Enhancing Motion in Text-to-Video Generation with Decomposed Encoding and Conditioning

Penghui Ruan[1,2], **Pichao Wang**[3*], **Divya Saxena**[1], **Jiannong Cao**[1†], **Yuhui Shi**[2†]
[1]The Hong Kong Polytechnic University, Hong Kong, China
[2]Southern University of Science and Technology, Shenzhen, China
[3]Amazon, Seattle, United States
penghui.ruan@connect.polyu.hk, pichaowang@gmail.com
{divsaxen, csjcao}@comp.polyu.edu.hk, shiyh@sustech.edu.cn

Abstract

Despite advancements in Text-to-Video (T2V) generation, producing videos with realistic motion remains challenging. Current models often yield static or minimally dynamic outputs, failing to capture complex motions described by text. This issue stems from the internal biases in text encoding, which overlooks motions, and inadequate conditioning mechanisms in T2V generation models. To address this, we propose a novel framework called DEcomposed MOtion (DEMO), which enhances motion synthesis in T2V generation by decomposing both text encoding and conditioning into content and motion components. Our method includes a content encoder for static elements and a motion encoder for temporal dynamics, alongside separate content and motion conditioning mechanisms. Crucially, we introduce text-motion and video-motion supervision to improve the model's understanding and generation of motion. Evaluations on benchmarks such as MSR-VTT, UCF-101, WebVid-10M, EvalCrafter, and VBench demonstrate DEMO's superior ability to produce videos with enhanced motion dynamics while maintaining high visual quality. Our approach significantly advances T2V generation by integrating comprehensive motion understanding directly from textual descriptions. Project page: https://PR-Ryan.github.io/DEMO-project/

1 Introduction

The field of Text-to-Video (T2V) generation [21, 46, 7, 8, 19, 56, 26, 74, 3, 59, 70] has seen significant advancements, especially with the advent of diffusion models. These models have demonstrated impressive capabilities in generating visually appealing videos from textual descriptions. However, a persistent challenge remains: generating videos with realistic and complex motions. Most existing T2V models produce outputs that resemble static animations or exhibit minimal camera movement, falling short of capturing the intricate motions described in textual inputs [21, 46, 7, 19, 56, 74].

This limitation arises from two primary challenges. The first challenge is the inadequate motion representation in text encoding. Current T2V models utilize large-scale visual-language models (VLMs), such as CLIP [40], as text encoders. These VLMs are highly effective at capturing static elements and spatial relationships but struggle with encoding dynamic motions. This is primarily due to their training focus, which biases them towards recognizing nouns and objects [35], while verbs and actions are less accurately represented [17, 69, 38]. The second challenge is the reliance on spatial-only text conditioning. Existing models often extend Text-to-Image (T2I) generation

*The work does not relate to author's position at Amazon.
†Corresponding authors.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

 https://doi.org/10.52202/079017-2241

techniques to T2V tasks [21, 46, 7, 19, 56, 74, 3], applying text information through spatial cross-attention on a frame-by-frame basis. While effective for generating high-quality static images, this approach is insufficient for videos, where motion is a critical component that spans both spatial and temporal dimensions. A holistic approach that integrates text information across these dimensions is essential for generating videos with realistic motion dynamics.

Recent efforts to address these challenges have involved incorporating additional control signals such as sketches [15], strokes [9, 23, 60, 51, 68], database samples [72], depth maps [33], and human poses [71, 6, 12], reference videos [62, 73, 34, 65], and bounding boxes [55] into the T2V generation process. These signals are derived either from reference videos or pre-trained motion generation models [36, 30]. While these approaches improve motion synthesis, they depend on external references or pre-trained models, which may not always be practical. Moreover, they introduce complexity and potential inefficiencies, as they require separate handling of additional data sources.

To address these challenges, we introduce Decomposed Motion (DEMO), a novel framework designed to enhance motion synthesis in T2V generation. DEMO adopts a comprehensive approach by decomposing both text encoding and conditioning processes into content and motion components. Addressing the first challenge, DEMO decomposes text encoding into content encoding and motion encoding processes. The content encoding focuses on object appearance and spatial layout, capturing static elements such as “a girl" and “the road" in the scenario “A girl is walking to the left on the road." Meanwhile, the motion encoding captures the essence of object movement and temporal dynamics, interpreting actions like “walking" and directional cues like “to the left." This separation allows the model to better understand and represent the dynamic aspects of the described scenes. Regarding the second challenge, DEMO decomposes the text conditioning process into content and motion dimensions. The content conditioning module integrates spatial embeddings into the video generation process on a frame-by-frame basis, ensuring that static elements are accurately depicted in each frame. In contrast, the motion conditioning module operates across the temporal dimension, infusing dynamic motion embeddings into the video. This separation enables the model to capture and reproduce complex motion patterns described in the text. Moreover, DEMO incorporates novel text-motion and video-motion supervision techniques to enhance the model’s understanding and generation of motion. Text-motion supervision aligns cross-attention maps with the temporal changes observed in ground truth videos, guiding the model to focus on motion information. Video-motion supervision constrains the predicted video latent to mimic the motion patterns of real videos, promoting the generation of coherent and realistic motion dynamics. These supervision techniques ensure that the model not only generates visually appealing videos but also renders the intricate motions described in the text.

To validate our framework, we conduct extensive experiments on several benchmarks, including MSR-VTT [66], UCF-101 [50], WebVid-10M [1], EvalCrafter [31], and VBench [24]. DEMO achieves substantial improvements in metrics related to motion dynamics and visual fidelity, indicating its superior capability to generate videos that are both visually appealing and dynamically accurate.

2 Related Work

T2V Generation. The T2V domain has made substantial strides, building on the progress in T2I generation. The first T2V model, VDM [21], introduces a space-time factorized U-Net for temporal modeling, training on both images and videos. For high-definition videos, models like ImagenVideo [19], Make-A-Video [46], LaVie [59], and Show-1 [70] use cascades of diffusion models with spatial and temporal super-resolution. MagicVideo [74], Video LDM [4], and LVDM [16] apply latent diffusion for video, working in a compressed latent space. VideoFusion [32] separates video noise into base and residual components. ModelScopeT2V uses 1D convolutions and attention to approximate 3D operations. Stable Video Diffusion (SVD) [3] divides the process into T2I pre-training, video pre-training, and fine-tuning and demonstrate the necessity of a well-curated high-quality pretraining dataset for developing a strong base model. Despite these advancements, the generated videos still exhibit limited motion dynamics, often appearing largely static with minimal motion, highlighting an ongoing challenge in achieving dynamic and realistic motions.

T2V Generation with Rich Motion. Generating video with rich motion is still an open challenge in the field of T2V generation. Existing works [15, 9, 23, 60, 51, 68, 72, 33, 71, 6, 12, 62, 73, 34,

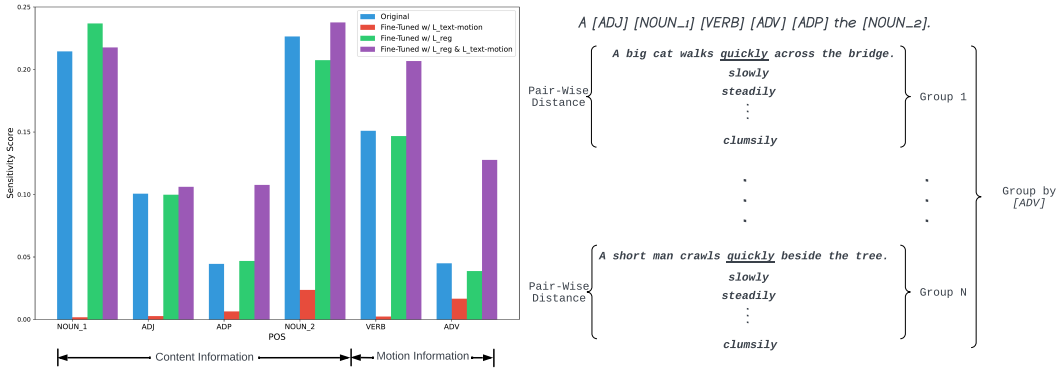

Figure 1: **Our Pilot Study**. We generated a set of prompts (262144 in total) following a fixed template, grouping them according to the different parts of speech (POS). These grouped texts are then passed into the CLIP text encoder, and we calculate the sensitivity as the average sentence distance within each group. As shown on the left-hand side, compared to POS representing content, CLIP is less sensitive to POS representing motion. (Results are consistent across different templates and different sets of words within each POS. Further details can be found in the appendix.)

65, 55] address this challenge by incorporating additional control signals that inherently contain rich motion information. Tune-A-Video [65] proposes spatial-temporal self-attention into the T2I backbone and trains the model on a single reference video. The model thus learns to generate new videos with motions specified by the reference video. Materzynska *et al.* [34] follow the idea of T2I customization [13, 42, 27] to fine-tune the model and a specific text token on a small set of reference videos. The model can then recontextualize with that learned token to generate new videos with specific motions. DreamVideo [62] further customizes both the appearances and motions given reference images and videos. MotionDirector [73] proposes a dual-path Low-Rank Adaptations [22] to decouple the motions and appearances residing in the reference videos. MotionCtrl [60] incorporates object trajectories and camera poses into the T2V generation by conditioning them in the convolution and temporal transformer layers, respectively. Contrasting with these approaches, DEMO prioritizes the generation of videos that exhibit significant motions derived solely from textual descriptions without relying on additional signals.

3 Method

Latent Video Diffusion Models (LVDMs). LVDMs build on the diffusion models [48, 20] by training a 3D U-Net as the noise predictor, where a VQ-VAE [37] or a VQ-GAN [11] is employed to compress the video into low-dimensional latent space. The 3D U-net consists of down-sample, middle, and up-sample blocks. Each of these blocks comprises multiple convolution layers augmented by spatial and temporal transformers. The spatial transformer consists of spatial self-attention, spatial cross-attention, and feed-forward layers. The temporal transformer consists of temporal self-attention and feed-forward layers. The 3D U-Net is trained with a text encoder to minimize the noise-prediction loss in the latent space given as follows:

$$\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{t,z_0,\epsilon\sim\mathcal{N}(0,1),p}[||\epsilon - \epsilon_\theta(z_t, t, \mathcal{E}(p))||_2^2] \tag{1}$$

where z is the video latent corresponding to x in the pixel space, t is the time step, $\mathcal{E}$ is a text encoder, p is a text prompt, and ϵ is noise sampled from Gaussian distribution. z_t is noisy z_0 after t steps diffusion forward process given by:

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon,\ \bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s \tag{2}$$

where α_t is a pre-defined noise schedule.

