Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation

Jay Zhangjie Wu[1] Yixiao Ge[2] Xintao Wang[2] Stan Weixian Lei[1] Yuchao Gu[1]
Yufei Shi[1] Wynne Hsu[4] Ying Shan[2] Xiaohu Qie[3] Mike Zheng Shou[1]*

[1]Show Lab, National University of Singapore [2]ARC Lab,[3]Tencent PCG
[4]School of Computing, National University of Singapore

https://tuneavideo.github.io

Figure 1: ***Tune-A-Video:*** A new method for T2V generation using one text-video pair and pretrained T2I models.

Abstract

To replicate the success of text-to-image (T2I) generation, recent works employ large-scale video datasets to train a text-to-video (T2V) generator. Despite their promising results, such paradigm is computationally expensive. In this work, we propose a new T2V generation setting—One-Shot Video Tuning, where only one text-video pair is presented. Our model is built on state-of-the-art T2I diffusion models pre-trained on massive image data. We make two key observations: 1) T2I models can generate still images that represent verb terms; 2) extending T2I models to generate mul-

*Corresponding Author.

tiple images concurrently exhibits surprisingly good content consistency. To further learn continuous motion, we introduce Tune-A-Video, which involves a tailored spatio-temporal attention mechanism and an efficient one-shot tuning strategy. At inference, we employ DDIM inversion to provide structure guidance for sampling. Extensive qualitative and numerical experiments demonstrate the remarkable ability of our method across various applications.

1. Introduction

The large-scale multimodal dataset [41], consisting of billions of text-image pairs crawled from the Internet, has enabled a breakthrough in Text-to-Image (T2I) generation [30, 35, 6, 42, 40]. To replicate this success in Text-to-Video (T2V) generation, recent works [42, 15, 18, 53, 47] have extended spatial-only T2I generation models to the spatio-temporal domain. These models generally adopt the standard paradigm of training on large-scale text-video datasets (*e.g.*, WebVid-10M [2]). Although this paradigm produces promising results for T2V generation, it requires extensive training on large hardware accelerators, which is expensive and time-consuming.

Humans possess the ability to create new concepts, ideas, or things by utilizing their existing knowledge and the information provided to them. For example, when presented a video with a textual description of "a man skiing on snow", we can imagine how a panda would ski on snow, drawing upon our knowledge of what a panda looks like. As T2I models pretrained with large-scale image-text data already capture knowledge of open-domain concepts, a intuitive question arises: *can they infer other novel videos from a single video example, like humans?* A new T2V generation setting is therefore introduced, namely, One-Shot Video Tuning, where only a single text-video pair is used to train a T2V generator. The generator is expected to capture essential motion information from the input video and synthesize novel videos with edited prompts.

Intuitively, the key to successful video generation lies in preserving the continuous motion of consistent objects. So we make the following observations on state-of-the-art T2I diffusion models [37] that motivate our method accordingly. (1) **Regarding motion:** T2I models are able to generate images that align well with the text, including the verb terms. For example, given the text prompt "a man is running on the beach", the T2I models produce the snapshot where a man is running (not walking or jumping), albeit not necessarily in a continuous manner (the first row of Fig. 2). This serves as evidence that T2I models can properly attend to verbs via cross-modal attention for static motion generation. (2) **Regarding consistent objects:** Simply extending the spatial self-attention in the T2I model from one image to multiple images produces consistent content across frames. Taking

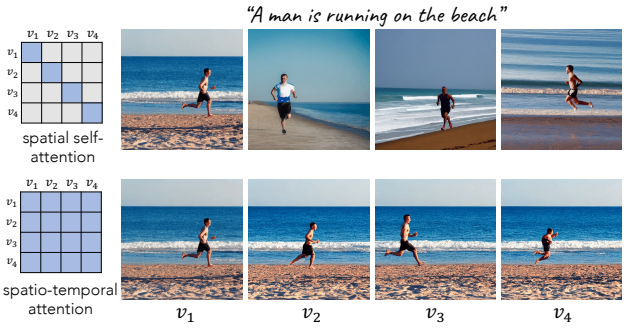


Figure 2: ***Observations on pretrained T2I models:*** 1) They can generate still images that accurately represent the verb terms. 2) Extending spatial self-attention to spatio-temporal attention produces consistent content across frames.

the same example, when we generate consecutive frames in parallel with extended spatio-temporal attention, the same man and the same beach can be observed in the resultant sequence though the motion is still not continuous (the second row of Fig. 2). This implies that the self-attention layers in T2I models are only driven by spatial similarities rather than pixel positions.

We implement our findings into a simple yet effective method called Tune-A-Video. Our method is based on a simple inflation of state-of-the-art T2I models over spatio-temporal dimension. However, using full attention in space-time inevitably leads to quadratic growth in computation. It is thus infeasible for generating videos with increasing frames. Additionally, employing a naive fine-tuning strategy that updates all the parameters can jeopardize the pre-existing knowledge of T2I models and hinder the generation of videos with new concepts. To tackle these problems, we introduce a sparse spatio-temporal attention mechanism that only visits the *first* and the *former* video frame, as well as an efficient tuning strategy that only updates the projection matrices in attention blocks. Empirically, these designs maintain consistent objects across all frames but lack continuous motion. Therefore, at inference, we further seek structure guidance from input video through DDIM inversion, which is a reverse process of DDIM sampling [43]. With the inverted latent as initial noise, we produce temporally-coherent videos featuring smooth movement. Notably, our method is inherently compatible with exiting personalized and conditional pretrained T2I models, such as DreamBooth [39] and T2I-Adapter [29], providing a personalized and controllable user interface.

We showcase remarkable results of Tune-A-Video across a wide range of applications for text-driven video generation (see Fig. 1). We compare our method against the state-of-the-art baselines through extensive qualitative and quantitative experiments, demonstrating its superiority. In summary, our key contributions are as follows:

- We introduce a new setting of One-Shot Video Tuning for T2V generation, which eliminates the burden of training with large-scale video datasets.
- We present Tune-A-Video, which is the first framework for T2V generation using pretrained T2I models.
- We propose efficient attention tuning and structural inversion that significant improve temporal consistency.
- We demonstrate remarkable results of our method through extensive experiments.

