

# DisMo: Disentangled Motion Representations for Open-World Motion Transfer

Thomas Ressler-Antal Frank Fundel\* Malek Ben Alaya\*  
 Stefan Andreas Baumann Felix Krause Ming Gui Björn Ommer

CompVis @ LMU Munich, Munich Center for Machine Learning (MCML)

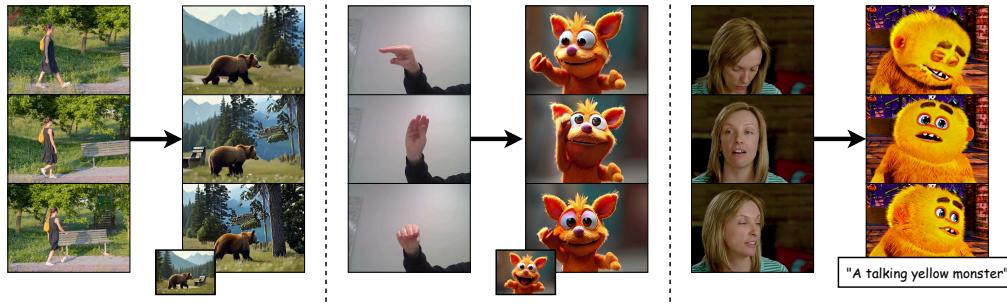


Figure 1: **Motion transfer examples enabled by our abstract motion representations.** We extract abstract motion representations from driving videos and transfer them onto new content, represented either by source images (**Left**) and (**Middle**), or text prompts (**Right**).

## Abstract

Recent advances in text-to-video (T2V) and image-to-video (I2V) models, have enabled the creation of visually compelling and dynamic videos from simple textual descriptions or initial frames. However, these models often fail to provide an explicit representation of motion separate from content, limiting their applicability for content creators. To address this gap, we propose *DisMo*, a novel paradigm for learning abstract motion representations directly from raw video data via an image-space reconstruction objective. Our representation is generic and independent of static information such as appearance, object identity, or pose. This enables open-world motion transfer, allowing motion to be transferred across semantically unrelated entities without requiring object correspondences, even between vastly different categories. Unlike prior methods, which trade off motion fidelity and prompt adherence, are overfitting to source structure or drifting from the described action, our approach disentangles motion semantics from appearance, enabling accurate transfer and faithful conditioning. Furthermore, our motion representation can be combined with any existing video generator via lightweight adapters, allowing us to effortlessly benefit from future advancements in video models. We demonstrate the effectiveness of our method through a diverse set of motion transfer tasks. Finally, we show that the learned representations are well-suited for downstream motion understanding tasks, consistently outperforming state-of-the-art video representation models such as V-JEPA in zero-shot action classification on benchmarks including Something-Something v2 and Jester. Project page: <https://compvis.github.io/DisMo>

\*Equal contribution.

## 1 Introduction

A striking feature of human visual intelligence is our innate ability to understand motion and to reason about it independently from static attributes such as appearance, position, and viewpoint, generalizing seamlessly across different instances within an object category and even beyond to entirely different objects and scenarios [Johansson, 1973]. Already at a young age, we are able to generalize motion despite stark deviations in environment and content to a different scene allowing already little children to imagine the plausible movement of fictitious characters. This capacity highlights a fundamental cognitive ability to represent and reason about motion as an abstract, invariant phenomenon.

Recent breakthroughs in generative video synthesis have enabled models to produce visually compelling and dynamically realistic videos from minimal inputs such as textual descriptions (text-to-video, T2V) or single initial frames (image-to-video, I2V) [Blattmann et al., 2023, Brooks et al., 2024, HaCohen et al., 2024, Kong et al., 2024, Veo-Team et al., 2024]. Despite the remarkable rendering quality achieved by these models, they lack explicit representations of underlying kinematics, inherently limiting their ability to precisely control or manipulate the motion of generated scenes.

Existing T2V and I2V approaches implicitly represent motion by modeling videos as sequences of static frames, which aligns poorly with human perception that inherently treats motion as a distinct and explicitly perceived entity rather than merely a tomography across time [Kandel et al., 1991]. Furthermore, precise textual descriptions of complex or abstract motion remain challenging, underscoring the need for novel representation approaches that go beyond verbal or purely frame-based encodings.

Explicit methods for representing dynamic scene content have traditionally relied on dense, pixel-level motion descriptors such as optical flow [Zhao et al., 2024, Chefer et al., 2025, Liang et al., 2024]. While optical flow effectively captures detailed per-pixel changes between frames, it remains strongly coupled to specific instance structures, locations, and viewpoints, thus lacking the invariance needed for robust motion generalization and transfer across diverse scenarios. Similarly, sparse representations such as trackers that follow key points through time are somewhat less dependent on appearance but still fundamentally require structural consistency across frames [Karaev et al., 2024]. Consequently, neither dense nor sparse methods readily support motion transfer between significantly different instances, object categories, and viewpoints.

Inspired by recent successes in text-to-image (T2I) synthesis, where explicit structural control mechanisms have greatly enhanced user-driven generation [Rombach et al., 2022, Zhang et al., 2023a, Stracke et al., 2024], we argue that generative video models similarly require novel mechanisms for explicitly representing and controlling motion separate of appearance and spatial configurations. Such explicit and abstract representations would significantly enhance user control over dynamically generated content.

In this work, we introduce a method for learning abstract motion representations that are invariant to static appearance and structure-based information, directly from raw video data using only a standard image reconstruction loss. Remarkably, our representation emerges entirely through end-to-end training regime without the need for specialized regularization, contrastive losses, or complex training setups. We leverage a simple yet effective flow-matching-based reconstruction objective, commonly used in visual generation methods, to obtain a representation that is largely invariant to factors typically encoded in the reconstruction loss, such as position, identity, category, and pose. The result is a motion representation robust enough to perform challenging and creative open-world motion transfers, enabling scenarios such as transferring the movement of a human to that of an ape, an imaginary cartoon character, or even completely unconventional transformations. This capability is achieved through an information bottleneck further by strong augmentations that makes it impossible to reconstruct the target using visual and appearance-based information alone, forcing the model to analyze changes between frames and encode them into an abstract embedding.

