AI-Based Multilingual Text-to-Video Generation for PIB Press Releases”

G. Greeshma
CSE: Computer Science & Engineering
Presidency University (Bng)
Bangalore ,India
grandhegreeshma05@gmail.com

Tabjula Bhavani
CSE: Computer Science & Engineering
Presidency University (Bng)
Bangalore ,India
bhavanitabjula@gmail.com

Bestha Narendra Kumar
CSE: Computer Science & Engineering
Presidency University (Bng)
Bangalore ,India
230804abhi@gmail.com

Abstract

Text-to-video (T2V) diffusion models have shown promising capabilities in synthesizing realistic videos from input text prompts. However, the input text description alone provides limited control over the precise objects movements and camera framing. In this work, we tackle the motion customization problem, where a reference video is provided as motion guidance. While most existing methods choose to fine-tune pre-trained diffusion models to reconstruct the frame differences of the reference video, we observe that such strategy suffer from content leakage from the reference video, and they cannot capture complex motion accurately. To address this issue, we propose Motion Matcher, a motion customization framework that fine-tunes the pretrained T2V diffusion model at the feature level. Instead of using pixel-level objectives, Motion Matcher compares high-level, spatio-temporal motion features to fine-tune diffusion models, ensuring precise motion learning. For the sake of memory efficiency and accessibility, we utilize a pretrained T2V diffusion model, which contains considerable prior knowledge about

Video motion, to compute these motion features. In our experiments, we demonstrate state-ofthe-art motion customization performances, validating the design of our framework.

1. Introduction

To control the rhythm of a movie scene, movie directors would carefully arrange the precise movements and positioning of both the actors and the camera for each shot (as known as staging/blocking). Similarly, to control the pacing and flow of AI-generated videos, users should have control over the dynamics and composition of videos produced by generative models. To this end, numerous motion control methods have been proposed to control moving object trajectories in videos generated by text-to-video (T2V) diffusion models. Motion customization, in particular, aims to control T2V diffusion models with the motion of a reference video .With the assistance of the reference video, users are able to specify the desired object movements and camera framing in detail. Formally speaking, given a reference video, motion customization aims to adjust a pre-trained T2V diffusion model, so the output videos sampled from the adjusted model follow the object movements and camera framing of the reference video (see Fig. 1 for an example). Given that motion is a high-level concept involving both spatial and temporal dimensions , motion customization is considered a non-trivial task.

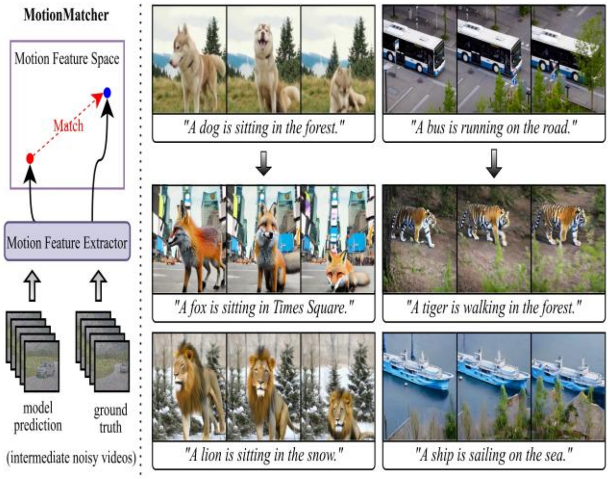
Recently, many motion customization methods have been proposed to eliminate the influence of visual appearance in the reference video. Among them, a standout strategy is fine-tuning the pre-trained T2V diffusion model to reconstruct the frame differences of the reference video. For instance, VMC [26] and SMA [36] use a motion distillation objective that reconstructs the residual frames of the reference video. MotionDirector [76] proposes an appearancedebiased objective that reconstructs the differences between an anchor frame and all other frames. However, we find that frame differences do not accurately represent motion. For example, two videos with the same motion, such as a red car and a blue car both driving leftward, can yield completely different frame differences because the pixel changes occur in different color channels in each video. Moreover, since frame differences only process videos at the pixel level, they cannot capture complex motion that requires a high-level understanding of video, such as rapid movements or movements in low-texture regions. In these cases, the strategy of reconstructing frame differences fails to reproduce the target motion.

To address this issue, we propose Motion Matcher, a novel fine-tuning framework for motion customization via motion feature matching. Instead of aligning pixel values or frame differences as in previous methods, Motion Matcher aligns the projected motion features extracted from a pre-trained feature extractor. Since these motion features are calculated with a sophisticated pre-trained model, they are capable of capturing complex motion that requires a high-level, spatio-temporal understanding of video. This effectively addresses the limitation of previous work, where frame differences fail to capture complex motion.