2. Related Work

Our work lies in the intersection of several fields: diffusion models and methods for image/video generation from text prompts, text-driven editing of a real image/video, and generative models trained on a single video. Here we provide a brief overview of the key accomplishments in each field, highlighting their connections and differences from our proposed method.

Text-to-Image diffusion models. Text-to-Image (T2I) generation has been studied extensively, in past years many of the models were based on transformers [36, 51, 50, 6, 9]. Several T2I generative models [30, 40, 10, 37] have recently adopted diffusion models [16]. GLIDE [30] proposes classifier-free guidance [17] in the diffusion model to improve image quality, while DALLE-2 [35] improves text-image alignments using CLIP [34] feature space. Imagen [40] uses cascaded diffusion models for high definition video generation, and subsequent works like VQ-diffusion [10] and Latent Diffusion Models (LDMs) [37] operate in the latent space of an autoencoder to improve training efficiency. Our method builds on LDMs, by inflating the 2D model to spatio-temporal domain in latent space.

Text-to-Video generative models. While there have been significant advancements in T2I generation, generating videos from text is still lagging behind due to the scarcity of high-quality, large-scale text-video datasets, and the inherent complexity of modeling temporal consistency and coherence. Early works [27, 32, 25, 23, 11, 24] primarily focus on generating videos in simple domains, such as moving digits or specific human actions. Recently, GODIVA [45] is the first model to utilize 2D VQ-VAE and sparse attention for T2V generation, which allows for more realistic scenes. NÜWA [48] expands upon GODIVA by presenting a unified representation for various generation tasks through a multitask learning approach. To further enhance T2V generation performance, CogVideo [19] is developed by incorporating additional temporal attention modules on top of a pre-trained T2I model, CogView2 [6].

To replicate the success of T2I diffusion models, Video Diffusion Models (VDM) [18] uses a space-time factorized U-Net with joint image and video data training. Imagen Video [15] improves VDM using cascaded diffusion models and v-prediction parameterization to generate high definition videos. Make-A-Video [42] and MagicVideo [53] share similar motivations and aim to transfer progress from T2I generation to T2V generation. Although current T2V generative models have shown impressive results, their success heavily rely on being trained using extensive video data. In contrast, we present a new framework for T2V generation via an efficient tuning of pre-trained T2I diffusion models on one text-video pair.

Text-driven video editing. Recent diffusion-based image editing models [26, 14, 5, 49, 21, 44] can process each individual frame in a video, but this produces inconsistency between frames due to the lack of temporal awareness in the model. Text2Live [3] allows some texture-based video editing using text prompts, but struggles to accurately reflect the intended edits due to its dependence on Layered Neural Atlases [20]. Moreover, generating a neural atlas typically takes about 10 hours, whereas our approach only requires a 10-minute training per video and can sample a video in just 1 minute. Two concurrent works, Dreamix [28] and Gen-1 [7], both utilize the video diffusion model (VDM) for video editing purposes. Although their impressive outcomes, it is worth noting that the VDMs are computationally demanding and necessitate large-scale captioned images and videos for training. Additionally, their training data and pre-trained models are not publicly accessible.

Generation from a single video. Single-video GANs [1, 12] generate new videos of similar appearance and dynamics to the input video. However, these GAN-based methods are limited in computation time (e.g., HPVAE-GAN [12] takes 8 days to train on a short video of 13 frames), and thus are impractical and unscalable to some extent. Patch nearest-neighbour methods [13] perform video generation of higher quality while reducing computation expense by orders of magnitude. However, they are limited in generalization, and therefore can only handle tasks where it is natural to "copy" parts of the input video. Lately, SinFusion [31] adapts diffusion models to single-video tasks, and enables autoregressive video generation with improved motion generalization capabilities; however, it is still incapable of producing videos that contains novel semantic contexts.

3. Method

Let $\mathcal{V} = \{v_i | i \in [1, m]\}$ be a video containing m frames, $\mathcal{P}$ be the source prompt describing $\mathcal{V}$. Our goal is to generate a novel video $\mathcal{V}^*$ driven by an edited text prompt $\mathcal{P}^*$. For example, consider a video and a source prompt "a man is skiing", and assume that the user wants to alter the color of the clothes, incorporate a cowboy hat to the skier, or even replace the skier with Spider Man while preserving the motion of the original video. The user can directly modify the

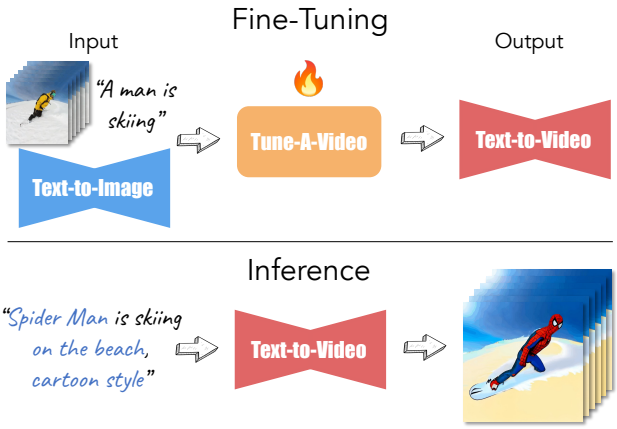


Figure 3: ***High-level overview of Tune-A-Video.*** Given a captioned video, we finetune a pre-trained T2I model (*e.g.*, Stable Diffusion) for T2V modeling. During inference, we generate novel videos that represent the edits in text prompt while preserving the temporal consistency of input video.

source prompt by further describing the appearance of the skier or replacing it with another word.

An intuitive solution is to train a T2V model on large-scale video datasets, but it is computationally expensive [42, 15, 7]. In this paper, we propose a new setting called One-Shot Video Tuning that achieves the same goal using a publicly available T2I model and a single text-video pair.

Next, we provide a short background of diffusion models in Sec. 3.1, followed by a detailed description of our method in Sec. 3.2 and Sec. 3.3. An overview of our approach is depicted in Fig. 3.