We evaluate these embeddings via zero-shot probing across multiple action classification datasets and ensure their representative performance. Additionally, we design these motion representations to be invariant to the choice of renderer, allowing seamless integration with state-of-the-art video diffusion models through lightweight adapters [Stracke et al., 2024], which reduces compute as expensive video pipelines do not need to be retrained. This orthogonal approach transforms existing generative video models from merely powerful renderers into highly controllable generative systems.

Our main contributions are as follows:

- We construct a powerful representation learning paradigm powered only by an image-space reconstruction objective and show how this leads to an abstract motion representation space that is trained for invariance to static content information such as appearance or structure.
- We leverage this invariance and condition a pre-trained generative video model on our learned motion representations, which, despite training in a matched setting is capable of zero-shot motion transfer.
- We show how this setup achieves state-of-the-art performance on open-world motion transfer with a high degree of transferability, especially in cross-category and cross-viewpoint settings.
- We show how our motion embeddings can be utilized in downstream action classification settings.

## 2 Related Work

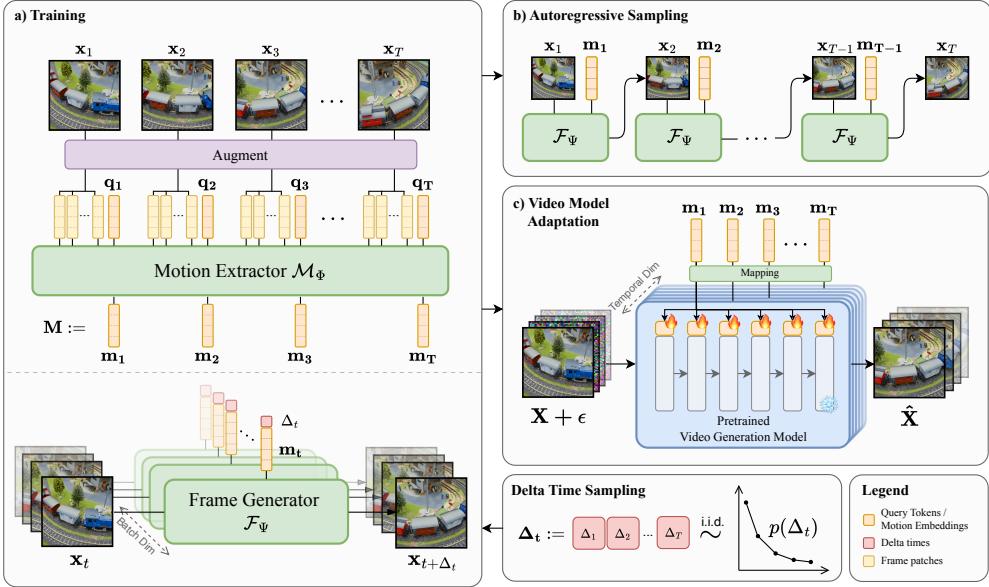
**Video Synthesis** Generative image and video models have experienced rapid advances in recent years, largely driven by diffusion-based training objectives [Ho et al., 2020, Song et al., 2020, Lipman et al., 2022]. Originally developed for the image domain, these powerful frameworks have been extended to the video domain, demonstrating impressive generative capabilities and realism. Brooks et al. [2024] demonstrated that large-scale video synthesis enables the realistic generation of complex spatiotemporal dynamics from simple text and image prompts. Subsequently, several state-of-the-art text-to-video (T2V) and image-to-video (I2V) diffusion models have further significantly advanced video synthesis capabilities [Blattmann et al., 2023, Yang et al., 2024, HaCohen et al., 2024, Kong et al., 2024, Veo-Team et al., 2024, Polyak et al., 2025] and made such models widely accessible.

Despite their success, current T2V/I2V methods primarily focus on synthesizing visually appealing videos from textual and image prompts but do not explicitly model the underlying motion separately from appearance. As a result, they inherently lack the capacity for precise or intuitive control over the generated motion, limiting their applicability in scenarios requiring detailed, user-specified dynamics.

**Motion Control in Video Generation** Enabling detailed motion control in generative video models remains a challenging and active research area. Recent methods typically achieve motion control by providing explicit low-level spatial or temporal constraints, such as bounding boxes [Jain et al., 2024, Wang et al., 2024a, Wu et al., 2025], trajectories [Wang et al., 2024b, Wu et al., 2024, Yin et al., 2023], or structural maps like optical flow and depth sequences [Zhang et al., 2023b, Esser et al., 2023]. Although effective within certain scenarios, these conditioning signals inherently tie motion control to precise spatial locations and/or structures in the source video, limiting their invariance and transferability. Consequently, these approaches struggle with scenarios involving substantial variations in viewpoint, appearance, or semantic category.

Parametric approaches offer another perspective by learning domain-specific motion representations, frequently used in specialized areas such as facial reenactment and talking-head generation [Siarohin et al., 2019, 2021, Wang et al., 2021, Zhao and Zhang, 2022, Wang et al., 2022, Hong et al., 2022, Yang et al., 2022, Gao et al., 2023]. These methods typically rely on pre-defined or learned keypoint-based *category-specific* representations (either 2D or 3D) to encode motion explicitly. For instance, First-Order Motion Models (FOMM) [Siarohin et al., 2019], Motion Representations for Articulated Animation (MRAA) [Siarohin et al., 2021], and Latent Image Animator (LIA) [Wang et al., 2022] have demonstrated impressive performance. However, due to their specialized nature and dependence on predefined structural cues, parametric models have to be specifically developed for the target domain/category and typically generalize poorly beyond their narrowly defined application, not allowing for advanced applications such as cross-category motion transfer.

In contrast to both explicit low-level conditioning and domain-specific parametric models, our approach aims to learn a highly abstract, invariant motion representation. By explicitly modeling motion independently of appearance, pose, viewpoint, and semantic category, we substantially enhance the flexibility and transferability required for diverse motion control scenarios.