3.1 Decomposed Text Encoding

As shown in our pilot study in Figure 1, the CLIP text encoder can distinguish different motions, but it is not as sensitive to motion as it is to content. Consequently, the text encoding focuses more

https://doi.org/10.52202/079017-2241

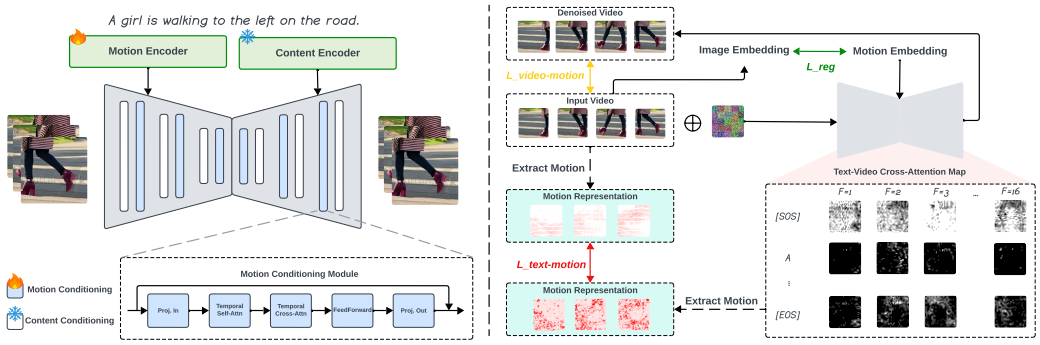

Figure 2: **Overview of DEMO Training**. As shown in the left-hand side, DEMO incorporate dual text encoding and text conditioning (for simplicity, other layers in the UNet are omitted). As shown in the right-hand side, during training, the $\mathcal{L}_{\text{text-motion}}$ is used to enhance motion encoding, the $\mathcal{L}_{\text{reg}}$ is used to avoid catastrophic forgetting, the $\mathcal{L}_{\text{video-motion}}$ is to enhance motion integration. The **snowflakes** and **flames** denote frozen and trainable parameters, respectively.

on content encoding rather than motion encoding. To preserve the generalization ability of the T2V generation model, we retain the original text encoder, referring to it as the content encoder (denoted as $\mathcal{E}_c$). Additionally, we introduce a new text encoder, referred to as the motion encoder (denoted as $\mathcal{E}_m$), which is specifically designed to capture object movement and temporal dynamics in textual descriptions (as shown in the left-hand side of Figure 2). We initialize our motion encoder from a CLIP text encoder and then fine-tune it using specialized text-motion supervision, as described below.

Text-Motion Supervision. Research [27, 28, 61] has shown that cross-attention maps represent the structure of visual content. The cross-attention operation can be viewed as a projection of text information into the visual structure domain. With this understanding, we aim to shift the text encoder's focus more toward motion information by constraining the temporal changes of cross-attention maps to closely mimic those observed in ground truth videos, as illustrated by the **red** line in Figure 2. Formally, given a noisy video latent z_t at time step t and a text prompt p, the cross-attention maps $\mathcal{A}^i \in \mathbb{R}^{H^i \times W^i \times F \times S}$, where H^i and W^i are the height and width of video latent at the ith cross-attention layer, F is the number of frames, S is the sequence length, for a cross-attention layer i are defined as follows:

$$\mathcal{A}^i = \frac{1}{N}\sum_{n}^{N} \text{softmax}\left(\frac{Q^{(n)}(K^{(n)})^T}{\sqrt{d_n}}\right) \tag{3}$$

$$Q^{(n)} = W_Q^{(n)} \cdot z_t, K^{(n)} = W_K^{(n)} \cdot \mathcal{E}_m(p) \tag{4}$$

where W_Q and W_K are projection matrices for query and key, $i \in \{1, 2, ...M\}$ is layer index, $n \in \{1, 2, ..., N\}$ represents each head in multi-head cross-attention, and d_n is the dimension of each head.

We empirically find that the cross-attention maps corresponding to the "[eos]" token, which aggregate the whole sentence's semantics, play a pivotal role in generating motion. This aligns with the understanding that motion is a global concept and cannot be captured by a single word. For instance, phrases like "A baby/dog is walking/running forwards/backward." demonstrate that different combinations of words can result in significantly different motions. Hence, we focus on the cross-attention maps related to the "[eos]" token and constrain them to mimic the motion patterns observed in the ground truth videos. This approach forms the basis of our text-motion loss, defined as follows:

$$\mathcal{L}_{\text{text-motion}} = -\mathbb{E}_{t, x_0, \epsilon \sim \mathcal{N}(0,1), p}\left[\frac{1}{M}\sum_{i=1}^{M} \cos(\phi(\mathcal{A}^i_{[eos]}), \phi(x_0))\right] \tag{5}$$

where ϕ is a function to extract motion dynamics from a video. In our case, we use optical flow to represent the motion dynamics (noting that optical flow is only used during training; during

inference, we use only the text prompt as input). In light of the potential scale differences between the cross-attention maps and video pixel values, we compute the cosine similarity between them. Additionally, for cross-attention at different spatial resolutions, we downsample the ground truth video to match the spatial resolution of the cross-attention maps.

Regularization. Recall that CLIP [40] is trained with a contrastive learning objective to match texts and images from a group of text-image pairs. However, directly fine-tuning the CLIP text encoder with Equation 1 and Equation 5, which differ significantly from the original contrastive learning objective, can easily lead to catastrophic forgetting [29]. To mitigate this, we introduce a regularization term in the fine-tuning objective to preserve its generalization ability. Specifically, we penalize the text embedding if it diverges from the corresponding image embedding, maintaining alignment with the original CLIP contrastive learning objective, as illustrated by the **green** line in Figure 2. The regularization loss is defined as follows:

$$\mathcal{L}_{\text{reg}} = -\mathbb{E}_{x_0,p}\left[cos(\mathcal{E}_m(p), \mathcal{E}^{img}(x_0^{F/2}))\right] \tag{6}$$

where $\mathcal{E}^{img}$ represents the CLIP image encoder. Given that there is only one text prompt for the entire video, we select medium frame $x_0^{F/2}$ and compute its image embedding.

3.2 Decomposed Text Conditioning

DEMO employs separate content conditioning and motion conditioning modules to incorporate content and motion information. To preserve the generative capabilities of our base model, we maintain the original text conditioning module, referred to here as the content conditioning module. We then strategically introduce a novel temporal transformer, referred to as the motion conditioning module (detailed structure shown in Figure 2), to incorporate motion information along the temporal axis. To encourage the motion conditioning module to generate and render motion dynamics, we train this module under video-motion supervision, as described below.

Video-Motion Supervision. Recall that at each diffusion denosing step t, we can obtain the predicted $\hat{z}_{0,t}$ at time step t, which is given by:

$$\hat{z}_{0,t}(t, z_t, \mathcal{E}_m(p), \mathcal{E}_c(p)) = \frac{z_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(z_t, t, \mathcal{E}_m(p), \mathcal{E}_c(p))}{\sqrt{\bar{\alpha}_t}} \tag{7}$$

This predicted $\hat{z}_{0,t}$ encapsulates the motion information in the video domain. We then prioritize the motion generation by constraining the predicted $\hat{z}_{0,t}$ to mimic the motion pattern in the real video, as illustrated by the **yellow** line in Figure 2. We define our video-motion loss as follows:

$$\mathcal{L}_{\text{video-motion}} = \mathbb{E}_{t,z_0,\epsilon\sim\mathcal{N}(0,1)}\|\Phi(z_0) - \Phi(\hat{z}_{0,t})\|_2^2 \tag{8}$$

where Φ is a function to extract motion features from a video. Given that $\mathcal{L}_{\text{diffusion}}$ is a pixel-wise denoising loss (whether raw pixel or latent pixel), choosing Φ as a general motion representation that is not in pixel space may lead to conflicting objectives due to the differing representation spaces. Instead, we choose Φ as the consecutive frame difference defined as follows:

$$\Phi(z_0) = z_0^{2:F} - z_0^{1:F-1} \tag{9}$$

where $z_0^{2:F}$ denotes the video latent from frame 2 to frame F, and $z_0^{1:F-1}$ denotes the video latent from frame 1 to frame $F-1$.

3.3 Joint Training

Our final loss is a weighted combination of $\mathcal{L}_{\text{text-motion}}$, $\mathcal{L}_{\text{reg}}$, $\mathcal{L}_{\text{video-motion}}$, and original diffusion loss $\mathcal{L}_{\text{diffusion}}$ as follows:

$$\mathcal{L} = \mathcal{L}_{\text{diffusion}} + \alpha\mathcal{L}_{\text{text-motion}} + \beta\mathcal{L}_{\text{reg}} + \gamma\mathcal{L}_{\text{video-motion}} \tag{10}$$

where α, β, and γ are scaling factors to balance different loss terms.