3.1. Preliminaries

Denoising diffusion probabilistic models (DDPMs). DDPMs [16] are latent generative models trained to recreate a fixed forward Markov chain $x_1, \dots, x_T$. Given the data distribution $x_0 \sim q(x_0)$, the Markov transition $q(x_t|x_{t-1})$ is defined as a Gaussian distribution with a variance schedule $\beta_t \in (0, 1)$, that is,

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

By the Bayes' rules and Markov property, one can explicitly express the conditional probabilities $q(x_t|x_0)$ and $q(x_{t-1}|x_t, x_0)$ as

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbb{I}), \quad t = 1, \dots, T,$$
$$q(x_{t-1}|x_t, x_0) = \mathcal{N}(x_{t-1}; \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t\mathbb{I}), \quad t = 1, \dots, T,$$
$$w.r.t. \quad \alpha_t = 1-\beta_t,\ \bar{\alpha}_t = \prod_{s=1}^{t}\alpha_s,\ \tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t,$$
$$\tilde{\mu}_t(x_t, x_0) = \frac{\sqrt{\bar{\alpha}_t}\beta_t}{1-\bar{\alpha}_t}x_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t.$$

To generate the Markov chain $x_1, \dots, x_T$, DDPMs leverage the reverse process with a prior distribution $p(x_T) = \mathcal{N}(x_T; 0, \mathbb{I})$ and Gaussian transitions

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)), \quad t = T, \dots, 1.$$

Learnable parameters θ are trained to guarantee that the generated reverse process is close to the forward process.

To this end, DDPMs follow the variational inference principle by maximizing the variational lower bound of the negative log-likelihood, which has a closed-form given the KL divergence among Guassian distributions. Empirically, these models can be interpreted as a sequence of weight-sharing denoising autoencoders $\epsilon_\theta(x_t, t)$, which are trained to predict a denoised variant of their input x_t. The objective can be simplified as $\mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\right]$.

Latent diffusion models (LDMs). LDMs [37] are newly introduced variants of DDPMs that operate in the latent space of an autoencoder. LDMs consist of two key components. First, an autoencoder [8, 45] is trained with patch-wise losses on a large collection of images, where an encoder $\mathcal{E}$ learns to compress images x into latent representations $z = \mathcal{E}(x)$, and a decoder $\mathcal{D}$ learns to reconstruct the latent back to pixel space, such that $\mathcal{D}(\mathcal{E}(x)) \approx x$. The second component is a DDPM that is trained to remove the noise added to the sampled data. For a text-guided LDM, the objective is given by: $\mathbb{E}_{z,\epsilon\sim\mathcal{N}(0,1),t,c}\left[\|\epsilon - \epsilon_\theta(z_t, t, c)\|_2^2\right]$., where $c = \psi(\mathcal{P}^*)$ is the embedding of textual condition $\mathcal{P}^*$.

3.2. Network Inflation

A T2I diffusion model (*e.g.*, LDM [37]) typically employs a U-Net [38], which is a neural network architecture based on a spatial downsampling pass followed by an upsampling pass with skip connections. It is composed of stacked 2D convolutional residual blocks and transformer blocks. Each transformer block consists of a spatial self-attention layer, a cross-attention layer, and a feed-forward network (FFN). The spatial self-attention leverages pixel locations in feature maps for similar correlation, while the cross-attention considers correspondence between pixels and conditional inputs (*e.g.*, text). Formally, given latent representation z_{v_i} of video frame v_i, the spatial self-attention mechanism [46] implements $\text{Attention}(Q, K, V) = \text{Softmax}(\frac{QK^T}{\sqrt{d}}) \cdot V$, with

$$Q = W^Q z_{v_i}, K = W^K z_{v_i}, V = W^V z_{v_i},$$

where W^Q, W^K, and W^V are learnable matrices that project the inputs to query, key and value, respectively, and and d is the output dimension of key and query features.

We extend a 2D LDM to the spatio-temporal domain. Similar to VDM [18], we inflate the 2D convolution layers to pseudo 3D convolution layers, with 3×3 kernels being replaced by $1 \times 3 \times 3$ kernels and append a temporal self-attention layer in each transformer block for temporal modeling. To enhance the temporal coherence,

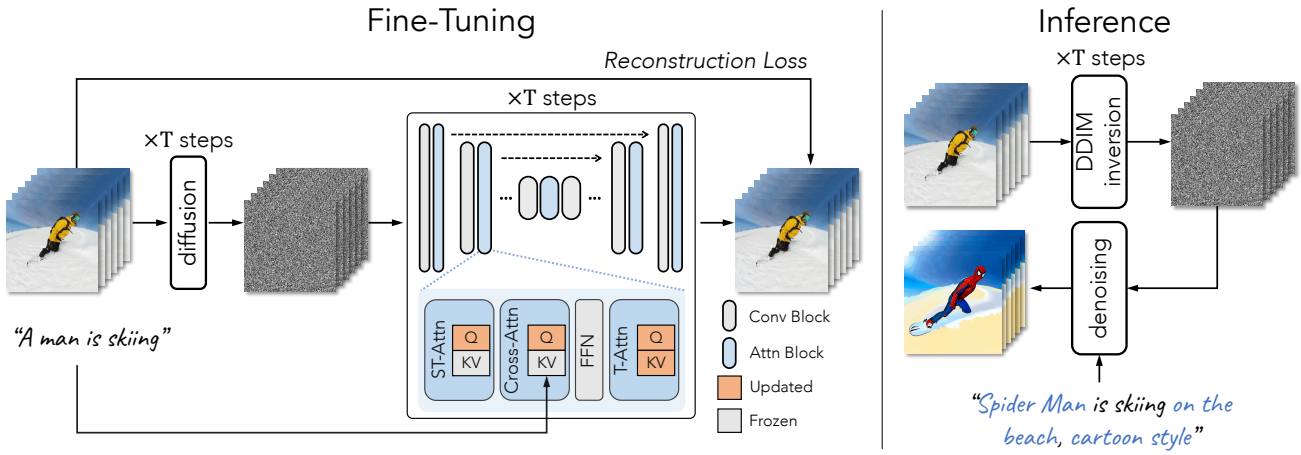


Figure 4: ***Pipeline of Tune-A-Video:*** Given a text-video pair (*e.g.*, "a man is skiing") as input, our method leverages the pretrained T2I diffusion models for T2V generation. During fine-tuning, we update the projection matrices in attention blocks using the standard diffusion training loss. During inference, we sample a novel video from the latent noise inverted from the input video, guided by an edited prompt (*e.g.*, "Spider Man is surfing on the beach, cartoon style").