Motion Matcher differs from traditional fine-tuning approaches. At each fine-tuning step, it starts off by using a feature extractor to compute the motion features of the output video and the motion features of the reconstruction ground truth video. Our feature matching objective then minimizes the L2 distance between the two feature vectors. However, since the output videos of T2V diffusion models are in latent space and at certain noise levels, the feature extractor must be able to process latent noisy videos. To obtain such a feature extractor, we take advantages of (1) pre-trained T2V diffusion models' ability in extracting features from noisy, latent videos and (2) the spatio-temporal information encoded in attention maps. We find that cross attention maps (CA) in pre-trained diffusion models contain information about camera framing, while temporal self attention maps (TSA) represent object movements. Therefore, we utilize them to represent motion features. Ultimately, the design of our framework is validated through detailed analysis and extensive experiments.

To summarize, our key contributions include:

- We propose Motion Matcher, a feature-level fine-tuning framework for motion customization. It leverages a pretrained feature extractor to map videos into a motion feature space, capturing high-level motion information. By aligning the motion features, the diffusion model learns to generate videos with the target motion.
- To extract features from *noisy latent videos*, we utilize the pre-trained diffusion model as a feature extractor, as it naturally processes such inputs.
- We identify two sources of motion cues—cross-attention maps and temporal self-attention maps—and use them to form the motion features.
- We demonstrate that Motion Matcher achieves state-ofthe-art performance through comprehensive experiments. It offers superior joint controllability of text and motion, advancing scene staging in AI-generated videos.


MotionMatcher
Motion Feature Space
Match
Motion Feature Extractor
model prediction
ground truth
(intermediate noisy videos)
"A dog is sitting in the forest."
"A bus is running on the road."
"A fox is sitting in Times Square."
"A tiger is walking in the forest."
"A lion is sitting in the snow."
"A ship is sailing on the sea."

† ‡

Figure 1. MotionMatcher can customize pre-traind T2V diffusion models with a user-provided reference video (top row). Once customized, the diffusion model is able to transfer the precise motion (including object movements and camera framing) in the reference video to a variety of scenes (middle and bottom rows).

2. Literature Review and Technology Analysis

Authors / Year	**Reff**	**Technique / Model**	**Contribution**	**Limitation**
[1] Arnab et al. (2021)	*ViViT: Video Vision Transformer*	Transformer for video representation	Introduced transformer-based architecture	Focused on classification/understanding, not generation

Authors / Year	**Reff**	**Technique / Model**	**Contribution**	**Limitation**
			for video understanding	
[2] Balaji et al. (2019)	*Conditional GAN for Text-to-Video*	GAN with discriminative filter	Early attempt at synthesizing short videos from text	Low resolution, limited temporal length
[3] Black & Anandan (1993)	*Optical Flow Framework*	Robust estimation of motion	Foundational work in motion estimation for videos	Classical approach, not generative
[4] Blattmann et al. (2023)	*Align Your Latents*	Latent diffusion for video	Achieved high-resolution and temporally consistent video synthesis	Very computationally expensive
[5] Bruhn et al. (2005)	*Lucas/Kanade + Horn/Schunck*	Combined optic flow methods	Hybrid approach improving accuracy of flow estimation	Focused on motion analysis, not direct video generation
[6] Chen et al. (2023)	*VideoCrafter1*	Open diffusion model for video	Open-source, high-quality video generation	Limited support for long videos and real-time generation
[7] Dosovitskiy et al. (2015)	*FlowNet*	CNN-based optical flow	First deep learning model for optical flow estimation	Restricted to flow estimation, not full video synthesis
[8] Epstein et al. (2023)	*Diffusion Self-Guidance*	Controllable diffusion	Added fine-grained control in generative models	Primarily image-focused, less on videos
[9] Fan et al. (2017)	*Point Set Generation Network*	3D reconstruction from images	Enabled 3D object generation useful for	Not designed for temporal video synthesis

Authors / Year	Reff	Technique / Model	Contribution	Limitation
			video realism	
[10] Ge et al. (2023)	*Long Video Generation (VQGAN + time-aware)*	Time-agnostic VQGAN	Generated longer videos with temporal consistency	High complexity, scaling challenges

3. Related work

3.1. Text-to-video generation

Text-to-video (T2V) generation models aim to synthesize videos that comply with user-provided text descriptions. Previously, a large number of T2V models have been proposed, including GANs ,autoregressive models ,and diffusion models [4, 17, 70].

Following the success of text-to-image (T2I) diffusion models ,researchers have also put considerable effort into training T2V diffusion models recently. To achieve this, a commonly used approach is inflating a pretrained T2I diffusion model by inserting temporal layers and finetuning the whole model on video data .On the other hand, models like Animate Diff [11] and VideoLDM [4] also insert additional temporal layers, but they only finetune the newly-added temporal layers for decoupling purposes. In contrast to the first approach, these models are typically limited to generating simple motion [73]. To ensure motion complexity, we adopt the former type of model as the base model in this work.