https://doi.org/10.52202/079017-2241

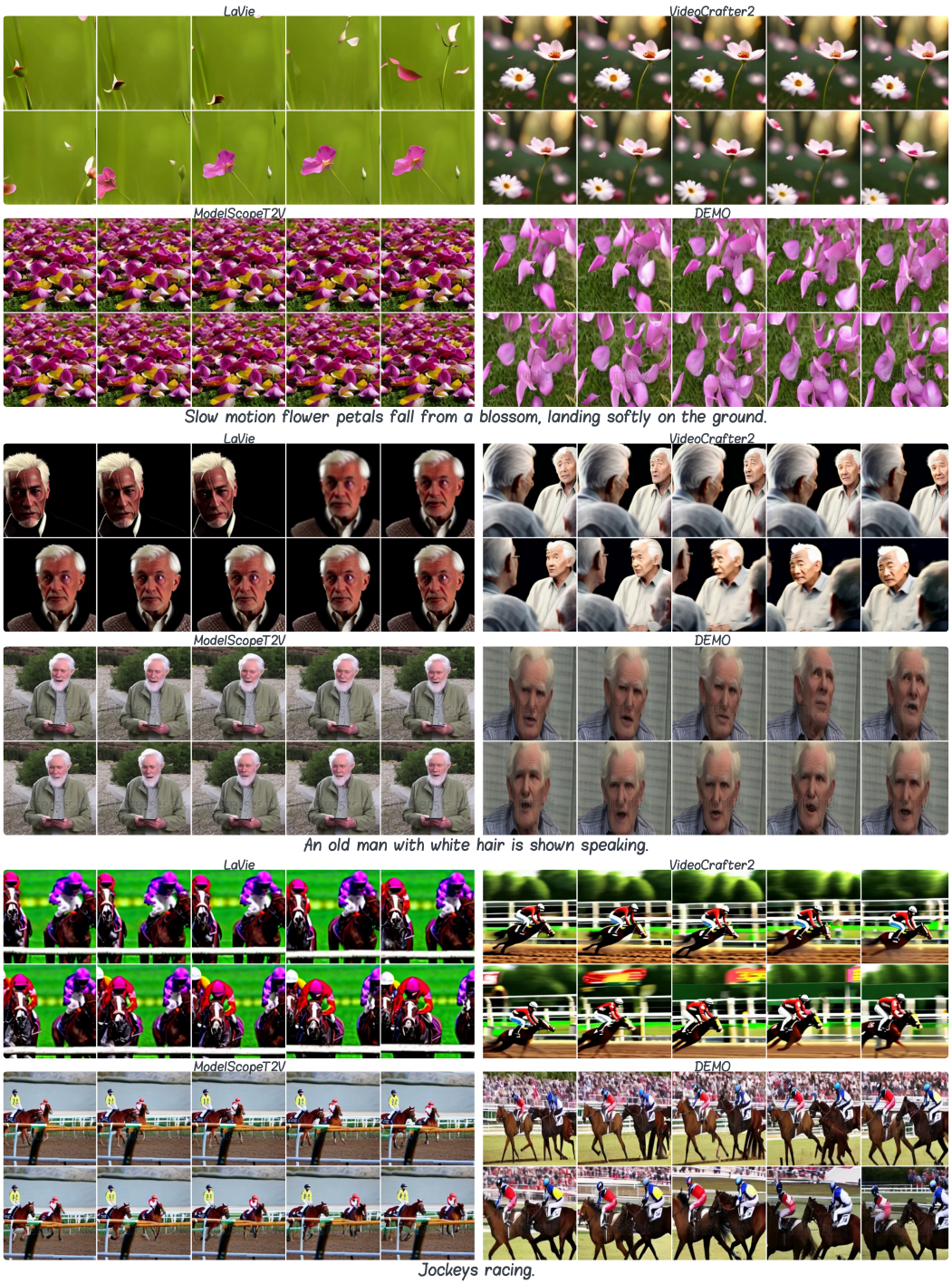

Figure 3: **Qualitative Comparison.** Each video is generated with 16 frames. We display frames 1, 2, 4, 6, 8, 10, 12, 14, 15, and 16, arranged in two rows from left to right. Full videos are available in the supplementary materials.

4 Experiments

4.1 Implementation Details

To rule out potential dataset bias, we use the same training dataset as our base model ModelScopeT2V. Specifically, we use WebVid-10M [1], a large-scale dataset of short videos with textual descriptions as our fine-tuning dataset. The training details and hyperparameters can be found in the Appendix.

4.2 Qualitative Evaluations

In this subsection, we conduct a qualitative comparison among LaVie [59], VideoCrafter2 [8], ModelScopeT2V [56], and DEMO. For a fair comparison, we use the same seed for each of these methods. The comparative analysis is illustrated in Figure 3, where we showcase examples generated by these methods. Upon examination, it is evident that these models are capable of producing high-quality videos. However, a notable distinction arises in the dynamic representation of motion within the generated videos. The ModelScopeT2V model, while visually appealing, predominantly generates static scenes. For instance, in the scenario described as "Slow motion flower petals fall from a blossom, landing softly on the ground" (the first example in Figure 3), the video generated by ModelScopeT2V captures the petals landing on the ground but lacks the motion of the petals falling. In contrast, DEMO significantly outperforms by capturing the essence of motion, producing a video where the petals fall slowly and gently to the ground. Similarly, LaVie demonstrates a similar issue, as illustrated in the third example, where the jockeys remain largely static. VideoCrafter2 exhibits relatively large motion dynamics but suffers from motion blur, as shown in the third example. Conversely, DEMO vividly captures the jockeys racing, thereby providing a more realistic representation. This underscores the advanced capability of DEMO to generate videos that not only visually represent a scene but also dynamically encapsulate the ongoing motion.

4.3 Quantitative Evaluations

Table 1: Results of zero-shot T2V generation on MSR-VTT (Evaluation protocol comparison can be found in the appendix).

Model	FID (↓)	FVD (↓)	CLIPSIM (↑)
MagicVideo [74]	-	1290	-
Make-A-Video [46]	13.17	-	0.3049
Show-1 [70]	13.08	538	0.3072
Video LDM [4]	-	-	0.2929
LaVie [59]	-	-	0.2949
PYoCo [14]	10.21-9.73	-	-
VideoFactory [58]	-	-	0.3005
EMU VIDEO [45]	-	-	-
SVD [3]	-	-	-
ModelScopeT2V[3] [56]	14.89	557	0.2941
ModelScopeT2V fine-tuned	13.80	536	0.2932
DEMO	**11.77**	**422**	**0.2965**

Table 2: Results of zero-shot T2V generation on UCF-101 (Evaluation protocol comparison can be found in the appendix).

Model	IS (↑)	FVD (↓)
MagicVideo [74]	-	655.00
Make-A-Video [46]	33.00	367.23
Show-1 [70]	35.42	394.46
Video LDM [4]	33.45	550.61
LaVie [59]	-	526.30
PYoCo [14]	47.76	355.19
VideoFactory [58]	-	410.00
EMU VIDEO [45]	42.70	606.20
SVD [3]	-	242.02
ModelScopeT2V [56]	**37.55**	628.17
ModelScopeT2V fine-tuned	37.21	612.53
DEMO	36.35	**547.31**

Zero-shot T2V Generation on MSR-VTT. We evaluate the performance of our model on the MSR-VTT [66] test set by calculating FID [18], FVD [54, 39], and CLIPSIM [63] metrics. For FID and FVD, in alignment with prior studies [56], we randomly sample 2048 videos and one prompt for each video from the test set. For CLIPSIM, we follow previous works [46, 64, 56] and use nearly 60k sentences from the entire test set to generate videos. As illustrated in Table 1, DEMO demonstrates notable advancements over the ModelScopeT2V baseline in terms of video quality metrics. Specifically, DEMO achieves an FID score of 11.77, showing marked improvement in individual frame quality compared to the baseline score of 14.89. For FVD, DEMO achieves a score of 422 compared to the baseline of 557, indicating improved overall video quality. It is important to note that the FVD is calculated using an I3D model pre-trained on the Kinetics-400 dataset [5] for action recognition. By computing the FVD over its logits, this metric not only reflects visual quality but also emphasizes motion quality in video generation. Additionally, DEMO improves the CLIPSIM

[3]Results reproduced from our own evaluation.

https://doi.org/10.52202/079017-2241

Table 3: Results of T2V generation on WebVid-10M (Val).

Model	FID (↓)	FVD (↓)	CLIPSIM (↑)
ModelScopeT2V	11.14	508	0.2986
ModelScopeT2V fine-tuned	10.53	461	0.2952
DEMO	**9.86**	**351**	**0.3083**

Table 4: Results of zero-shot T2V generation on EvalCrafter.

Model	Video Quality			Motion Quality		
	VQA_A (↑)	VQA_T (↑)	IS (↑)	Action Score (↑)	Motion AC-Score (↑)	Flow Score (↑)
ModelScopeT2V	15.12	**16.88**	14.60	75.88	44	2.51
ModelScopeT2V fine-tuned	15.89	16.39	14.92	74.23	40	2.72
DEMO w/o $\mathcal{L}_{\text{video-motion}}$	18.78	15.12	17.13	76.20	48	3.11
DEMO	**19.28**	15.65	**17.57**	**78.22**	**58**	**4.89**

score from 0.2941 to 0.2965, further demonstrating its superior ability to generate high-quality videos that are well-aligned with their textual descriptions.

Zero-shot T2V Generation on UCF-101. For UCF-101 [50], we report the IS [43] and FVD on the 101 action classes. For IS and FVD, we follow previous works [4, 14] to generate 100 videos for each of the 101 classes. We directly use the class names as prompts. As shown in Table 2, compared with baseline ModelScoprT2V, we improve the FVD from 628.17 to 547.31. However, we observed a slight decrease in IS, which may be attributed to the limited textual information provided by UCF-101 class names, such as "baby crawling" and "cliff diving." These prompts primarily suggest motion, and our model, optimized to emphasize this motion, may have over-focused on this limited information. This overemphasis potentially limited the diversity of generated content, lowering the IS.

T2V Generation on WebVid-10M (Val). For WebVid-10M [1], we perform T2V generation on the validation set. As shown in Table 3, we evaluate the FID, FVD, and CLIPSIM, where we randomly sample 5K text-video pairs from the validation set. Our model achieves an FID score of 9.86, an FVD score of 351, and a CLIPSIM score of 0.3083. These outcomes underscore our framework's substantial enhancement of video quality.

Zero-shot T2V Generation on EvalCrafter. EvalCrafter [31] provides 700 diverse prompts across categories like human, animal, objects, and landscape, each with a scene, style, and camera movement description. For our evaluation, we generate one video for each of the 700 text prompts. As shown in Table 4, we have obtained significant improvement over the baseline ModelScopeT2V in both video quality and motion quality. In terms of video quality, DEMO enhances both the Video Quality Assessment for Aesthetics (VQA_A) and the IS, albeit with a slight decrease in the Video Quality Assessment for Technical Quality (VQA_T). For motion quality, EvalCrafter uses three metrics: Action-Score, Flow-Score, and Motion AC-Score. The Action-Score, based on the VideoMAE V2 model [57] and MMAction2 toolbox, measures action recognition accuracy on Kinetics-400 classes, with higher scores indicating better human action recognition. Flow-Score and Motion AC-Score, derived from RAFT model [53] optical flows, evaluate motion dynamics. The Flow-Score measures the general motion dynamics by calculating the average magnitude of optical flow in the video, while the Motion AC-Score assesses how well the motion dynamics align with the text prompt. For motion quality, our model surpasses the baseline across all metrics (Action-Score, Flow-Score, and Motion AC-Score), showcasing DEMO's superior ability to generate videos characterized by better motion quality and higher motion dynamics.

Zero-shot T2V Generation on VBench. VBench [24] is a comprehensive benchmark to evaluate video quality. In our evaluation of VBench, we focus specifically on motion quality. We report on

Table 5: Results of zero-shot T2V generation on VBench.