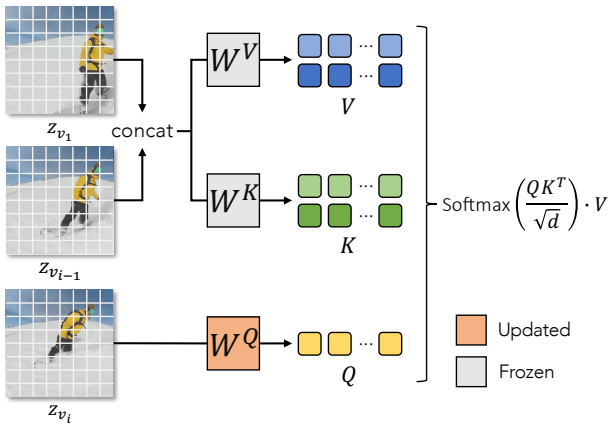


Figure 5: ***Illustration of our ST-Attn:*** Latent features of frame v_i, previous frames v_{i-1} and v_1 are projected to query Q, key K and value V. Output is a weighted sum of the values, weighted by the similarity between the query and key features. We highlight the updated parameter W^Q.

we further extend the spatial self-attention mechanism to the spatio-temporal domain. There are alternative options for spatio-temporal attention (ST-Attn) mechanism, including full attention and causal attention which also capture spatio-temporal consistency. However, such straightforward choices are actually not feasible in generating videos with increasing frames due to their high computational complexity. Specifically, given m frames and N sequences for each frame, the complexity for both full attention and causal attention is $\mathcal{O}((mN)^2)$. It is not affordable if we need to generate long videos with a large value of m.

Here, we propose to use a sparse version of causal attention mechanism, where the attention matrix are computed between frame z_{v_i} and two previous frames z_{v_1} and $z_{v_{i-1}}$, remaining low computational complexity at $\mathcal{O}(2m(N)^2)$. Specifically, we derive query feature from frame z_{v_i}, key and value features from the *first* frame z_{v_1} and the *former* frame $z_{v_{i-1}}$, and implement $\text{Attention}(Q, K, V)$ with

$$Q = W^Q z_{v_i}, K = W^K \left[z_{v_1}, z_{v_{i-1}}\right], V = W^V \left[z_{v_1}, z_{v_{i-1}}\right],$$

where $[\cdot]$ denotes concatenation operation. Note that the projection matrices W^Q, W^K, and W^V are shared across space and time. See Fig. 5 for a visual depiction.

3.3. Fine-Tuning and Inference

Model fine-tuning. We now finetune our network on the given input video for temporal modeling. The spatio-temporal attention (ST-Attn) is designed to model temporal consistency by querying relevant positions in previous frames. Therefore, we propose to fix parameters W^K and W^V, and only update W^Q in ST-Attn layers. In contrast, we finetune the entire temporal self-attention (T-Attn) layers as they are newly added. Moreover, we propose to refine the text-video alignment by updating the query projection in cross-attention (Cross-Attn). In practice, finetuning the attention blocks is computationally efficient compared to full tuning [39], and meanwhile retains the original property of pre-trained T2I diffusion models. We use the same training objective in standard LDMs [37]. Fig. 4 illustrates the fine-tuning process with the trainable parameters highlighted.

Structure guidance via DDIM inversion. Finetuning the attention layers is essential to ensure spatial consistency across all frames. However, it does not offer much control over pixel shifts, resulting in stagnant videos in the loop. To tackle this problem, we incorporate structure guidance from the source video during the inference stage. Specifically, we

Figure 6: ***Sample results of our method.***

obtain a latent noise of source video $\mathcal{V}$ through DDIM inversion with no textual condition. This noise serves as the starting point for DDIM sampling, which is guided by an edited prompt $\mathcal{T}^*$. The output video $\mathcal{V}^*$ is then given by

$$\mathcal{V}^* = \mathcal{D}(\text{DDIM-samp}(\text{DDIM-inv}(\mathcal{E}(\mathcal{V})), \mathcal{T}^*)).$$

Note that for the same input video, we only need to perform DDIM inversion once. Our experiments demonstrate its effectiveness in accurately conveying the structural movements from the source video to the generated videos.

4. Applications of Tune-A-Video

We showcase several applications of our Tune-A-Video for text-driven video generation and editing.

Object editing. One of the major applications of our method is to modify the object through the editing of text prompts. This allows replacing, adding, or removing objects with ease. Fig. 6 shows some examples. We can replace "a man" with "Spider Man" or "Wonder Woman", "a rabbit" with "a cat" or "a puppy", or even switch out "a watermelon" for "a cheeseburger", simply by modifying the corresponding words. We can add an object such as "a cowboy hat" or "sunglasses" by further describing it in the prompt. To remove an object, we can easily delete the corresponding phrase—for example, the watermelon.

Background change. Our method also enables users to change the video background (*i.e.*, the place where the object is), while preserving the consistency of the object's movements. For example, we can modify the background of the skiing man in Fig. 6 to be "on the beach" or "at sunset", by adding a new location/time description, and change the countryside road view in Fig. 7 to sea view, by replacing an existing location description.

Style transfer. Thanks to the open-domain knowledge of pretrained T2I models, our method transfer videos into a variety of styles that are difficult to learn solely from video data [42]. For example, we transform real-world videos into comic styles (Fig. 6), or Van Gogh style (Fig. 10), by appending the global style descriptor to the prompt.

Personalized and controllable generation. Our method can be easily integrated with personalized T2I models (*e.g.*, DreamBooth [39], which takes 3-5 images as input and returns a personalized T2I model), by directly finetuing on them. For instance, we can use a DreamBooth personalized for "Modern Disney Style" or "Mr Potato Head" to create videos of a specific style or subject (Fig. 11). Our method can also be integrated with conditional T2I models like T2I-Adapter [29] and ControlNet [52], to enable diverse controls on the generated videos at no extra training cost. For example, we can further edit the motion using a sequence of human pose as control (*e.g.*, dancing in Fig. 1).

Figure 7: ***Qualitative comparison between evaluated methods.*** Zoom in for best view.