3.2. Motion control in T2V generation

To enable detailed control over camera framing and object movements in T2V generation, recent research has explored trajectory-based [59, 63, 65, 72], box-based [25, 33, 57, 61], and reference-based motion control. Trajectory-based and box-based motion control are typically achieved by conditioning T2V diffusion models on additional motion signal and training them on large video datasets [57, 59, 63, 72], or by directly manipulating attention maps at the inference stage [25, 33, 61]. However, these approaches require users to explicitly define the trajectories of moving objects within frames, which is usually laborious and provides limited control over the entire scene. In contrast, reference-based motion control can specify the target motion more comprehensively via a reference video [26, 31, 36, 71, 76]. In this work, we focus on motion customization, which is considered reference-based motion control.

3.3. Motion customization of T2V diffusion models

Recently, motion customization has emerged as a new area of research. It adapts the pre-trained T2V diffusion model to generate videos that replicate the camera framing and object movements of a user-provided reference video. To avoid learning visual appearance, VMC [26] and SMA [36] fine-tune the pre-trained T2V diffusion model by aligning the residual frames of the output video with the residual frames of the reference video. MotionDirector [76] proposes a dual-path fine-tuning method to avoid learning visual appearance and simultaneously utilizes an objective that matches frame differences. However, since frame differences do not accurately represent motion, these methods struggle to replicate complex motion.

Another strategy is using diffusion guidance [8, 14, 34] to achieve controllable generation. Specifically, DMT [71] employs the intermediate spatio-temporal features in diffusion models as a guidance signal, whereas MotionClone [31] uses intermediate temporal attention maps for guidance. Despite being training-free, these methods need to compute additional gradients during inference, resulting in a lengthy sampling process. Moreover, as noted in [37, 47], the large guidance weights used in diffusion guidance can lead to the generation of out-of-distribution samples.

While other motion customization approaches exist, they address different tasks. For instance, DreamVideo [60] and Customize-A-Video [42] focus solely on replicating object movements without preserving the camera framing, whereas MotionMaster [21] deals exclusively with camera movements. In contrast, our method provides control over both object movements and camera framing.

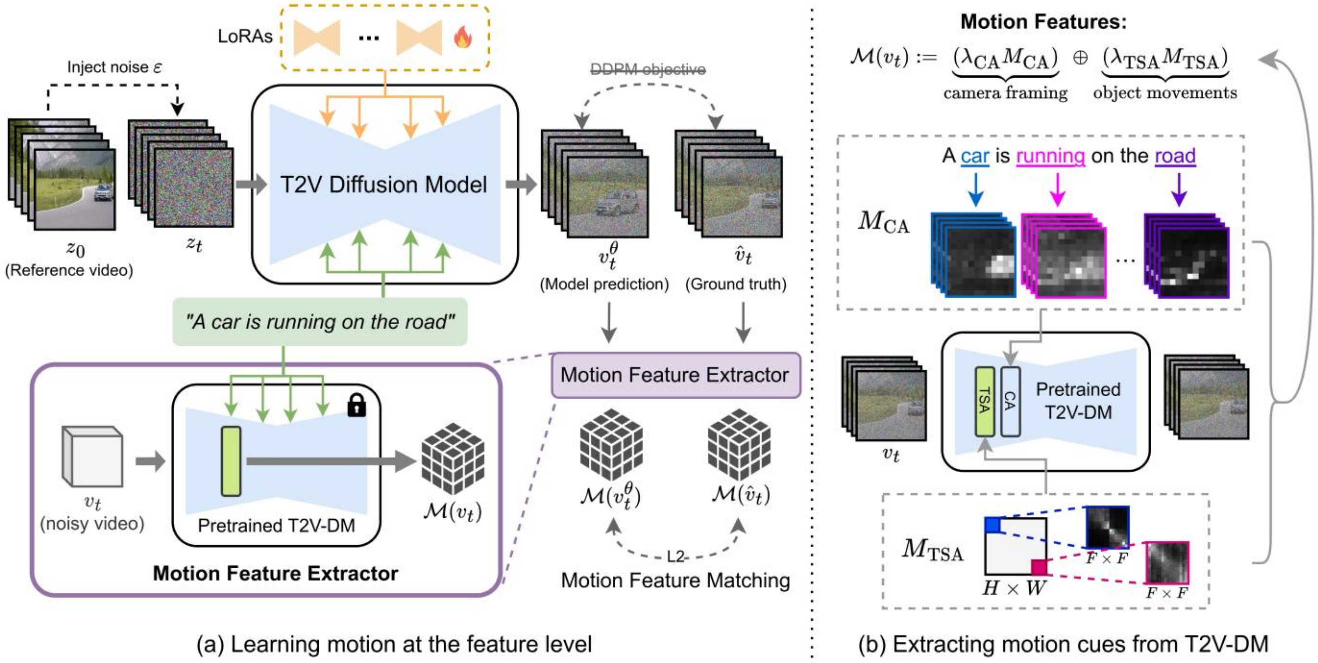
4. Method

Problem formulation To control scene staging in AIgenerated videos, we tackle the problem of motion customization, specifically as defined in DMT [71]. Given a reference video z_0 and a text prompt y associated with it, we aim to adjust a pre-trained T2V diffusion model ϵ_θ, so that the output videos sampled from the adjusted model replicate both the *object movements* and *camera framing* in z_0.