Model	Motion Dynamics (↑)	Human Action (↑)	Temporal Flickering (↑)	Motion Smoothness(↑)
ModelScopeT2V	62.50	90.40	96.02	96.19
ModelScopeT2V fine-tuned	63.75	90.40	**96.35**	**96.38**
DEMO	**68.90**	**90.60**	94.63	96.09

four key metrics: Motion Dynamics, Human Action, Temporal Flickering, and Motion Smoothness. As shown in Table 5, DEMO significantly improves motion dynamics from 62.50 to 68.90. However, we observed only a slight improvement in human action recognition. This indicates that while our model enhances the richness and complexity of motion, it provides limited benefit in improving the accuracy of human action representation. Additionally, we note slight decreases in temporal flickering and motion smoothness. This observation aligns with findings from the VBench paper, which suggest that increased motion dynamics can conflict with temporal flickering and motion smoothness.

4.4 Ablation Studies

Impact of $\mathcal{L}_{\text{reg}}$ and $\mathcal{L}_{\text{text-motion}}$. As shown in Figure 1, we compute the sensitivity of our motion encoder with different loss combinations. The red columns indicate the motion encoder with $\mathcal{L}_{\text{text-motion}}$ only completely loses its ability to distinguish different tokens, either motion or content, indicating a serious catastrophic forgetting where the model loses its original knowledge. The green columns show that fine-tuning the motion encoder with $\mathcal{L}_{\text{reg}}$ only preserves the model's generalization ability but does not increase the motion sensitivity. In contrast, the purple columns demonstrate that when training the motion encoder with both $\mathcal{L}_{\text{reg}}$ and $\mathcal{L}_{\text{text-motion}}$, the model gain increased sensitivity to tokens representing motion without losing sensitivity to tokens representing content.

Impact of $\mathcal{L}_{\text{video-motion}}$. To validate the effectiveness of our video-motion loss, we perform an ablation study on the EvalCrafter dataset. As shown in Table 4, without $\mathcal{L}_{\text{video-motion}}$, the model shows a slight improvement in motion quality compared to the baseline. This is because the motion encoder provides the model with enriched motion information for generation. However, without explicitly constraining the model to mimic realistic motion, it may still focus on generating high-quality individual frames rather than coherent video sequences with rich motion dynamics. By introducing video-motion loss, the model achieves significantly higher motion quality, demonstrating the importance of this loss in guiding the model in producing videos with enhanced motion dynamics.

Table 6: Ablation study on additional parameters in motion encoder.

Benchmark	Metric	ModelScopeT2V	ModelScopeT2V fine-tuned	ModelScopeT2V + motion encoder	DEMO
MSRVTT	FID (↓)	14.89	13.80	13.98	**11.77**
	FVD (↓)	557	536	552	**422**
	CLIPSIM (↑)	0.2941	0.2932	0.2935	**0.2965**
UCF-101	IS (↑)	37.55	37.21	**37.66**	36.35
	FVD (↓)	628.17	612.53	601.25	**547.31**
WebVid-10M	FID (↓)	11.14	10.53	10.45	**9.86**
	FVD (↓)	508	461	458	**351**
	CLIPSIM (↑)	0.2986	0.2952	0.2967	**0.3083**
EvalCrafter	VQA_A (↑)	15.12	15.89	16.21	**19.28**
	VQA_T (↑)	**16.88**	16.39	16.34	15.65
	IS (↑)	14.60	14.92	15.02	**17.57**
	Action Score (↑)	75.88	74.23	75.20	**78.22**
	Motion AC-Score (↑)	44	40	46	**58**
	Flow Score (↑)	2.51	2.72	2.44	**4.89**
Vbench	Motion Dynamics (↑)	62.50	63.75	63.50	**68.90**
	Human Action (↑)	90.40	90.40	90.20	**90.60**
	Temporal Flickering (↑)	96.02	**96.35**	95.45	94.63
	Motion Smoothness (↑)	96.19	**96.38**	96.22	96.09

Impact of additional parameters in motion encoder. To rule out the effect of additional parameters introduce by motion encoder, we evaluated the effect of training with a CLIP text encoder on the overall model performance. We then compared three different variations: (1) the original ModelScopeT2V, (2) a fine-tuned version of ModelScopeT2V without additional motion encoder parameters, and (3) ModelScopeT2V with the motion encoder while maintaining its original training loss. As shown in Table 6, we observed that the performance of the model with the additional motion encoder parameters is comparable to the fine-tuned version without these extra parameters. This suggests that, without specific supervision or additional constraints, the effect of the added text encoder parameters is marginal. However, the DEMO model consistently outperforms all variations, demonstrating the effectiveness of our method in improving both video quality and text-video alignment.

Efficiency Analysis. To validate the efficiency of our proposed methods, we trained the baseline model for the same number of iterations and compared its performance with ours. As shown in Tables 1, 2, 3, 4, and 5, continuing to fine-tune the model results in only marginal improvements in

https://doi.org/10.52202/079017-2241

Figure 4: **Limitations**. DEMO does not support creating videos containing sequential motions specified by text. As shown in the example, two motions,“a man standing in a kitchen and talking" and “a mixer and a carton of milk are shown", appear simultaneously.

video quality. Additionally, we observed a slight degradation in CLIPSIM, indicating that further training may not benefit text-video alignment.

5 Limitations and Future Work

Despite DEMO’s efficiency in enhancing motion synthesis without relying on additional signals, it faces significant challenges in generating different motions sequentially, as illustrated in Figure 4. These challenges likely stem from the text encoder’s difficulty in comprehending the order of actions and the motion generation model’s limited capability to generate different motions. A potential solution to this issue involves annotating each frame with a specific prompt and training the model on video clips of varying lengths rather than a fixed duration. We consider exploring this direction in our future work.

6 Broader Impacts

Our model achieves higher visual fidelity and motion quality, which can benefit various fields such as content creation and visual simulation. However, our model is fine-tuned on web data, specifically WebVid-10M [1]. As a result, the model may not only learn how to generate videos but also inadvertently learn societal biases present in the web data, which may include inappropriate or NSFW content. Potential post-processing steps, such as applying a video classifier to filter out undesirable content, could help mitigate this issue.

7 Conclusion

In this paper, we have presented DEMO, an innovative framework crafted to advance motion synthesis in T2V generation. By separating text encoding and text conditioning into distinct content and motion dimensions, DEMO facilitates the creation of static scenes and their dynamic evolution. To encourage our model to focus on motion encoding and motion generation, we propose novel text-motion and video-motion supervision. Our extensive evaluations across various benchmarks have illustrated DEMO’s capability to significantly improve motion synthesis, showcasing its potential within the field. In future work, we plan to augment T2V datasets with more detailed descriptions and delve into advanced motion embedding techniques. By focusing on these areas, we aim to advance the frontiers of research in this dynamic and rapidly evolving domain.

8 Acknowledgement

This work is partially supported by National Science Foundation of China (No. 62250710682), Shenzhen Fundamental Research Program (No. JCYJ20200109141235597), NSFC/RGC Collaborative Research Scheme (No. CRS_PolyU501/23), HK RGC Theme-based Research Scheme (No. PolyU T43-513/23-N) and Research Grants Council of the Hong Kong Special Administrative Region, China (No. PolyU15205924). We also acknowledge the support from Research Institute for Artificial Intelligence of Things, The Hong Kong Polytechnic University, and Center for Computational Science and Engineering at Southern University of Science and Technology.

References

[1] Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1708–1718. IEEE, Montreal, QC, Canada (Oct 2021). https://doi.org/10.1109/ICCV48922.2021.00175, https://ieeexplore.ieee.org/document/9711165/

[2] Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Zhang, Q., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., Karras, T., Liu, M.Y.: eDiff-I: Text-to-Image Diffusion Models with an Ensemble of Expert Denoisers (Mar 2023), http://arxiv.org/abs/2211.01324, arXiv:2211.01324 [cs]

[3] Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., Jampani, V., Rombach, R., Ai, S.: Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets

[4] Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models (Apr 2023), http://arxiv.org/abs/2304.08818, arXiv:2304.08818 [cs]

[5] Carreira, J., Zisserman, A.: Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset (Feb 2018), http://arxiv.org/abs/1705.07750, arXiv:1705.07750 [cs]

[6] Chang, D., Shi, Y., Gao, Q., Fu, J., Xu, H., Song, G., Yan, Q., Yang, X., Soleymani, M.: MagicDance: Realistic Human Dance Video Generation with Motions & Facial Expressions Transfer (Nov 2023), http://arxiv.org/abs/2311.12052, arXiv:2311.12052 [cs]

[7] Chen, H., Xia, M., He, Y., Zhang, Y., Cun, X., Yang, S., Xing, J., Liu, Y., Chen, Q., Wang, X., Weng, C., Shan, Y.: VideoCrafter1: Open Diffusion Models for High-Quality Video Generation (Oct 2023), http://arxiv.org/abs/2310.19512, arXiv:2310.19512 [cs]

[8] Chen, H., Zhang, Y., Cun, X., Xia, M., Wang, X., Weng, C., Shan, Y.: VideoCrafter2: Overcoming Data Limitations for High-Quality Video Diffusion Models (Jan 2024), http://arxiv.org/abs/2401.09047, arXiv:2401.09047 [cs]

[9] Chen, T.S., Lin, C.H., Tseng, H.Y., Lin, T.Y., Yang, M.H.: Motion-Conditioned Diffusion Model for Controllable Video Synthesis (Apr 2023), http://arxiv.org/abs/2304.14404, arXiv:2304.14404 [cs]

[10] Dai, X., Hou, J., Ma, C.Y., Tsai, S., Wang, J., Wang, R., Zhang, P., Vandenhende, S., Wang, X., Dubey, A., Yu, M., Kadian, A., Radenovic, F., Mahajan, D., Li, K., Zhao, Y., Petrovic, V., Singh, M.K., Motwani, S., Wen, Y., Song, Y., Sumbaly, R., Ramanathan, V., He, Z., Vajda, P., Parikh, D.: Emu: Enhancing Image Generation Models Using Photogenic Needles in a Haystack (Sep 2023), http://arxiv.org/abs/2309.15807, arXiv:2309.15807 [cs]

[11] Esser, P., Rombach, R., Ommer, B.: Taming Transformers for High-Resolution Image Synthesis (Jun 2021), http://arxiv.org/abs/2012.09841, arXiv:2012.09841 [cs]

[12] Feng, M., Liu, J., Yu, K., Yao, Y., Hui, Z., Guo, X., Lin, X., Xue, H., Shi, C., Li, X., Li, A., Kang, X., Lei, B., Cui, M., Ren, P., Xie, X.: DreaMoving: A Human Video Generation Framework based on Diffusion Models (Dec 2023), http://arxiv.org/abs/2312.05107, arXiv:2312.05107 [cs]

