A Recipe for Scaling up Text-to-Video Generation with Text-free Videos

Xiang Wang[1*] Shiwei Zhang[2†] Hangjie Yuan[3] Zhiwu Qing[1] Biao Gong[2] Yingya Zhang[2]
Yujun Shen[4] Changxin Gao[1] Nong Sang[1†]

[1]Key Laboratory of Image Processing and Intelligent Control,
School of Artificial Intelligence and Automation, Huazhong University of Science and Technology
[2]Alibaba Group [3]Zhejiang University [4]Ant Group

{wxiang,qzw,cgao,nsang}@hust.edu.cn, {zhangjin.zsw,yingya.zyy}@alibaba-inc.com
hj.yuan@zju.edu.cn, {a.biao.gong,shenyujun0302}@gmail.com

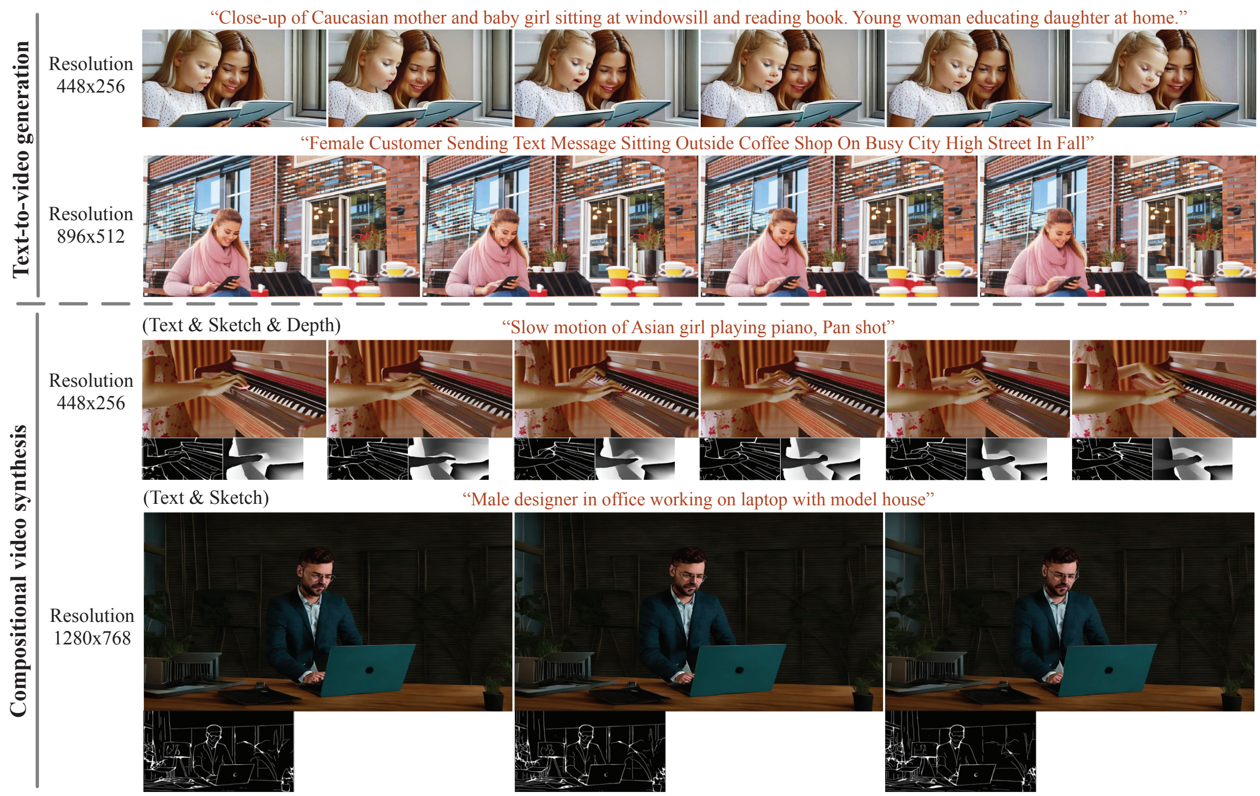

Figure 1. **Example video results** generated by the proposed `TF-T2V` on text-to-video generation and compositional video synthesis tasks *without training on any video-text pairs.*

Abstract

Diffusion-based text-to-video generation has witnessed impressive progress in the past year yet still falls behind text-to-image generation. One of the key reasons is the limited scale of publicly available data (e.g., 10M video-text pairs in WebVid10M vs. 5B image-text pairs in LAION), considering the high cost of video captioning. Instead, it could be far easier to collect unlabeled clips from video platforms like YouTube. Motivated by this, we come up with a novel text-to-video generation framework, termed `TF-T2V`, which can directly ***learn with text-free videos****. The rationale behind is to separate the process of text decoding from that of temporal modeling. To this end, we employ a content branch and a motion branch, which are jointly optimized with weights shared. Following such a pipeline, we study the effect of doubling the scale of training set (i.e., video-only WebVid10M) with some randomly collected text-free videos and are encouraged to observe the performance improvement (FID from 9.67 to 8.19 and FVD from 484 to 441), demonstrating the scalability of*

∗ Intern at Alibaba Group. † Corresponding authors.

our approach. We also find that our model could enjoy sustainable performance gain (FID from 8.19 to 7.64 and FVD from 441 to 366) after reintroducing some text labels for training. Finally, we validate the effectiveness and generalizability of our ideology on both native text-to-video generation and compositional video synthesis paradigms. Code and models will be publicly available at here.

1. Introduction

Video generation aims to synthesize realistic videos that possess visually appealing spatial contents and temporally coherent motions. It has witnessed unprecedented progress in recent years with the advent of deep generative techniques [22, 53], especially with the emergence of video diffusion models [4, 34, 40, 54, 60, 67, 78]. Pioneering approaches [28, 33, 67] utilize pure image diffusion models or fine-tuning on a small amount of video-text data to synthesize videos, leading to temporally discontinuous results due to insufficient motion perception [39, 79]. To achieve plausible results, current text-to-video methods like VideoLDM [4] and ModelScopeT2V [54] usually insert temporal blocks into latent 2D-UNet [43] and train the model on expansive video-text datasets, *e.g.*, WebVid10M [2]. To enable more controllable generation, VideoComposer [58] proposes a compositional paradigm that incorporates additional conditions (*e.g.*, depth, sketch, motion vectors, *etc.*) to guide synthesis, allowing customizable creation.