4.1. Preliminary: Text-to-video diffusion models

Text-to-video (T2V) diffusion models are probabilistic generative models that synthesize videos by gradually denoising a sequence of randomly sampled Gaussian noise frames (in latent space), guided by a textual condition y.

Architecture To model temporal information, T2V diffusion models typically inflate a pre-trained text-to-image (T2I) diffusion model by inserting temporal layers. These temporal layers are made up of feedforward networks and temporal self-



(a) Learning motion at the feature level

(b) Extracting motion cues from T2V-DM

Figure 2. Overview of MotionMatcher. (a) We fine-tune the pre-trained T2V diffusion model (T2V-DM) using the *motion feature matching* objective. Unlike the standard *pixel-level* DDPM loss, we align the motion features of the predicted noisy video v_t^θ with those of the ground truth noisy video $\hat{v}_t$. To extract motion features from *noisy latent videos*, we use a pre-trained T2V-DM (frozen) as a feature extractor. (b) We leverage the cross-attention (CA) maps and temporal self-attention (TSA) maps in the pre-trained T2V diffusion model to extract motion cues. The final motion features are the combination of the CA maps and TSA maps.

4.2. Learning motion at the feature level

Identifying motion in video requires a *high-level* understanding of both the spatial and temporal aspects of the video, so using the standard *pixel-level* DDPM reconstruction loss (Eq. (3)) for motion customization cannot accurately learn motion, and may introduce irrelevant information, such as content and visual appearance.

To this end, we introduce the *motion feature matching* objective, where a deep feature extractor M is used to extract motion information from videos at a high level.

Instead of directly aligning the predicted noisy video v_t^θ with the ground truth $\hat{v}_t$ at the pixel level, we align their high-level motion features (extracted by M):

where M is a motion feature extractor for *noisy latent videos*, and w'_t is the time-dependent weight in Eq. (3). As illustrated in Fig. 2(a), this *motion feature matching* objective aims to minimize the L2 discrepancy between the two videos in the motion feature space, ensuring that the motion in output video matches the motion in the reference video.

However, designing the motion feature extractor M in Eq. (4) is non-trivial, as it needs to extract features from *noisy latent videos*. First of all, most feature extractors, such as ViViT [1], EfficientNet [52], DenseNet-201 [22], and ResNet-50 [12], are trained on clean visual data, so we cannot directly applied them to noisy videos. Secondly, since the videos $\hat{v}_t$ and v_t^{θ} in Eq. (4) are in latent space, our feature extractor must be designed to process *latent videos* directly. Otherwise, we would need to decode them back into pixel-space videos before applying off-the-shelf feature extractors. This would incur substantial computational and memory overhead during training, due to both backpropagation through the large VAE decoder and the cost of processing “full-resolution” videos.

Here we claim that the pre-trained T2V diffusion model serve as a proper feature extractor for *noisy latent videos*. Firstly, recent work has shown both theoretically and experimentally that pre-trained diffusion models are capable of extracting high-level semantics and structural information from visual data, making them a “unified feature extractor” [64, 67]. Secondly, since diffusion models are trained on *noisy latent inputs*, using them as feature extractors for *noisy latent videos* helps prevent a training-inference gap. For these reasons, MotionMatcher leverages the pretrained T2V diffusion model as the motion feature extractor M.

4.3. Extracting motion cues from diffusion models

In this section, we identify the locations within the intermediate layers of diffusion models from which motion-specific features can be extracted.

Extracting cues for camera framing Recent studies have shown that the cross-attention (CA) maps in diffusion models closely reflect the spatial arrangement of objects within the frame [25, 33, 44, 66, 69]. Building on this, we leverage the CA maps from T2V diffusion models to describe the composition of each video frame (see Fig. 2(b)), thereby determining the camera framing throughout the video (*e.g.*, shot size and composition).

Formally speaking, CA maps are calculated by first reshaping the intermediate 3D activations $\Phi \in \mathrm{R}^{H\times W\times F\times D}$ into the shape $(H \times W \times F) \times D$, where F, H, W, and D denote the number of frames, height, width, and depth of the activations. Cross-attention is then performed between the activations Φ and word embeddings $\tau(y)$ as follows :

(5)

where τ denotes the text encoder used in the T2V diffusion model, and y is the text prompt given by the user. In $M_{\mathrm{CA}} \in [0,1]_{F\times H\times W\times|c|}$, each element $(M_{\mathrm{CA}})_{i,j,k,l}$ represents the correlation between the spatial-temporal coordinate (i,j,k) and the l’th word in the text prompt. As shown in Fig. 3, M_{CA} highlights the region within the frame that corresponds to an object. It focuses on structural information and eliminates visual appearance.

