiasong Feng[1,2], **Ao Ma**[1,3], **Jing Wang**[1,4], **Ke Cao**[1,5] and **Zhanjie Zhang**[1,6,‡]
[1] 360 AI Research
[2] Beijing University of Technology
[3] Wuhan University
[4] Sun Yat-sen University
[5] University of Science and Technology of China
[6] Zhejiang University
{maaoaoma, zhangzhanj}@126.com

Abstract

Synthesizing motion-rich and temporally consistent videos remains a challenge in artificial intelligence, especially when dealing with extended durations. Existing text-to-video (T2V) models commonly employ spatial cross-attention for text control, equivalently guiding different frame generations without frame-specific textual guidance. Thus, the model's capacity to comprehend the temporal logic conveyed in prompts and generate videos with coherent motion is restricted. To tackle this limitation, we introduce **FancyVideo**, an innovative video generator that improves the existing text-control mechanism with the well-designed **C**ross-frame **T**extual **G**uidance **M**odule (CTGM). Specifically, CTGM incorporates the Temporal Information Injector (TII) and Temporal Affinity Refiner (TAR) at the beginning and end of cross-attention, respectively, to achieve frame-specific textual guidance. Firstly, TII injects frame-specific information from latent features into text conditions, thereby obtaining cross-frame textual conditions. Then, TAR refines the correlation matrix between cross-frame textual conditions and latent features along the time dimension. Extensive experiments comprising both quantitative and qualitative evaluations demonstrate the effectiveness of FancyVideo. Our approach achieves state-of-the-art T2V generation results on the EvalCrafter benchmark and facilitates the synthesis of dynamic and consistent videos. Note that the T2V process of FancyVideo essentially involves a text-to-image step followed by T+I2V. This means it also supports the generation of videos from user images, i.e., the image-to-video (I2V) task. A significant number of experiments have shown that its performance is also outstanding.

‡ Corresponding authors.

1 Introduction

With the advancement of the diffusion model, the text-to-image (T2I) generative models [Blattmann *et al.*, 2023b; Ho *et al.*, 2022; Luo *et al.*, 2023; Ma *et al.*, 2024; Liu *et al.*, 2025] can produce high-resolution and photo-realistic images by complex text prompts, resulting in various applications. Currently, many studies [Wang *et al.*, 2024; Guo *et al.*, 2023a] explore the text-to-video (T2V) generative model due to the great success of T2I models. However, building a powerful T2V model remains challenging as it requires maintaining temporal consistency while generating coherent motions simultaneously. Moreover, due to limited memory, most diffusion-based T2V models [Wang *et al.*, 2024; Guo *et al.*, 2023a; Zhang *et al.*, 2024a; Guo *et al.*, 2023b; Chen *et al.*, 2023; Menapace *et al.*, 2024] can only produce fewer than 16 frames of video per sampling without extra assistance (i.e., super-resolution).

The existing T2V models [Zhang *et al.*, 2024a; Guo *et al.*, 2023b; Chen *et al.*, 2023; Menapace *et al.*, 2024] typically employ spatial cross-attention between text conditions and latent features for achieving text control generation. However, as shown in Fig. 2(I), this manner shares the same text condition across different frames, thus lacking the specific textual guidance tailored to each frame. Consequently, these T2V models struggle to comprehend the temporal logic of text prompts and produce videos with coherent motion. Taking AnimateDiff [Guo *et al.*, 2023b] as an example, in Fig. 1, we exhibit its generated video and visualize the [verb]-focused region (which is closely associated with the video motion) based on the attention map from the cross-attention module. Ideally, these regions should transition smoothly over time and align with the semantics of motion instructions. However, as observed in the upper right of the figure, the [verb]-focused region remains nearly identical across different frames due to the consistent textual guidance between frames. Meanwhile, the video exhibits poor motion in the upper left of the figure.

Furthermore, we perform a similar visual analysis for the longer video (e.g., 64 frames) generation and find that this problem is more prominent, as illustrated in the lower part of Fig. 1. Therefore, we believe this approach hampers the