**Figure 2: Method Overview.** (a) During training, our motion extractor  $\mathcal{M}_\theta$  receives augmented frames from a video  $\mathbf{X}$ , along with additional motion query tokens  $\mathbf{Q}$ . These are then individually passed to the frame generator  $\mathcal{F}_\psi$ , alongside the corresponding source frame  $\mathbf{x}_t$ , from which it learns to reconstruct a frame at a future timestep  $t + \Delta_t$ . (b) To transfer a motion sequence  $\mathbf{M}$  onto another target image, we can directly utilize the trained Frame Generator  $\mathcal{F}_\psi$  autoregressively as a low-cost option. (c) For high-quality motion transfer, we adapt pre-trained off-the-shelf video generation models. A motion sequence is first embedded using a mapping network, before being introduced to the frozen video model. The processed motion sequence is arranged such that each token at timestep  $t$  in the pre-trained backbone receives conditioning only from the temporally corresponding motion embedding  $\mathbf{m}_t$ .

### 3 Method

Recent approaches in video modeling commonly utilize generative diffusion or flow-based objectives applied either directly on video frames [Ho et al., 2022] or on latent representations extracted from pretrained autoencoders [Blattmann et al., 2023, Brooks et al., 2024, HaCohen et al., 2024, Kong et al., 2024, Veo-Team et al., 2024]. While these have been successful for high-fidelity synthesis tasks, our goal differs fundamentally: instead of directly generating videos, we aim to learn abstract, motion representations that are invariant to the appearance of the scene. Specifically, we define motion as the abstract temporal dynamics that describe *how* scenes and the objects therein evolve, independent from their actual appearance, style, spatial configuration, or even semantic category. The representations can subsequently serve as a powerful, generalizable conditioning signal for controlling motion in downstream video generation or discriminative tasks.

#### 3.1 Preliminaries

**Flow Matching** Flow matching [Lipman et al., 2022] is a generative modeling framework that learns distributions via a continuous-time flow field that smoothly transforms a simple prior distribution—typically a Gaussian—into a complex target data distribution. Given data samples  $\mathbf{z} \sim p_{\text{data}}(\mathbf{z})$ , they learn a vector field  $\mathbf{v}_\psi(\mathbf{z}_\tau, \tau)$  with learnable parameters  $\psi$  describing the flow of the distribution at intermediate times  $\tau \in [0, 1]$ . This vector field is learned by minimizing the MSE loss

$$\mathcal{L}(\psi) = \mathbb{E}_{\tau, \mathbf{z}_0, \mathbf{z}_1} \|\mathbf{v}_\psi(\mathbf{z}_\tau, \tau) - \mathbf{u}(\mathbf{z}_\tau, \tau)\|_2^2, \quad (1)$$

where  $\mathbf{z}_\tau = \tau \mathbf{z}_1 + (1 - \tau) \mathbf{z}_0$ ,  $\mathbf{z}_1 \sim p_{\text{data}}$ ,  $\mathbf{u}(\mathbf{z}_\tau, \tau) = \mathbf{z}_1 - \mathbf{z}_0$  is the target flow field, and  $\mathbf{z}_0$  is drawn from the prior distribution, which is typically chosen to be Gaussian  $\mathbf{z}_0 \sim \mathcal{N}(0, \mathbf{I})$ . Samples from this learned distribution can be generated by drawing from the prior and solving the learned ODE.

### 3.2 Learning Transferable Motion Representations

To learn abstract motion representations, we introduce *DisMo*, a method consisting of two primary components learned jointly: a motion extractor  $\mathcal{M}_\theta$  and a motion-conditioned frame generator  $\mathcal{F}_\psi$ . The motion extractor computes meaningful motion embeddings from raw input videos, while the frame generator leverages these embeddings alongside source frames to reconstruct target frames. The overall framework is trained end-to-end using a simple yet powerful generative image-space reconstruction objective, modeling the distribution of frames conditioned on a source frame and learned motion embedding. The learned motion embedding inherently encodes temporal dynamics disentangled from frame contents and appearance. An overview of the architecture and training paradigm is depicted in Fig. 2.

**Motion Extraction.** Consider an input video  $\mathbf{X} = \{\mathbf{x}_t\}_{t=1}^T$  consisting of  $T$  frames  $\mathbf{x}_t$ . From this, our goal is to extract a sequence of motion embeddings  $\mathbf{M} = \{\mathbf{m}_t\}_{t=1}^T$ . We formally define the motion extractor as  $\mathcal{M}_\theta : \mathbf{X} \rightarrow \mathbf{M}$  with parameters  $\theta$ . This motion extractor jointly processes the given frames and extracts a sequence of motion embeddings for all timesteps  $t = \{1, \dots, T\}$ . Each motion embedding encodes local temporal dynamics while ignoring static visual information.

**Dual-Stream Frame Generation.** The frame generator  $\mathcal{F}_\psi$ , parametrized by  $\psi$ , reconstructs target frames conditioned on a motion embedding and a source frame, and is responsible for providing the learning signal for the motion extractor. Formally, the generator learns the conditional distribution

$$\mathcal{F}_\psi : \mathbf{x}_{t+\Delta_t} \sim p(\mathbf{x}_{t+\Delta_t} | \mathbf{x}_t, \mathbf{m}_t, \Delta_t) \quad (2)$$

describing the future destination frame  $\mathbf{x}_{t+\Delta_t}$  at timestep  $t + \Delta_t$ . As conditioning, the source frame  $\mathbf{x}_t$ , the motion embedding at the source time  $\mathbf{m}_t$  extracted using our motion extractor  $\mathcal{M}_\theta$  and the distance  $\Delta_t$  to the target frame are given.  $\Delta_t$  can either be a constant or can be sampled independently for each prediction. Effectively, this provides the frame generator with two streams of available information: content information captured by the source frame  $\mathbf{x}_t$ , and information that is embedded into the motion representation  $\mathbf{m}_t$ . Due to the information bottleneck induced on the motion embedding via the limited embedding size, this representation is encouraged to store only high-level, invariant temporal information. Essentially, this represents the residual knowledge not encoded by  $\mathbf{x}_t$  that supports the frame generator in predicting the target frame  $\mathbf{x}_{t+\Delta_t}$ .

