Video Diffusion Models are Training-free Motion Interpreter and Controller

Zeqi Xiao[1], **Yifan Zhou**[1], **Shuai Yang**[2], **Xingang Pan**[1]
[1]S-Lab, Nanyang Technological University,
[2]Wangxuan Institute of Computer Technology, Peking University
{zeqi001, yifan006}@e.ntu.edu.sg
williamyang@pku.edu.cn, xingang.pan@ntu.edu.sg

Abstract

Video generation primarily aims to model authentic and customized motion across frames, making understanding and controlling the motion a crucial topic. Most diffusion-based studies on video motion focus on motion customization with training-based paradigms, which, however, demands substantial training resources and necessitates retraining for diverse models. Crucially, these approaches do not explore how video diffusion models encode cross-frame motion information in their features, lacking interpretability and transparency in their effectiveness. To answer this question, this paper introduces a novel perspective to *understand*, *localize*, and *manipulate* motion-aware features in video diffusion models. Through analysis using Principal Component Analysis (PCA), our work discloses that robust motion-aware feature already exists in video diffusion models. We present a new MOtion FeaTure (MOFT) by eliminating content correlation information and filtering motion channels. MOFT provides a distinct set of benefits, including the ability to encode comprehensive motion information with clear interpretability, extraction without the need for training, and generalizability across diverse architectures. Leveraging MOFT, we propose a novel training-free video motion control framework. Our method demonstrates competitive performance in generating natural and faithful motion, providing architecture-agnostic insights and applicability in a variety of downstream tasks.

1 Introduction

Video generation has experienced notable advancements in recent years, particularly in the realm of video diffusion models, such as text-to-video (T2V) generation [15; 5; 45; 43] and image-to-video (I2V) generation [2; 14; 6]. Apart from producing high-quality content in individual frames, capturing authentic and customized motion across frames is a crucial feature of video generation. Thus, understanding and controlling the motion play pivotal roles in video generation.

Most methods [46; 50; 42; 15; 44] that study video motion focus on motion customization, i.e. allowing users to specify a moving direction [46] or a point-drag command [50]. These methods typically adopt training-based paradigms [46; 50; 42; 15] that introduce motion conditions and train additional modules to ensure that the output videos adhere to these conditions. Despite their progress, these approaches require significant training resources and need retraining for different models, and their effectiveness often remains black-box. More critically, they do not address a fundamental question: *How do video diffusion models encode cross-frame motion information within their features?*

Project page at this URL.

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

(a) Reference (b) Similarity heatmap of MOFT on various videos.

(c) MOFT as guidance for motion control

Figure 1: **Characteristics of MOtion FeaTure (MOFT).** (a-b) Rich Motion Information: We extract MOFT at the red point in the reference video in (a) and draw similarity heatmaps in (b) across various videos (yellow indicates higher similarity). The heatmap aligns well with the motion flow in the bottom left. (c) MOFT serves as guidance for controlling motion direction in the light-masked region, with the motion direction signal illustrated by red arrows in the first image.

Understanding the encoding of motion information is crucial for two reasons: a) it offers architecture-agnostic insights, meaning that such knowledge can be applied across different models and their checkpoints, an important consideration given the rapid evolution of video diffusion models; and b) it supports various downstream applications. For instance, the DIffusion FeaTure [41] demonstrates how diffusion features can encapsulate rich semantic information, enabling applications like correspondence extraction [17; 23] and image/video editing [9; 30; 8].

To this end, this paper introduces a novel perspective to *understand*, *localize*, and *manipulate* motion-aware features in video diffusion models. We first establish that removing content correlation information helps to pronounce motion information in video diffusion features. By applying Principal Component Analysis (PCA) [47] on these diffusion features, we observe a strong correlation between the principal components and video motions. Further explorations reveal that certain channels of the features play a more significant role in determining motion direction than others. Based on these observations, we present a straightforward strategy to extract motion information embedded in the features, termed MOtion FeaTure (MOFT). Through content correlation removal and motion channel filter, MOFT establishes impressive correspondence on videos with the same motion direction, as illustrated in Fig. 1 (a-b). Importantly, this strategy proves to be generalizable across various text-to-video or image-to-video generation models [15; 14; 43; 4; 2] (Fig. 4), such as AnimatedDiff [15], ModelScope [43], and Stable Video Diffusion [2].

Building upon the motion-aware MOFT, we propose a pipeline for video motion control in a training-free manner, without the modification of model parameters. The approach leverages compositional loss functions for content manipulation [9; 11; 31; 1; 19]. Specifically, we design loss functions to optimize noisy latents in the denoising process with reference MOFT, which can be synthesized via direction signal or extracted from reference videos. Furthermore, our pipeline can be extended for point-drag manipulation. With MOFT guidance to generate coarse motion in the early denoising stages, fine-grained point-drag manipulation with DIFT [41] guidance becomes feasible for videos. Various experiments showcase the effectiveness of MOFT in controlling the motions of diverse scenarios across different video diffusion models without the need for any training. Remarkably, our training-free method even outperforms some data-driven methods in achieving natural and faithful motion. Our main contributions are summarized as follows:

- We perform a deep analysis of motion information embedded in video generation models. Our work discloses that robust motion-aware feature already exists in video diffusion models.
- Through our analysis, we present MOtion FeaTure (MOFT) that effectively captures motion information. MOFT has several advantages: a) it encodes rich motion information with high

interpretability; b) it can be extracted in a training-free way; and c) it is generalizable to various architectures.

- We propose a novel training-free video motion control framework based on MOFT. Our method demonstrates competitive performance with natural and faithful motion. Unlike previous training-based methods that need independent training for each different architecture and checkpoint, our method is readily applicable to different architectures and checkpoints.

2 Related Works

Video Diffusion Models. The field of video generation has witnessed substantial progress in recent years, particularly in the domain of video diffusion models. Noteworthy contributions include advancements in text-to-video (T2V) generation [20; 15; 5; 45; 43; 32] which aim to generate high-fidelity videos that align with textual descriptions. Besides, image-to-video (I2V) [2; 14; 6] takes image conditions as inputs and generates videos aligned with the image. Beyond the production of high-quality content within individual frames, the capability to capture authentic and customized motion across frames stands out as a significant feature in the realm of video generation.