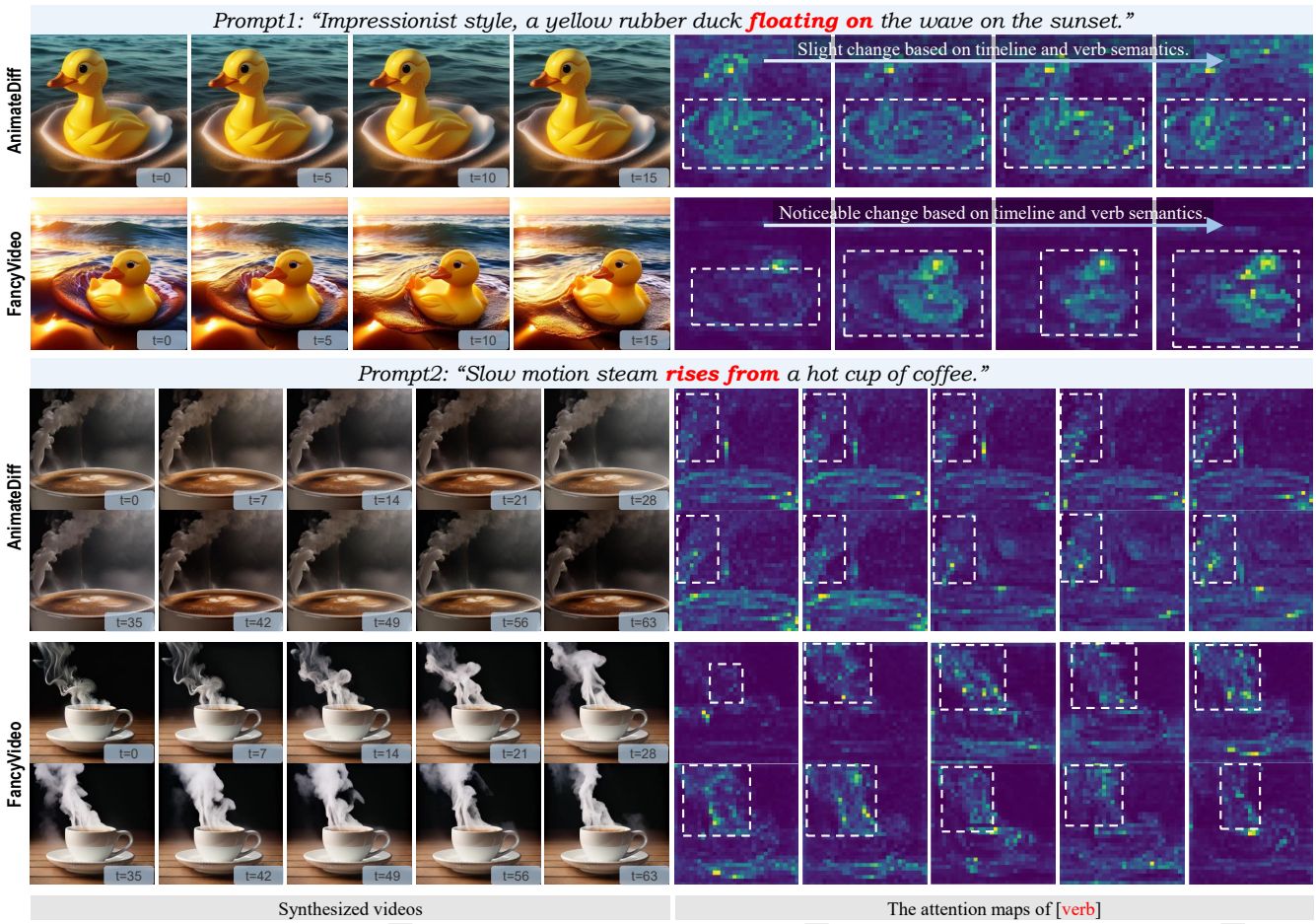

Figure 1: The generated videos and the attention maps of [verb] belong to FancyVideo and AnimateDiff. We present the 16-frame video (top) and longer 64-frame video (bottom). Due to the inadequate time-specific textual guidance in the AnimateDiff, the [verb] focused region remains almost constant, resulting in a lack of motion in the video. In contrast, **FancyVideo** effectively alleviates this issue through cross-frame textual guidance. The [verb] focused region changes based on the timeline and semantics, thereby generating motion-rich videos.

advancement of video dynamics and consistency and is sub-optimal for video generation tasks based on text prompts.

To this end, we present a novel T2V model named **FancyVideo**, capable of comprehending complex spatial-temporal relationships within text prompts. By employing a cross-frame textual guidance strategy, FancyVideo can generate more dynamic and plausible videos in a sampling process. Specifically, to boost the model's capacity for understanding spatial-temporal information in text prompts, we optimize the spatial cross-attention through the proposed **C**ross-frame **T**extual **G**uidance **M**odule (CTGM), comprising a Temporal Information Injector (TII) and Temporal Affinity Refiner (TAR). As illustrated in Fig. 2(II), TII injects temporal information from latent features into text conditions, building cross-frame textual conditions. Then, TAR refines the affinity between frame-specific text embedding and video along time dimension, adjusting the temporal logic of textual guidance. Through the cooperative interaction between TII and TAR, FancyVideo fully captures the motion logic embedded within images and text. Consequently, its motion token-focused area shifts logically with frames, as illustrated in the lower right part of Fig. 1. This characteristic enables FancyVideo to produce dynamic videos, as displayed in the lower left part of the figure. Experiments demonstrate that FancyVideo successfully generates dynamic and consistent videos, achieving the SOTA results on the EvalCrafter [Liu *et al.*, 2023] benchmark and the competitive performance on UCF-101 [Soomro *et al.*, 2012] and MSR-VTT [Xu *et al.*, 2016]. Additionally, FancyVideo supports generating videos from user-input images, i.e., the image-to-video task. We have also conducted extensive experiments to demonstrate the superiority of our method.

Contributions. **1)** We introduce FancyVideo, the pioneering endeavor as far as our knowledge extends, delving into cross-frame textual guidance for the T2V task. This approach offers a fresh perspective to enhance current text-control methodologies. **2)** We propose the Cross-frame Textual Guidance Module (CTGM), which constructs cross-frame textual conditions and subsequently guides the modeling of latent features with robust temporal plausibility. It can effectively

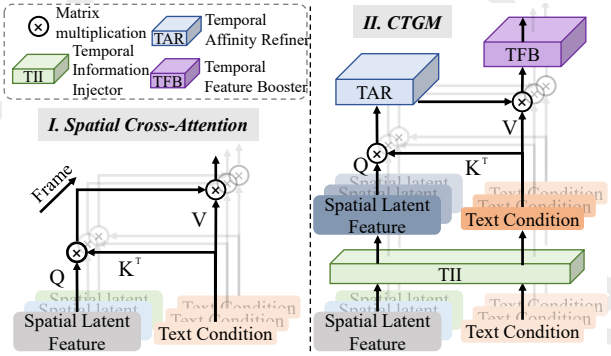

Figure 2: The structure of spatial cross-attention and CTGM.

enhance the motion and consistency of video. **3)** We demonstrate that incorporating cross-frame textual guidance represents an effective approach for achieving high-quality video generation. Our experiments showcase that this approach attains state-of-the-art results on both quantitative and qualitative evaluations.

2 Related Work

Text to Video Generation. Generative models like GANs [Wang *et al.*, 2020; Munoz *et al.*, 2021; Gur *et al.*, 2020], auto-regressive models [Wang *et al.*, 2019; Yan *et al.*, 2021], and implicit neural representations [De Luigi *et al.*, 2023] have been explored for video generation. Recently, diffusion models [Rombach *et al.*, 2022; Zhang *et al.*, 2024b; Zhang *et al.*, 2024c] have advanced text-to-image quality. Stable Diffusion [Rombach *et al.*, 2022] uses a VAE [Kingma and Welling, 2013] latent space to reduce cost [Jiang *et al.*, 2023]. T2V models [Wu *et al.*, 2023a] add temporal layers to T2I models but often lack frame-to-frame consistency. We propose cross-frame textual guidance to improve temporal coherence.