Despite this, the progress in text-to-video generation still falls behind text-to-image generation [42, 43]. One of the key reasons is the limited scale of publicly available video-text data, considering the high cost of video captioning [83]. Instead, it could be far easier to collect text-free video clips from media platforms like YouTube. There are some works sharing similar inspiration, Make-A-Video [50] and Gen-1 [12] employ a two-step strategy that first leverages a large ($\sim$1B parameters) diffusion prior model [42] to convert text embedding into image embedding of CLIP [41] and then enters it into an image-conditioned generator to synthesize videos. However, the separate two-step manner may cause issues such as error accumulation [13], increased model size and latency [42, 69], and does not support text-conditional optimization if extra video-text data is available, leading to sub-optimal results. Moreover, the characteristics of scaling potential on video generation are still under-explored.

In this work, we aim to train a single unified video diffusion model that allows text-guided video generation by exploiting the widely accessible text-free videos and explore its scaling trend. To achieve this, we present a novel two-branch framework named `TF-T2V`, where a content branch is designed for spatial appearance generation, and a motion branch specializes in temporal dynamics synthesis. More specifically, we utilize the publicly available image-text datasets such as LAION-5B [48] to learn text-guided and image-guided spatial appearance generation. In the motion branch, we harness the video-only data to conduct image-conditioned video synthesis, allowing the temporal modules to learn intricate motion patterns without relying on textual annotations. Paired video-text data, if available, can also be incorporated into co-optimization. Furthermore, unlike previous methods that impose training loss on each frame individually, we introduce a temporal coherence loss to explicitly enforce the learning of correlations between adjacent frames, enhancing the continuity of generated videos. In this way, the proposed `TF-T2V` achieves text-to-video generation by assembling contents and motions with a unified model, overcoming the high cost of video captioning and eliminating the need for complex cascading steps.

Notably, `TF-T2V` is a plug-and-play paradigm, which can be integrated into existing text-to-video generation and compositional video synthesis frameworks as shown in Fig. 1. Different from most prior works that rely heavily on video-text data and train models on the widely-used watermarked and low-resolution (around 360P) WebVid10M [2], `TF-T2V` opens up new possibilities for optimizing with text-free videos or partially paired video-text data, making it more scalable and versatile in widespread scenarios, such as high-definition video generation. To study the scaling trend, we double the scale of the training set with some randomly collected text-free videos and are encouraged to observe the performance improvement, with FID from 9.67 to 8.19 and FVD from 484 to 441. Extensive quantitative and qualitative experiments collectively demonstrate the effectiveness and scaling potential of the proposed `TF-T2V` in terms of synthetic continuity, fidelity, and controllability.

2. Related Work

In this section, we provide a brief review of relevant literature on text-to-image generation, text-to-video generation, and compositional video synthesis.

Text-to-image generation. Recently, text-to-image generation has made significant strides with the development of large-scale image-text datasets such as LAION-5B [48], allowing users to create high-resolution and photorealistic images that accurately depict the given natural language descriptions. Previous methods [16, 26, 49] primarily focus on synthesizing images by adopting generative adversarial networks (GANs) to estimate training sample distributions. Distinguished by the promising stability and scalability, diffusion-based generation models have attracted increasing attention [27, 42–45]. Diffusion models utilize iterative steps to gradually refine the generated image, resulting in improved quality and realism. Typically, Imagen [45] and GLIDE [38] explore text-conditional diffusion models and boost sample quality by applying classifier-free guidance [19]. DALL·E 2 [42] first leverages an image prior

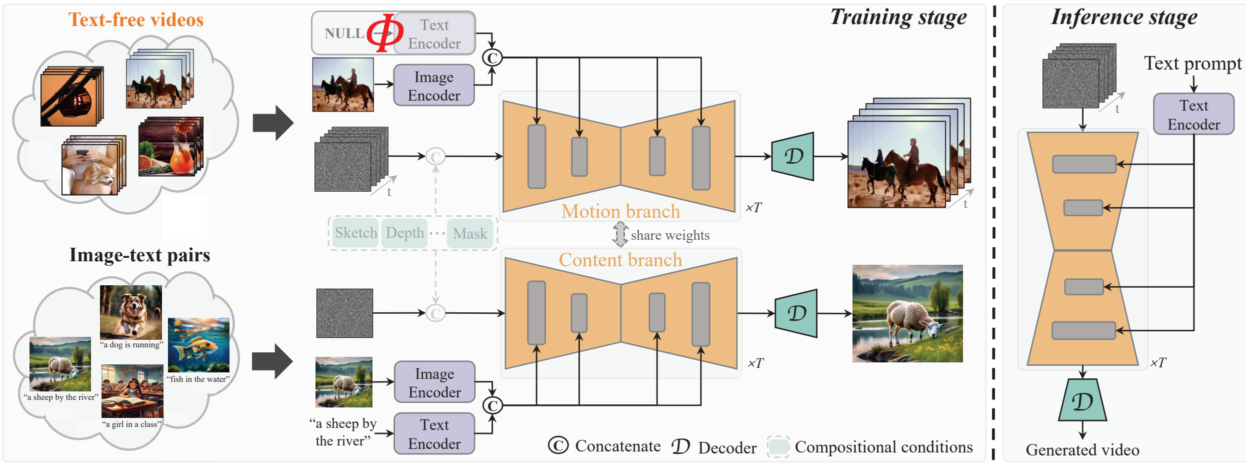

Figure 2. **Overall pipeline** of `TF-T2V`, which consists of two branches. In the content branch, paired image-text data is leveraged to learn text-conditioned and image-conditioned spatial appearance generation. The motion branch supports the training of motion dynamic synthesis by feeding text-free videos (or partially paired video-text data if available). During the training stage, both branches are optimized jointly. Notably, `TF-T2V` can be seamlessly integrated into the compositional video synthesis framework by incorporating composable conditions. In inference, `TF-T2V` enables text-guided video generation by taking text prompts and random noise sequences as input.

to bridge multi-modal embedding spaces and then learns a diffusion decoder to synthesize images in the pixel space. Stable Diffusion [43] introduces latent diffusion models that conduct iterative denoising processes at the latent level to save computational costs. There are also some works that generate customized and desirable images by incorporating additional spatial control signals [24, 36, 77].

Text-to-video generation. This task poses additional challenges compared to text-to-image generation due to the temporal dynamics involved in videos. Various early techniques have been proposed to tackle this problem, such as recurrent neural networks combined with GANs [3, 51, 53, 61, 64] or transformer-based autoregressive models [22, 73]. With the subsequent advent of video diffusion models pretrained on large-scale video-text datasets [2, 63, 71], video content creation has demonstrated remarkable advances [1, 4, 6–9, 14, 15, 17, 18, 21, 23, 28, 31–33, 35, 37, 39, 56, 57, 62, 65, 67, 69, 74–76]. Imagen Video [21] learns cascaded pixel-level diffusion models to produce high-resolution videos. Following [42], Make-A-Video [50] introduces a two-step strategy that first maps the input text to image embedding by a large (∼1B parameters) diffusion prior model and then embeds the resulting embedding into an image-conditional video diffusion model to synthesize videos in pixel space. VideoLDM [4] and ModelScopeT2V [54] extend 2D-UNet into 3D-UNet by injecting temporal layers and operate a latent denoising process to save computational resources. In this paper, we present a single unified framework for text-to-video generation and study the scaling trend by harnessing widely accessible text-free videos.