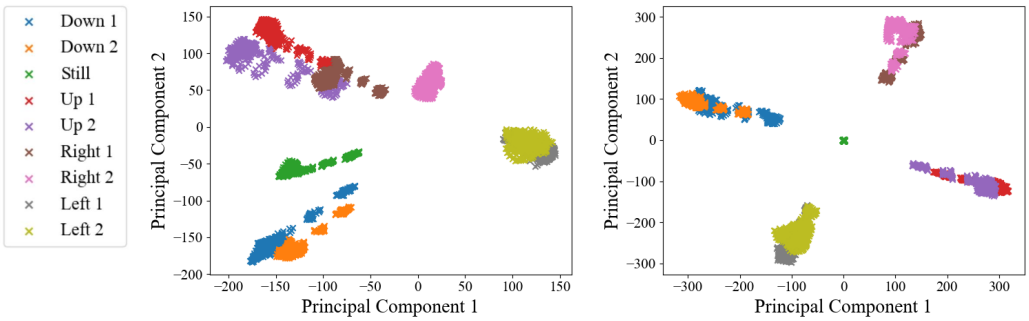
Diffusion Feature Understanding. The analysis and comprehension of diffusion features [41; 29; 10; 27; 37] have garnered increasing attention. A comprehensive understanding of diffusion features not only yields architecture-agnostic insights applicable across diverse models and checkpoints but also enhances various downstream applications. For instance, DIffusion FeaTure (DIFT) [41] demonstrates that diffusion features embed impressive semantic correspondence and can be extracted with a simple strategy. This strategy proves effective across various architectures, spanning image diffusion models [34] to video diffusion models [15; 43]. Its versatility facilitates a range of applications, including correspondence extraction [17; 23] and image/video editing [9; 30; 8]. Recently, Freecontrol [29] applied PCA on diffusion features and extracted semantic basics for training-free spatial control. Its method can be generalized to any conditional input and any model. Similarly, video diffusion models encode rich motion information within the features. However, less effort has been made to analyze it.

Video Motion Control. Considerable efforts have been dedicated to tailoring video motion according to user preferences [46; 44; 14; 50; 42; 16; 7; 13; 48]. For example, MotionCtrl [46] facilitates precise control over camera poses and object motion, allowing for fine-grained motion manipulation. VideoComposer [44] introduces motion control through the incorporation of additional motion vectors, while DragNUWA [50] proposes a method for video generation that relies on an initial image, provided point trajectories, and text prompts. These methodologies typically rely on training-based paradigms, incorporating motion conditions during training and requiring additional modules to ensure that the resulting videos adhere to these specified conditions. Despite their advancements, these approaches demand substantial training resources, necessitating retraining for different models, and often exhibit a black-box nature in terms of their effectiveness. In contrast, this paper introduces a novel pipeline for controlling video motion using an interpretable motion-aware feature. Notably, this approach is training-free and can be generalized across various architectural frameworks, offering a more versatile and resource-efficient solution.

3 MOtion FeaTures (MOFT)

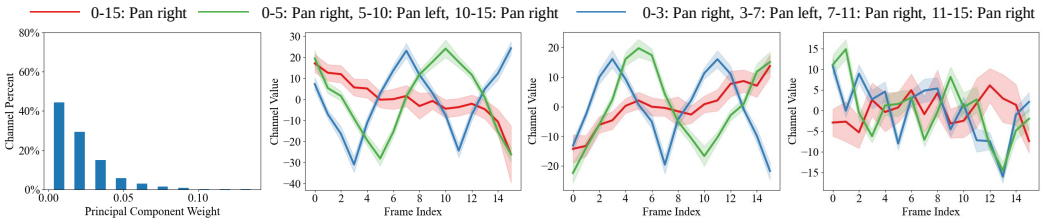
In this section, we first analyze how video diffusion models encode cross-frame motion information, then provide the strategy to extract motion features from pre-trained video diffusion models.

Similar to [41; 49; 29], our analysis focuses on diffusion features extracted from the intermediate blocks of diffusion models. We denote them as $\mathcal{X} \in \mathbb{R}^{H \times W \times F \times D}$, where H, W, F and D are dimensions of height, width, frames, and channels, respectively. As proved by prior works [15; 51; 46], cross-frame features play a crucial role in video motion control. For example, AnimateDiff [15] trains temporal self-attention LoRAs [21] that operate on the temporal dimension to control the global motion direction. Consequently, we argue that the temporal dimension encapsulates rich motion information. However, extracting motion information from diffusion features is non-trivial, as they also contain other information such as semantic and structural correlation.



(a) PCA of vanilla feature (b) PCA of feature after correlation removal

Figure 2: **Visualization of PCA on video diffusion features.** The left side indicates the frame-wise panning direction, with each color representing a specific direction pattern. We apply PCA to diffusion features extracted from videos with different motion directions and plot their projections on the leading two principle components. (a) The result does not exhibit a distinguishable correlation with motion direction. (b) Features are clearly separated by their motion direction.



(a) PC weight hist. on channels (b) Value of a motion channel (c) Value of a motion channel (d) Value of non-motion channel

Figure 3: **Cross-frame Channel Value.** (a) We plot the histogram of the weight of $\mathcal{P}_1$. It reveals that only a few channels significantly contribute to determining the principal components. (b-c) The motion channels exhibit a pronounced correlation with motion direction trends. (d) In contrast, the non-motion channels show little correspondence with motion direction.

3.1 Content Correlation Removal

Inspired by VideoFusion [28] which uses shared noise to model content correlation across frames and residual noise to model dynamic difference, we hypothesize that we can filter out the content correlation by eliminating similar information across frames:

$$\mathcal{X}^{\text{norm}} = \mathcal{X} - \frac{1}{F}\sum_{i=1}^{F}\mathcal{X}_i, \tag{1}$$

where $\mathcal{X}_i$ indicates the i^{th} frame of feature $\mathcal{X}$. The shared latents, to which we refer as content correlation information, encompass shared aspects such as semantic content and appearance. In contrast, the residual latents primarily capture motion information, which also can be interpreted as deformation in structure.

To validate the hypothesis, we apply Principal Component Analysis (PCA) [47] on $\mathcal{X}$ and $\mathcal{X}^{\text{norm}}$. Specifically, we create a series of videos with the entire scene moving horizontally or vertically, resulting in a set of features $\{\mathcal{X}^1, \mathcal{X}^2, ..., \mathcal{X}^n\}$ extracted from videos in the process of DDIM [39] inversion. In this subsection, we omit the choice of video model architecture and feature selection for simplicity. We analyze and project the D-dimensional features of the first frame on the leading two principal components ($\mathcal{P}_1$ and $\mathcal{P}_2$). As shown in Fig. 2 (a), the result of the vanilla feature does not exhibit a distinguishable correlation with motion direction. In contrast, as shown in Fig. 2 (b), normalized features are successfully separated by their motion direction. It reveals that the normalization operation removes content correlation information and emphasizes motion information.

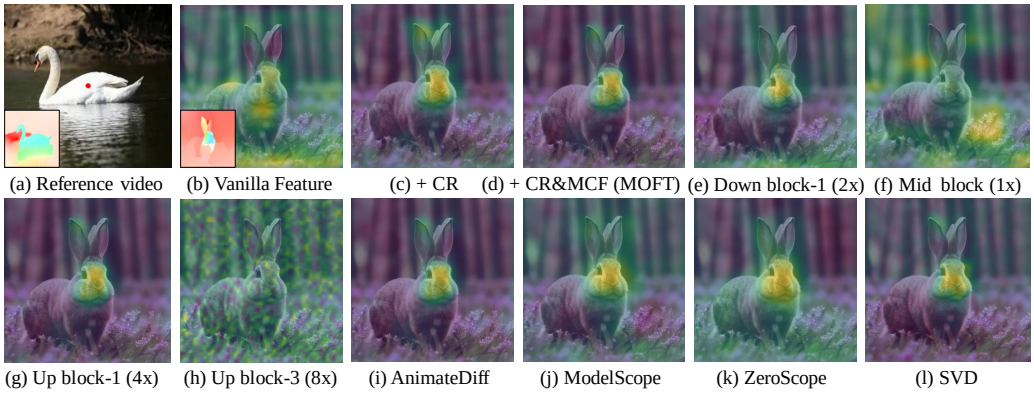


Figure 4: **Similarity heatmap between feature of the source point and target features**. Given the red source point in (a), we plot the similarity heatmap on target videos. Yellow indicates regions with higher similarity. We normalize all similarity to 0-1 for better illustration. (b-d) Similarity heatmap of features with different designs. "CR" indicates "content removal". "MCF" indicates motion channel filter. (e-h) Similarity heatmap of MOFT in different layers in the U-Net. (2x) means relative spatial resolution scale 2. (i-l) Similarity heatmap of MOFT in different video generation models.