We use this frame generator  $\mathcal{F}_\psi$  to jointly learn both it and the motion extractor  $\mathcal{M}_\theta$  end-to-end by (implicitly) maximizing the likelihood [Kingma and Gao, 2023] of the reconstructed destination frames  $\mathbf{x}_{t+\Delta_t}$ , conditioned on respective source frames and their corresponding motion embeddings:

$$\mathbb{E}_{\mathbf{X} \sim p_{\text{data}}} [\log(p_\psi(\mathbf{x}_{t+\Delta_t} | \mathbf{x}_t, \mathcal{M}_\theta(\mathbf{X}, t)))] \quad (3)$$

We practically implement this via a flow matching [Lipman et al., 2022] objective (cf. Eq. (1))

$$\mathcal{L}(\theta, \psi) = \mathbb{E}_{\mathbf{X} \sim p_{\text{data}}, \tau, \mathbf{z}_0 \sim \mathcal{N}(0, \mathbf{I}), (t, t+\Delta_t)} \|\mathbf{v}_\psi(\mathbf{z}_\tau, \mathbf{x}_t, \mathcal{M}_\theta(\mathbf{X}, t), \tau) - \mathbf{u}(\mathbf{z}_\tau, \tau)\|_2^2, \quad (4)$$

where  $\mathbf{z}_\tau = \tau \mathbf{x}_{t+\Delta_t} + (1 - \tau) \mathbf{z}_0$ , and  $\mathbf{v}_\psi$  is the vector field learned by the frame generator  $\mathcal{F}_\psi$ .

**Augmentation Pipeline.** To ensure that motion embeddings remain invariant to appearance, we additionally employ an augmentation pipeline inspired by self-supervised learning methods [He et al., 2022, Tong et al., 2022, Chen et al., 2020]. Appearance augmentations include random photometric transformations (e.g., hue, contrast, brightness adjustments) and geometric transformations (e.g., cropping, rotation, translation, shearing, aspect ratio changes), applied uniformly across all frames to avoid confusing these changes with actual motion. By applying these augmentations during training, we further encourage the learned motion embeddings to capture high-level, transferable dynamics rather than video-specific details.

### 3.3 Motion-Guided Video Generation

Once the motion extractor and frame generator are trained, they can be directly used for autoregressive motion transfer. Given a source video, we extract a sequence of motion embeddings and, starting

from a target initial frame, we condition the frame generator on these embeddings at each time step to predict the next frame in an autoregressive manner (see Fig. 2b). While effective, this setup is limited by the temporal scope captured in a single motion embedding and the capacity of the frame generator. To address these limitations and enable longer, more coherent video generation, we further adapt off-the-shelf video generation model by conditioning them on sequences of motion embeddings. This approach allows us to leverage the strong generative priors of large video models while guiding their synthesis with our learned motion representation.

To adapt a pre-trained video backbone, we utilize LoRAdapters [Stracke et al., 2024], adaptive low-rank adapters [LoRAs, Hu et al., 2022] that allow introducing additional conditioning into a frozen backbone. We extend a pre-trained spatio-temporal video model with conditional LoRAs that receive a sequence of motion embeddings during the generation process as input. Importantly, the conditioning is designed such that each spatiotemporal token gets influenced only by the motion embedding that aligns with its temporal position. This makes the approach efficient, yet effective. This fine-tuning regime is illustrated in Fig. 2c. Our generated motion embeddings are generally applicable to a wide range of off-the-shelf video models. We demonstrate that DisMo can be adapted to various video models in the Appendix (see Table E), while benefitting from their advantages.

## 4 Experiments

We evaluate our learning paradigm and show the effectiveness of the learned motion representations for motion transfer tasks. We further show that the motion embeddings are disentangled from appearance and structure-specific information, preventing unwanted information leakage during transfer tasks. Finally, we demonstrate that our motion representations encode higher-level temporal dynamics that is useful in zero-shot action classification settings.

**Implementation Details.** We implement our motion extractor as a 3D vision transformer [Dosovitskiy et al., 2021, 86M params] with an initial frame embedding stage consisting of a DINOv2-B [Oquab et al., 2023] model, and adapt our frame generator from a pretrained diffusion transformer [Peebles and Xie, 2023, 675M params]. We train the two components using AdamW [Loshchilov and Hutter, 2019] for 530k steps at a batch size of 32 on open-world videos from K-710 [Li et al., 2022], SSv2 [Goyal et al., 2017], Moments in Time [Monfort et al., 2019a] and OpenVid-1m [Nan et al., 2025] to learn general-purpose motion representations. Unless otherwise specified, we use LTX-Video-2B [HaCohen et al., 2024] as our video generation backbone that we adapted following the procedure described in Section 3.2. Evaluations with other video model backbones, additional implementation details and ablations of our main design choices and components are provided in the Appendix.

### 4.1 Open-World Motion Transfer

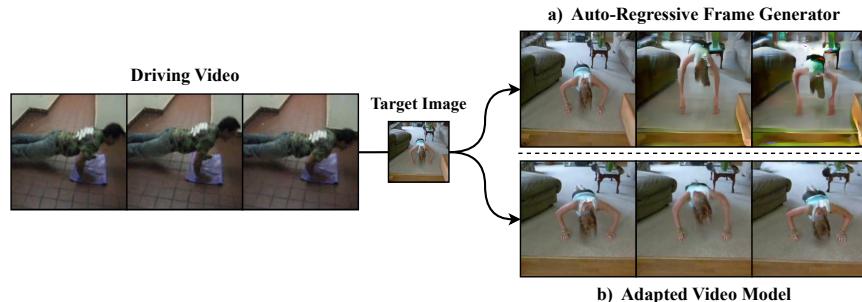


Figure 3: **Qualitative motion transfer comparison between (a) our auto-regressive frame generator and (b) an adapted video model.** We transfer motion extracted from a driving video onto a new target image using each model. While both approaches manage to transfer high-level motion semantics in a view- and appearance-invariant manner, the adapted video model achieves higher generation fidelity.