Compositional video synthesis. Traditional text-to-video methods solely rely on textual descriptions to control the video generation process, limiting desired fine-grained customization such as texture, object position, motion patterns, *etc*. To tackle this constraint and pursue higher controllability, several controllable video synthesis methods [8, 9, 12, 29, 58, 68, 72, 79, 81] have been proposed. These methods utilize additional control signals, such as depth or sketch, to guide the generation of videos. By incorporating extra structured guidance, the generated content can be precisely controlled and customized. Among these approaches, VideoComposer [58] stands out as a pioneering and versatile compositional technique. It integrates multiple conditioning signals including textual, spatial and temporal conditions within a unified framework, offering enhanced controllability, compositionality, and realism in the generated videos. Despite the remarkable quality, these methods still rely on high-quality video-text data to unleash powerful and customizable synthesis. In contrast, our method can be directly merged into existing controllable frameworks to customize videos by exploiting text-free videos.

3. Method

We first provide a brief introduction to the preliminaries of the video diffusion model. Then, we will elaborate on the mechanisms of `TF-T2V` in detail. The overall framework of the proposed `TF-T2V` is displayed in Fig. 2.

3.1. Preliminaries of video diffusion model

Diffusion models involve a forward diffusion process and a reverse iterative denoising stage. The forward process of diffusion models is gradually imposing random noise to clean data x_0 in a Markovian chain:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_{t-1}}x_{t-1}, \beta_t I), t = 1, ..., T \quad (1)$$

where $\beta_t \in (0, 1)$ is a noise schedule and T is the total time step. When T is sufficiently large, *e.g.* $T = 1000$, the resulting x_T is nearly a random Gaussian distribution $\mathcal{N}(0, I)$. The role of diffusion model is to denoise x_T and learn to iteratively estimate the reversed process:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sum\nolimits_\theta(x_t, t)) \quad (2)$$

We usually train a denoising model $\hat{x}_\theta$ parameterized by θ to approximate the original data x_0 and optimize the following v-prediction [21, 46] problem:

$$\mathcal{L}_{base} = \mathbb{E}_\theta[||v - \hat{x}_\theta(x_t, t, c)||_2^2] \quad (3)$$

where c is conditional information such as textual prompt, and v is the parameterized prediction objective. In representative video diffusion models [4, 54, 58], the denoising model $\hat{x}_\theta$ is a latent 3D-UNet [4, 54] modified from its 2D version [43] by inserting additional temporal blocks, which is optimized in the latent feature space by applying a variational autoencoder [11], and Eq. (3) is applied on each frame of the input video to train the whole model.

3.2. TF-T2V

The objective of `TF-T2V` is to learn a text-conditioned video diffusion model to create visually appealing and temporally coherent videos with text-free videos or partially paired video-text data. Without loss of generality, we first describe the workflow of our `TF-T2V` in the scenario where only text-free video is used. With merely text-free videos available for training, it is challenging to guide content creation by textual information since there lacks text-visual correspondence. To tackle this issue, we propose to resort to web-scale and high-quality image-text datasets [47, 48], which are publicly accessible on the Internet. However, this raises another question: *how can we leverage the image-text data and text-free videos in a unified framework?*

Recalling the network architecture in 3D-UNet, the spatial modules mainly focus on appearance modeling, and the temporal modules primarily aim to operate motion coherence. The intuition is that we can utilize image-text data to learn text-conditioned spatial appearance generation and adopt high-quality text-free videos to guide consistent motion dynamic synthesis. In this way, we can perform text-to-video generation in a single model to synthesize high-quality and consistent videos during the inference stage. Based on this, the proposed `TF-T2V` consists of two branches: a content branch for spatial appearance generation and a motion branch for motion dynamic synthesis.

3.2.1 Spatial appearance generation

Like previous text-to-image works [43, 77], the content branch of `TF-T2V` takes a noised image $I_{image} \in H \times W \times C$ as input, where H, W, C are the height, width, and channel dimensions respectively, and employs conditional signals (*i.e.*, text and image embeddings) to offer semantic guidance for content generation. This branch primarily concentrates on optimizing the spatial modules in the video diffusion model and plays a crucial role in determining appealing visual quality. In order to ensure that each condition can also control the created content separately, we randomly drop text or image embeddings with a certain probability during training. The text and image encoders from CLIP [41] are adopted to encode embeddings.

3.2.2 Motion dynamic synthesis

The pursuit of producing highly temporally consistent videos is a unique hallmark of video creation. Recent advancements [4, 54, 57, 58] in the realm of video synthesis usually utilize large-scale video-text datasets such as WebVid10M [2] to achieve coherent video generation. However, acquiring large-scale video-text pairs consumes extensive manpower and time, hindering the scaling up of video diffusion models. To make matters worse, the widely used WebVid10M is a watermarked and low-resolution (around 360P) dataset, resulting in unsatisfactory video creation that cannot meet the high-quality video synthesis requirements.

To mitigate the above issues, we propose to leverage high-quality text-free videos that are easily accessible on video media platforms, *e.g.*, YouTube and TikTok. To fully excavate the abundant motion dynamics within the text-free videos, we train a image-conditioned model. By optimizing this image-to-video generation task, the temporal modules in the video diffusion model can learn to perceive and model diverse motion dynamics. Specifically, given a noised video $I_{video} \in F \times H \times W \times C$, where F is the temporal length, the motion branch of `TF-T2V` learns to recover the undisturbed video guided by the image embedding. The image embedding is extracted from the center frame of the original video by applying CLIP's image encoder [41].

Since large-scale image-text data used for training contains abundant movement intentions [30], `TF-T2V` can achieve text-to-video generation by assembling spatial appearances involving motion trends and predicted motion dynamics. When extra paired video-text data is available, we conduct both text-to-video and image-to-video generation based on video-text pairs to train `TF-T2V` and further enhance the perception ability for desirable textual control.

In addition, we notice that previous works apply the training loss (*i.e.*, Eq. (3)) on each frame of the input video individually without considering temporal correlations between frames, suffering from incoherent appearances and motions. Inspired by the early study [25, 55, 59, 80] finding that the difference between two adjacent frames usually contains motion patterns, *e.g.*, dynamic trajectory, we thus

Table 1. **Quantitative comparison** with state-of-the-art methods for text-to-video task on MSR-VTT in terms of FID, FVD, and CLIPSIM.