Note that the human pose sequence can be automatically detected from real-world videos using an off-the-shelf pose estimation model [4]. The compatibility of our method with personalized and conditional T2I models offers more possibilities for users to create the video content they desire.

5. Experiments

5.1. Implementation Details

Our development is based on Latent Diffusion Models [37] (a.k.a Stable Diffusion) and the public pretrained weights[1]. We sample 32 uniform frames at resolution of 512×512 from input video, and finetune the models with our method for 500 steps on a learning rate 3×10^{-5} and a batch size 1. At inference, we use DDIM sampler [43] with classifier-free guidance [17] in our experiments. For a single video, it takes about 10 minutes for finetuning, and about 1 minute for sampling on a NVIDIA A100 GPU.

5.2. Baseline Comparisons

Dataset. To evaluate our approach, we use 42 representative videos taken from DAVIS dataset [33]. We automatically produce the video footage using an off-the-shelf captioning model [22], and manually design 140 edited prompts across our applications in Sec. 4. More details on our benchmark are provided in Sec. A.

Baselines. We compare our method against three baselines: 1) *CogVideo* [19]: a T2V model trained on a dataset of 5.4 million captioned videos, and is capable of generating videos directly from text prompts in a zero-shot manner. 2) *Plug-and-Play* [44]: a cutting-edge image editing model that can edit each frame of a video individually. 3) *Text2LIVE* [3]: a recent approach for text-guided video editing that employs layered neural atlases [20].

Qualitative results. We present a visual comparison of our approach against several baselines in Fig. 7. We observe that while CogVideo can produce videos that reflect the general concept in the text, the output videos varies a lot in quality and it cannot take a video as input. Plug-and-Play, on the other hand, successfully edits each video frame individually, but lacks frame consistency as the temporal context is neglected (*e.g.*, the appearance of the Porsche car is not consistent across frames). Text2LIVE, while capable of producing temporally smooth videos, struggles to ac-

[1] https://huggingface.co/CompVis/stable-diffusion-v1-4

Table 1: ***Quantitative comparison with evaluated baselines.*** * indicates Tune-A-Video *vs*. CogVideo, ** indicates Tune-A-Video *vs*. Plug-and-Play.

Method	Frame Consisitency		Textual alignment	
	CLIP Score	User Preference	CLIP Score	User Preference
CogVideo	90.64	12.14	23.91	15.00
Plug-and-Play	88.89	37.86	27.56	23.57
Tune-A-Video	**92.40**	**87.86* / 62.14****	**27.58**	**85.00* / 76.43****

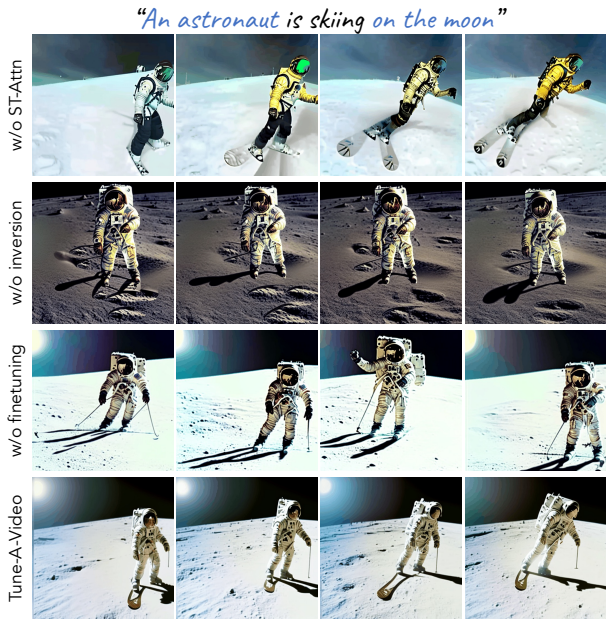


Figure 8: ***Ablation study.*** Input video is shown in Fig .6.

curately represent the edited prompt (*e.g.*, the Porsche car still appears in the shape of the original jeep car). This may be due to its reliance on layered neural atlases, which restricts its editing ability. In contrast, our method generates temporally-coherent videos that preserve structural information from the input video and align well with edited words and details. Additional qualitative comparison can be found in Fig. 12.

Quantitative results. We quantify our method against baselines through automatic metrics and user study, and report frame consistency and textual faithfulness in Tab. 1.

Automatic metrics. For frame consistency, we compute CLIP [34] image embeddings on all frames of output videos and report the average cosine similarity between all pairs of video frames. To measure textual faithfulness, we compute average CLIP score between all frames of output videos and corresponding edited prompts. Our results indicate that CogVideo produces consistent video frames but struggle to represent the textual description, whereas Plug-and-Play achieves high textual faithfulness but failed to generate consistent content. In contrast, our method outperforms baselines in both metrics.

Figure 9: ***Limitations:*** Our method might produce unpleasant results when the input video contains multiple objects and exhibits occlusions. For example, the two pandas at the bottom being mixed together.

User study. For frame consistency, we present two videos generated by our method and a baseline in random order and ask the raters "which one has better temporal consistency?". For textual faithfulness, we additionally show the textual description and ask the raters "which video better aligns with the textual description?". We recruit 5 participants to annotate each example and use a majority vote for the final result. Additional details are provided in Appendix (Sec. B). We observe that CogVideo and Plug-and-Play are less preferred due to frame-wise and frame-text inconsistency, whereas our method achieves higher user preference in both aspects.

5.3. Ablation Study

We conduct an ablation study to assess the importance of the spatio-temporal attention (ST-Attn) mechanism, DDIM inversion, and finetuning in our Tune-A-Video. Each design is individually ablated to analyze its impact. The results, presented in Fig. 8, show that the model *w/o ST-Attn* displays significant content discrepancies (evident from the skier's clothing color). In contrast, the model *w/o inversion* maintains consistent content but fails to replicate the motion (*i.e.*, skiing) in the input video. Thanks to the ST-Attn and inversion, model *w/o finetuning* still suffices consistent content across frames. However, the motion in consecutive frames is not smooth, resulting in flickering videos. Additional video examples of ablation study can be found in Fig. 13. These results indicate that all of our key designs contribute to the successful results of our method.