3.2 Motion Channel Filter

Principal components can not only reduce dimension but also reflect the importance of each dimension by the projection weights. We visualize the projection weights of $\mathcal{P}_1 \in \mathbb{R}^{D\times 1}$ in Fig 3 (a). It reveals that only a few channels significantly contribute to determining the principal components, indicating these channels encode richer motion information. We term them Motion Channels.

To further explore the relationship between these channels and the motion in videos, we create videos panning in different directions at various frames and visualize the channel with the highest two projection weights in $\mathcal{P}_1$. As depicted in Fig. 3 (b-c), the value trend is closely associated with the panning direction of the video. Specifically, in Fig. 3 (b), the motion channel value decreases during a rightward pan and increases during a leftward pan. In contrast, a channel with low projection weight does not exhibit much correspondence (Fig. 3 (d)). These observations indicate that we can extract motion-aware features by filtering these motion channels.

3.3 MOFT Extraction

With the above explorations, we introduce a straightforward strategy for extracting motion information from video diffusion models, which we term Motion Feature (MOFT). Our method includes two designs: content correlation removal and motion channel filter. The process can be represented as follows

$$\mathcal{M} = (\mathcal{X}_{[j]} - \frac{1}{F}\sum_{i=1}^{F}\mathcal{X}_{i,[j]}), \quad j \in \mathcal{C}, \tag{2}$$

where $\mathcal{M}$ is the extracted MOFT, i operates on the frame dimension, and j operates on channel dimension. $\mathcal{C}$ is the channel index set of motion channels.

We illustrate how content correlation removal and motion channel filter improve the motion correspondence in Fig. 4 (a-d). Vanilla video features demonstrate weak alignment with the reference motion. The proposed content correlation removal significantly improves the alignment. Further application of the motion channel filter enhances focus on the motion area (*e.g.*, the rabbit head), yielding higher correspondence.

We conduct an additional ablation study and visualize the impact of selecting different video diffusion features from various blocks within the U-Net of AnimateDiff [15]. Fig. 4 (e-h) intuitively reveals that features with relative medium resolutions achieve better motion correspondence. To this end, we select the features after upper block 1.

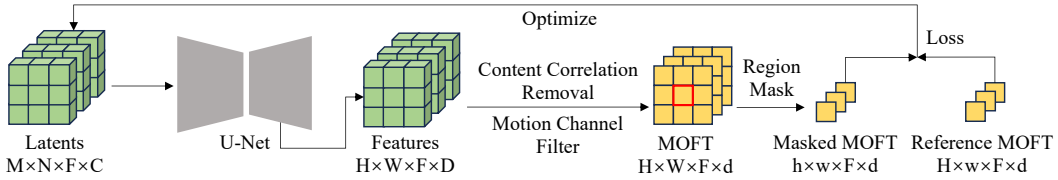


Figure 5: **Motion Control Pipeline**. We use reference MOFT as guidance and optimize latents to alter the sampling process. In one denoising step, we get the intermediate features and extract MOFT from it with content correlation removal and motion channel filter. We optimize the latents to alter the sampling process with the loss of masked MOFT and reference MOFT.

While the above analysis is based on AnimateDiff [15], the property of MOFT holds in different base video models [15; 43; 4; 2] (Fig. 4 (i-l)), demonstrating that MOFT is versatile across different video generation frameworks, consistently achieving reliable motion alignment.

While MOFT is reminiscent of optical flow, which also describes motion, a key limitation of optical flow is that it cannot be directly extracted from video diffusion models during the denoising process and hence cannot serve as the guidance for motion control. In contrast, MOFT is available even at early denoising steps and is naturally suitable for motion control, as we will discuss in the next section.

4 MOFT Guidance

With the motion-aware MOFT, we propose a pipeline for video motion control in a training-free manner (Sec. 4.1). Furthermore, our pipeline can be extended for point-drag manipulation (Sec. 4.2).

4.1 Motion Control

We design a pipeline to control motion in the generated videos in a training-free way, as depicted in Fig. 5. Following [36; 49], we optimize latents to alter the sampling process. The loss function $\mathcal{L}^c$ is

$$\mathcal{L}^c = \frac{1}{|\mathcal{R}|} \sum_{(i,j)\in\mathcal{R}} ||\mathcal{M}_{i,j} - \mathcal{M}^r_{i,j}||, \tag{3}$$

where $\mathcal{M}$ is the MOFT we extract during the denoising phase, $\mathcal{M}^r$ is the reference MOFT feature, and $\mathcal{R}$ is the position set of the region that we want to control motion. We provide two possible ways to construct the reference MOFT $\mathcal{M}^r$: 1) Extract MOFT from reference videos. We perform DDIM inversion [39] on reference videos and extract MOFT in the inversion stage. 2) Synthesize MOFT based on the statistic regularity. As shown in Fig. 3 (b-c), frame-wise motion channel values exhibit high correspondence with frame-wise motion. We can fit it into a piecewise linear function, where each piece function ranges from statistic minimum to statistic maximum. In this way, we can flexibly modulate frame-wise reference motion as guidance. The detailed process is shown in Alg. 1

Algorithm 1: Optimization Process

Input: Noisy latents z at timestep t, region mask $\mathcal{R}$, reference MOFT $\mathcal{M}^r$, the network $\mathcal{N}$, Motion Channel Mask $\mathcal{C}$, learning rate η

Output: Optimized latents $\hat{z}$

begin
- Get intermediate feature $\mathcal{X}$ from the network $\mathcal{N}$;
- Given $\mathcal{X}$, $\mathcal{C}$, extract MOFT $\hat{\mathcal{M}}$ by Eq. 2;
- Given $\mathcal{M}^r$, $\mathcal{M}$, and $\mathcal{R}$, compute the loss $\mathcal{L}$ by Eq. 3;
- Optimize $\hat{z}$ by updating $\hat{z} \leftarrow z - \eta\nabla\mathcal{L}$;
- **return** $\hat{z}$;

end

4.2 Point-Drag Manipulation

Point-drag manipulation is designed to precisely relocate points within image and video frames to reach specific target points. In the image domain, this manipulation method often relies on motion

(a) Reference (b) Step 800 (c) Step 400 (d) Step 0 (e) Reference (f) Step 800 (g) Step 400 (h) Step 0

Figure 6: **Effects of DIFT and MOFT on different denoising time steps.** Given the source point in (a) (for DIFT) and (e) (for MOFT), we plot the similarity heat map of DIFT (b-d) and MOFT (f-h) of different denoising steps. Yellow indicates higher similarity. The red point in (b-d) indicates the position with highest similarity. It suggests that MOFT can provide more valid information than DIFT at the early denoising stages.

supervision and point-tracking [33; 36], ensuring the precise tracking of point trajectories to achieve the desired target points. In the video domain, however, we can directly optimize whole point trajectories by setting targets in each frame. The loss function for optimizing point trajectories $\tau = p_1, p_2, ...p_F$ is:

$$\mathcal{L}^p = \sum_{i=2}^{F} ||\mathcal{D}(p_i) - \text{sg}(\mathcal{D}(p_1))||, \tag{4}$$

where $\mathcal{D}$ is the diffusion feature (DIFT) and sg is the "stop gradient" operation.