Method	Zero-shot	Parameters	FID (↓)	FVD (↓)	CLIPSIM (↑)
Nüwa [66]	No	-	47.68	-	0.2439
CogVideo (Chinese) [22]	Yes	15.5B	24.78	-	0.2614
CogVideo (English) [22]	Yes	15.5B	23.59	1294	0.2631
MagicVideo [82]	Yes	-	-	1290	-
Make-A-Video [50]	Yes	9.7B	13.17	-	**0.3049**
ModelScopeT2V [54]	Yes	1.7B	11.09	550	0.2930
VideoComposer [58]	Yes	1.9B	10.77	580	0.2932
Latent-Shift [1]	Yes	1.5B	15.23	-	0.2773
VideoLDM [4]	Yes	4.2B	-	-	0.2929
PYoCo [14]	Yes	-	9.73	-	-
`TF-T2V` (WebVid10M)	Yes	1.8B	9.67	484	0.2953
`TF-T2V` (WebVid10M+Internal10M)	Yes	1.8B	**8.19**	**441**	0.2991

propose a temporal coherence loss that utilizes the frame difference as an additional supervisory signal:

$$\mathcal{L}_{coherence} = \mathbb{E}_{\theta}[\sum_{j=1}^{F-1}||(v_{j+1}-v_j)-(o_{j+1}-o_j)||_2^2] \quad (4)$$

where o_j and v_j are the predicted frame and corresponding ground truth. This loss term measures the discrepancy between the predicted frame differences and the ground truth frame differences of the input parameterized video. By minimizing Eq. (4), `TF-T2V` helps to alleviate frame flickering and ensures that the generated videos exhibit seamless transitions and promising temporal dynamics.

3.2.3 Training and inference

In order to mine the complementary advantages of spatial appearance generation and motion dynamic synthesis, we jointly optimize the entire model in an end-to-end manner. The total loss can be formulated as:

$$\mathcal{L}_{total} = \mathcal{L}_{base} + \lambda \mathcal{L}_{coherence} \quad (5)$$

where $\mathcal{L}_{base}$ is imposed on video and image together by treating the image as a "single frame" video, and λ is a balance coefficient that is set empirically to 0.1.

After training, we can perform text-guided video generation to synthesize temporally consistent video content that aligns well with the given text prompt. Moreover, `TF-T2V` is a general framework and can also be inserted into existing compositional video synthesis paradigm [58] by incorporating additional spatial and temporal structural conditions, allowing for customized video creation.

4. Experiments

In this section, we present a comprehensive quantitative and qualitative evaluation of the proposed `TF-T2V` on text-to-video generation and composition video synthesis.

4.1. Experimental setup

Implementation details. `TF-T2V` is built on two typical open-source baselines, *i.e.*, ModelScopeT2V [54] and VideoComposer [58]. DDPM sampler [20] with $T = 1000$ steps is adopted for training, and we employ DDIM [52] with 50 steps for inference. We optimize `TF-T2V` using AdamW optimizer with a learning rate of 5e-5. For input videos, we sample 16 frames from each video at 4FPS and crop a 448×256 region at the center as the basic setting. Note that we can also easily train high-definition video diffusion models by collecting high-quality text-free videos (see examples in the Appendix). LAION-5B [48] is utilized to provide image-text pairs. Unless otherwise stated, we treat WebVid10M, which includes about 10.7M video-text pairs, as a text-free dataset to train `TF-T2V` and do not use any textual annotations. To study scaling trends, we gathered about 10M high-quality videos without text labels from internal data, termed the Internal10M dataset.

Table 2. **Human preference results** on text-to-video generation.

Method	Text alignment	Visual quality	Temporal coherence
ModelScopeT2V [54]	83.5%	74.0%	81.3%
`TF-T2V`	**86.5%**	**87.0%**	**92.5%**

Metrics. (i) To evaluate text-to-video generation, following previous works [4, 54], we leverage the standard Fréchet Inception Distance (FID), Fréchet Video Distance (FVD), and CLIP Similarity (CLIPSIM) as quantitative evaluation metrics and report results on MSR-VTT dataset [70]. (ii) For controllability evaluation, we leverage depth error, sketch error, and end-point-error (EPE) [10] to verify whether the generated videos obey the control of input conditions. Depth error measures the divergence between the input depth conditions and the eliminated depth of the synthesized video. Similarly, sketch error examines the sketch control. EPE evaluates the flow consistency between the reference video and the generated video. In addition, human evaluation is also introduced to validate our method.

4.2. Evaluation on text-to-video generation

Tab. 1 displays the comparative quantitative results with existing state-of-the-art methods. We observe that `TF-T2V` achieves remarkable performance under various metrics.