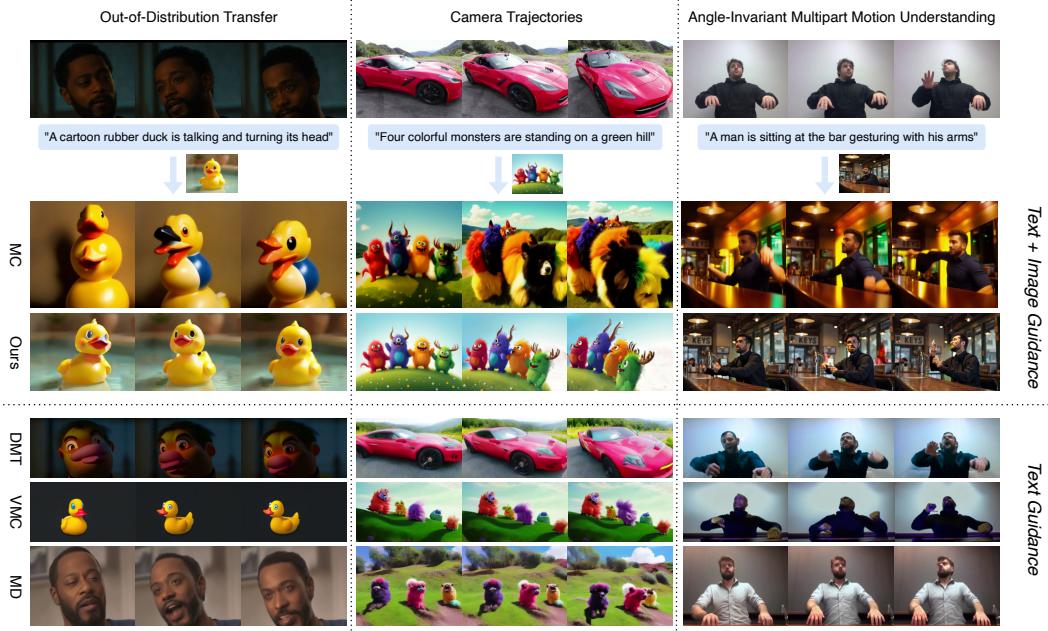


Figure 4: **Motion transfer examples in different settings.** We compare different motion transfer methods on three settings: **(Left)** Inter-category motion transfer. **(Middle)** An example showcasing camera motion. **(Right)** Composed motion transfer.

We qualitatively compare our auto-regressive frame generator and the adapted high-fidelity video model in a motion transfer scenario in Fig. 3. While both successfully transfer higher-level motion from the source video onto the target image, the adapted video model yields a substantial improvement in generation quality. We therefore rely solely on the adapted model in the following experiments.

Specifically, we evaluate it on general and out-of-distribution motion transfer scenarios and compare it against four state-of-the-art approaches: VMC [Jeong et al., 2024], DMT [Yatim et al., 2024], MotionClone [Ling et al., 2024] and MotionDirector [Zhao et al., 2024]. VMC and MotionDirector are fine-tuning-based methods that require extensive retraining per motion concept, whereas DMT and MotionClone are training-free approaches that leverage aggregated priors from a pre-trained video diffusion model for a test-time optimization objective.

Unlike many motion transfer methods, such as VMC and DMT, which rely solely on text-based appearance guidance, our method can optionally incorporate a start frame conditioning. This provides more nuanced control over the content, appearance, and structure of the target entities. However, our approach also performs effectively using only textual prompts.

To comprehensively assess method performance, we build upon prior evaluation strategies [Zhao et al., 2024, Jeong et al., 2024, Yatim et al., 2024, Park et al., 2024], assembling a diverse set of videos and text prompts primarily from the DAVIS dataset [Pont-Tuset et al., 2017]. Additionally, we include challenging examples from OpenVid-1M [Nan et al., 2025] and other sources. Our evaluation set features varied scenarios such as stationary and moving cameras, in-domain and out-of-domain motion transfers, and creatively challenging out-of-distribution cases.

We quantitatively compare methods using three established metrics: *Motion Fidelity*: Measured using a tracking-based metric inspired by Chamfer distance, evaluating alignment between generated and driving video tracklets [Yatim et al., 2024]. *Prompt Adherence*: Evaluated by computing the cosine similarity between CLIP embeddings of generated frames and corresponding textual prompts. *Temporal Consistency*: Assessed by measuring the alignment of CLIP embeddings between consecutive frames. *Driving Video Similarity*: assessed by measuring alignment of CLIP embeddings of source and generated video frames. Our evaluation (see Table 1) shows that DisMo consistently matches or outperforms other state-of-the-art methods across all four metrics. Notably, the evaluated models exhibit inherent trade-offs: DMT and MotionDirector achieve high motion fidelity but frequently overfit on the exact structure of source videos, compromising textual adherence, while

Table 1: **Quantitative Motion Transfer Evaluation.** We compare with four other motion transfer methods with respect to motion fidelity, text adherence, temporal consistency, and similarity to the source video during our generative transfer experiments. *DisMo* performs best on all four metrics, while other approaches struggle with the trade-off between motion fidelity and condition adherence.

Method	Motion Fidelity ↑	Prompt Adherence ↑	Temporal Consistency ↑	Driving Video Similarity ↓
VMC* [Jeong et al., 2024]	0.57	0.26	0.94	0.59
DMT† [Yatim et al., 2024]	0.70	0.24	0.93	0.66
MotionClone† [Ling et al., 2024]	0.63	<b>0.27</b>	0.91	0.59
MotionDirector* [Zhao et al., 2024]	0.70	0.16	0.92	0.82
<i>DisMo</i> ‡ ( <i>Ours</i> )	<b>0.75</b>	<b>0.27</b>	<b>0.95</b>	<b>0.55</b>

\*Per-sample finetuning

†Inference-time optimization

‡General pre-trained

VMC and MotionClone maintain strong textual adherence at the expense of accurate motion transfer. Meanwhile, our approach performs best on both metrics, not suffering from this trade-off. Qualitative examples in Fig. 4 further illustrate this. Other methods either fail to transfer the object/camera motion faithfully or fail to adhere to the appearance of the prompt/start frame. *DisMo*’s ability to represent motion abstractly allows greater flexibility and generalization compared to existing approaches.

Furthermore, perceptual aspects such as motion transfer quality are best assessed through human evaluation in addition to automated metrics. To this end, we conducted a user study comparing several state-of-the-art models, including our own (*DisMo*), to evaluate their performance in terms of realism, prompt adherence, and motion transfer quality. As shown in Table 2, while MotionDirector achieves high motion transfer quality (25.96%), it exhibits a notable decline in prompt adherence (7.47%), likely due to visual inconsistencies and content bleeding. In contrast, our method, *DisMo*, attains comparable motion transfer quality (24.71%) while substantially outperforming MotionDirector in both realism and prompt adherence. Moreover, *DisMo* achieves these results with markedly higher inference efficiency than existing methods. We refer the reader to Section E in the supplementary for further details.