[13] Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion (Aug 2022), http://arxiv.org/abs/2208.01618, arXiv:2208.01618 [cs]

[14] Ge, S., Nah, S., Liu, G., Poon, T., Tao, A., Catanzaro, B., Jacobs, D., Huang, J.B., Liu, M.Y., Balaji, Y.: Preserve Your Own Correlation: A Noise Prior for Video Diffusion Models (Aug 2023), http://arxiv.org/abs/2305.10474, arXiv:2305.10474 [cs]

[15] Guo, Y., Yang, C., Rao, A., Agrawala, M., Lin, D., Dai, B.: SparseCtrl: Adding Sparse Controls to Text-to-Video Diffusion Models (Nov 2023), http://arxiv.org/abs/2311.16933, arXiv:2311.16933 [cs]

[16] He, Y., Yang, T., Zhang, Y., Shan, Y., Chen, Q.: Latent Video Diffusion Models for High-Fidelity Long Video Generation (Mar 2023), http://arxiv.org/abs/2211.13221, arXiv:2211.13221 [cs]

[17] Hendricks, L.A., Nematzadeh, A.: Probing Image-Language Transformers for Verb Understanding (Jun 2021), http://arxiv.org/abs/2106.09141, arXiv:2106.09141 [cs]

[18] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium (Jan 2018), http://arxiv.org/abs/1706.08500, arXiv:1706.08500 [cs, stat]

[19] Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., Salimans, T.: Imagen Video: High Definition Video Generation with Diffusion Models (Oct 2022), http://arxiv.org/abs/2210.02303, arXiv:2210.02303 [cs]

[20] Ho, J., Jain, A., Abbeel, P.: Denoising Diffusion Probabilistic Models (Dec 2020), http://arxiv.org/abs/2006.11239, arXiv:2006.11239 [cs, stat]

[21] Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video Diffusion Models (Jun 2022), http://arxiv.org/abs/2204.03458, arXiv:2204.03458 [cs]

[22] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-Rank Adaptation of Large Language Models (Oct 2021), http://arxiv.org/abs/2106.09685, arXiv:2106.09685 [cs]

[23] Huang, H.P., Su, Y.C., Sun, D., Jiang, L., Jia, X., Zhu, Y., Yang, M.H.: Fine-grained Controllable Video Generation via Object Appearance and Context (Dec 2023), http://arxiv.org/abs/2312.02919, arXiv:2312.02919 [cs]

[24] Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., Wang, Y., Chen, X., Wang, L., Lin, D., Qiao, Y., Liu, Z.: VBench: Comprehensive Benchmark Suite for Video Generative Models (Nov 2023), http://arxiv.org/abs/2311.17982, arXiv:2311.17982 [cs]

[25] Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization (Jan 2017), http://arxiv.org/abs/1412.6980, arXiv:1412.6980 [cs]

[26] Kondratyuk, D., Yu, L., Gu, X., Lezama, J., Huang, J., Hornung, R., Adam, H., Akbari, H., Alon, Y., Birodkar, V., Cheng, Y., Chiu, M.C., Dillon, J., Essa, I., Gupta, A., Hahn, M., Hauth, A., Hendon, D., Martinez, A., Minnen, D., Ross, D., Schindler, G., Sirotenko, M., Sohn, K., Somandepalli, K., Wang, H., Yan, J., Yang, M.H., Yang, X., Seybold, B., Jiang, L.: VideoPoet: A Large Language Model for Zero-Shot Video Generation

[27] Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-Concept Customization of Text-to-Image Diffusion

[28] Li, S., van de Weijer, J., Hu, T., Khan, F.S., Hou, Q., Wang, Y., Yang, J.: StyleDiffusion: Prompt-Embedding Inversion for Text-Based Editing (Aug 2023), http://arxiv.org/abs/2303.15649, arXiv:2303.15649 [cs]

[29] Li, Y., Liu, X., Kag, A., Hu, J., Idelbayev, Y., Sagar, D., Wang, Y., Tulyakov, S., Ren, J.: TextCraftor: Your Text Encoder Can be Image Quality Controller (Mar 2024), http://arxiv.org/abs/2403.18978, arXiv:2403.18978 [cs]

[30] Liang, J., Fan, Y., Zhang, K., Timofte, R., Van Gool, L., Ranjan, R.: MoVideo: Motion-Aware Video Generation with Diffusion Models (Nov 2023), http://arxiv.org/abs/2311.11325, arXiv:2311.11325 [cs, eess]

[31] Liu, Y., Cun, X., Liu, X., Wang, X., Zhang, Y., Chen, H., Liu, Y., Zeng, T., Chan, R., Shan, Y.: EvalCrafter: Benchmarking and Evaluating Large Video Generation Models (Oct 2023), http://arxiv.org/abs/2310.11440, arXiv:2310.11440 [cs]

[32] Luo, Z., Chen, D., Zhang, Y., Huang, Y., Wang, L., Shen, Y., Zhao, D., Zhou, J., Tan, T.: VideoFusion: Decomposed Diffusion Models for High-Quality Video Generation (Oct 2023), http://arxiv.org/abs/2303.08320, arXiv:2303.08320 [cs]

[33] Lv, J., Huang, Y., Yan, M., Huang, J., Liu, J., Liu, Y., Wen, Y., Chen, X., Chen, S.: GPT4Motion: Scripting Physical Motions in Text-to-Video Generation via Blender-Oriented GPT Planning (Nov 2023), http://arxiv.org/abs/2311.12631, arXiv:2311.12631 [cs]

[34] Materzynska, J., Sivic, J., Shechtman, E., Torralba, A., Zhang, R., Russell, B.: Customizing Motion in Text-to-Video Diffusion Models (Dec 2023), http://arxiv.org/abs/2312.04966, arXiv:2312.04966 [cs]

[35] Momeni, L., Caron, M., Nagrani, A., Zisserman, A., Schmid, C.: Verbs in Action: Improving verb understanding in video-language models (Apr 2023), http://arxiv.org/abs/2304.06708, arXiv:2304.06708 [cs]

[36] Ni, H., Shi, C., Li, K., Huang, S.X., Min, M.R.: Conditional Image-to-Video Generation with Latent Flow Diffusion Models (Mar 2023), http://arxiv.org/abs/2303.13744, arXiv:2303.13744 [cs]

[37] Oord, A.v.d., Vinyals, O., Kavukcuoglu, K.: Neural Discrete Representation Learning (May 2018), http://arxiv.org/abs/1711.00937, arXiv:1711.00937 [cs]

[38] Park, J.S., Shen, S., Farhadi, A., Darrell, T., Choi, Y., Rohrbach, A.: Exposing the Limits of Video-Text Models through Contrast Sets. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 3574–3586. Association for Computational Linguistics, Seattle, United States (2022). https://doi.org/10.18653/v1/2022.naacl-main.261, https://aclanthology.org/2022.naacl-main.261

[39] Parmar, G., Zhang, R., Zhu, J.Y.: On Aliased Resizing and Surprising Subtleties in GAN Evaluation (Jan 2022), http://arxiv.org/abs/2104.11222, arXiv:2104.11222

[40] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision (Feb 2021), http://arxiv.org/abs/2103.00020, arXiv:2103.00020 [cs]

[41] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-Resolution Image Synthesis with Latent Diffusion Models (Apr 2022), http://arxiv.org/abs/2112.10752, arXiv:2112.10752 [cs]

[42] Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation (Mar 2023), http://arxiv.org/abs/2208.12242, arXiv:2208.12242 [cs]

[43] Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved Techniques for Training GANs (Jun 2016), http://arxiv.org/abs/1606.03498, arXiv:1606.03498 [cs]

[44] Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: LAION-5B: An open large-scale dataset for training next generation image-text models (Oct 2022), http://arxiv.org/abs/2210.08402, arXiv:2210.08402 [cs]

 https://doi.org/10.52202/079017-2241

[45] Sheynin, S., Polyak, A., Singer, U., Kirstain, Y., Zohar, A., Ashual, O., Parikh, D., Taigman, Y.: Emu Edit: Precise Image Editing via Recognition and Generation Tasks (Nov 2023), http://arxiv.org/abs/2311.10089, arXiv:2311.10089 [cs]

[46] Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., Parikh, D., Gupta, S., Taigman, Y.: Make-A-Video: Text-to-Video Generation without Text-Video Data (Sep 2022), http://arxiv.org/abs/2209.14792, arXiv:2209.14792 [cs]

[47] Smith, L.N., Topin, N.: Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates (May 2018), http://arxiv.org/abs/1708.07120, arXiv:1708.07120 [cs, stat]

[48] Sohl-Dickstein, J., Weiss, E.A., Maheswaranathan, N., Ganguli, S.: Deep Unsupervised Learning using Nonequilibrium Thermodynamics (Nov 2015), http://arxiv.org/abs/1503.03585, arXiv:1503.03585 [cond-mat, q-bio, stat]

[49] Song, J., Meng, C., Ermon, S.: Denoising Diffusion Implicit Models (Oct 2022), http://arxiv.org/abs/2010.02502, arXiv:2010.02502 [cs]

[50] Soomro, K., Zamir, A.R., Shah, M.: UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild (Dec 2012), http://arxiv.org/abs/1212.0402, arXiv:1212.0402 [cs]

[51] Su, S., Guo, L., Gao, L., Shen, H., Song, J.: MotionZero:Exploiting Motion Priors for Zero-shot Text-to-Video Generation (Nov 2023), http://arxiv.org/abs/2311.16635, arXiv:2311.16635 [cs]

[52] Sun, K., Pan, J., Ge, Y., Li, H., Duan, H., Wu, X., Zhang, R., Zhou, A., Qin, Z., Wang, Y., Dai, J., Qiao, Y., Wang, L., Li, H.: JourneyDB: A Benchmark for Generative Image Understanding (Oct 2023), http://arxiv.org/abs/2307.00716, arXiv:2307.00716 [cs]

[53] Teed, Z., Deng, J.: RAFT: Recurrent All-Pairs Field Transforms for Optical Flow (Aug 2020), http://arxiv.org/abs/2003.12039, arXiv:2003.12039 [cs]

[54] Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards Accurate Generative Models of Video: A New Metric & Challenges (Mar 2019), http://arxiv.org/abs/1812.01717, arXiv:1812.01717 [cs, stat]