However, direct application of this method results in poor video motion control because DIFT struggles with semantic correspondence at early denoising steps, as shown in Fig. 6 Row 1. Since spatial and temporal structures are already determined at early steps, DIFT's effectiveness is limited. Conversely, MOFT provides relatively distinguished motion information in early denoising stage performance (Fig. 6 Row 2), suggesting a strategy of using MOFT for initial coarse motion control and DIFT for precise point-drag manipulation. Please refer to Supplementary Material for details.

5 Experiments

5.1 Implementation details

If not specified, the default video generation models of the following experiments are implemented in AnimateDiff [15] (T2V) and SparseCtrl [14] (I2V). For T2V generation, we first generate a normal video as the editing source, then apply motion direction and region mask to the video for motion control. To preserve consistency with the source video, we apply (1) region gradient clip, and (2) shared key and value. Details of these techniques and video results can be found in the Supplementary Material. Our results are at a resolution of 512x512 and 16 frames unless otherwise specified. We use DDIM with 25 denoising steps for each sample. It takes approximately 3 minutes to generate one sample on an RTX 3090 GPU.

5.2 Qualitative Results

We showcase qualitative outcomes in Fig. 7. The figure illustrates the successful animation of videos by our method, guided by diverse control signals while preserving a natural and authentic sense of motion. Additionally, we exhibit the results of applying our motion control technique to alternative video generation models, such as ModelScope [43] and ZeroScope [4], employing the same control strategy (see Fig. 8). These results highlight the generalizability of MOFT across various video generation models. We also showcase the application of our method on point-drag manipulation (Sec. 4.2) in Fig. 9, where we successfully move the starting points to the targets.

5.3 Motion Feature Design

This subsection experiments on the effectiveness of motion feature designs with two metrics: *a) Motion Fidelity.* Following [49], we use Motion Fidelity to assess the fidelity of our results in the alignment of synthesis guidance or reference guidance. We use off-the-shelf tracking method [24] to estimate the tracklets $\mathcal{T} = \{\tau_1, ..., \tau_M\}$ in the generated videos. For guidance, we manually construct synthesized tracklets for synthesis guidance and use estimated tracklets for reference guidance, we

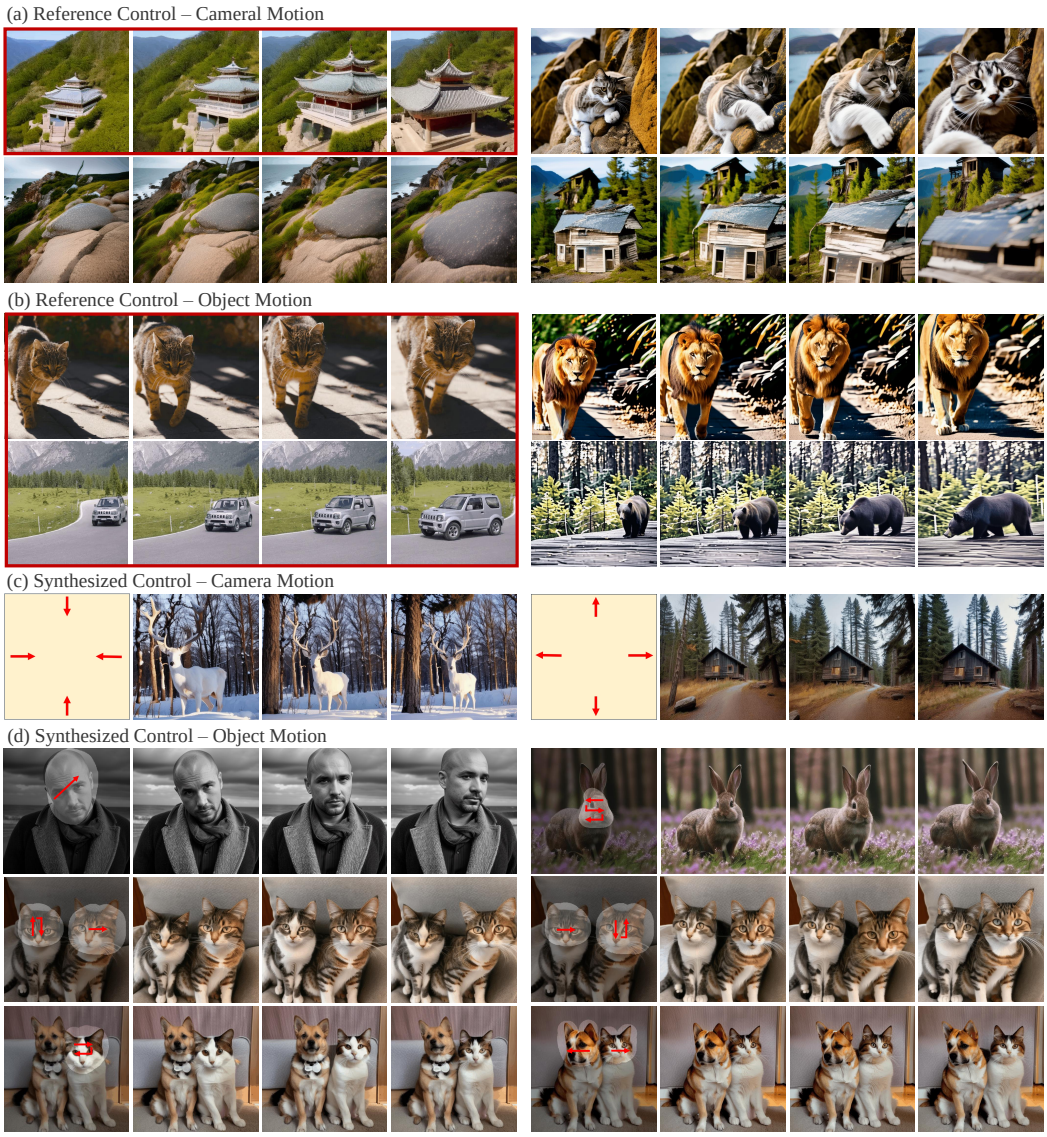


Figure 7: **Qualitative results.** We illustrate several animation clips with different reference or synthesized motion control signals. The red boxes in (a-b) stand for reference videos. *We highly recommend readers refer to the supplementary material for a better visual experience.*

denote them both as $\widetilde{\mathcal{T}} = \{\widetilde{\tau}_1, ..., \widetilde{\tau}_N\}$ for simplicity. The motion fidelity score is defined as follows:

$$\frac{1}{m}\sum_{\widetilde{\tau}\in\widetilde{\mathcal{T}}}\max_{\tau\in\mathcal{T}}\mathbf{corr}(\tau,\widetilde{\tau}) + \frac{1}{n}\sum_{\tau\in\mathcal{T}}\max_{\widetilde{\tau}\in\widetilde{\mathcal{T}}}\mathbf{corr}(\tau,\widetilde{\tau}). \tag{5}$$