Image-conditioned Video Generation. To bridge the gap between text and video, recent work leverages images for clearer video generation. SVD [Blattmann *et al.*, 2023a] treats images as noisy latent inputs, while MoonShot [Zhang *et al.*, 2024a] improves semantic consistency using a CLIP encoder. Though effective, these I2V methods rely on input images. Hierarchical approaches [Zeng *et al.*, 2023; Chen *et al.*, 2023] use images as keyframes to extend video length with fewer constraints. These methods, though I2V-capable, are essentially T2V. FancyVideo adopts a hierarchical design with cross-frame textual guidance, enabling more frames per iteration and faster inference.

3 Method

3.1 Preliminaries

Latent Diffusion Models. LDMs [Sohl-Dickstein *et al.*, 2015; Ho *et al.*, 2020] enhance efficiency by running diffusion in the VAE-compressed latent space [Kingma and Welling, 2013] instead of pixel space. The forward process adds Gaussian noise ($\epsilon \sim \mathcal{N}(0, I)$) to the latent code $\mathbf{z}$, yielding:

$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\mathbf{z} + \sqrt{1-\bar{\alpha}_t}\epsilon, \tag{1}$$

where $\bar{\alpha}_t$ denotes a noise scheduler with timestep t. For the inverse process, it trains a denoising model (f_θ) with the objective:

$$\mathbb{E}_{\mathbf{z}\sim p(z),\epsilon\sim\mathcal{N}(0,I),t}\left[\|\boldsymbol{y} - f_\theta(\mathbf{z}_t, \mathbf{c}, t)\|^2\right], \tag{2}$$

where $\mathbf{c}$ represents the condition and target $\boldsymbol{y}$ can be noise ϵ, denoising input $\mathbf{z}$ or $\boldsymbol{v}$-prediction ($\boldsymbol{v} = \sqrt{\bar{\alpha}_t}\epsilon - \sqrt{1-\bar{\alpha}_t}\mathbf{z}$) in [Salimans and Ho, 2022]. In this paper, we adopt the $\boldsymbol{v}$-prediction as the supervision.

Zero terminal-SNR Noise Schedule. Previous studies proposed zero terminal SNR [Lin *et al.*, 2024] to handle the signal-to-noise ratio (SNR) difference between the testing and training phase, which hinders the generation quality. At training, due to the residual signal left by the noise scheduler, the SNR is still not zero at the terminal timestep T. However, the sampler lacks realistic data when sampling from random gaussian noise during the test, resulting in a zero SNR. This train-test discrepancy is unreasonable and an obstacle to generating high-quality videos. Therefore, following the [Lin *et al.*, 2024; Girdhar *et al.*, 2023], we scale up the noise schedule and set $\bar{\alpha_T} = 0$ to fix this problem.

3.2 Model Architecture

Fig. 3 illustrates the overall architecture of FancyVideo. The model is structured as a pseudo-3D UNet, which integrates frozen spatial blocks, sourced from a text-to-image model, along with Cross-frame Textual Guidance Modules (CTGM) and temporal attention blocks. The model takes three features as input: noisy latent $\mathcal{Z}_n \in \mathbb{R}^{f\times h\times w\times c}$, where h and w indicate the height and width of the latent, f signifies the number of frames, and c denotes the channels of the latent; mask indicator $\mathcal{M} \in \mathbb{R}^{f\times h\times w\times 1}$, with elements set to 1 for the first frame and 0 for all other frames; image indicator $\mathcal{I} \in \mathbb{R}^{f\times h\times w\times c}$, with initial image as the first frame and 0 for all other frames. The denoising input $\mathcal{Z}$ is formed by concatenating $\mathcal{Z}_n$, $\mathcal{M}$ and $\mathcal{I}$ along the channel dimension, represented as $\mathcal{Z} = [\mathcal{Z}_n; \mathcal{M}; \mathcal{I}] \in \mathbb{R}^{f\times h\times w\times(2c+1)}$. Within each spatial block, we first incorporate prior knowledge of the motion score as embeddings. In each subsequent cross-attention layer, CTGM is employed to capture the intricate dynamics described in the text prompts. Afterward, we apply temporal attention blocks to enhance the temporal relationships across various patches.

Motion Embedding

To achieve more controllable video generation in terms of motion amplitude, we introduce motion score information calculated by the RAFT [Teed and Deng, 2020] alongside the timestep information. Specifically, we calculate a motion score for the training samples in the dataset within a range of 0.1 to 10. The score are then encoded into motion features through a motion embedding layer. By controlling the motion score, we can generate videos with stronger motion. However, simply adjusting the score may lead to unrealistic motion. We use CTGM to prevent these issues.