[55] Wang, J., Zhang, Y., Zou, J., Zeng, Y., Wei, G., Yuan, L., Li, H.: Boximator: Generating Rich and Controllable Motions for Video Synthesis (Feb 2024), http://arxiv.org/abs/2402.01566, arXiv:2402.01566 [cs]

[56] Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., Zhang, S.: ModelScope Text-to-Video Technical Report (Aug 2023), http://arxiv.org/abs/2308.06571, arXiv:2308.06571 [cs]

[57] Wang, L., Huang, B., Zhao, Z., Tong, Z., He, Y., Wang, Y., Wang, Y., Qiao, Y.: VideoMAE V2: Scaling Video Masked Autoencoders with Dual Masking (Apr 2023), http://arxiv.org/abs/2303.16727, arXiv:2303.16727 [cs]

[58] Wang, W., Yang, H., Tuo, Z., He, H., Zhu, J., Fu, J., Liu, J.: VideoFactory: Swap Attention in Spatiotemporal Diffusions for Text-to-Video Generation (Jun 2023), http://arxiv.org/abs/2305.10874, arXiv:2305.10874 [cs]

[59] Wang, Y., Chen, X., Ma, X., Zhou, S., Huang, Z., Wang, Y., Yang, C., He, Y., Yu, J., Yang, P., Guo, Y., Wu, T., Si, C., Jiang, Y., Chen, C., Loy, C.C., Dai, B., Lin, D., Qiao, Y., Liu, Z.: LAVIE: High-Quality Video Generation with Cascaded Latent Diffusion Models (Sep 2023), http://arxiv.org/abs/2309.15103, arXiv:2309.15103 [cs]

[60] Wang, Z., Yuan, Z., Wang, X., Chen, T., Xia, M., Luo, P., Shan, Y.: MotionCtrl: A Unified and Flexible Motion Controller for Video Generation (Dec 2023), http://arxiv.org/abs/2312.03641, arXiv:2312.03641 [cs]

[61] Wang, Z., Sha, Z., Ding, Z., Wang, Y., Tu, Z.: TokenCompose: Grounding Diffusion with Token-level Supervision (Dec 2023), http://arxiv.org/abs/2312.03626, arXiv:2312.03626 [cs]

[62] Wei, Y., Zhang, S., Qing, Z., Yuan, H., Liu, Z., Liu, Y., Zhang, Y., Zhou, J., Shan, H.: DreamVideo: Composing Your Dream Videos with Customized Subject and Motion (Dec 2023), http://arxiv.org/abs/2312.04433, arXiv:2312.04433 [cs]

[63] Wu, C., Huang, L., Zhang, Q., Li, B., Ji, L., Yang, F., Sapiro, G., Duan, N.: GODIVA: Generating Open-DomaIn Videos from nAtural Descriptions (Apr 2021), http://arxiv.org/abs/2104.14806, arXiv:2104.14806 [cs]

[64] Wu, C., Liang, J., Ji, L., Yang, F., Fang, Y., Jiang, D., Duan, N.: N\"UWA: Visual Synthesis Pre-training for Neural visUal World creAtion (Nov 2021), http://arxiv.org/abs/2111.12417, arXiv:2111.12417 [cs]

[65] Wu, J.Z., Ge, Y., Wang, X., Lei, W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation (Mar 2023), http://arxiv.org/abs/2212.11565, arXiv:2212.11565 [cs]

[66] Xu, J., Mei, T., Yao, T., Rui, Y.: MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5288–5296. IEEE, Las Vegas, NV, USA (Jun 2016). https://doi.org/10.1109/CVPR.2016.571, http://ieeexplore.ieee.org/document/7780940/

[67] Xue, H., Hang, T., Zeng, Y., Sun, Y., Liu, B., Yang, H., Fu, J., Guo, B.: Advancing High-Resolution Video-Language Representation with Large-Scale Video Transcriptions (Jul 2022), http://arxiv.org/abs/2111.10337, arXiv:2111.10337 [cs]

[68] Yin, S., Wu, C., Liang, J., Shi, J., Li, H., Ming, G., Duan, N.: DragNUWA: Fine-grained Control in Video Generation by Integrating Text, Image, and Trajectory (Aug 2023), http://arxiv.org/abs/2308.08089, arXiv:2308.08089 [cs]

[69] Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., Zou, J.: When and why vision-language models behave like bags-of-words, and what to do about it? (Mar 2023), http://arxiv.org/abs/2210.01936, arXiv:2210.01936 [cs]

[70] Zhang, D.J., Wu, J.Z., Liu, J.W., Zhao, R., Ran, L., Gu, Y., Gao, D., Shou, M.Z.: Show-1: Marrying Pixel and Latent Diffusion Models for Text-to-Video Generation (Oct 2023), http://arxiv.org/abs/2309.15818, arXiv:2309.15818 [cs]

[71] Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., Liu, Z.: MotionDiffuse: Text-Driven Human Motion Generation with Diffusion Model (Aug 2022), http://arxiv.org/abs/2208.15001, arXiv:2208.15001 [cs]

[72] Zhang, M., Guo, X., Pan, L., Cai, Z., Hong, F., Li, H., Yang, L., Liu, Z.: ReMoDiffuse: Retrieval-Augmented Motion Diffusion Model (Apr 2023), http://arxiv.org/abs/2304.01116, arXiv:2304.01116 [cs]

[73] Zhao, R., Gu, Y., Wu, J.Z., Zhang, D.J., Liu, J., Wu, W., Keppo, J., Shou, M.Z.: MotionDirector: Motion Customization of Text-to-Video Diffusion Models (Oct 2023), http://arxiv.org/abs/2310.08465, arXiv:2310.08465 [cs]

[74] Zhou, D., Wang, W., Yan, H., Lv, W., Zhu, Y., Feng, J.: MagicVideo: Efficient Video Generation With Latent Diffusion Models (May 2023), http://arxiv.org/abs/2211.11018, arXiv:2211.11018 [cs]

 https://doi.org/10.52202/079017-2241

9 Appendix

9.1 Training Details and Hyperparameters

As shown in Table 7, we train DEMO using the Adam optimizer [25] with a OneCycle scheduler [47]. Specifically, the learning rate varies within the range of [0.00001, 0.00005], while the momentum oscillates between 0.85 and 0.99. We train our model using Deepspeed framework with stage 2 zero optimization and cpu offloading. DEMO is trained on 4 NVIDIA Tesla A100 GPUs with a batch size of 24 per GPU. DEMO takes 256×256 images as inputs and utilizes a VQGAN with a compression rate of 8 to encode images into a latent space of 32× 32. DEMO is trained with 1000 diffusion steps. We set the classifier-free guidance scale as 9 with the probability of 0.1 randomly dropping the text during training. For inference, we use the DDIM sampler [49] with 50 inference steps.

Table 7: Training Hyperparameters

Hyperparameter		DEMO
U-net	LDM	✓
	Compression Rate	8
	Latent Shape	32×32×16
	Channels	320
	Channel Multiplier	1,2,4,4
	Attention Resolutions	16, 8, 4
	Head Channels	32
	# of Parameters	1.68B
	Dropout Rate	0.1
Motion Encoder	Architecture	CLIP ViT-H/14
	Token Length	77
	Token Dimension	1024
	# of Parameters	354.03M
Motion Conditioning	Proj In & Proj Out	Linear
	Normalization	GroupNorm 32
	Activation	GEGLU
	Channels	320
	Attention Resolutions	16, 8, 4
	Head Channels	32
	# of Parameters	238.32M
Training	DDPM Time Steps	0, 1000
	Optimizer	Adam
	Learning Rate	0.00001, 0.00005
	Scheduler	OneCycle Scheduler
	Classifier-free Guidance Scale	9
	Loss Weight α	0.1
	Loss Weight β	0.3
	Loss Weight γ	0.1
	Optical Flow Estimator	Raft [53]
Inference	DDIM Sampling Steps	50

9.2 Pliot Study Details

To test the sensitivity of the motion encoder to parts of speech (POS) representing content and motion information, we generated a set of prompts following the template: A [ADJ][NOUN_1][VERB][ADV][ADP] the [NOUN_2]. We then grouped these prompts according to their respective POS categories. Next, we calculated the pairwise sentence similarity within each group using the "[eot]" token to determine sentence similarity. The average similarity within each group, as well as across different groups, was reported. This setup groups different words with the same POS under the same context, thereby eliminating potential biases introduced by the context.

Table 8: Training dataset of current T2V models.

Model	Base Model	Training Dataset
MagicVideo [74]	LDM [41]	WebVid-10M [1] + 10M from HD-VILA-100M [67] + Interal 7M
Make-A-Video [46]	-	2.3B from Laion-5B [44] + WebVid-10M [1] + 10M from HD-VILA-100M[67]
Video LDM [4]	LDM [41]	RDS for driving/WebVid-10M [1] for T2V
ModelScopeT2V [56]	LDM [41]	2.3B from Laion-5B [44] + WebVid-10M [1]
Show-1 [70]	DeepFloyd[4]+ ModelScopeT2V [56]	WebVid-10M [1]
LaVie [59]	Stable Diffusion 1.4 [41]	Laion5B [44] + WebVid-10M [1] + Vimeo-25M [59]
PyoCo [14]	eDiff-I [2]	1.2B text-image dataset + 22.5M text-video dataset
VideoFactory [58]	LDM [41]	HD-VG-130M [58] + WebVid-10M [1]
EMU VIDEO [45]	Emu [10]	34M licensed text-video dataset
SVD [3]	Stable Diffusion 2.1 [41]	LVD-F [3] (152M) + 250K pre-captioned video clips of high visual fidelity.
DEMO	ModelScopeT2V [56]	WebVid-10M [1]

Table 9: Comparison of different evaluation protocol on MSR-VTT.