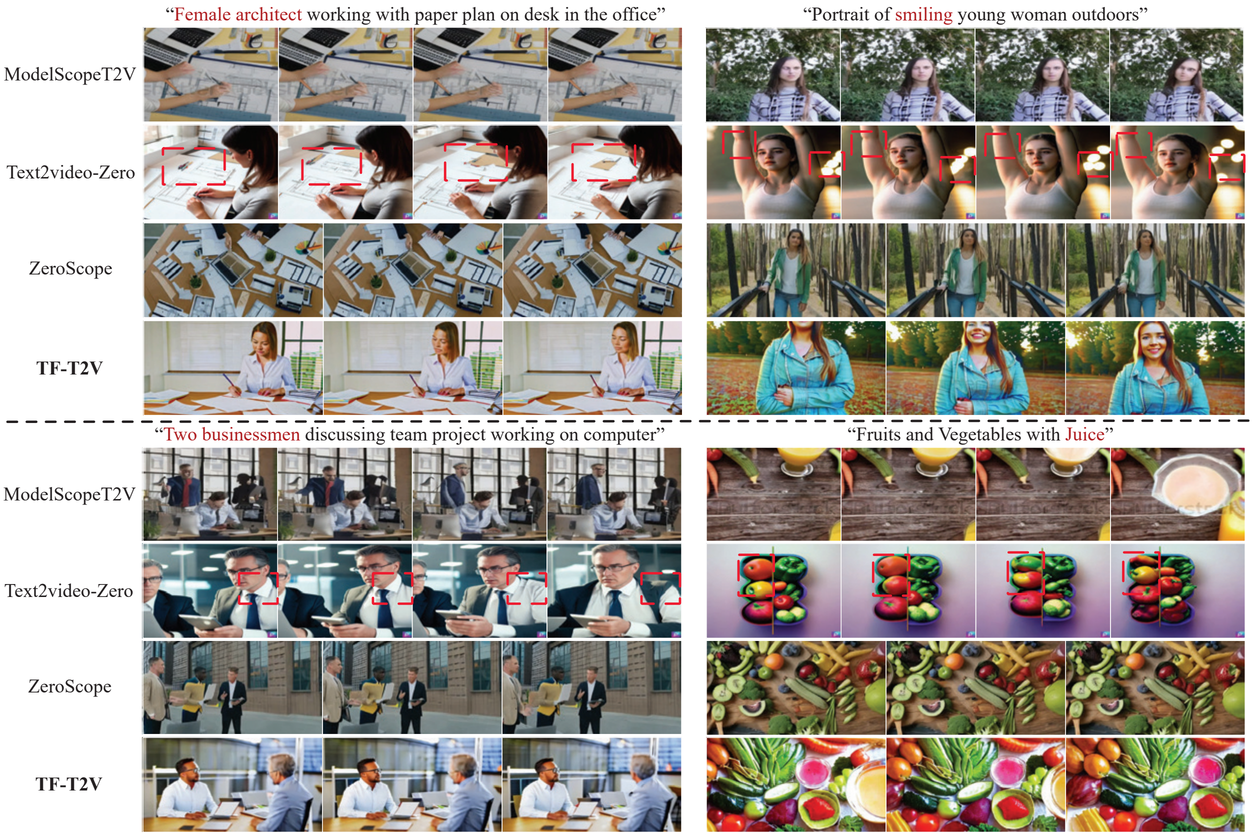

Figure 3. **Qualitative comparison on text-to-video generation**. Three representative open-source text-to-video approaches are compared, including ModelScopeT2V [54], Text2video-Zero [28] and ZeroScope [5]. Please refer to the Appendix for videos and more comparisons.

Table 3. **Evaluation of structure control** based on depth signals.

Method	Condition	Depth error (↓)
VideoComposer [58]	Text	0.382
VideoComposer [58]	Text and depth	0.217
`TF-T2V`	Text and depth	**0.209**

Table 4. **Evaluation of structure control** based on sketch signals.

Method	Condition	Sketch error (↓)
VideoComposer [58]	Text	0.1854
VideoComposer [58]	Text and sketch	0.1161
`TF-T2V`	Text and sketch	**0.1146**

Table 5. **Evaluation of motion control** based on motion vectors.

Method	Condition	EPE (↓)
VideoComposer [58]	Text	4.13
VideoComposer [58]	Text and motion vector	1.98
`TF-T2V`	Text and motion vector	**1.88**

Table 6. **Human evaluations** on compositional video synthesis.

Method	Structure alignment	Visual quality	Temporal coherence
VideoComposer [58]	79.0%	66.0%	77.5%
`TF-T2V`	**89.0%**	**79.5%**	**84.5%**

Notably, `TF-T2V` trained on WebVid10M and Internal10M obtains higher performance than the counterpart on WebVid10M, revealing promising scalable capability. We show the qualitative visualizations in Fig. 3. From the results, we can find that compared with previous methods, `TF-T2V` obtains impressive video creation in terms of both temporal continuity and visual quality. The human assessment in Tab. 2 also reveals the above observations. The user study is performed on 100 randomly synthesized videos.

4.3. Evaluation on compositional video synthesis

We compare the controllability of `TF-T2V` and VideoComposer on 1,000 generated videos in terms of depth control (Tab. 3), sketch control (Tab. 4) and motion control (Tab. 5). The above experimental evaluations highlight the effectiveness of `TF-T2V` by leveraging text-free videos. In Fig. 4 and 5, we show the comparison of `TF-T2V` and existing methods on compositional video generation. We notice that `TF-T2V` exhibits high-fidelity and consistent video generation. In addition, we conduct a human evaluation on 100 randomly sampled videos and report the results in Tab. 6. The preference assessment provides further evidence of the superiority of the proposed `TF-T2V`.

4.4. Ablation study

Effect of temporal coherence loss. To enhance temporal consistency, we propose a temporal coherence loss. In Tab. 7, we show the effectiveness of the proposed tem-

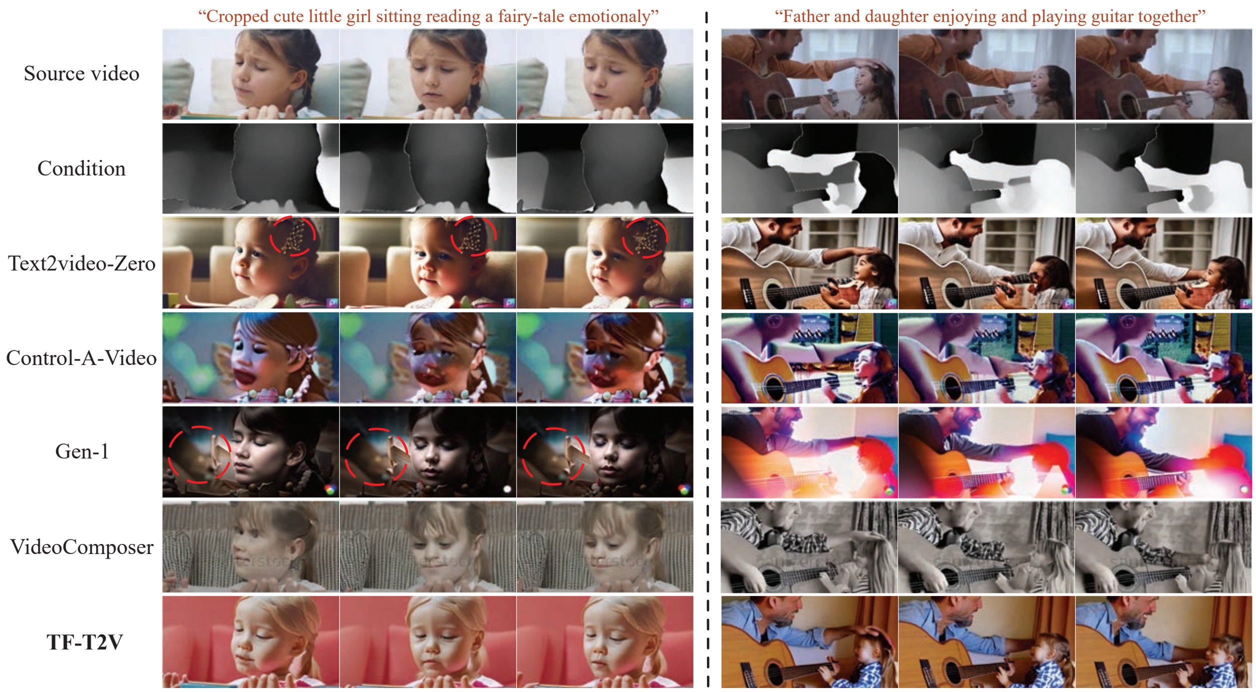

Figure 4. **Qualitative comparison on compositional depth-to-video generation**. The videos are generated by taking textual prompts and structural guidance as conditions. Compared with existing methods, `TF-T2V` yields more structural compliance and high-fidelity results.