Extracting cues for object movements Since cross attention maps cannot describe motion that does not involve spatial shifts (*e.g.*, rotation and non-rigid motion), it is crucial to extract additional cues to represent such object movements. Since we discover that the temporal self-attention (TSA) maps in T2V diffusion models can capture detailed object movements, we also incorporate them into the motion features (see Fig. 2(b)).

To compute temporal self-attention (TSA) maps M_{TSA}, we begin by reshaping the model’s intermediate 3D activations $\Phi \in \mathrm{R}^{H\times W\times F\times D}$ into the shape $(H \times W) \times F \times D$.

For each particular spatial coordinate (i,j), we compute the self-attention weights between frames as follows:

where i and j denote the spatial coordinates. Specifically, each element $(M_{\mathrm{TSA}})_{i,j,k,l}$ of the TSA map $M_{\mathrm{TSA}} \in [0,1]^{H\times W\times F\times F}$ represents the degree of relevance between the k’th and l’th frames at the spatial coordinate

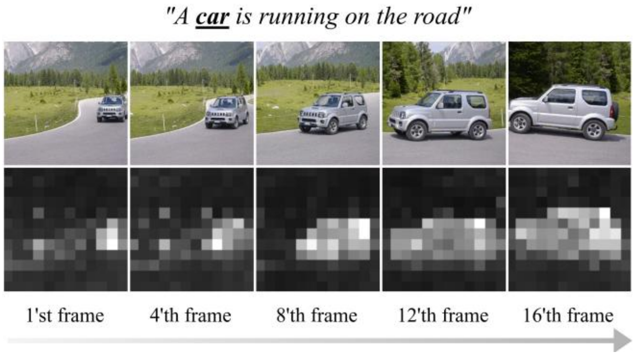
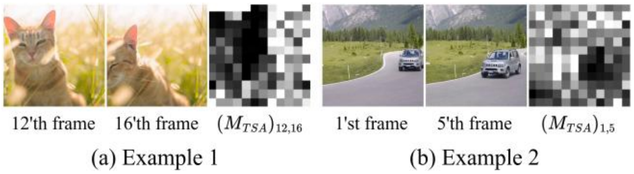


Figure 3. Example of cross-attention maps. We visualize the cross-attention map M_{CA}, computed between the activations in T2V diffusion models and the text prompt y. Here we obtain the CA map by adding noise to the video and using the pre-trained frame. diffusion model as a feature extractor. The extracted CA maps reveal the placement and shot sizes of the object associated with the word "car" in each video



(a) Example 1

(b) Example 2

Figure 4. Example of temporal self-attention maps. We visualize the temporal self-attention map M_{CA}, computed between two different frames. Here we obtain the TSA map by adding noise to the video and using the pre-trained diffusion model as a feature extractor. The extracted TSA maps describe the dynamics of the video in detail.

(i,j), capturing the dynamics of the video. As visualized in Fig. 4, the darker regions, which indicate low correlation between frames, correspond closely to areas where significant changes occur between the two frames. Therefore, by collecting the TSA maps for all $F \times F$ frame pairs, we can capture the inter-frame dynamics in detail.

With the cross-attention maps capturing camera framing, and the temporal self-attention maps reflecting object movements, we combine both to form the motion features:

(6)

where λ_{CA} and λ_{TSA} are weights that control the contributions of each component.

3.4. Motion-aware LoRA fine-tuning

After extracting the motion features, we fine-tune the pretrained T2V diffusion model using the *motion feature matching* objective in Eq. (4). By aligning the M_{CA} component, we ensure that the *camera framing* in the generated video matches that of the reference video, and align-