Model	FID (↓)	FID-CLIP (↓)	FVD (↓)	CLIPSIM (↑)	Evaluation Protocol
MagicVideo [74]	36.50	-	1290	-	Text prompt on test set; unknown number.
Make-A-Video [46]	-	13.17	-	0.3049	FID and CLIPSIM are evaluated on 59794 videos with text prompt from test set.
Show-1 [70]	-	13.08	538	0.3072	FID and FVD are evaluated with 2048 videos generated on test set. CLIPSIM is evaluate on 59794 videos with prompts.
Video LDM [4]	-	-	-	0.2929	CLIPSIM is calculate on 2990 videos with prompts from test set.
LaVie [59]	-	-	-	0.2949	CLIPSIM is calculate on 2990 videos with prompts from test set.
PYoCo [14]	25.39-22.14	10.21-9.73	-	-	The same as Make-A-Video.
VideoFactory [58]	-	-	-	0.3005	CLIPSIM is calculate on 2990 videos with prompts from test set.
ModelScopeT2V [56]		14.89	557	0.2941	FID and FVD are evaluated with 2048 videos generated on test set. CLIPSIM is evaluate on 59794 videos with prompts.
ModelScopeT2V fine-tuned		13.80	536	0.2932	Same as ModelScopeT2V
DEMO		**11.77**	**422**	**0.2965**	Same as ModelScopeT2V

We define the sensitivity of our motion encoder as one minus this similarity. The full set of different words within each POS category is defined as follows:

$$
\begin{aligned}
\text{ADJ} &= \{\text{"big"}, \text{"small"}, \text{"tall"}, \text{"short"}, \text{"fat"}, \text{"thin"}, \text{"young"}, \text{"old"}\} \\
\text{NOUN}_1 &= \{\text{"cat"}, \text{"dog"}, \text{"horse"}, \text{"child"}, \text{"man"}, \text{"woman"}, \text{"bird"}, \text{"fish"}\} \\
\text{VERB} &= \{\text{"walk"}, \text{"run"}, \text{"jump"}, \text{"crawl"}, \text{"eat"}, \text{"swim"}, \text{"fly"}, \text{"climb"}\} \\
\text{ADV} &= \{\text{"quickly"}, \text{"slowly"}, \text{"suddenly"}, \text{"steadily"}, \text{"cautiously"}, \\
&\quad \text{"briskly"}, \text{"gracefully"}, \text{"clumsily"}\} \\
\text{ADP} &= \{\text{"across"}, \text{"over"}, \text{"through"}, \text{"beside"}, \text{"against"}, \text{"under"}, \text{"above"}, \text{"near"}\} \\
\text{NOUN}_2 &= \{\text{"river"}, \text{"bridge"}, \text{"mountain"}, \text{"tree"}, \text{"house"}, \text{"lake"}, \text{"field"}, \text{"forest"}\}
\end{aligned}
$$

Given these six categories with eight words each, we have a total of $8^6 = 262144$ prompts. It is noteworthy that we did not observe significant differences when using different templates or different sets of words within each POS. The results were consistent across different setups, and we selected these prompts to try to make these prompts meaningful.

9.3 Detailed Training and Evaluation for T2V Models

As shown in Table 8, existing T2V models are trained using diverse datasets and strategies, leading to various evaluation standards across different datasets, as detailed in Table 9 and Table 10. Here,

[4]https://github.com/deep-floyd/IF

https://doi.org/10.52202/079017-2241

Table 10: Comparison of different evaluation protocols on UCF-101.

Model	IS (↑)	FVD (↓)	Evaluation Protocol
MagicVideo [74]	-	655.00	Evaluated on videos generated with class labels; unknown number.
Make-A-Video [46]	33.00	367.23	One template sentence per class label; 100 videos per class.
Show-1 [70]	35.42	394.46	One template sentence per class label; 20 videos per prompt for IS; FVD on 2048 sampled videos.
Video LDM [4]	33.45	550.61	Class label only; 100 videos per class.
LaVie [59]	-	526.30	Class label only; 100 videos per class.
PYoCo [14]	47.76	355.19	One template sentence per class label; 20 videos per prompt for IS; FVD on 2048 sampled videos.
VideoFactory [58]	-	410.00	One template sentence per class label; 100 videos per class.
EMU VIDEO [45]	42.70	606.20	One template sentence per class label; 100 videos per class.
SVD [3]	-	242.02	FVD on 13,320 videos using class labels only.
ModelScopeT2V [56]	**37.49**	630.23	100 videos per class using class labels only.
ModelScopeT2V fine-tuned	37.21	612.53	100 videos per class using class labels only.
DEMO	36.35	**547.31**	100 videos per class using class labels only.

we detail and justify our evaluation standards. For our evaluation on MSR-VTT, we follow the base model's approach to compute the CLIPSIM on the entire MSR-VTT dataset. For FID computation, CLIP-ViT/B 32 is used to extract the frame features. For FID and FVD, we randomly sample 2048 videos, following the ModelScopeT2V paper's methodology. For our evaluation on UCF-101, to eliminate bias introduced by template sentences (as done in several previous works), we directly use the class labels to compute the IS and FVD scores.

9.4 Extended Quantitative Evaluations

To evaluate the generalization of our methods, we applied DEMO on ZeroScope, we report the performance as follows:

Table 11: Quantitative results on ZeroScope.

Benchmark	**Metric**	**ZeroScope**	**DEMO+ZeroScope**
MSRVTT	FID (↓)	14.57	13.59
	FVD (↓)	812	543
	CLIPSIM (↑)	0.2833	0.2945
UCF-101	IS (↑)	37.22	37.01
	FVD (↓)	744	601
WebVid-10M	FID (↓)	11.34	10.03
	FVD (↓)	615	479
	CLIPSIM (↑)	0.2846	0.2903
EvalCrafter	VQA_A (↑)	27.76	33.02
	VQA_T (↑)	33.87	37.28
	IS (↑)	14.20	15.28
	ActionScore (↑)	67.78	72.55
	MotionAC-Score (↑)	44	62
	FlowScore (↑)	1.10	5.25
Vbench	MotionDynamics (↑)	42.72	70.28
	HumanAction (↑)	67.36	88.34
	TemporalFlickering (↓)	97.39	94.83
	MotionSmoothness (↑)	97.92	95.72

9.5 Extended Qualitative Evaluations

In this section, we provide extended qualitative comparison between our method and the baseline.

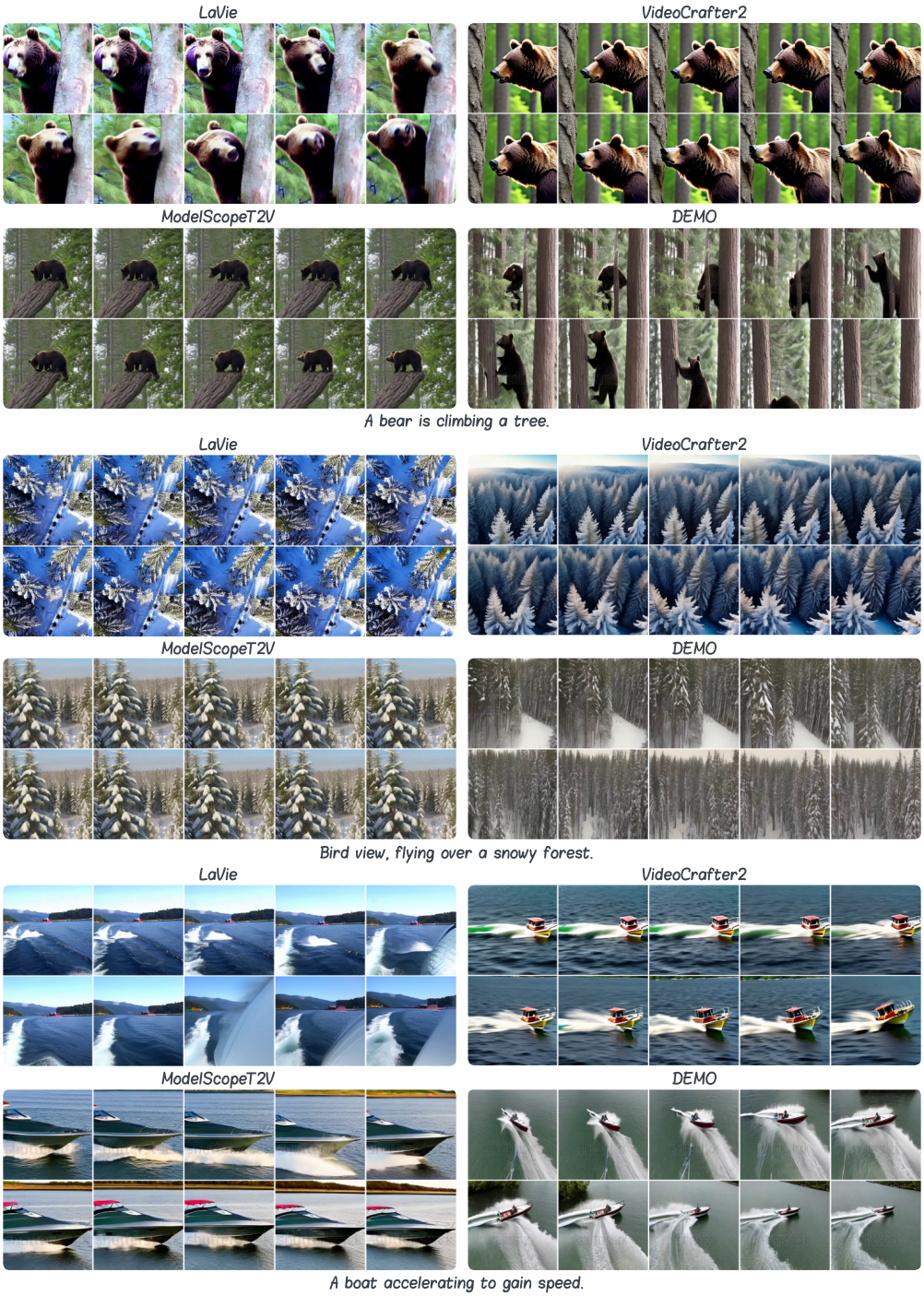

Figure 5: **Extended qualitative comparison.** Each video is generated with 16 frames. We display frames 1, 2, 4, 6, 8, 10, 12, 14, 15, and 16, arranged in two rows from left to right. Full videos are available in the supplementary materials.