6. Limitations and Future Work

Fig. 9 presents a failure case of our method when the input video contains multiple objects and exhibits occlusion. This may be due to the inherent limitation of the T2I model in handling multiple objects and object interactions. A potential solution is to use additional conditional information, such as depth, to enable the model to differentiate between different objects and their interactions. This avenue of research is left as future work.

7. Conclusion

In this paper, we introduce a new task for T2V generation called One-Shot Video Tuning. This task involves training a T2V generator using only a single text-video pair and pretrained T2I models. We present Tune-A-Video, a simple yet effective framework for text-driven video generation and editing. To generate continuous videos, we propose an efficient tuning strategy and structural inversion that enable generating temporally-coherent videos. Extensive experiments demonstrate the remarkable results of our method spanning a wide range of applications.

References

[1] Rajat Arora and Yong Jae Lee. Singan-gif: Learning a generative video model from a single gif. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1310–1319, 2021. 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 *ICCV*, pages 1728–1738, 2021. 2

[3] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, pages 707–723. Springer, 2022. 3, 7

[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 7

[5] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022. 3

[6] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *arXiv preprint arXiv:2204.14217*, 2022. 2, 3

[7] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. *arXiv preprint arXiv:2302.03011*, 2023. 3, 4

[8] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 4

[9] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, pages 89–106. Springer, 2022. 3

[10] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *CVPR*, pages 10696–10706, 2022. 3

[11] Tanmay Gupta, Dustin Schwenk, Ali Farhadi, Derek Hoiem, and Aniruddha Kembhavi. Imagine this! scripts to compositions to videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 598–613, 2018. 3

[12] 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. 3

[13] Niv Haim, Ben Feinstein, Niv Granot, Assaf Shocher, Shai Bagon, Tali Dekel, and Michal Irani. Diverse generation from a single video made possible. In *European Conference on Computer Vision*, pages 491–509. Springer, 2022. 3

[14] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 3

[15] 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. 2, 3, 4

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 3, 4

[17] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 3, 7

[18] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv:2204.03458*, 2022. 2, 3, 4

[19] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 3, 7, 12

[20] Yoni Kasten, Dolev Ofri, Oliver Wang, and Tali Dekel. Layered neural atlases for consistent video editing. *ACM Transactions on Graphics (TOG)*, 40(6):1–12, 2021. 3, 7

[21] 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. *arXiv preprint arXiv:2210.09276*, 2022. 3

[22] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 7, 12

[23] Yitong Li, Martin Min, Dinghan Shen, David Carlson, and Lawrence Carin. Video generation from text. In *AAAI*, volume 32, 2018. 3

[24] Yue Liu, Xin Wang, Yitian Yuan, and Wenwu Zhu. Cross-modal dual learning for sentence-to-video generation. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1239–1247, 2019. 3

[25] Tanya Marwah, Gaurav Mittal, and Vineeth N Balasubramanian. Attentive semantic video generation using captions. In *ICCV*, pages 1426–1434, 2017. 3

[26] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided

image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2021. 3

[27] Gaurav Mittal, Tanya Marwah, and Vineeth N Balasubramanian. Sync-draw: Automatic video generation using deep recurrent attentive architectures. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1096–1104, 2017. 3

[28] 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

[29] 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. 2, 6

[30] Alex 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. *arXiv preprint arXiv:2112.10741*, 2021. 2, 3

[31] Yaniv Nikankin, Niv Haim, and Michal Irani. Sinfusion: Training diffusion models on a single image or video. *arXiv preprint arXiv:2211.11743*, 2022. 3

[32] Yingwei Pan, Zhaofan Qiu, Ting Yao, Houqiang Li, and Tao Mei. To create what you tell: Generating videos from captions. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1789–1798, 2017. 3

[33] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 7, 12

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 8

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

[36] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, pages 8821–8831. PMLR, 2021. 3

[37] 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, 5, 7

[38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 4

[39] 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. *arXiv preprint arXiv:2208.12242*, 2022. 2, 5, 6

[40] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 2, 3

[41] 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. *arXiv preprint arXiv:2210.08402*, 2022. 2

[42] 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. *arXiv preprint arXiv:2209.14792*, 2022. 2, 3, 4, 6

[43] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 7

[44] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. *arXiv preprint arXiv:2211.12572*, 2022. 3, 7, 12

[45] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 3, 4

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[47] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399*, 2022. 2

[48] 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 *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI*, pages 720–736. Springer, 2022. 3

[49] Chen Henry Wu and Fernando De la Torre. Unifying diffusion models' latent space, with applications to cyclediffusion and guidance. *arXiv preprint arXiv:2210.05559*, 2022. 3

[50] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2021. 3

[51] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022. 3

[52] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 6

[53] 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. 2, 3

Appendix

A. Dataset Details

We select 42 videos from the DAVIS dataset [33], covering a range of categories including animals, vehicles, and humans. The selected video items are listed in Tab. 2. To obtain video footage, we use BLIP-2 [22] for automated captions. We then manually design three edited prompts for each video, resulting 140 edited prompts in total. These edited prompts include object editing, background changes, and style transfers, as described in Sec. 4.

Table 2: ***Names of videos selected from DAVIS dataset.***

bear	blackswan	boat
breakdance-flare	camel	car-roundabout
car-shadow	car-turn	cows
dog	dog-agility	drift-turn
elephant	flamingo	girl-dog
gold-fish	golf	guitar-violin
hike	hockey	horsejump-high
horsejump-low	kid-football	kite-surf
lab-coat	libby	lions
longboard	lucia	mallard-water
man-bike	mbike-santa	mbike-trick
motorbike	parkour	rhino
running	scooter-gray	snowboard
swing	tandem	tennis

B. User Study Details

We conduct a user study on our dataset of 140 edited prompts to compare our method against two baselines: Plug-and-Play [44] and CogVideo [19]. The comparison results are shown in Tab. 1. The participants of the user study are mainly students and colleagues in university. We ask 5 raters to evaluate each edited prompt by comparing two videos generated by two different methods (shown in random order) and answering two following questions:

1. Which video has higher consistency? Please select the one that looks more smooth as a video.
2. Which video matches the text better? Please select the one that better represents the given text description.

C. Additional Results

Fig. 10 and Fig. 11 showcase additional video examples of our methods, Fig. 12 provides additional comparison with baselines, and Fig. 13 gives additional results of ablation study.