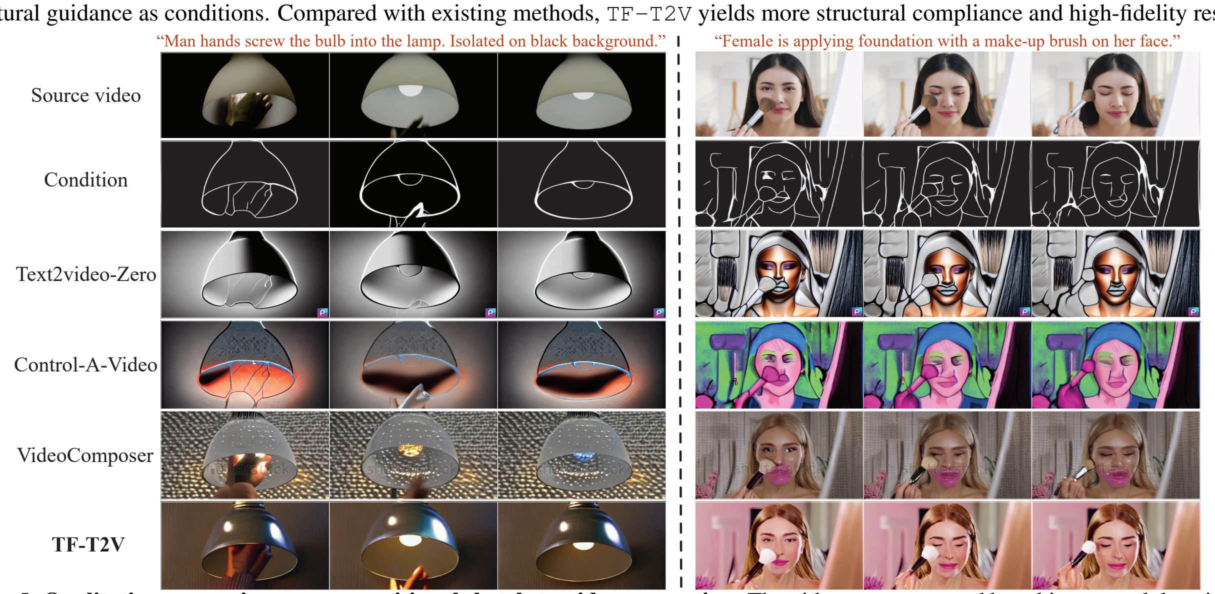

Figure 5. **Qualitative comparison on compositional sketch-to-video generation**. The videos are generated by taking textual descriptions and structural guidance as conditions. Compared with other methods, `TF-T2V` produces more realistic and consistent results.

poral coherence loss in terms of frame consistency. The metric results are obtained by calculating the average CLIP similarity of two consecutive frames in 1,000 videos. We further display the qualitative comparative results in Fig. 6 and observe that temporal coherence loss helps to alleviate temporal discontinuity such as color shift.

Table 7. **Text-to-video evaluation** on frame consistency.

Method	Frame consistency (%) ↑
w/o temporal coherence loss	89.71
`TF-T2V`	**91.06**

4.5. Evaluation on semi-supervised setting

Through the above experiments and observations, we verify that text-free video can help improve the continuity and quality of generated video. As previously stated, `TF-T2V` also supports the combination of annotated video-text data and text-free videos to train the model, *i.e.*, the semi-supervised manner. The annotated text can provide additional fine-grained motion signals, enhancing the align-

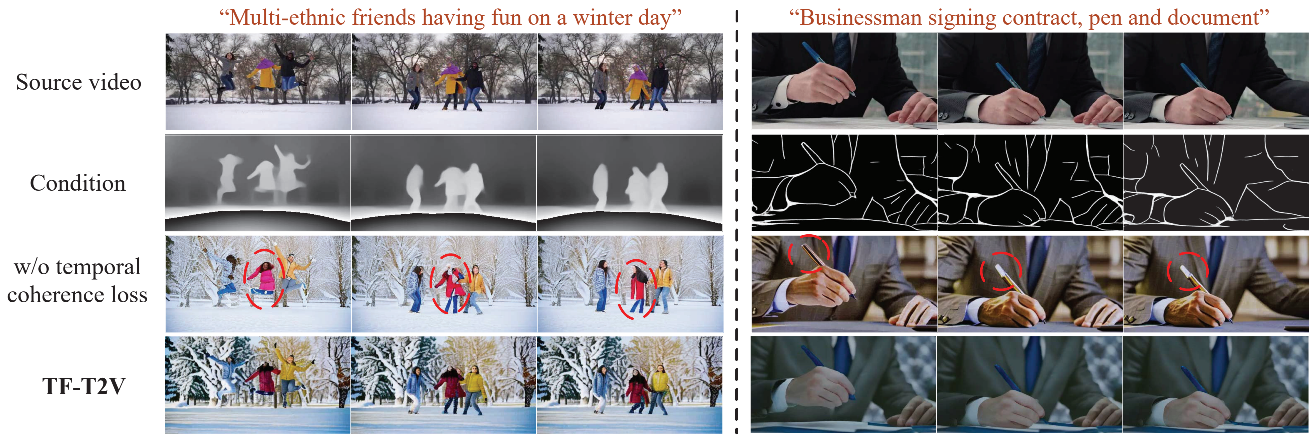

Figure 6. **Qualitative ablation study**. The videos are generated by taking textual descriptions and structural guidance as conditions.