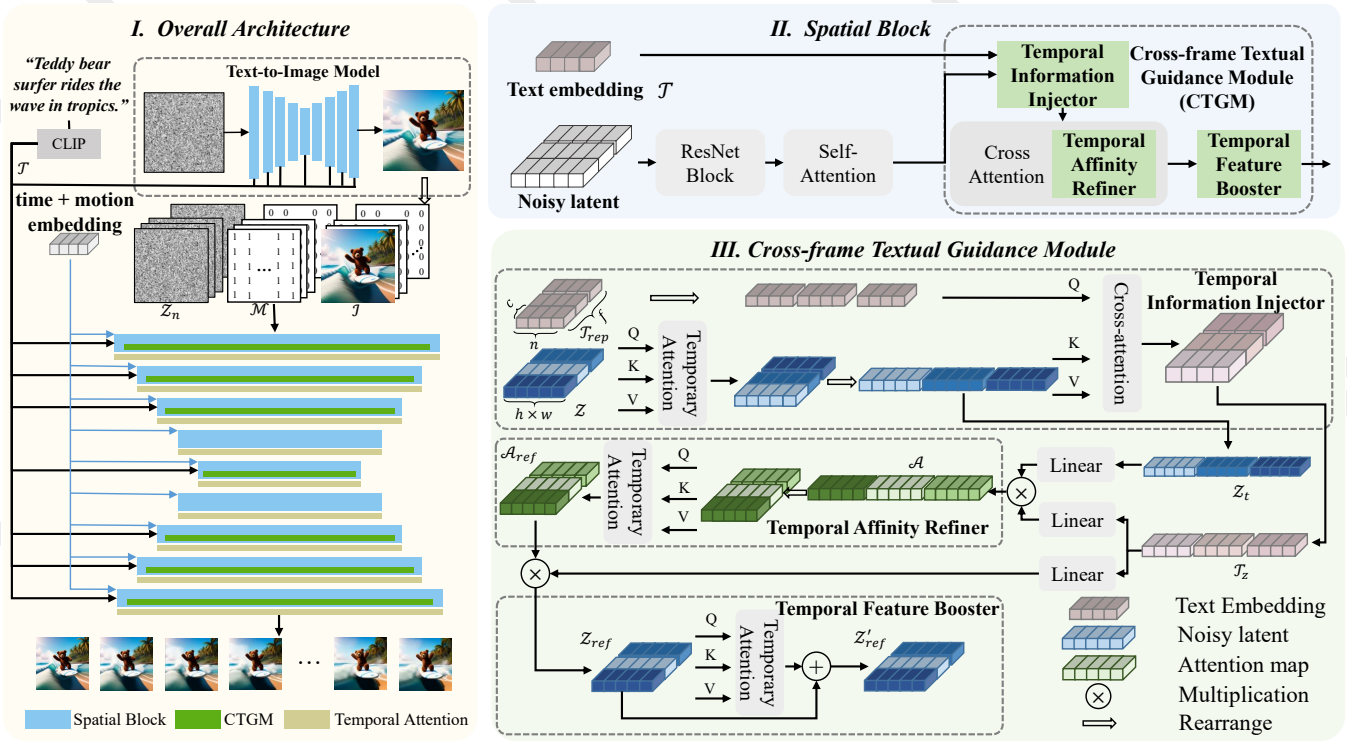

Figure 3: The overall architecture of our method. FancyVideo is a T+I2V model that concatenates noise latent, mask indicator, and image indicator as input. We insert our Cross-frame Textual Guidance Module (CTGM) into each spatial block. CTGM consists of three components: Temporal Information Injector, Temporal Affinity Refiner, and Temporal Feature Booster. These components are inserted at the beginning, middle, and end of cross-attention, respectively.

Cross-frame Textual Guidance Module

CTGM advances the existing text control method through two sub-modules: Temporal Information Injector (TII) and Temporal Affinity Refiner (TAR) as depicted in Fig. 3(III). Before engaging in cross-attention, TII initially extracts temporal latent feature $\mathcal{Z}_t$ and then incorporates temporal information into text embedding $\mathcal{T}_{rep}$ based on $\mathcal{Z}_t$, obtaining cross-frame textual condition $\mathcal{T}_z$. Subsequently, TAR refines the affinity between $\mathcal{Z}_t$ and $\mathcal{T}_z$ along the time axis, enhancing the temporal coherence of textual guidance. The computation process of the CTGM can be formalized as:

$$\mathcal{Z}_t, \mathcal{T}_z = \mathrm{TII}(\mathcal{Z}, \mathcal{T}_{rep}), \tag{3}$$

$$\mathcal{Z}_{ref} = \mathrm{Softmax}(\frac{\mathrm{TAR}(W_q\mathcal{Z}_t, W_k\mathcal{T}_z)}{\sqrt{d_k}}W_v(\mathcal{T}_z), \tag{4}$$

where W_q, W_k, and W_v represent the linear layers for query, key, and value in original cross-attention, respectively. The hyper-parameter d_k is acquired from the query dimensions. $\mathrm{TII}(\cdot,\cdot)$ and $\mathrm{TAR}(\cdot)$ denotes the functions of TII and TAR. In the end, we get refined noisy latent feature Z_{ref}. A detailed description of these three modules is provided as follows.

Temporal Information Injector. In previous work [Guo *et al.*, 2023b; Girdhar *et al.*, 2023], the text embedding $\mathcal{T}_{rep}$ is repeated equally f times, resulting in $\mathcal{T}_{rep} \in \mathbb{R}^{f\times n\times c}$, n denoting the length of the embedding vector. We inject temporal information into the embedding before performing spatial cross-attention, thereby enabling distinct focal points on the text within different frames. In Temporal Information Injector (TII), we initially reshape the noisy latent $\mathcal{Z}$ from $\mathbb{R}^{f\times h\times w\times c}$ to $\mathbb{R}^{(hw)\times f\times c}$ and apply temporal self-attention to acquire $\mathcal{Z}_t$. Then, we conduct spatial cross-attention, using the repeated text embedding $\mathcal{T}_{rep}$ as queries and the noisy latent $\mathcal{Z}_t \in \mathbb{R}^{f\times(hw)\times c}$ as both keys and values, resulting in the text embedding $\mathcal{T}_z$ with frame-specific temporal information. The formalization of the TII module can be expressed as follows:

$$\begin{aligned}\mathcal{Z}_t, \mathcal{T}_z &= \mathrm{TII}(\mathcal{Z}, \mathcal{T}_{rep}) \\ &= \mathrm{SelfAttn_t}(\mathcal{Z}), \\ &\quad \mathrm{CrossAttn_s}(\mathrm{SelfAttn_t}(\mathcal{Z}), \mathcal{T}_{rep})\end{aligned} \tag{5}$$

where $\mathrm{SelfAttn_t}$ denotes temporal self-attention and $\mathrm{CrossAttn_s}$ denotes spatial cross-attention. Through TII, we obtain the noisy latent $\mathcal{Z}_t$ with temporal information and the latent-aligned text embedding $\mathcal{T}_z$.