The correlation between two tracklets $\mathbf{corr}(\tau, \widetilde{\tau})$ is computed as:

$$\mathbf{corr}(\tau,\widetilde{\tau}) = \frac{1}{F}\sum_{k=1}^{F}\frac{v_k^x\cdot\widetilde{v}_k^x + v_k^y\cdot\widetilde{v}_k^y}{\sqrt{(v_k^x)^2+(v_k^y)^2}\cdot\sqrt{(\widetilde{v}_k^x)^2+(\widetilde{v}_k^y)^2}}, \tag{6}$$

where (v_k^x, v_k^y), $(\widetilde{v}_k^x, \widetilde{v}_k^y)$ are the k^{th} frame displacement of tracklets τ, $\widetilde{\tau}$, respectively. *b) Image Quality.* We follow [22; 25] that uses an image quality predictor trained on the SPAQ dataset [12] to evaluate frame-wise quality regarding distortion like noise, blur, or over-exposure. We collect a total of 270 prompt-motion direction pairs for experiments.

Table 1 summarizes our results. The vanilla feature shows poor motion fidelity and image quality due to extraneous information disrupting motion control. Removing content correlation significantly

Figure 8: **Qualitative results on Modelscope [43] and ZeroScope [4].**

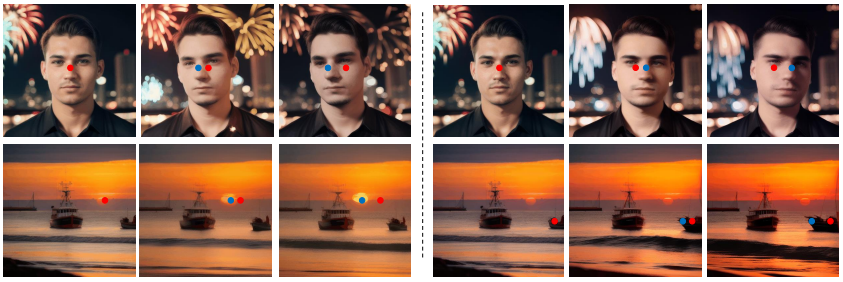


Figure 9: **Qualitative results of point-drag manipulation.** Red points indicate starting points. Blue points indicate target points of the corresponding frames. We display three frames per video clip.

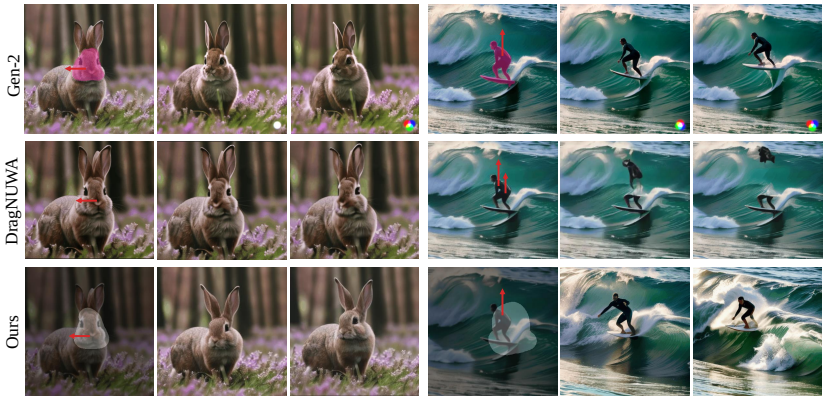


Figure 10: **Motion quality comparisons.** Gen-2 [35] and Ours accept editing region and motion direction as the control signal. DragNUWA [50] accepts point trajectories as the control signal.

improves both metrics, yielding results comparable to the Space-Motion Map (SMM) feature [49], likely because SMM also removes content correlation through frame-wise differences. MOFT guidance achieves the highest motion fidelity, with only a minor loss in image quality compared to the original unguided generation.

5.4 Point-drag Manipulation

We conducted additional experiments to assess the efficacy of incorporating motion control in point-based manipulation. In this comparison, we introduce DragNUWA [50], a potent data-driven method, for reference. We follow [33; 36] to use the *Mean Distance* between edited points and target points to evaluate the drag precision. Specifically, we still use [24] to estimate the tracklets $\mathcal{T} = \{\tau_1, ..., \tau_{,=M}\}$ of given small region. We average these tracklets into $\overline{\tau}$ and calculate the mean distance with target tracklet τ^t. We normalize the final distance into [0,1], with 0 indicating no mean distance error. We collect a total of 40 image-motion direction pairs for experiments. As indicated in Table 2, applying only DIFT guidance results in poor drag precision. By comparison, incorporating our MOFT yields substantial improvements, effectively narrowing the performance gap with the training-based DragNUWA. The finding is coherent with our analysis in Sec. 4.2.

Table 1: Experiments on Motion Feature Design

Guide type	Methods	Motion Fidelity (↑)	Imaging Quality (↑)
None	Origin	-	**0.697**
Reference Guidance	SMM feature [49]	70.2	0.681
	Vanilla Feature	31.1	0.512
	+ CR	67.1	0.671
	+ CR & MCF (Ours)	<u>82.5</u>	<u>0.693</u>
Synthesis Guidance	Ours	**84.0**	<u>0.693</u>

Table 2: Drag Precision

Name	Mean Distance (↓)
DragNUWA [50]	**0.075**
DIFT [41]	0.437
+ MOFT (Ours)	<u>0.175</u>

Table 3: User Preference

Methods	Faithfulness (↑)	Naturalness (↑)
DragNUWA [50]	2.50	2.08
Gen-2 MB [35]	**3.37**	<u>2.90</u>
Ours	<u>3.21</u>	**3.49**

5.5 User study

We conducted a survey to investigate users' preferences regarding videos generated with motion control. Employing a blind rating protocol, participants were randomly exposed to videos generated by Gen-2 Motion Brush [35], DragNUWA [50], and our proposed method. Participants were instructed to rate from 1 to 5 (worst to best) on two metrics: *1) Motion Faithfulness* to measure how well the motion aligns with the control signal. *2) Motion Naturalness* to evaluate the naturalness and realism of the motion. We collect human feedback from 26 people on 56 video clips. As depicted in Table 3 and Fig. 10, it is evident that Gen-2 MB excels in achieving highly faithful motion control at the cost of motion naturalness. Gen-2 MB and DragNUWA tend to generate stiff and unrealistic motions. In contrast, our proposed methods demonstrate competitive motion faithfulness while simultaneously preserving the natural and authentic quality of motion.

6 Limitations add Future Works

While our approach has yielded appealing results, some limitations require future studies:

1) Presently, our approach lacks support for motion control in real videos. Primarily, this limitation stems from the lack of research on video inversion techniques over video diffusion models. We have observed significant alterations in content when employing initial noise from DDIM inversion [39] on real videos. Future research focused on video inversion holds promise for resolving this issue.