Table 2: **Human Evaluation of Motion Transfer.** We compare with four other motion transfer methods with respect to realism, prompt matching, and motion transfer quality. *DisMo* achieves the highest scores in realism and prompt matching by a wide margin, while remaining competitive in motion transfer quality.

Method	Realism (%)	Prompt Matching (%)	Motion Transfer Quality (%)
DMT [Yatim et al., 2024]	10.93	9.60	17.73
MotionDirector [Zhao et al., 2024]	10.98	7.47	<b>25.96</b>
VMC [Jeong et al., 2024]	<u>20.04</u>	<u>26.13</u>	16.98
MotionClone [Ling et al., 2024]	19.91	19.42	14.62
<i>DisMo</i> ( <i>Ours</i> )	<b>38.13</b>	<b>37.38</b>	<u>24.71</u>

## 4.2 Appearance and Structural Disentanglement

**Identity Reconstruction from Motion Representations.** To assess the extent to which our motion representations disentangle person-specific appearance from action dynamics, we designed a controlled identity classification experiment. Ideally, motion representations should capture dynamic patterns while minimizing appearance information. To evaluate this property, we utilize the Invariant Action Recognition Dataset (IARD) [Tacchetti et al., 2019], which contains controlled recordings of five actors performing five different actions from multiple viewpoints. We extract features from videos of different actors performing a subset of actions and evaluate identity classification on a disjoint action excluded from the training set, enabling analysis of identity-invariant motion representations. A  $k$ -Nearest Neighbors classifier ( $k = 20$ ) was used to predict the identity of the actor based solely on the extracted representations. Higher classification accuracy indicates greater retention of appearance-specific information, while lower accuracy suggests stronger invariance to identity. We compare our method against VideoMAE, VideoMAEv2, and V-JEPA in Table 3.

The results reveal clear differences in the amount of identity information each model retains: Our method achieves 23.82% accuracy, close to the 20% random baseline for five-way classification, indicating minimal encoding of identity features. In contrast, VideoMAEv2 reaches 58.94%, while V-JEPA and VideoMAE achieve 96.23% and 99.14%, respectively. These findings suggest that our approach disentangles motion from appearance more effectively, making it particularly well-suited for identity-invariant motion understanding and transfer.

We further ablate action classification accuracy on IARD in a similar setting. Specifically, we set up a  $k$ NN-based action classification experiment and adopt an identity-based split: videos from one individual are held out for testing, while videos from the remaining four individuals are used for training. This evaluation protocol offers a more realistic and challenging assessment by requiring generalization across identities, thereby measuring a model’s ability to capture action dynamics independently of person-specific cues. As shown in Table 3, our representations achieve the best performance on this task, demonstrating that they are both structured and semantically meaningful.

**Table 3: Identity Disentanglement.** We evaluate disentanglement from appearance by performing identity classification on IARD, which contains annotations for both actions and identities. Low identity classification performance indicates high disentanglement. Our motion representations encode significantly less appearance information, while performing best on action classification.

Model	Action Accuracy $\uparrow$	Identity Accuracy $\downarrow$	vs. Random (20%)
VideoMAE [Tong et al., 2022]	73.44	99.14	+79.14
VideoMAEv2 [Wang et al., 2023]	80.25	58.94	+38.94
V-JEPA [Bardes et al., 2024]	82.03	96.23	+76.23
<i>DisMo</i> ( <i>Ours</i> )	<b>90.74</b>	<b>23.82</b>	<b>+3.82</b>

**Appearance and Structure Invariance during Motion Transfer.** We further evaluate the invariance of motion representations under appearance and structural changes in a motion transfer setting. Starting with an original driving video, we construct two augmented versions: one with photometric transformations (e.g., changes in brightness, contrast, and color) and one with geometric transformations (e.g., rotations, cropping, and spatial shifts). We then perform motion transfer using all three driving videos, the original and the two augmented versions, while keeping the same source prompt and seed. Ideally, an invariant motion representation should produce minimal differences between these outputs. Hence, we compare the generated videos driven by the two augmented views against the one generated using the unaltered video. We compute the CLIP distance [Radford et al., 2021] as a semantic, SSIM [WangZhou et al., 2004] as a structural, LPIPS [Zhang et al., 2018] as a patch-wise as well as the L1 distance as a pixel-wise similarity measurement. As shown in Table 3, our method demonstrates significantly higher invariance towards these transformations compared to other motion transfer methods on all metrics, indicating a high degree of appearance and structural disentanglement. We also notice the positive effect that the augmentations provide during training. While a model trained without augmentations nevertheless demonstrates substantially higher invariance compared to other baselines, a model trained with augmentations performs notably better, especially with regard to geometric transformations.

**Table 4: Generation Similarity under Driving Video Augmentations.** We measure invariance between generated videos driven by unaltered and augmented input videos using both photometric and geometric transformations. *DisMo* showcases a significantly higher degree of invariance towards these alternations compared to the baselines. This effect is reinforced when applying augmentations during training.

Method	Photometric Augmentations				Geometric Augmentations			
	CLIP $\uparrow$	SSIM $\uparrow$	L1 $\downarrow$	LPIPS $\downarrow$	CLIP $\uparrow$	SSIM $\uparrow$	L1 $\downarrow$	LPIPS $\downarrow$
DMT [Yatim et al., 2024]	0.66	0.51	0.17	0.59	0.66	0.48	0.18	0.62
MotionDirector [Zhao et al., 2024]	0.62	0.46	0.15	0.55	0.60	0.44	0.16	0.60
VMC [Jeong et al., 2024]	0.71	0.44	0.17	0.58	0.71	0.40	0.20	0.62
MotionClone [Ling et al., 2024]	0.76	0.52	0.13	0.46	0.72	0.43	0.18	0.58
<i>DisMo</i> (w/o augmentations)	<u>0.89</u>	<u>0.70</u>	<u>0.09</u>	<u>0.26</u>	<u>0.88</u>	<u>0.65</u>	<u>0.11</u>	<u>0.33</u>
<i>DisMo</i> ( <i>Ours</i> )	<b>0.90</b>	<b>0.72</b>	<b>0.08</b>	<b>0.25</b>	<b>0.90</b>	<b>0.71</b>	<b>0.08</b>	<b>0.26</b>

### 4.3 Action Classification

*DisMo*'s motion embeddings, while particularly useful for generative motion transfer tasks, demonstrate significant generalization potential for discriminative downstream applications due to their abstract representation of dynamic properties. To investigate their generalization and effectiveness, we evaluate *DisMo* embeddings on a discriminative task, namely action classification.