Temporal Affinity Refiner. To dynamically allocate attention to text embedding across different frames, we design the Temporal Affinity Refiner (TAR) to refine the attention map of spatial cross-attention. In spatial cross-attention, the noisy latent serves as the query, while the text embedding serves as both the key and value. The attention map $\mathcal{A} \in \mathbb{R}^{f\times(hw)\times n}$, compute as $\mathcal{A} = (W_q\mathcal{Z}_t)(W_k\mathcal{T}_z)^T/\sqrt{d_k}$, reflects the affinity between the text and patches. Then, TAR applies temporal

self-attention to the attention map $\mathcal{A} \in \mathbb{R}^{(hw)\times f\times n}$, obtaining the refined attention map $\mathcal{A}_{ref}$, which can be represented as:

$$\mathcal{A}_{ref} = \mathrm{TAR}(\mathcal{A}) = \mathrm{SelfAttn_t}(\mathcal{A}) \tag{6}$$

With the TAR, $\mathcal{A}_{\text{ref}}$ establishes a more logical temporal connection in the affinity matrix. It can perform more dynamic action while ensuring no additional video distortion occurs. Finally, the cross-attention process is completed with the refined attention map as $\mathcal{Z}_{ref} = \mathrm{Softmax}(\mathcal{A}_{ref})(W_v\mathcal{T}_z)$.

4 Experiments

In the quantitative experiments, FancyVideo utilizes the T2I base model to generate images as the first frame. In the qualitative experiments, for aesthetic purposes and to remove watermarks, an external model is used to generate a beautiful first frame.

4.1 Qualitative Evaluation

We choose AnimateDiff [Guo *et al.*, 2023b], DynamiCrafter [Xing *et al.*, 2023], and two commercialized products, Pika [PikaLabs, 2024] and Gen2 [Runway, 2024], for a composite qualitative analysis. It is worth noting that in the quantitative experiments, the first frame of FancyVideo is generated by SDXL to achieve a more aesthetically pleasing result and to minimize the appearance of watermark (although subsequent frames may still exhibit it).

As shown in Fig. 4, our approach exhibits superior performance, outperforming previous methods regarding temporal consistency and motion richness. In contrast, AnimateDiff, DynamiCrafter, and Gen2 generate videos with less motion. Pika struggles to produce object-consistent and high-quality video frames. Remarkably, our method can accurately understand the motion instructions in the text prompt (e.g., "A teddy bear walking ... beautiful sunset." and "A teddy bear running ... City." case).

4.2 Quantitative Evaluation

For a comprehensive comparison with the SOTA methods, we adopt three popular benchmarks (e.g., EvalCrafter [Liu *et al.*, 2023], UCF-101 [Soomro *et al.*, 2012], and MSR-VTT [Xu *et al.*, 2016]) and human evaluation to evaluate the quality of video generation. Among them, EvalCrafter is a relatively comprehensive benchmark for video generation currently. UCF-101 and MSR-VTT are benchmarks commonly used in previous methods [Girdhar *et al.*, 2023; Zhang *et al.*, 2023]. Meanwhile, human evaluation can compensate for the inaccuracies in existing text-conditioned video generation evaluation systems.

EvalCrafter Benchmark. EvalCrafter [Liu *et al.*, 2023] quantitatively evaluates the quality of text-to-video generation from four aspects (including Video Quality, Text-video Alignment, Motion Quality, and Temporal Consistency). Each dimension contains multiple subcategories of indicators shown in the Table. 1. As discussed in community [Liu and Cun, 2024], the authors acknowledge that the original manner of calculating the comprehensive metric was inappropriate. For a more intuitive comparison, we introduce a comprehensive metric for every aspect by considering each sub-indicators numerical scale and positive-negative attributes.

In detail, we compare the performance of the previous video generation SOTA methods (e.g., Pika [PikaLabs, 2024], Gen2 [Runway, 2024], Show-1 [Zhang *et al.*, 2023], Lumiere [Bar-Tal *et al.*, 2024], DynamiCrafter [Xing *et al.*, 2023], and AnimateDiff [Guo *et al.*, 2023b]) and exhibit in Table. 1. Our method demonstrates outstanding performance beyond existing methods at the Video Quality and Text-video Alignment aspect. Although Show-1 has the best Motion Quality (81.56), its Video Quality is poor (only 85.08). That indicates that it cannot generate high-quality videos with reasonable motion. However, our method has the second highest Motion Quality (72.99) and the best Video Quality (177.72), achieving the trade-off between quality and motion. The above results indicate the superiority of FancyVideo and its ability to generate temporal-consistent and motion-accurate video.

UCF-101 & MSR-VTT. Following the prior work [Zhang *et al.*, 2023], we evaluate the zero-shot generation performance on UCF-101 [Soomro *et al.*, 2012] and MSR-VTT [Xu *et al.*, 2016] as shown in Table. 2. We use Fréchet Video Distance (FVD) [Unterthiner *et al.*, 2019], Inception Score (IS) [Wu *et al.*, 2021], Fréchet Inception Distance (FID) [Heusel *et al.*, 2017], and CLIP similarity (CLIPSIM) as evaluation metrics and compared some current SOTA methods. FancyVideo achieves competitive results, particularly excelling in IS and CLIPSIM with scores of 43.66 and 0.3076, respectively. Besides, previous studies [Ho *et al.*, 2022; Girdhar *et al.*, 2023; Wu *et al.*, 2023b] have pointed out that these metrics do not accurately reflect human perception and are affected by the gap between the distribution of training and test data and the image's low-level detail.