2) Our current approach does not allow for precise motion scale guidance in motion control. While there are strategies to roughly control motion scales, such as adding up control weights for larger motion scales or implementing gradient clips for smaller ones, achieving high precision in motion scale manipulation requires further investigation.

7 Conclusion

In summary, our analysis reveals a robust motion-aware feature in video diffusion models, leading to the development of a training-free MOtion FeaTure (MOFT). MOFT encodes rich, interpretable motion information, is extracted without training, and is applicable across diverse architectures. We introduce a novel training-free video motion control framework based on MOFT, demonstrating competitive performance with natural motion. Importantly, our approach is versatile, easily adaptable to various architectures and checkpoints without the need for independent training.

Acknowledgements. This research is supported by MOE AcRF Tier 1 (RG97/23) and is also supported under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

References

[1] Bansal, A., Chu, H.M., Schwarzschild, A., Sengupta, S., Goldblum, M., Geiping, J., Goldstein, T.: Universal guidance for diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 843–852 (2023)

[2] Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)

[3] Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., Zheng, Y.: Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22560–22570 (2023)

[4] Cerspense: Zeroscope. `https://huggingface.co/cerspense/zeroscope_v2_576w` (2023)

[5] Chen, H., Xia, M., He, Y., Zhang, Y., Cun, X., Yang, S., Xing, J., Liu, Y., Chen, Q., Wang, X., et al.: Videocrafter1: Open diffusion models for high-quality video generation. arXiv preprint arXiv:2310.19512 (2023)

[6] Chen, H., Zhang, Y., Cun, X., Xia, M., Wang, X., Weng, C., Shan, Y.: Videocrafter2: Overcoming data limitations for high-quality video diffusion models. arXiv preprint arXiv:2401.09047 (2024)

[7] Chen, X., Liu, Z., Chen, M., Feng, Y., Liu, Y., Shen, Y., Zhao, H.: Livephoto: Real image animation with text-guided motion control. arXiv preprint arXiv:2312.02928 (2023)

[8] Deng, Y., Wang, R., Zhang, Y., Tai, Y.W., Tang, C.K.: Dragvideo: Interactive drag-style video editing. arXiv preprint arXiv:2312.02216 (2023)

[9] Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)

[10] Du, X., Kolkin, N., Shakhnarovich, G., Bhattad, A.: Generative models: What do they know? do they know things? let's find out! arXiv preprint arXiv:2311.17137 (2023)

[11] Epstein, D., Jabri, A., Poole, B., Efros, A., Holynski, A.: Diffusion self-guidance for controllable image generation. Advances in Neural Information Processing Systems **36** (2024)

[12] Fang, Y., Zhu, H., Zeng, Y., Ma, K., Wang, Z.: Perceptual quality assessment of smartphone photography. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3677–3686 (2020)

[13] Geyer, M., Bar-Tal, O., Bagon, S., Dekel, T.: Tokenflow: Consistent diffusion features for consistent video editing. arXiv preprint arXiv:2307.10373 (2023)

[14] Guo, Y., Yang, C., Rao, A., Agrawala, M., Lin, D., Dai, B.: Sparsectrl: Adding sparse controls to text-to-video diffusion models. arXiv preprint arXiv:2311.16933 (2023)

[15] Guo, Y., Yang, C., Rao, A., Wang, Y., Qiao, Y., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv:2307.04725 (2023)

[16] He, H., Xu, Y., Guo, Y., Wetzstein, G., Dai, B., Li, H., Yang, C.: Cameractrl: Enabling camera control for text-to-video generation. arXiv preprint arXiv:2404.02101 (2024)

[17] Hedlin, E., Sharma, G., Mahajan, S., Isack, H., Kar, A., Tagliasacchi, A., Yi, K.M.: Unsupervised semantic correspondence using stable diffusion. Advances in Neural Information Processing Systems **36** (2024)

[18] Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)

[19] Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)

[20] Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. Advances in Neural Information Processing Systems **35**, 8633–8646 (2022)

[21] 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. arXiv preprint arXiv:2106.09685 (2021)

[22] Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. arXiv preprint arXiv:2311.17982 (2023)

[23] Ju, Y., Hu, K., Zhang, G., Zhang, G., Jiang, M., Xu, H.: Robo-abc: Affordance generalization beyond categories via semantic correspondence for robot manipulation. arXiv preprint arXiv:2401.07487 (2024)

[24] Karaev, N., Rocco, I., Graham, B., Neverova, N., Vedaldi, A., Rupprecht, C.: Cotracker: It is better to track together. arXiv preprint arXiv:2307.07635 (2023)

[25] Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5148–5157 (2021)

[26] Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)

[27] Luo, G., Darrell, T., Wang, O., Goldman, D.B., Holynski, A.: Readout guidance: Learning control from diffusion features. arXiv preprint arXiv:2312.02150 (2023)

[28] 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. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10209–10218 (2023)

[29] Mo, S., Mu, F., Lin, K.H., Liu, Y., Guan, B., Li, Y., Zhou, B.: Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition. arXiv preprint arXiv:2312.07536 (2023)

[30] Mou, C., Wang, X., Song, J., Shan, Y., Zhang, J.: Dragondiffusion: Enabling drag-style manipulation on diffusion models. arXiv preprint arXiv:2307.02421 (2023)

[31] Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)

[32] OpenAI: Video generation models as world simulators. `https://openai.com/research/video-generation-models-as-world-simulators` (2024)

[33] Pan, X., Tewari, A., Leimkühler, T., Liu, L., Meka, A., Theobalt, C.: Drag your gan: Interactive point-based manipulation on the generative image manifold. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–11 (2023)

[34] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)

[35] runway: Gen-2. `https://research.runwayml.com/gen2` (2023)

[36] Shi, Y., Xue, C., Pan, J., Zhang, W., Tan, V.Y., Bai, S.: Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. arXiv preprint arXiv:2306.14435 (2023)

[37] Si, C., Huang, Z., Jiang, Y., Liu, Z.: Freeu: Free lunch in diffusion u-net. arXiv preprint arXiv:2309.11497 (2023)

[38] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. PMLR (2015)

[39] Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)

[40] Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)

[41] Tang, L., Jia, M., Wang, Q., Phoo, C.P., Hariharan, B.: Emergent correspondence from image diffusion. Advances in Neural Information Processing Systems **36** (2024)

[42] Wang, J., Zhang, Y., Zou, J., Zeng, Y., Wei, G., Yuan, L., Li, H.: Boximator: Generating rich and controllable motions for video synthesis. arXiv preprint arXiv:2402.01566 (2024)

[43] Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., Zhang, S.: Modelscope text-to-video technical report. arXiv preprint arXiv:2308.06571 (2023)

[44] Wang, X., Yuan, H., Zhang, S., Chen, D., Wang, J., Zhang, Y., Shen, Y., Zhao, D., Zhou, J.: Videocomposer: Compositional video synthesis with motion controllability. Advances in Neural Information Processing Systems **36** (2024)

[45] Wang, Y., Chen, X., Ma, X., Zhou, S., Huang, Z., Wang, Y., Yang, C., He, Y., Yu, J., Yang, P., et al.: Lavie: High-quality video generation with cascaded latent diffusion models. arXiv preprint arXiv:2309.15103 (2023)