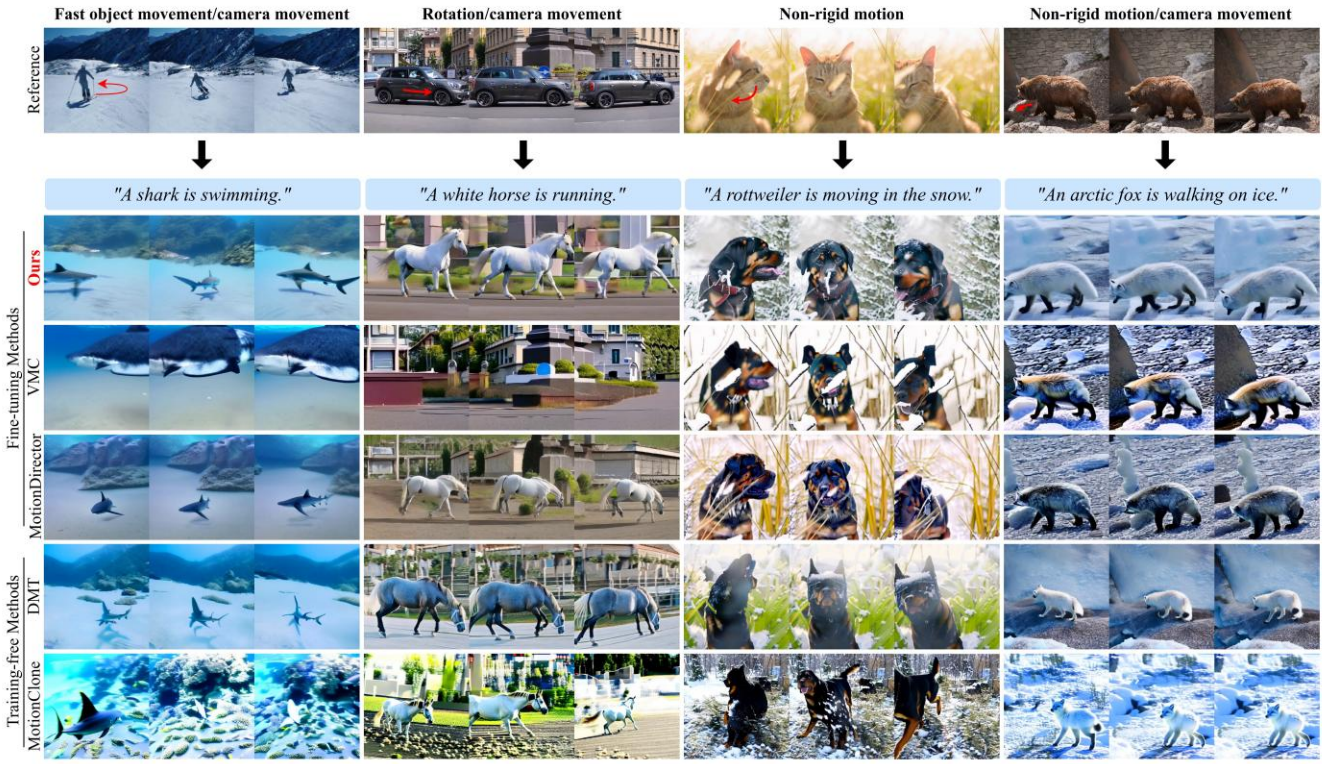

Fast object movement/camera movement
Rotation/camera movement
Non-rigid motion
Non-rigid motion/camera movement
Reference
"A shark is swimming."
"A white horse is running."
"A rottweiler is moving in the snow."
"An arctic fox is walking on ice."
Fine-tuning Methods
Ours
VMC
MotionDirector
Training-free Methods
DMT
MotionClone

Figure 5. Qualitative comparisons. Compared to existing methods such as VMC [26], MotionDirector [76], DMT [71], and MotionClone [31], our approach demonstrates superior text alignment and video quality, achieving high-fidelity motion transfer from reference videos to new scenes.

ing M_{TSA} ensures that the *dynamics* in the generated video align with those of the reference video.

To preserve the model's pre-trained knowledge while fine-tuning, we apply low-rank adaptations (LoRAs) [20] to fine-tune the model with fewer trainable parameters:

(7)

where $\Delta\theta$ is a low-rank parameter increment. Having these motion-aware LoRAs, MotionMatcher is capable of synthesizing videos that are guided by both the textual description and the motion in the user-provided reference video.

5. Experiments

5.1. Experiment setup

Dataset To evaluate MotionMatcher's ability to transfer motion from a reference video to a new scene, we collect a dataset of 42 video-text pairs. These videos encompass a wide range of motion types, such as fast object movement, rotation, non-rigid motion, and camera movement. We also ensure that the scenes in the editing text prompts are distinct from the scene in the reference video while remaining compatible with its motion.

Implementation details For a fair comparison, we use Zeroscope [50] as the base T2V diffusion model across all methods, given its ability to model complex motion and widespread usage in previous work [36, 71, 76]. We finetune the model with LoRA [20] for 400 steps at a learning rate of 0.0005. To extract motion features, we obtain attention maps M_{CA} and M_{TSA} from down block.2, with weights λ_{CA} and λ_{TSA} both set to 2000. These hyperparameters are chosen to balance control over camera framing and object movements. After extracting features from intermediate layers, we stop the forward pass to avoid unnecessary computation. For further implementation details, please refer to the supplementary material.

Baselines We compare our method against four recent approaches to motion customization, including two finetuning methods—VMC [26] and MotionDirector [76]— and two training-free methods—DMT [71] and MotionClone [31]. Detailed descriptions of these methods are provided in Sec. 2.3.