The strengths of *DisMo*'s motion embeddings are particularly apparent in zero-shot action classification tasks on datasets where motion is a primary discriminative cue. We benchmark our method against state-of-the-art video representation learning approaches on datasets that emphasize temporal dynamics and motion patterns, making them ideal for evaluating motion-centric representations. Specifically, we evaluate on ARID [Xu et al., 2021], which tests robustness under low-light conditions; IARD [Taccetti et al., 2019], designed to assess identity-invariant motion understanding; Jester [Materzynska et al., 2019], which comprises short hand gesture clips for evaluating fine-grained local motion sensitivity; and Something-Something V2 [Goyal et al., 2017], a large-scale dataset focused on nuanced temporal interactions.

**Table 5: Zero-shot Action Classification.** We compare *DisMo* with state-of-the-art baselines in frozen evaluation with a  $k$ NN probe on multiple downstream action classification datasets. Extracted features are mean-pooled before being passed to the  $k$ NN classifier. All models are evaluated at resolution 224<sup>2</sup> using a single clip from the video. Compared to other video baselines, *DisMo* exhibits a consistent improvement across all downstream datasets, showcasing the superiority of abstract motion representations on zero-shot frozen evaluation over competing video representations.

Method	Arch.	Params.	Samples seen	ARID $\uparrow$	Jester $\uparrow$	SSv2 $\uparrow$	IARD $\uparrow$
VideoMAE [Tong et al., 2022]	ViT-L/16	343M	410M	17.29	20.11	7.06	73.44
VideoMAE v2 [Wang et al., 2023]	ViT-L/16	304M	1.08B	32.61	43.83	16.56	80.25
V-JEPA [Bardes et al., 2024]	ViT-L/16	200M	270M	25.16	30.84	21.11	82.03
<i>DisMo</i> ( <i>Ours</i> )	DisMo-B	172M	17M <sup>†</sup>	<b>57.29</b>	<b>56.66</b>	<b>22.19</b>	<b>90.74</b>

<sup>†</sup> The number of samples reported here refers only to video clips and does not include images used to pre-train the DINOv2 and DiT backbones.

Our model demonstrates strong performance across a variety of zero-shot action classification benchmarks, with a notable lead in the k-NN setting (see Table 5). On datasets like ARID, Jester, and SSv2, which all require understanding temporal motion patterns and generalize across diverse visual conditions, *DisMo* consistently outperforms state-of-the-art baselines. For instance, *DisMo* improves k-NN accuracy by a large margin on ARID (+24.7 points over V-JEPA) and achieves the highest scores on Jester and SSv2, indicating robust motion-sensitive representations. In the identity-conditioned IARD setting, *DisMo* achieves the highest k-NN accuracy (90.74%), outperforming all baselines, reflecting a cleaner, less entangled representation, without any supervision or adaptation.

Our identity reconstruction analysis (Section 4.2) further supports this interpretation. We observed that *DisMo* encodes substantially less identity information than other models, which often reach near-perfect identity reconstruction. This property of disentangling motion from appearance enables better generalization across individuals and tasks, as the representations focus on action dynamics rather than performer-specific features. These results collectively underscore the promise of motion-centric representation learning for building generalizable and robust video models.

## 5 Conclusion

In this work, we introduce *DisMo*, a novel paradigm for learning abstract motion representations from video data using a purely reconstruction-based objective. We demonstrate that these representations are inherently decoupled from appearance, making them well-suited for motion transfer. *DisMo* outperforms existing open-world transfer methods on qualitative and quantitative evaluations. Furthermore, we demonstrate that the learned embeddings are also effective for downstream tasks such as action classification, outperforming state-of-the-art video representation models in a zero-shot setting. While our analysis focuses on zero-shot probing, exploring more advanced techniques such as linear or attentive probing could offer deeper insights into the structure and capabilities of the learned motion space, which we leave for future work.

## Acknowledgements

This project has been supported by the project “GeniusRobot” (01IS24083) funded by the Federal Ministry of Research, Technology and Space (BMFTR), the bidt project KLIMA-MEMES, the German Federal Ministry for Economic Affairs and Energy (BMWE) within the project “NXT GEN AI METHODS – Generative Methoden für Perzeption, Prädiktion und Planung”, and the Horizon Europe project ELLIOT (GA No. 101214398). The authors gratefully acknowledge the Gauss Center for Supercomputing for providing compute through the NIC on JUWELS/JUPITER at JSC and the HPC resources supplied by the NHR@FAU Erlangen.

## References

- Gunnar Johansson. Visual perception of biological motion and a model for its analysis. *Perception & psychophysics*, 14:201–211, 1973.
- Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.
- Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. URL <https://openai.com/research/video-generation-models-as-world-simulators>.
- Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.
- Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanyvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.
- Veo-Team, :, Agrim Gupta, Ali Razavi, Andeep Toor, Ankush Gupta, Dumitru Erhan, Eleni Shaw, Eric Lau, Frank Belletti, Gabe Barth-Maron, Gregory Shaw, Hakan Erdogan, Hakim Sidahmed, Henna Nandwani, Hernan Moraldo, Hyunjik Kim, Irina Blok, Jeff Donahue, José Lezama, Kory Mathewson, Kurtis David, Matthieu Kim Lorrain, Marc van Zee, Medhini Narasimhan, Miaosen Wang, Mohammad Babaeizadeh, Nelly Papalampidi, Nick Pezzotti, Nilpa Jha, Parker Barnes, Pieter-Jan Kindermans, Rachel Hornung, Ruben Villegas, Ryan Poplin, Salah Zaiem, Sander Dieleman, Sayna Ebrahimi, Scott Wisdom, Serena Zhang, Shlomi Fruchter, Signe Nørly, Weizhe Hua, Xinchen Yan, Yuqing Du, and Yutian Chen. Veo 2. 2024. URL <https://deepmind.google/technologies/veo/veo-2/>.
- Eric R. Kandel, James H. Schwartz, and Thomas M. Jessell, editors. *Principles of Neural Science*. Elsevier, New York, third edition, 1991.
- Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jia-Wei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. Motiondirector: Motion customization of text-to-video diffusion models. In *European Conference on Computer Vision*, pages 273–290. Springer, 2024.
- Hila Chefer, Uriel Singer, Amit Zohar, Yuval Kirstain, Adam Polyak, Yaniv Taigman, Lior Wolf, and Shelly Sheynin. Videojam: Joint appearance-motion representations for enhanced motion generation in video models. *arXiv preprint arXiv:2502.02492*, 2025.
- Jingyun Liang, Yuchen Fan, Kai Zhang, Radu Timofte, Luc Van Gool, and Rakesh Ranjan. Movideo: Motion-aware video generation with diffusion model. In *European Conference on Computer Vision*, pages 56–74. Springer, 2024.
- Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. In *European Conference on Computer Vision*, pages 18–35. Springer, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023a.