[46] Wang, Z., Yuan, Z., Wang, X., Chen, T., Xia, M., Luo, P., Shan, Y.: Motionctrl: A unified and flexible motion controller for video generation. arXiv preprint arXiv:2312.03641 (2023)

[47] Wold, S., Esbensen, K., Geladi, P.: Principal component analysis. Chemometrics and intelligent laboratory systems **2**(1-3), 37–52 (1987)

[48] Yang, S., Zhou, Y., Liu, Z., Loy, C.C.: Rerender a video: Zero-shot text-guided video-to-video translation. In: SIGGRAPH Asia 2023 Conference Papers. pp. 1–11 (2023)

[49] Yatim, D., Fridman, R., Tal, O.B., Kasten, Y., Dekel, T.: Space-time diffusion features for zero-shot text-driven motion transfer. arXiv preprint arXiv:2311.17009 (2023)

[50] 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. arXiv preprint arXiv:2308.08089 (2023)

[51] 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. arXiv preprint arXiv:2310.08465 (2023)

[52] Zheng, Z., Peng, X., Yang, T., Shen, C., Li, S., Liu, H., Zhou, Y., Li, T., You, Y.: Open-sora: Democratizing efficient video production for all (March 2024), `https://github.com/hpcaitech/Open-Sora`

8 Supplementory Material

In this section, we provide more analysis and results. It is highly recommended to refer to the attached webpage for better visual illustrations.

8.1 Preliminary for Latent Optimization

Diffusion models learn to recover clean images x from random noise $z_T \sim \mathcal{N}(0, I)$ with a sequential denoising process [18; 38; 40]. [34] proposed the latent diffusion model (LDM), which maps data into a lower-dimensional space via a variational auto-encoder (VAE) [26] and models the distribution of the latent embeddings instead. At the diffusion step t, random noise ϵ_t is added to x, giving a noisy image $z_t = \alpha_t x + \sigma_t \epsilon_t$, with α_t and σ_t the time-dependent parameters. The estimation of the denoised image is equivalent to predicting the noise ϵ_t.

Our latent optimization strategy is motivated by [36; 49] that uses intermediate features to supervise the latent optimization process, which can be formulated as

$$z_t^{new} = z_t - \eta \frac{\partial \mathcal{L}}{\partial z_t}, \tag{7}$$

where η is the learning rate and $\mathcal{L}$ is the loss function.

8.2 More Analysis and Details

Video Consistency Preservation. Since one of our applications is to control the motion in source-generated videos, it is important to preserve the consistency between the source video and the target video. To this end, we introduce two techniques: Shared K&V and Masked Gradient Clip. We visualize their qualitative comparison in Fig. 12.

Shared K&V. As proved by many previous works [36; 30; 3], inserting Key (K) & Value (V) of spatial attention from reference branch to target branch can help to preserve content information of reference generation. As shown in Fig. 11, we adopt this method to our motion control pipeline.

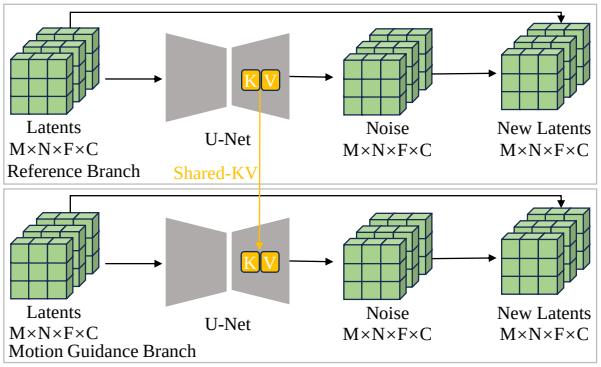


Figure 11: **Motion Control Pipeline with Shared K&V.** We apply the origin designing process in the reference branch while applying motion guidance in the motion guidance branch. During denoising, we insert the K&V of the reference branch to the motion guidance branch for content preservation.

As shown in Fig. 12 (a-c), shared K&V contributes to the consistency of the whole video. The generated video with vanilla motion guidance (Fig. 12 (b)) adds additional contents (i.e. a hat on the man's head) while adding shared K&V (Fig. 12 (c)) stays consistent with the original generation.

Masked Gradient Clip. Since we do not want to change much content out of the masked region during motion guidance optimization, we simply clip the guidance gradient g out of the masked region, which is

$$g^{\text{clip}} = \begin{cases} g, & (i,j,k) \in \mathcal{R} \& k \in \mathcal{F} \\ 0, & \text{else}, \end{cases} \tag{8}$$

(a) Origin generation

(b) Vanilla Motion Guide

(c) + Shared K&V

(d) + Gradient Clip for the first 8 frames

(e) + Gradient Clip for all frames

Figure 12: **Qualitative comparison on video consistency preservation.** We compare the generated results w./wo. our introduced techniques. The control signal is shown in the first image of (a), with the red arrow indicating the motion control direction and the light region indicating the control region. We highlight the noticeable region with red boxes. It reveals that Shared K&V contributes to the consistency of the whole video. Gradient Clip adds consistency out of masked regions but meanwhile reduces motion scale.

where i, j, k are indices of height, width, and frame, respectively. $\mathcal{R}$ is the spatial mask region index set. $\mathcal{F}$ is the frame set. As shown in Fig. 12 (d-e), gradient clipping adds consistency to the background content. While applying gradient clipping to more frames increases consistency, it also results in a smaller motion scale. Thus, applying gradient clipping involves a trade-off. In practice, we apply gradient clipping to the first 8 frames.

Timestep Choice for Motion Extracted from Video. As shown in Fig. 13, the trends and ranges of motion channels with the same motion direction are similar among different denoising timesteps, while the curve nearing the end of denoising steps is smoother and refined. To this end, we use the MOFT at the beginning of the inversion stage as the guidance for all control timesteps.

Motion Channel Filter Number. We ablate the effects to filter different numbers of motion channels in Tab. 4. The key finding is that preserving only a few channels most sensitive to motion can enhance both motion faithfulness and naturalness, as the effects of irrelevant information in other channels are removed. However, when we further reduce the channel number to the top 1%, both motion

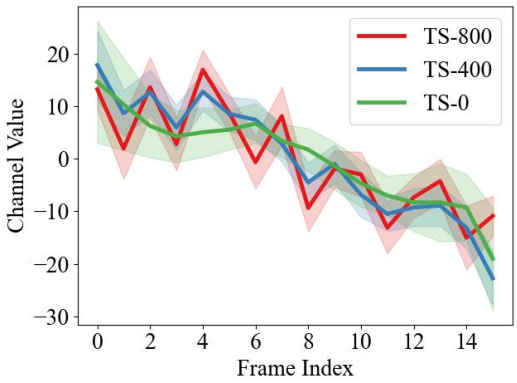


Figure 13: **Motion channel value for different denoising timestep.** TS-800 indicates denoising step 800.

Table 4: Motion Channel Ablation.