https://doi.org/10.52202/079017-2241

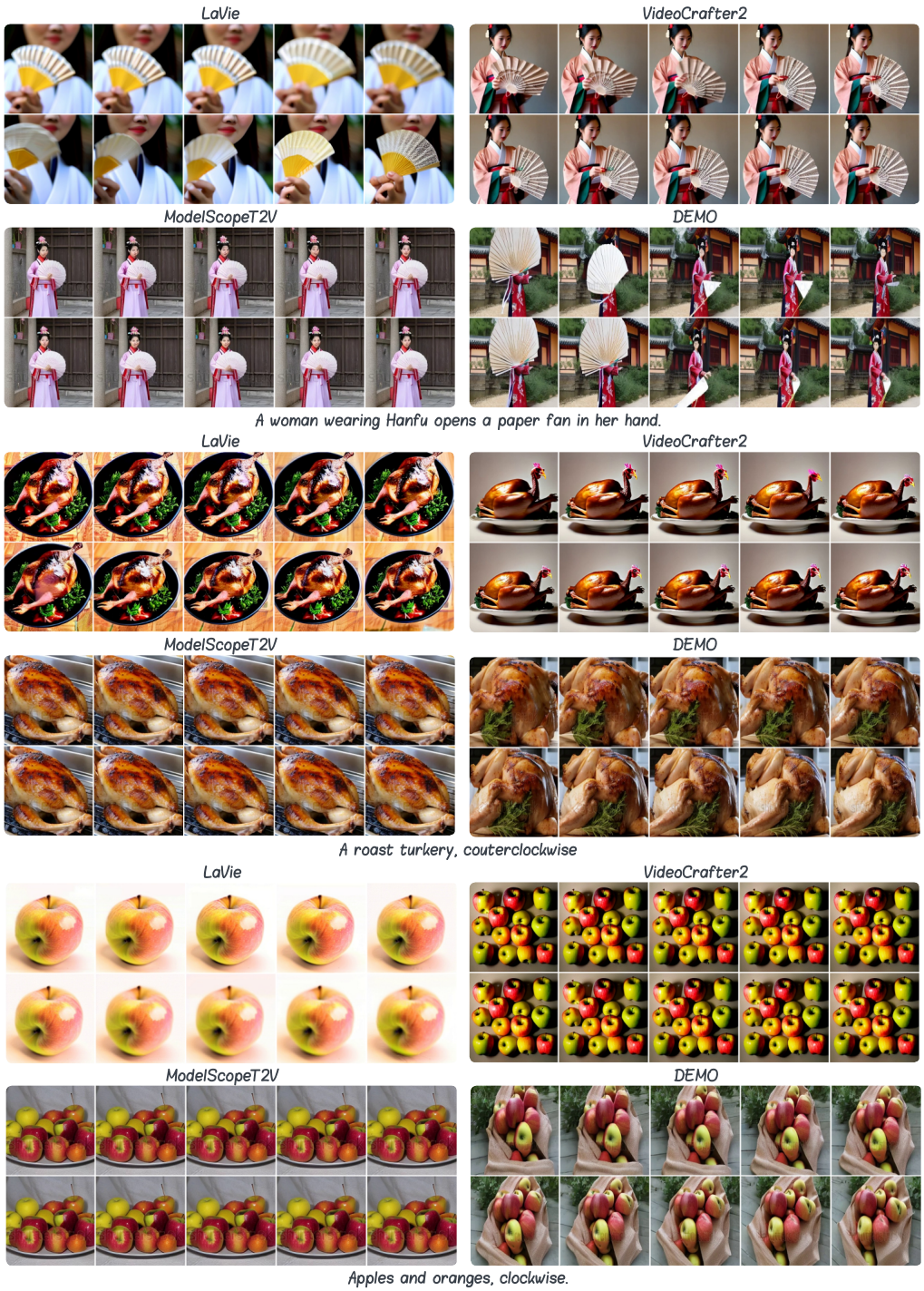

Figure 6: **Extended qualitative comparison.** Each video is generated with 16 frames. We display frames 1, 2, 4, 6, 8, 10, 12, 14, 15, and 16, arranged in two rows from left to right. Full videos are available in the supplementary materials.

https://doi.org/10.52202/079017-2241

Figure 7: **Extended qualitative comparison.** Each video is generated with 16 frames. We display frames 1, 2, 4, 6, 8, 10, 12, 14, 15, and 16, arranged in two rows from left to right. Full videos are available in the supplementary materials.

https://doi.org/10.52202/079017-2241

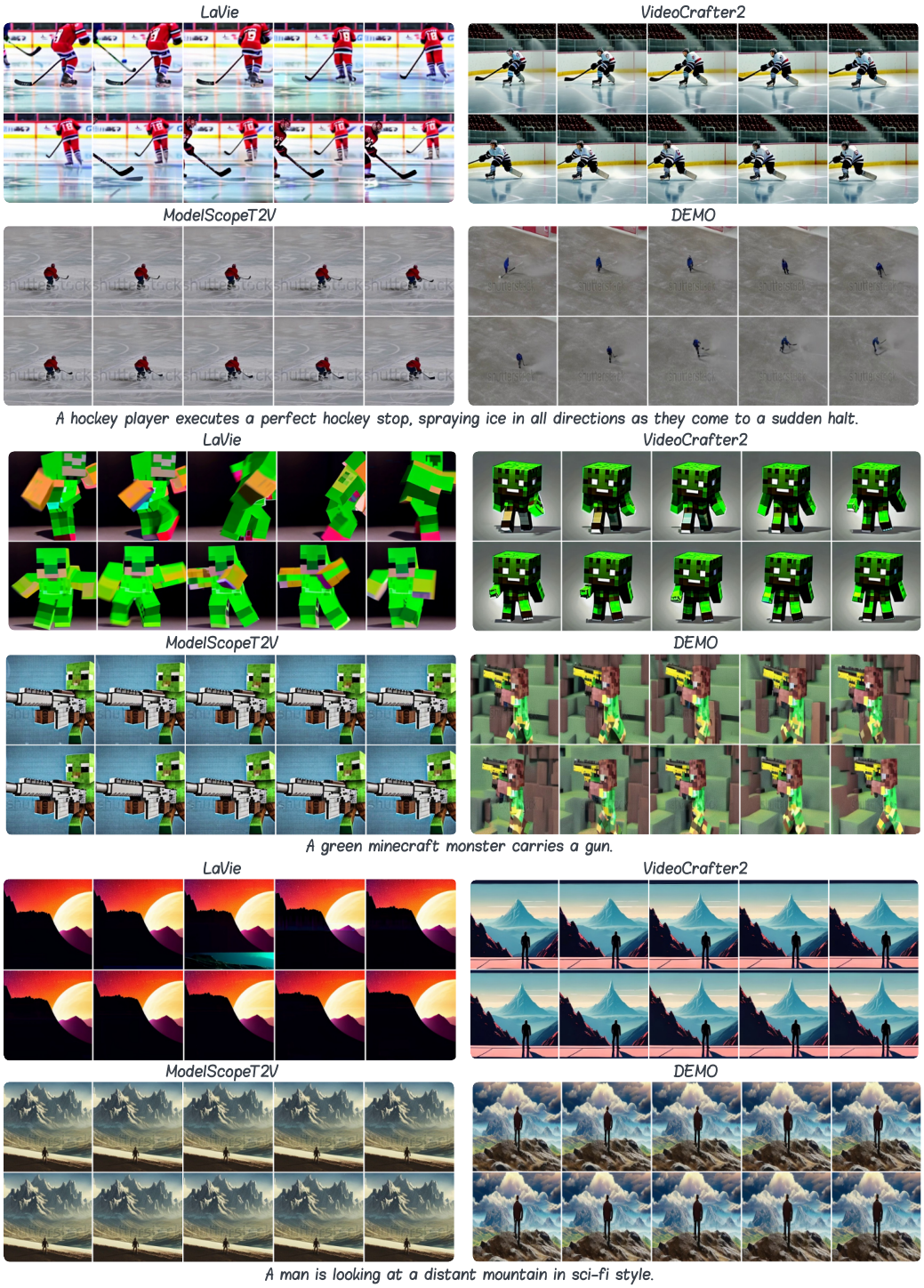

Figure 8: **Extended qualitative comparison.** Each video is generated with 16 frames. We display frames 1, 2, 4, 6, 8, 10, 12, 14, 15, and 16, arranged in two rows from left to right. Full videos are available in the supplementary materials.

9.6 Human Evaluation

To further assess the qualitative performance of our proposed method, we conducted a user study comparing our approach (DEMO) with several state-of-the-art video generation models. We randomly

selected 50 prompts from EvalCrafter [31], ensuring diversity across scenes, styles, and objects. For each comparison, 15 annotators evaluated the generated videos in terms of three main criteria: text-video alignment, visual quality, and motion quality. The study compared our method with ModelScopeT2V, LaVie, and VideoCrafter2.

The participants were asked to select their preferred video between the two models for each prompt. The comparative results are summarized in Table 12. Specifically, DEMO consistently outperformed ModelScopeT2V, LaVie, and VideoCrafter2, particularly in terms of motion quality, where it achieved a preference rate of 74% over ModelScopeT2V. Additionally, DEMO was favored in text-video alignment and visual quality by 62% and 66%, respectively. However, when compared to LaVie and VideoCrafter2, DEMO showed a lower performance in visual quality, which can be attributed to differences in training datasets. LaVie and VideoCrafter2 use higher-quality video and image datasets, such as Vimeo-25M [59] and JDB [52], respectively, while DEMO and ModelScopeT2V are trained on the WebVid10M dataset, which is lower in visual quality.

Furthermore, we conducted an additional user study to evaluate the effectiveness of our proposed video-motion supervision term, $\mathcal{L}_{\text{video-motion}}$. The results indicated that our method with motion supervision outperformed the version without motion supervision, achieving win rates of 58%, 56%, and 72% in text-video alignment, visual quality, and motion quality, respectively. These findings highlight the significant improvements brought by the video-motion supervision in generating smoother and more realistic motion dynamics.

Table 12: User Study Results: Comparison between DEMO and Other Models

Methods	Text-Video Alignment	Visual Quality	Motion Quality
DEMO vs ModelScopeT2V	62%	66%	74%
DEMO vs LaVie	56%	46%	62%
DEMO vs VideoCrafter2	60%	42%	52%
DEMO vs DEMO w/o $\mathcal{L}_{\text{video-motion}}$	58%	56%	72%

In summary, the user study provides strong evidence that DEMO improves motion quality without sacrificing text-video alignment. Despite the lower visual quality compared to models trained on high-quality datasets, our method demonstrates its strength in motion generation, a key aspect of video realism.

https://doi.org/10.52202/079017-2241

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: As shown in abstract and the fourth paragraph in the introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: As shown in limitations 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper does not include theoretical results.

   Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide code for our proposed method in supplementary materials.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

https://doi.org/10.52202/079017-2241

Justification: We provide the environment requirements and the command to run the experiments in the supplementary materials.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The training details are shown in appendix 9.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: Error bars are not reported because it would be too computationally expensive.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide computation resources in appendix 9.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted conform the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: As shown in broader impacts 6.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

https://doi.org/10.52202/079017-2241

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We credited the models and dataset we used in the paper and we list their licenses here.
    ModelScopeT2V: cc-by-nc-4.0,
    Open-CLIP: “https://github.com/mlfoundations/open_clip/blob/main/LICENSE"
    WebVid-10M: custom license: “https://github.com/m-bain/webvid/blob/main/TERMS.md".

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-2241