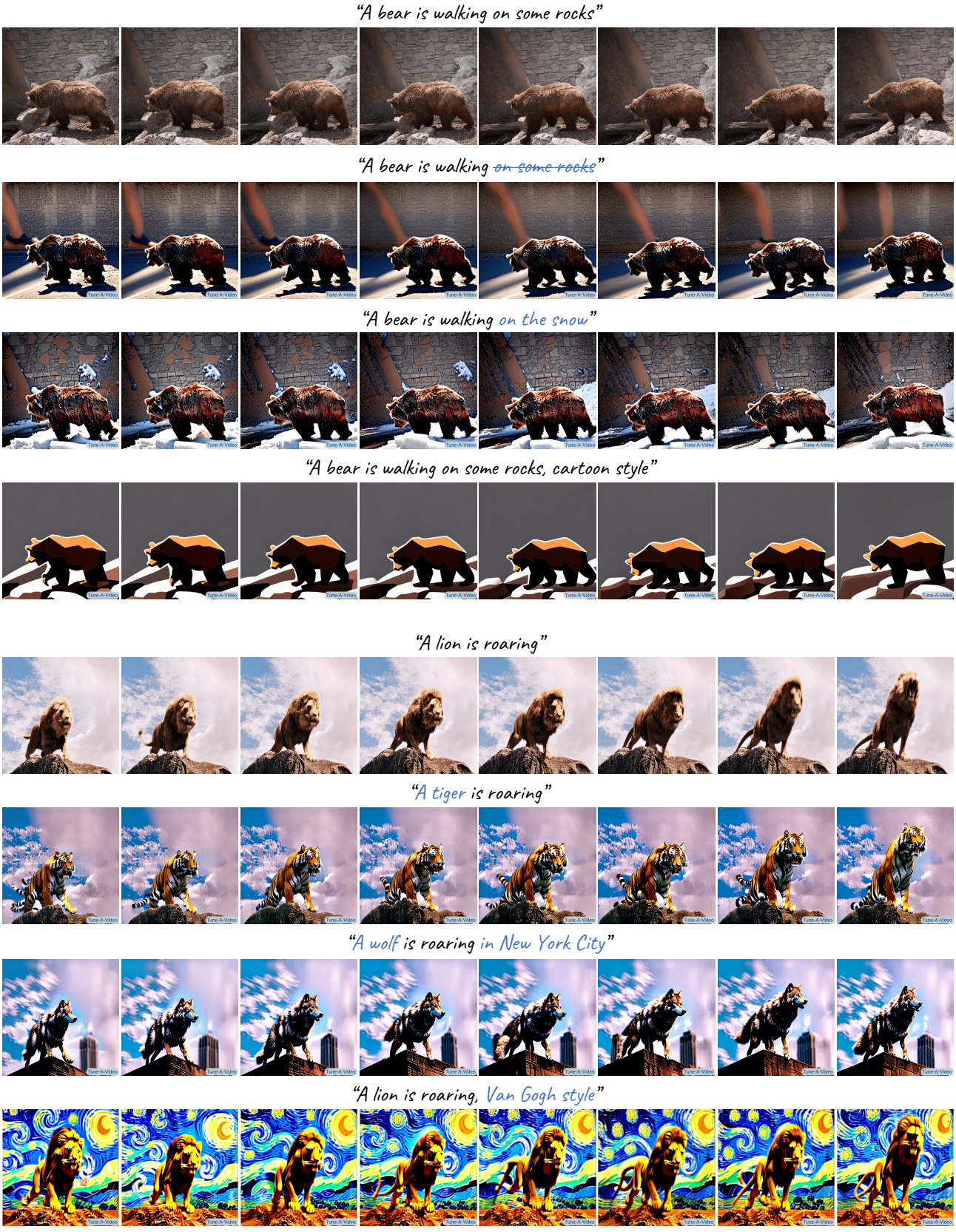


Figure 10: ***Additional sample results of our method (1/2).***

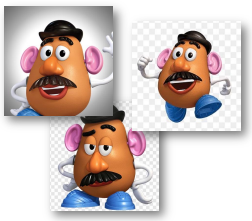
"Modern Disney Style"

"A bear is playing guitar"

"A rabbit is playing guitar, modern disney style"

"A handsome prince is playing guitar, modern disney style"

"A magic princess with sunglasses is playing guitar on the stage, modern disney style"



"Mr. Potato Head"

"A bear is playing guitar"

"Mr. Potato Head, made of Lego, is playing guitar on the snow"

"Mr. Potato Head, wearing sunglasses, is playing guitar on the beach"

"Mr. Potato Head is playing guitar in starry night, Van Gogh style"