Human Evaluation. Inspired by EvalCrafter [Liu *et al.*, 2023], we introduce a multi-candidate ranking protocol with four aspects: video quality, text-video alignment, motion quality, and temporal consistency. In this protocol, participants rank the results of multiple candidate models for each aspect. Each candidate model receives a score based on its ranking. For instance, if there are N candidate models ranked by video quality, the first model gets $N-1$ points, the second gets $N-2$ points, and so on, with the last model receiving 0 points. Adhering to this protocol, we selected 108 samples from the EvalCrafter validation set and gathered judgments from 100 individuals. As depicted in Fig. 5, our method significantly outperforms text-to-video conversion methods, including AnimateDiff [Guo *et al.*, 2023b], Pika [PikaLabs, 2024], and Gen2 [Runway, 2024], across all four aspects. FancyVideo demonstrates exceptional motion quality while preserving superior text-video consistency. Additionally, we conducted a similar comparison of four image-to-video methods, including DynamiCrafter [Xing *et al.*, 2023], Pika, and Gen2, as shown in Fig. 5.

4.3 Ablation Studies

In this section, we conduct extensive experiments and exhibit detailed visual comparisons on the EvalCrafter benchmark [Liu *et al.*, 2023] to thoroughly explore the effect of critical designs in CTGM. The ablation includes three key modules

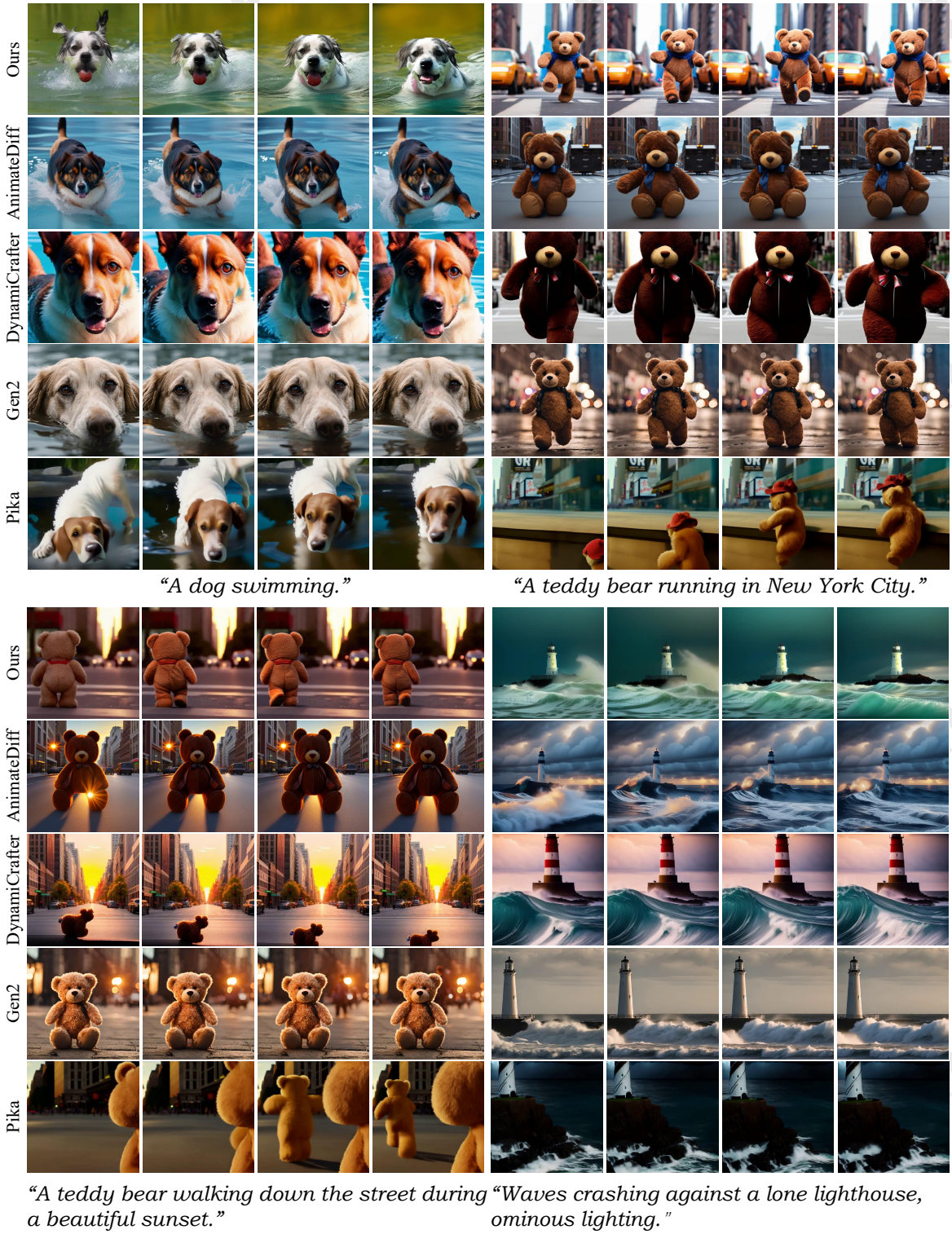

Figure 4: Qualititive analysis. We compare the video generation results from AnimateDiff [Guo *et al.*, 2023b], DynamiCrafter [Xing *et al.*, 2023], Pika [PikaLabs, 2024], Gen-2 [Runway, 2024], and our FancyVideo.