5.2. Evaluation metrics

We use four automatic metrics to evaluate the effectiveness of motion customization: (1) CLIP-T: To measure text alignment, we calculate the average CLIP [39] cosine similarity between the text prompt and all output frames. (2) Frame consistency: We compute the average CLIP cosine similarity between each pair of consecutive frames to assess frame consistency. (3) ImageReward: We calculate the average ImageReward [68] score for each frame, which evaluates both text alignment and image quality based on human preference. (4) Motion discrepancy: To quantify motion similarity between reference videos and generated videos, we leverage CoTracker3 [27], a state-of-theart point tracker that densely tracks the motion trajectories of 2D points throughout a video. Specifically, we use CoTracker3 to generate N 2D point trajectories for the reference video, denoted as $\hat{T}_0, \hat{T}_1, \cdots, \hat{T}_N \in R^{F\times2}$, and N 2D point trajectories for the generated video, denoted as $T_0, T_1, \cdots, T_N \in R^{F\times2}$. To measure the similarity between these two

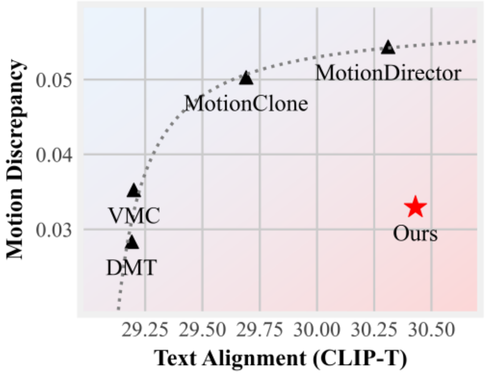


Figure 7. Illustration of the trade-off between text controllability and motion controllability. The quantitative comparison shows that our framework is preferable due to better text alignment and lower motion discrepancy.

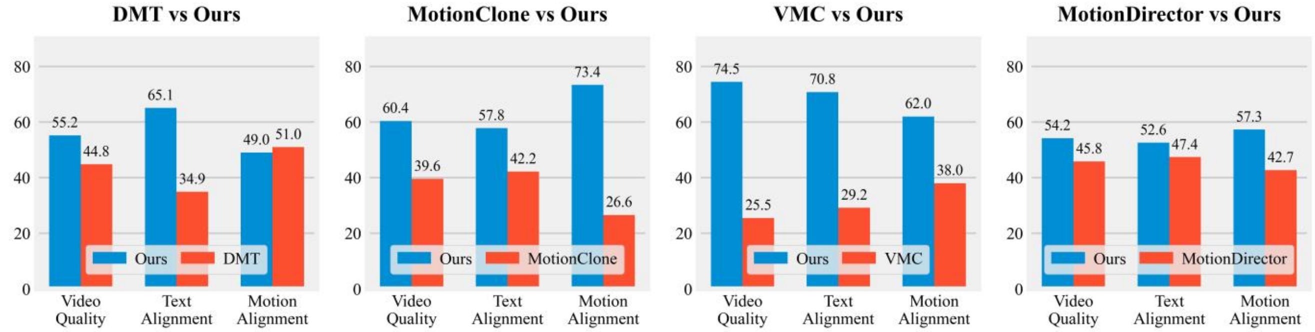


Figure 6. Human user study. The results show that human raters prefer our method over existing approaches in terms of video quality, text alignment, and motion alignment.

Methods	CLIP-T (↑)	ImageReward (↑)			Frame Consistency (↑)	Motion Discrepancy (↓)
DMT*		-0.0742	97.13	**0.0284**		
MotionClone*		-0.1133	96.91	0.0503		
VMC		-0.0162	97.19	0.0544		
	29.20			0.0330	96.89	0.0353
MotionDirector	30.31					
Ours	30.43	0.2301			97.20	

Table 1. Quantitative evaluation. Our method outperforms baseline approaches in text alignment, frame consistency, and overall human preference as measured by ImageReward [68]. Note that * denotes diffusion guidance-based methods.

sets of $F \times 2$ dimensional vectors, we use the Chamfer distance, a metric commonly used to assess the similarity between two sets of points in point cloud generation [9, 32, 53, 75]. Accordingly, the *motion discrepancy* score is defined as:

5.3. Main results

Quantitative results The quantitative results are reported in Tab. 1. Our method outperforms all baseline approaches in metrics such as CLIP-T, frame consistency, and ImageReward, demonstrating its superiority in preserving the prior knowledge in the base model during fine-tuning.

We also visualize the trade-off between text controllability and motion controllability in Fig. 7. As shown, our method provides significantly better joint controllability of both text and motion than existing motion customization approaches.

Qualitative results In Fig. 5, we present qualitative comparisons with baseline approaches across various types of motion. In the first example, only our method successfully reproduces the fast displacement in the reference video, confirming the effectiveness of our motion feature extractor in capturing complex motion. In the second example, VMC and MotionClone misposition the object within the frame, whereas MotionDirector and DMT fail to generate realistic videos complying with the text prompt. In contrast, our method faithfully follows the text prompt and places the object correctly. In the third and forth examples, our method also exhibits superior visual and motion quality.