Figure 11: ***Additional sample results of our method (2/2).***

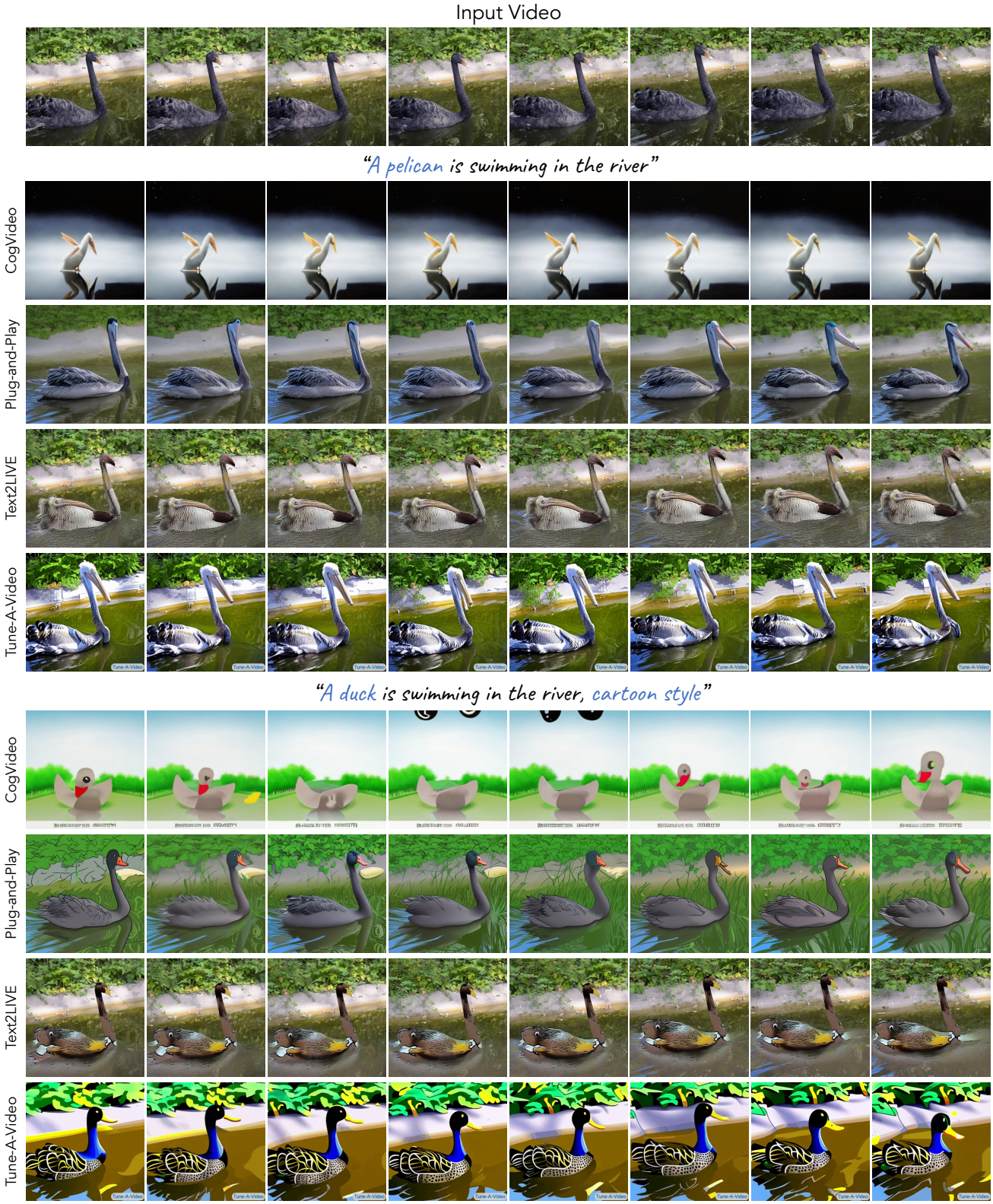


Figure 12: ***Additional qualitative comparsion between evaluated methods.***

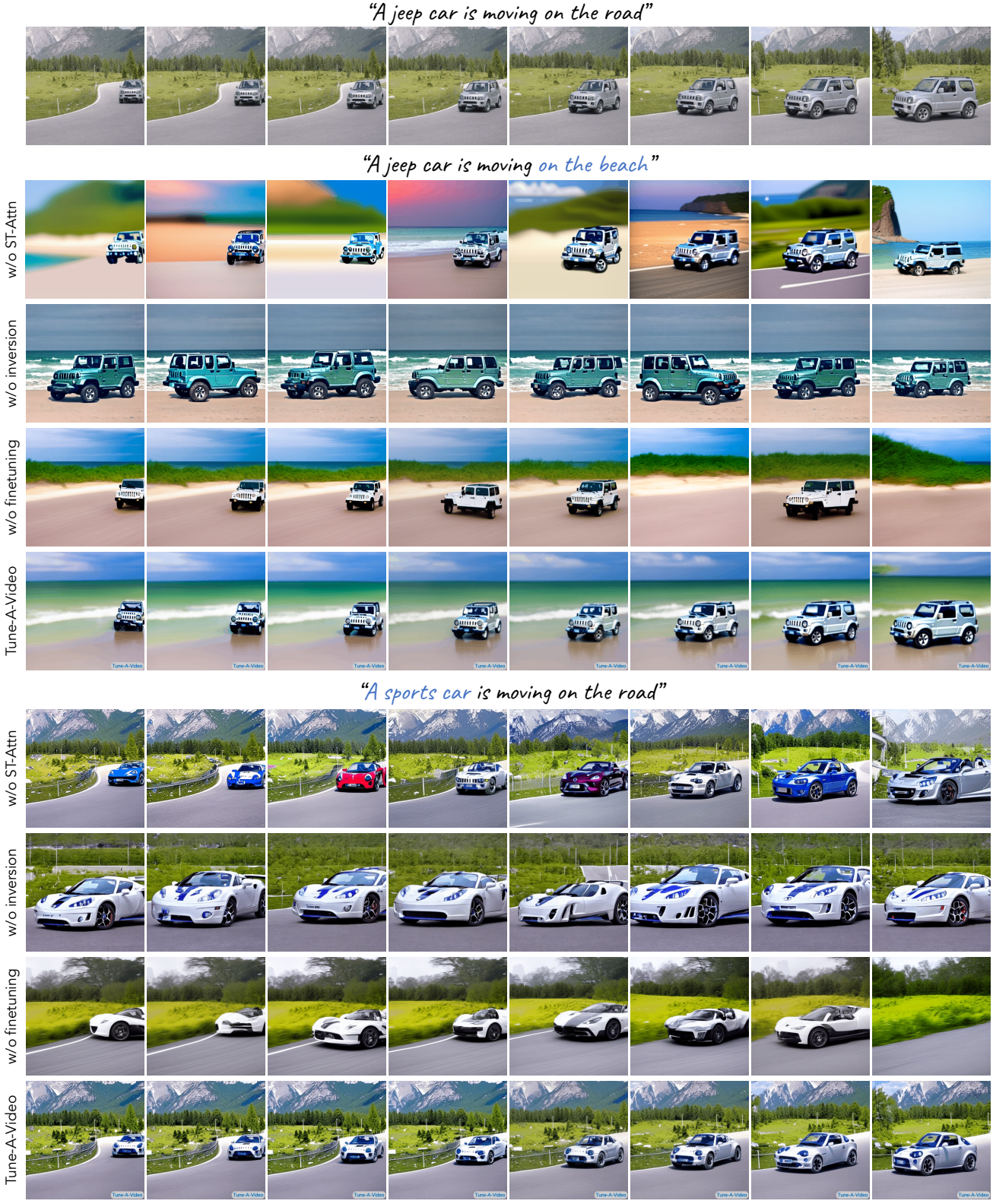


Figure 13: ***Additional ablation study.***