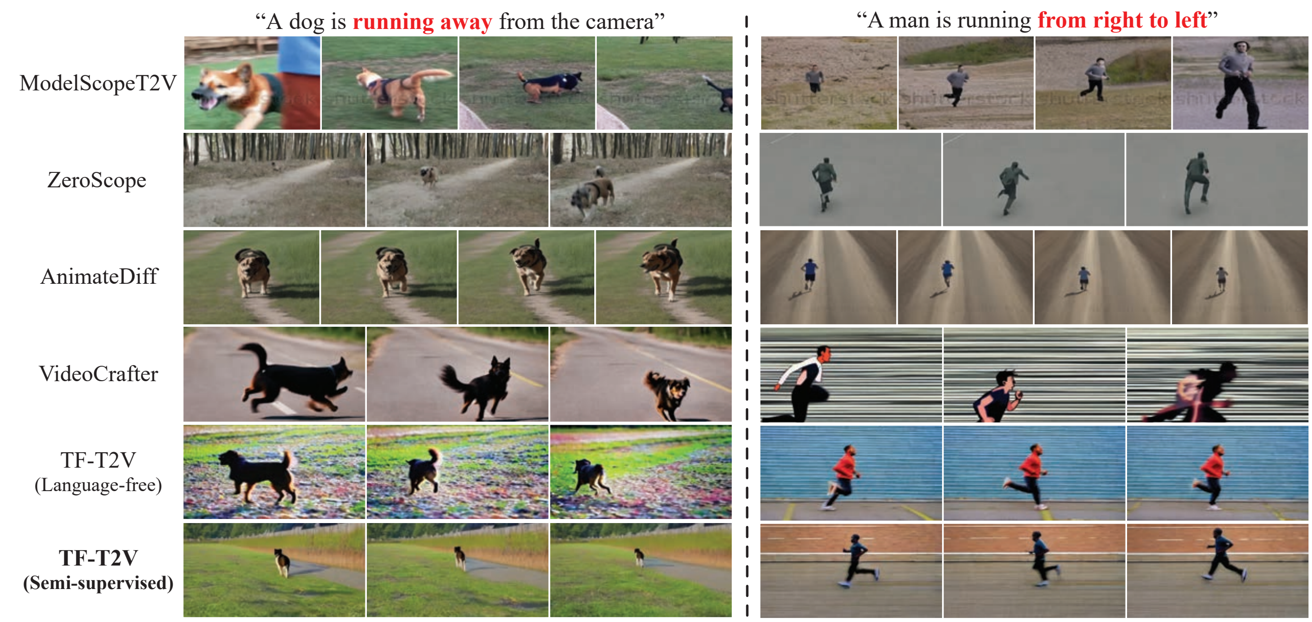

Figure 7. **Qualitative evaluation** on text-to-video generation with temporally-correlated text prompts involving the evolution of movement.

Table 8. **Quantitative experiments on text-to-video generation**. TF−T2V-Semi means the semi-supervised setting where labeled WebVid10M and text-free Internal10M are adopted.

Method	FID (↓)	FVD (↓)	CLIPSIM (↑)
ModelScopeT2V [54]	11.09	550	0.2930
TF−T2V	8.19	441	0.2991
TF−T2V-*Semi*	**7.64**	**366**	**0.3032**

ment of generated videos and the provided prompts involving desired motion evolution. We show the comparison results in Tab. 8 and find that the semi-supervised manner reaches the best performance, indicating the effectiveness of harnessing text-free videos. Notably, TF−T2V-Semi outperforms ModelScopeT2V trained on labeled WebVid10M, possessing good scalability. Moreover, the qualitative evaluations in Fig. 7 show that existing methods may struggle to synthesize text-aligned consistent videos when textual prompts involve desired temporal evolution. In contrast, TF−T2V in the semi-supervised setting exhibits excellent text-video alignment and temporally smooth generation.

5. Conclusion

In this paper, we present a novel and versatile video generation framework named TF−T2V to exploit text-free videos and explore its scaling trend. TF−T2V effectively decomposes video generation into spatial appearance generation and motion dynamic synthesis. A temporal coherence loss is introduced to explicitly constrain the learning of correlations between adjacent frames. Experimental results demonstrate the effectiveness and potential of TF−T2V in terms of fidelity, controllability, and scalability.

Acknowledgements. This work is supported by the National Natural Science Foundation of China under grant U22B2053 and Alibaba Research Intern Program.

References

[1] Jie An, Songyang Zhang, Harry Yang, Sonal Gupta, Jia-Bin Huang, Jiebo Luo, and Xi Yin. Latent-shift: Latent diffusion with temporal shift for efficient text-to-video generation. *arXiv preprint arXiv:2304.08477*, 2023. 3, 5

[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 *ICCV*, pages 1728–1738, 2021. 2, 3, 4

[3] Yogesh Balaji, Martin Renqiang Min, Bing Bai, Rama Chellappa, and Hans Peter Graf. Conditional GAN with discriminative filter generation for text-to-video synthesis. In *IJCAI*, page 2, 2019. 3

[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 *CVPR*, pages 22563–22575, 2023. 2, 3, 4, 5

[5] Cerspense. Zeroscope: Diffusion-based text-to-video synthesis. https://huggingface.co/cerspense/zeroscope_v2_576w, 2023. 6

[6] Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. In *ICCV*, pages 23206–23217, 2023. 3

[7] Wenhao Chai, Xun Guo, Gaoang Wang, and Yan Lu. Stablevideo: Text-driven consistency-aware diffusion video editing. In *ICCV*, pages 23040–23050, 2023.

[8] Tsai-Shien Chen, Chieh Hubert Lin, Hung-Yu Tseng, Tsung-Yi Lin, and Ming-Hsuan Yang. Motion-conditioned diffusion model for controllable video synthesis. *arXiv preprint arXiv:2304.14404*, 2023. 3

[9] Weifeng Chen, Jie Wu, Pan Xie, Hefeng Wu, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. Control-a-video: Controllable text-to-video generation with diffusion models. *arXiv preprint arXiv:2305.13840*, 2023. 3

[10] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *ICCV*, pages 2758–2766, 2015. 5

[11] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming Transformers for high-resolution image synthesis. In *CVPR*, pages 12873–12883, 2021. 4

[12] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *ICCV*, pages 7346–7356, 2023. 2, 3

[13] Rafail Fridman, Amit Abecasis, Yoni Kasten, and Tali Dekel. Scenescape: Text-driven consistent scene generation. *arXiv preprint arXiv:2302.01133*, 2023. 2

[14] Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. In *ICCV*, pages 22930–22941, 2023. 3, 5

[15] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023. 3

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *NeurIPS*, 27, 2014. 2

[17] 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. 3

[18] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *arXiv preprint arXiv:2211.13221*, 2022. 3

[19] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020. 5

[21] 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. 3, 4

[22] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via Transformers. In *ICLR*, 2023. 2, 3, 5

[23] 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. *arXiv preprint arXiv:2309.14494*, 2023. 3

[24] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. *ICML*, 2023. 3

[25] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J Black. Towards understanding action recognition. In *ICCV*, pages 3192–3199, 2013. 4

[26] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up GANs for text-to-image synthesis. In *CVPR*, pages 10124–10134, 2023. 2

[27] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *CVPR*, pages 6007–6017, 2023. 2

[28] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439*, 2023. 2, 3, 6

[29] Ariel Lapid, Idan Achituve, Lior Bracha, and Ethan Fetaya. Gd-vdm: Generated depth for better diffusion-based video generation. *arXiv preprint arXiv:2306.11173*, 2023. 3

[30] Jiangtong Li, Li Niu, and Liqing Zhang. Action-aware embedding enhancement for image-text retrieval. In *AAAI*, pages 1323–1331, 2022. 4

[31] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. *arXiv preprint arXiv:2303.04761*, 2023. 3

[32] Haoyu Lu, Guoxing Yang, Nanyi Fei, Yuqi Huo, Zhiwu Lu, Ping Luo, and Mingyu Ding. Vdt: An empirical study on video diffusion with Transformers. *arXiv preprint arXiv:2305.13311*, 2023.