These results conclude that our method preserves *the most* pre-trained knowledge during fine-tuning, while providing *the strongest* controllability for complex motion. For more results, please refer to Fig. 1 and the appendix.

6. Ablation study

We conduct an ablation study to examine the impact of incorporating M_{CA} and M_{TSA} in motion features. As illustrated in Fig. 8, without cross-attention maps M_{CA}, the model struggles to correctly position all the element of the scene. Meanwhile, removing temporal self-attention maps M_{TSA} reduces the precision of fine-grained dynamics. The quantitative results in Tab. 2 further validate the importance of both attention maps in controlling motion. These results confirm that both the *camera framing*, informed by M_{CA}, and *inter-frame dynamics*, informed by M_{TSA}, are essential for capturing overall motion.

6.1. Human user study

For a more accurate evaluation, we conduct a user study comparing our method with existing approaches based on human preferences. Following previous work [71, 76], we adopt the Two-alternative Forced Choice (2AFC) protocol. In the survey, the participants are presented with one video generated by our method and another video generated by a baseline approach. They are asked to compare the videos across three key aspects of motion customization: (1) Video quality: the degree to which the output video appears realistic and visually appealing, (2) Text alignment: how well the output video matches the text prompt, and (3) Motion alignment: the similarity in motion between the output video and the reference video. Ultimately, we collected 192 human evaluations per baseline and metric, totaling 2,304 human evaluations. These responses were gathered from 24 participants recruited via the Prolific platform.

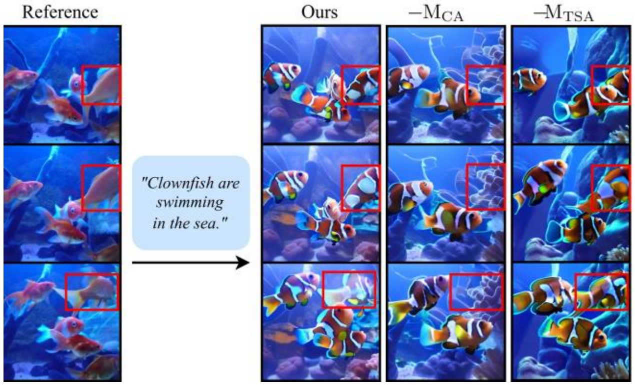


Figure 8. Qualitative results for ablation study. Without utilizing cross-attention maps M_{CA} in motion features, the model fails to capture all the fish in the video, whereas in the absence of temporal self-attention maps M_{TSA}, the model struggles to accurately replicate the fine-grained motion details. In contrast, our method successfully preserves both the scene composition and the interframe dynamics of the reference video.

	CLIP-T (↑)	ImageReward (↑)	Motion Discrep. (↓)
−CA	30.08	0.1252	0.0360
−TSA	30.67	0.4650	0.0693
Ours	30.43	0.2301	0.0330

Table 2. Ablation study. Our method, which utilizes both M_{CA} and M_{TSA}, achieves the lowest motion discrepancy score.

As shown in Fig. 6, human users prefer our method over existing approaches in all aspects. These results further confirm the superiority of our method.

7. Conclusion

We presented MotionMatcher, a feature-level fine-tuning framework for motion customization. MotionMatcher transforms the *pixel-level* DDPM objective into the *motion feature matching* objective, aiming to learn the target motion at the *feature level*. To extract motion features, MotionMatcher leverages the pre-trained T2V diffusion model as a deep feature extractor and identify valuable motion cues from two attention mechanisms within the model, representing both object movements and camera framing in videos. In the experiments, MotionMatcher demonstrated superior joint controllability of text and motion to prior approaches. These results suggest that MotionMatcher enhances control over scene staging in AI-generated videos, benefiting real-world applications in computer-generated imagery (CGI). For a discussion of MotionMatcher's limitations, please refer to the supplementary material.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Luciˇ c, and Cordelia Schmid.´ Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846,

2021. 4

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

[3] Michael J Black and Padmanabhan Anandan. A framework for the robust estimation of optical flow. In *1993 (4th) International Conference on Computer Vision*, pages 231–236. IEEE, 1993. 2

[4] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 2, 3

[5] Andres Bruhn, Joachim Weickert, and Christoph Schn´ orr.¨ Lucas/kanade meets horn/schunck: Combining local and global optic flow methods. *International journal of computer vision*, 61:211–231, 2005. 2

[6] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023. 2

[7] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015. 2

[8] Dave Epstein, Allan Jabri, Ben Poole, Alexei Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. *Advances in Neural Information Processing Systems*, 36:16222–16239, 2023. 3

[9] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 605–613, 2017. 7

[10] Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic vqgan and timesensitive.

-Dr.Jayavadivel pavi
Assistant Professor
School of Computer Science and Enginnering
Presidency University