Channel number	Faithfulness (↑)	Naturalness (↑)
top 100%	67.5	0.671
top 50%	70.5	0.681
top 10%	82.7	0.692
top 5%	83.2	**0.694**
top 4%	**84.0**	<u>0.693</u>
top 3%	<u>83.6</u>	<u>0.693</u>
top 1%	76.3	0.677

faithfulness and naturalness significantly decrease due to the loss of some motion-sensitive channels. In practice, we choose the top 4% of motion channels.

Point-Drag Manipulation Ablation. Following the main paper Sec. 4.2, we use MOFT for initial coarse motion control and DIFT for precise point-drag manipulation. The compositional loss function is

$$\mathcal{L}_t = w_t^c \mathcal{L}^c + w_t^p \mathcal{L}^p, \begin{cases} w_t^c > 0, w_t^p = 0, & \text{if} \quad t >= t_1 \\ w_t^c > 0, w_t^p > 0, & \text{if} \quad t_1 > t >= t_2 \\ w_t^c = 0, w_t^p > 0, & \text{if} \quad t_2 > t >= t_3 \\ w_t^c = 0, w_t^p = 0, & \text{if} \quad t_3 > t \end{cases} \tag{9}$$

where w_t^c and w_t^p are time-dependent weights under the threshold t_1, t_2 and t_3. In practice, the total denoising step is 25. We set $t_1 = 19$, $t_2 = 18$, $t_3 = 5$. We further ablate the effectiveness of the design in Fig. 14. Applying only DIFT results in limited motion. Using only MOFT produces motion but lacks precise point control. By combining DIFT and MOFT, we achieve precise point-drag control.

8.3 More Visualization

PCA video examples. In Fig. 15, we provide some of the video examples that we use to conduct PCA. We manually move the whole picture following specified motion directions to synthesis videos.

Qualitative Results. We provide some qualitative results in Fig. 16. More animated results can be found on the attached webpage.

8.4 More Results on Open-Sora [52]

In Fig. 17(a), we demonstrate that PCA can clearly separate videos with different motions based on their diffusion features from Open-Sora [52], an open-source video generation model capable of producing long videos. In Fig. 17(b), we show that our methods can be applied to higher resolutions (768×768) and longer videos (205 frames on Open-Sora).

Figure 14: **Qualitative results of point-drag manipulation ablation.** Red points indicate starting points. Blue points indicate ending points. We only display three frames per animation clip.

Figure 15: **Videos for PCA.** We manually move the whole picture following specified motion directions to synthesis videos.

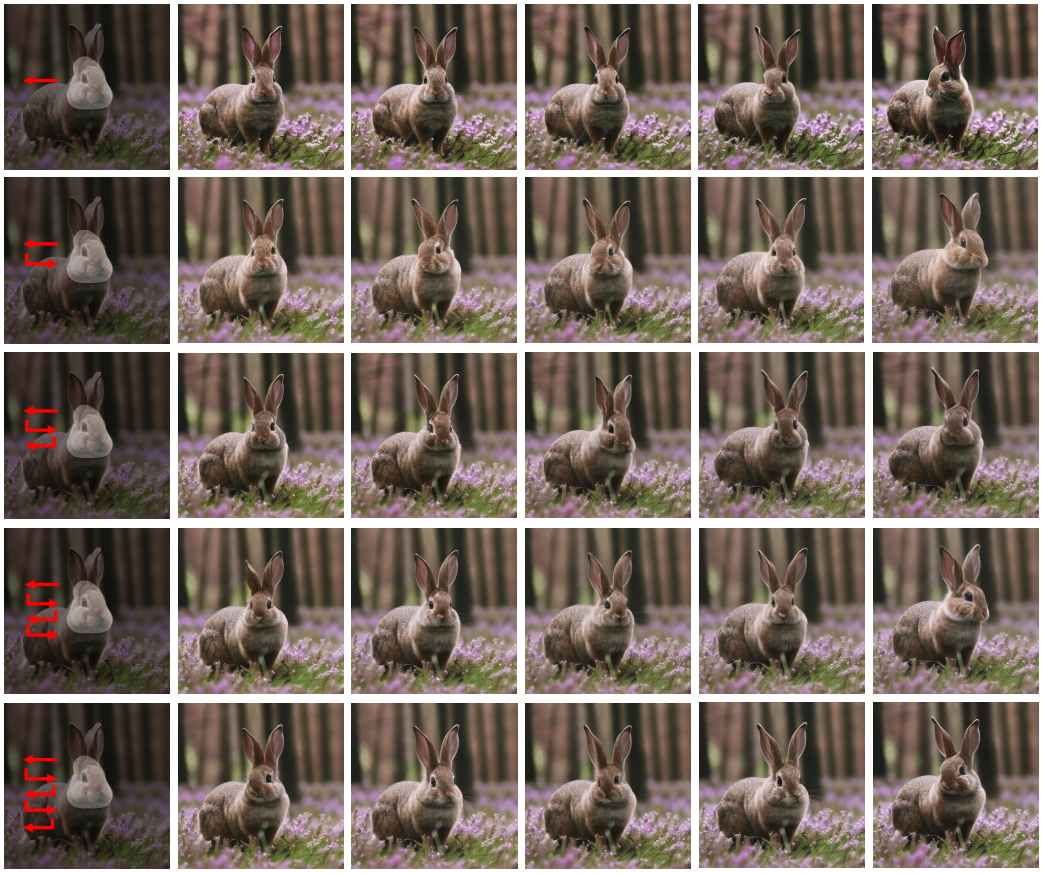
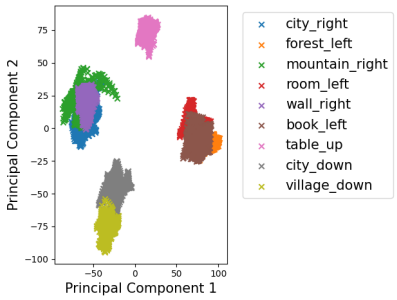
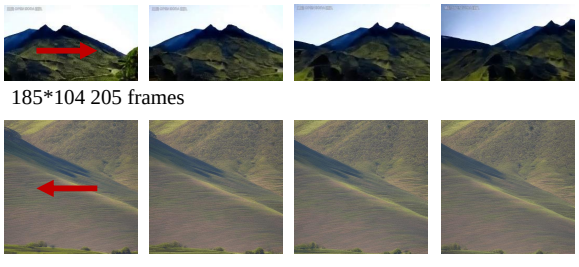


Figure 16: **More qualitative results.**



(a) PCA on Open-Sora



(b) Results on longer videos and higher resolutions

Figure 17: **More results on Open-Sora.**

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The claims accurately reflect our contribution and scope.

   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: We discuss our limitations in the supplementary material.

   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: This paper does not provide theoretical results. The assumptions in the paper are verified by experiments.

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 the necessary details for reproducibility in Sec. 5.1 and Supp. .

   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: [No]

Justification: We will release it later.

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: We provide the necessary details for reproducibility.

   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: [No]

   Justification: Our experiments are not accompanied by error bars

   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 it in Sec. 5.1.

   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: Our research is aligned with 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: [NA]

    Justification: There is no societal impact of the work performed.

    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

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: The 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 properly credit the code and models used in the paper.

    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: The 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: [Yes]

Justification: Our experiments include human feedback. We include the instructions and details in Sec. 5.5.

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: [No]

Justification: Our experiments do not have risks.

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.