Dimensions	Metrics	Pika	Gen2	Show-1	Lumiere	DynamiCrafter	AnimateDiff	FancyVideo
Video Quality	VQAA(↑)	59.09	59.44	23.19	40.06	74.56	65.94	85.78
	VQAT(↑)	64.96	76.51	44.24	32.93	59.48	52.02	74.56
	IS(↑)	14.81	14.53	17.65	17.64	18.37	16.54	17.38
	Comprehensive(↑)	138.86	150.48	85.08	90.63	152.41	134.50	**177.72**
Text-Video Alignment	CLIP-Score(↑)	20.46	20.53	20.66	20.36	20.80	19.70	20.85
	BLIP-BLEU(↑)	21.14	22.24	23.24	22.54	20.93	20.67	21.33
	SD-Score(↑)	68.57	68.58	68.42	67.93	67.87	66.13	68.14
	Detection-Score(↑)	58.99	64.05	58.63	50.01	64.04	51.19	66.66
	Color-Score(↑)	34.35	37.56	48.55	38.72	45.65	42.39	51.09
	Count-Score(↑)	51.46	53.31	44.31	44.18	53.53	22.40	59.19
	OCR Score(↓)	84.31	75.00	58.97	71.32	60.29	45.21	64.85
	Celebrity ID Score(↓)	45.31	41.25	37.93	44.56	26.35	42.26	25.76
	Comprehensive(↑)	325.35	350.02	366.91	327.86	386.18	335.01	**396.65**
Motion Quality	Action Score(↑)	71.81	62.53	81.56	72.12	72.22	61.94	72.99
	Motion AC-Score(→)	44	44	50	42	46	32	52
	Flow-Score(→)	0.50	0.70	2.07	6.99	0.96	2.403	1.7413
	Comprehensive(↑)	71.81	62.53	**81.56**	72.12	72.22	61.94	72.99
Temporal Consistency	CLIP-Temp(↑)	99.97	99.94	99.77	99.74	99.75	99.85	99.84
	Warping Error(↓)	0.0006	0.0008	0.0067	0.0162	0.0054	0.0177	0.0051
	Face Consistency(↑)	99.62	99.06	99.32	98.94	99.34	99.63	99.31
	Comprehensive(↑)	**199.59**	199.00	199.09	198.68	199.09	199.48	199.15

Table 1: Quantitative evaluation on the EvalCrafter. The best and second performing metrics are highlighted in **bold** and underline. Comprehensive denotes the composite metrics for these dimensions.

Method	Data	UCF-101			MSR-VTT	
		FVD(↓)	IS(↑)	FID(↓)	FVD(↓)	CLIPSIM (↑)
Emu Video	34M	606.20	42.70	-	-	-
AnimateDiff	10M	584.85	37.01	61.24	628.57	0.2881
DynamiCrafter	10M	404.50	41.97	**32.35**	**219.31**	0.2659
Show-1	10M	394.46	35.42	-	538.00	0.3072
Lumiere	10M	**332.49**	37.54	-	550.00	0.2939
FancyVideo	10M	412.64	**43.66**	47.01	333.52	**0.3076**

Table 2: Quantitative evaluation on the UCF-101 [Soomro *et al.*, 2012] and MSR-VTT [Xu *et al.*, 2016] .The best and second performing metrics are highlighted in **bold** and underline respectively.

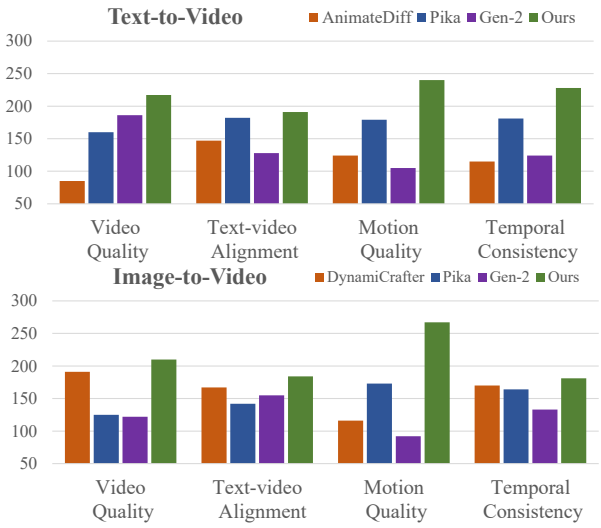

Figure 5: Human Evaluation Comparison. FancyVideo stands out significantly compared to other text-to-video and image-to-video generators in terms of Motion Quality and Temporal Consistency.

TAR	TII	Video Quality	Text-Video Alignment	Motion Quality	Temporal Consistency
		163.15	361.92	66.99	198.83
✓		172.44	379.40	71.24	199.08
	✓	173.82	380.24	71.84	199.04
✓	✓	177.72	396.65	72.99	199.15

Table 3: Ablation studies on the core component of FancyVideo

(TII and TAR), each boosting video quality. As shown in Table 3, TAR significantly improves both metrics, highlighting the importance of temporal attention refinement. Adding TII further enhances performance by refining latent features and enabling frame-level text control.

5 Conclusion

In this work, we present a novel video-generation method named FancyVideo, which optimizes common text control mechanisms (e.g., spatial cross-attention) from the cross-frame textual guidance. It improves cross-attention with a well-designed Cross-frame Textual Guidance Module (CTGM), implementing the temporal-specific textual condition guidance for video generation. A comprehensive qualitative and quantitative analysis shows it can produce more dynamic and consistent videos. This characteristic becomes more noticeable as the number of frames increases. Our method achieves state-of-the-art results on the EvalCrafter benchmark and human evaluations.

Contribution Statement

Jiasong Feng, Ao Ma, Jing Wang, and Ke Cao contributed equally to this research. Ao Ma served as the project leader.

References

[Bar-Tal *et al.*, 2024] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Yuanzhen Li, Tomer Michaeli, et al. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945*, 2024.

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

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

[Chen *et al.*, 2023] Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Seine: Short-to-long video diffusion model for generative transition and prediction. In *The Twelfth International Conference on Learning Representations*, 2023.

[De Luigi *et al.*, 2023] Luca De Luigi, Adriano Cardace, Riccardo Spezialetti, Pierluigi Zama Ramirez, Samuele Salti, and Luigi Di Stefano. Deep learning on implicit neural representations of shapes. *arXiv preprint arXiv:2302.05438*, 2023.

[Girdhar *et al.*, 2023] Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Emu video: Factorizing text-to-video generation by explicit image conditioning. *arXiv preprint arXiv:2311.10709*, 2023.

[Guo *et al.*, 2023a] Xun Guo, Mingwu Zheng, Liang Hou, Yuan Gao, Yufan Deng, Chongyang Ma, Weiming Hu, Zhengjun Zha, Haibin Huang, Pengfei Wan, et al. I2v-adapter: A general image-to-video adapter for video diffusion models. *arXiv preprint arXiv:2312.16693*, 2023.

[Guo *et al.*, 2023b] 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.