[33] 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. In *CVPR*, pages 10209–10218, 2023. 2, 3

[34] Yifeng Ma, Shiwei Zhang, Jiayu Wang, Xiang Wang, Yingya Zhang, and Zhidong Deng. Dreamtalk: When expressive talking head generation meets diffusion probabilistic models. *arXiv preprint arXiv:2312.09767*, 2023. 2

[35] Eyal Molad, Eliahu Horwitz, Dani Valevski, Alex Rav Acha, Yossi Matias, Yael Pritch, Yaniv Leviathan, and Yedid Hoshen. Dreamix: Video diffusion models are general video editors. *arXiv preprint arXiv:2302.01329*, 2023. 3

[36] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 3

[37] Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In *CVPR*, pages 18444–18455, 2023. 3

[38] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*, pages 16784–16804. PMLR, 2022. 2

[39] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. In *ICCV*, 2023. 2, 3

[40] Zhiwu Qing, Shiwei Zhang, Jiayu Wang, Xiang Wang, Yujie Wei, Yingya Zhang, Changxin Gao, and Nong Sang. Hierarchical spatio-temporal decoupling for text-to-video generation. *arXiv preprint arXiv:2312.04483*, 2023. 2

[41] 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 *ICML*, pages 8748–8763. PMLR, 2021. 2, 4

[42] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 2, 3

[43] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2, 3, 4

[44] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, pages 22500–22510, 2023.

[45] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 35:36479–36494, 2022. 2

[46] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022. 4

[47] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 4

[48] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 35:25278–25294, 2022. 2, 4, 5

[49] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in GANs. In *CVPR*, pages 1532–1540, 2021. 2

[50] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *ICLR*, 2023. 2, 3, 5

[51] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. StyleGAN-v: A continuous video generator with the price, image quality and perks of StyleGAN2. In *CVPR*, pages 3626–3636, 2022. 3

[52] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 5

[53] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. MocoGAN: Decomposing motion and content for video generation. In *CVPR*, pages 1526–1535, 2018. 2, 3

[54] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 2, 3, 4, 5, 6, 8

[55] Limin Wang, Zhan Tong, Bin Ji, and Gangshan Wu. Tdn: Temporal difference networks for efficient action recognition. In *CVPR*, pages 1895–1904, 2021. 4

[56] Wen Wang, Kangyang Xie, Zide Liu, Hao Chen, Yue Cao, Xinlong Wang, and Chunhua Shen. Zero-shot video editing using off-the-shelf image diffusion models. *arXiv preprint arXiv:2303.17599*, 2023. 3

[57] Wenjing Wang, Huan Yang, Zixi Tuo, Huiguo He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. Videofactory: Swap attention in spatiotemporal diffusions for text-to-video generation. *arXiv preprint arXiv:2305.10874*, 2023. 3, 4

[58] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *NeurIPS*, 2023. 2, 3, 4, 5, 6

[59] Xiang Wang, Shiwei Zhang, Zhiwu Qing, Changxin Gao, Yingya Zhang, Deli Zhao, and Nong Sang. Molo: Motion-

augmented long-short contrastive learning for few-shot action recognition. In *CVPR*, pages 18011–18021, 2023. 4

[60] Xiang Wang, Shiwei Zhang, Han Zhang, Yu Liu, Yingya Zhang, Changxin Gao, and Nong Sang. Videolcm: Video latent consistency model. *arXiv preprint arXiv:2312.09109*, 2023. 2

[61] Yaohui Wang, Piotr Bilinski, Francois Bremond, and Antitza Dantcheva. G3an: Disentangling appearance and motion for video generation. In *CVPR*, pages 5264–5273, 2020. 3

[62] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023. 3

[63] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023. 3

[64] Yuhan Wang, Liming Jiang, and Chen Change Loy. Styleinv: A temporal style modulated inversion network for unconditional video generation. In *ICCV*, pages 22851–22861, 2023. 3

[65] Yujie Wei, Shiwei Zhang, Zhiwu Qing, Hangjie Yuan, Zhiheng Liu, Yu Liu, Yingya Zhang, Jingren Zhou, and Hongming Shan. Dreamvideo: Composing your dream videos with customized subject and motion. *arXiv preprint arXiv:2312.04433*, 2023. 3

[66] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. Nüwa: Visual synthesis pre-training for neural visual world creation. In *ECCV*, pages 720–736. Springer, 2022. 5

[67] 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 *ICCV*, pages 7623–7633, 2023. 2, 3

[68] 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. *arXiv preprint arXiv:2306.00943*, 2023. 3

[69] Zhen Xing, Qi Dai, Han Hu, Zuxuan Wu, and Yu-Gang Jiang. Simda: Simple diffusion adapter for efficient video generation. *arXiv preprint arXiv:2308.09710*, 2023. 2, 3

[70] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, pages 5288–5296, 2016. 5

[71] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In *CVPR*, 2022. 3

[72] 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. 3

[73] Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, et al. Magvit: Masked generative video Transformer. In *CVPR*, pages 10459–10469, 2023. 3

[74] Sihyun Yu, Kihyuk Sohn, Subin Kim, and Jinwoo Shin. Video probabilistic diffusion models in projected latent space. In *CVPR*, pages 18456–18466, 2023. 3

[75] Hangjie Yuan, Shiwei Zhang, Xiang Wang, Yujie Wei, Tao Feng, Yining Pan, Yingya Zhang, Ziwei Liu, Samuel Albanie, and Dong Ni. Instructvideo: Instructing video diffusion models with human feedback. *arXiv preprint arXiv:2312.12490*, 2023.

[76] 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. 3

[77] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 3, 4

[78] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023. 2

[79] 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. 2, 3

[80] Zhang Zhang and Dacheng Tao. Slow feature analysis for human action recognition. *TPAMI*, 34(3):436–450, 2012. 4

[81] Min Zhao, Rongzhen Wang, Fan Bao, Chongxuan Li, and Jun Zhu. Controlvideo: Adding conditional control for one shot text-to-video editing. *arXiv preprint arXiv:2305.17098*, 2023. 3

[82] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022. 5

[83] Junbao Zhuo, Xingyu Zhao, Shuhui Wang, Huimin Ma, and Qingming Huang. Synthesizing videos from images for image-to-video adaptation. In *ACMMM*, pages 8294–8303, 2023. 2