[Gur *et al.*, 2020] Shir Gur, Sagie Benaim, and Lior Wolf. Hierarchical patch vae-gan: Generating diverse videos from a single sample. *Advances in Neural Information Processing Systems*, 33:16761–16772, 2020.

[Heusel *et al.*, 2017] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

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

[Ho *et al.*, 2022] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

[Jiang *et al.*, 2023] Zeyinzi Jiang, Chaojie Mao, Ziyuan Huang, Ao Ma, Yiliang Lv, Yujun Shen, Deli Zhao, and Jingren Zhou. Res-tuning: A flexible and efficient tuning paradigm via unbinding tuner from backbone. *Advances in Neural Information Processing Systems*, 36:42689–42716, 2023.

[Kingma and Welling, 2013] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[Lin *et al.*, 2024] Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5404–5411, 2024.

[Liu and Cun, 2024] Yaofang Liu and Xiaodong Cun. evalcrafter github project. https://github.com/evalcrafter/evalcrafter, 2024.

[Liu *et al.*, 2023] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. *arXiv preprint arXiv:2310.11440*, 2023.

[Liu *et al.*, 2025] Shanyuan Liu, Bo Cheng, Yuhang Ma, Liebucha Wu, Ao Ma, Xiaoyu Wu, Dawei Leng, and Yuhui Yin. Bridge diffusion model: Bridge chinese text-to-image diffusion model with english communities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 5541–5549, 2025.

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

[Ma *et al.*, 2024] Yuhang Ma, Shanyuan Liu, Ao Ma, Xiaoyu Wu, Dawei Leng, and Yuhui Yin. Hico: Hierarchical controllable diffusion model for layout-to-image generation. *Advances in Neural Information Processing Systems*, 37:128886–128910, 2024.

[Menapace *et al.*, 2024] Willi Menapace, Aliaksandr Siarohin, Ivan Skorokhodov, Ekaterina Deyneka, Tsai-Shien Chen, Anil Kag, Yuwei Fang, Aleksei Stoliar, Elisa Ricci, Jian Ren, et al. Snap video: Scaled spatiotemporal transformers for text-to-video synthesis. *arXiv preprint arXiv:2402.14797*, 2024.

[Munoz *et al.*, 2021] Andres Munoz, Mohammadreza Zolfaghari, Max Argus, and Thomas Brox. Temporal shift gan for large scale video generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3179–3188, 2021.

[PikaLabs, 2024] PikaLabs. Pika lab discord server. https://www.pika.art/, 2024.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[Runway, 2024] Runway. Gen2 discord server. https://research.runwayml.com/gen2/, 2024.

[Salimans and Ho, 2022] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

[Sohl-Dickstein *et al.*, 2015] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

[Soomro *et al.*, 2012] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. A dataset of 101 human action classes from videos in the wild. *Center for Research in Computer Vision*, 2(11):1–7, 2012.

[Teed and Deng, 2020] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020.

[Unterthiner *et al.*, 2019] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019.

[Wang *et al.*, 2019] Tsun-Hsuan Wang, Yen-Chi Cheng, Chieh Hubert Lin, Hwann-Tzong Chen, and Min Sun. Point-to-point video generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10491–10500, 2019.

[Wang *et al.*, 2020] Yaohui Wang, Piotr Bilinski, Francois Bremond, and Antitza Dantcheva. Imaginator: Conditional spatio-temporal gan for video generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1160–1169, 2020.

[Wang *et al.*, 2024] Jing Wang, Ao Ma, Jiasong Feng, Dawei Leng, Yuhui Yin, and Xiaodan Liang. Qihoo-t2x: An efficiency-focused diffusion transformer via proxy tokens for text-to-any-task. *arXiv e-prints*, pages arXiv–2409, 2024.

[Wu *et al.*, 2021] Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. Godiva: Generating open-domain videos from natural descriptions. *arXiv preprint arXiv:2104.14806*, 2021.

[Wu *et al.*, 2023a] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023.

[Wu *et al.*, 2023b] Tianxing Wu, Chenyang Si, Yuming Jiang, Ziqi Huang, and Ziwei Liu. Freeinit: Bridging initialization gap in video diffusion models. *arXiv preprint arXiv:2312.07537*, 2023.

[Xing *et al.*, 2023] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Xintao Wang, Tien-Tsin Wong, and Ying Shan. Dynamicrafter: Animating open-domain images with video diffusion priors. *arXiv preprint arXiv:2310.12190*, 2023.

[Xu *et al.*, 2016] 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.

[Yan *et al.*, 2021] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021.

[Zeng *et al.*, 2023] Yan Zeng, Guoqiang Wei, Jiani Zheng, Jiaxin Zou, Yang Wei, Yuchen Zhang, and Hang Li. Make pixels dance: High-dynamic video generation. *arXiv preprint arXiv:2311.10982*, 2023.

[Zhang *et al.*, 2023] 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. *arXiv preprint arXiv:2309.15818*, 2023.

[Zhang *et al.*, 2024a] David Junhao Zhang, Dongxu Li, Hung Le, Mike Zheng Shou, Caiming Xiong, and Doyen Sahoo. Moonshot: Towards controllable video generation and editing with multimodal conditions. *arXiv preprint arXiv:2401.01827*, 2024.

[Zhang *et al.*, 2024b] Zhanjie Zhang, Quanwei Zhang, Huaizhong Lin, Wei Xing, Juncheng Mo, Shuaicheng Huang, Jinheng Xie, Guangyuan Li, Junsheng Luan, Lei Zhao, et al. Towards highly realistic artistic style transfer via stable diffusion with step-aware and layer-aware prompt. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 7814–7822, 2024.

[Zhang *et al.*, 2024c] Zhanjie Zhang, Quanwei Zhang, Wei Xing, Guangyuan Li, Lei Zhao, Jiakai Sun, Zehua Lan, Junsheng Luan, Yiling Huang, and Huaizhong Lin. Artbank: Artistic style transfer with pre-trained diffusion model and implicit style prompt bank. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7396–7404, 2024.