Nick Stracke, Stefan Andreas Baumann, Joshua Susskind, Miguel Angel Bautista, and Björn Ommer. Ctrloralter: Conditional loradapter for efficient 0-shot control and altering of t2i models. In *ECCV*, pages 87–103. Springer, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, David Yan, Dhruv Choudhary, Dingkang Wang, Geet Sethi, Guan Pang, Haoyu Ma, Ishan Misra, Ji Hou, Jialiang Wang, Kiran Jagadeesh, Kunpeng Li, Luxin Zhang, Mannat Singh, Mary Williamson, Matt Le, Matthew Yu, Mitesh Kumar Singh, Peizhao Zhang, Peter Vajda, Quentin Duval, Rohit Girdhar, Roshan Sumbaly, Sai Saketh Rambhatla, Sam Tsai, Samaneh Azadi, Samyak Datta, Sanyuan Chen, Sean Bell, Sharadh Ramaswamy, Shelly Sheynin, Siddharth Bhattacharya, Simran Motwani, Tao Xu, Tianhe Li, Tingbo Hou, Wei-Ning Hsu, Xi Yin, Xiaoliang Dai, Yaniv Taigman, Yaqiao Luo, Yen-Cheng Liu, Yi-Chiao Wu, Yue Zhao, Yuval Kirstain, Zecheng He, Zijian He, Albert Pumarola, Ali Thabet, Artsiom Sanakoyeu, Arun Mallya, Baishan Guo, Boris Araya, Breena Kerr, Carleigh Wood, Ce Liu, Cen Peng, Dimitry Vengertsev, Edgar Schonfeld, Elliot Blanchard, Felix Juefei-Xu, Fraylie Nord, Jeff Liang, John Hoffman, Jonas Kohler, Kaolin Fire, Karthik Sivakumar, Lawrence Chen, Licheng Yu, Luya Gao, Markos Georgopoulos, Rashel Moritz, Sara K. Sampson, Shikai Li, Simone Parmeggiani, Steve Fine, Tara Fowler, Vladan Petrovic, and Yuming Du. Movie gen: A cast of media foundation models, 2025. URL <https://arxiv.org/abs/2410.13720>.

Yash Jain, Anshul Nasery, Vibhav Vineet, and Harkirat Behl. Peekaboo: Interactive video generation via masked-diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8079–8088, 2024.

Jiawei Wang, Yuchen Zhang, Jiaxin Zou, Yan Zeng, Guoqiang Wei, Liping Yuan, and Hang Li. Boximator: Generating rich and controllable motions for video synthesis. *arXiv preprint arXiv:2402.01566*, 2024a.

Jianzong Wu, Xiangtai Li, Yanhong Zeng, Jiangning Zhang, Qianyu Zhou, Yining Li, Yunhai Tong, and Kai Chen. Motionbooth: Motion-aware customized text-to-video generation. *Advances in Neural Information Processing Systems*, 37:34322–34348, 2025.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024b.

Weijia Wu, Zhuang Li, Yuchao Gu, Rui Zhao, Yefei He, David Junhao Zhang, Mike Zheng Shou, Yan Li, Tingting Gao, and Di Zhang. Draganything: Motion control for anything using entity representation. In *European Conference on Computer Vision*, pages 331–348. Springer, 2024.

- Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*, 2023.
- Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077*, 2023b.
- Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7346–7356, 2023.
- Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in neural information processing systems*, 32, 2019.
- Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13653–13662, 2021.
- Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10039–10049, 2021.
- Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3657–3666, 2022.
- Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. Latent image animator: Learning to animate images via latent space navigation. *arXiv preprint arXiv:2203.09043*, 2022.
- Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu. Depth-aware generative adversarial network for talking head video generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3397–3406, 2022.
- Kewei Yang, Kang Chen, Daoliang Guo, Song-Hai Zhang, Yuan-Chen Guo, and Weidong Zhang. Face2face  $\rho$ : Real-time high-resolution one-shot face reenactment. In *European conference on computer vision*, pages 55–71. Springer, 2022.
- Yue Gao, Yuan Zhou, Jinglu Wang, Xiao Li, Xiang Ming, and Yan Lu. High-fidelity and freely controllable talking head video generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5609–5619, 2023.
- Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models, 2022. URL <https://arxiv.org/abs/2210.02303>.
- Diederik Kingma and Ruiqi Gao. Understanding diffusion objectives as the elbo with simple data augmentation. *Advances in Neural Information Processing Systems*, 36:65484–65516, 2023.
- Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.
- Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.
- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607. PMLR, 2020.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL <https://openreview.net/forum?id=YicbFdNTTy>.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision, April 2023. URL <http://arxiv.org/abs/2304.07193>. arXiv:2304.07193 [cs].

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL <https://arxiv.org/abs/1711.05101>.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Limin Wang, and Yu Qiao. UniformerV2: Spatiotemporal learning by arming image vits with video uniformer, 2022. URL <https://arxiv.org/abs/2211.09552>.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The "something something" video database for learning and evaluating visual common sense. In *CVPR*, pages 5842–5850, 2017.

Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfruend, Carl Vondrick, and Aude Oliva. Moments in time dataset: one million videos for event understanding, 2019a. URL <https://arxiv.org/abs/1801.03150>.

Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. In *ICLR*, 2025. URL <https://openreview.net/forum?id=j7kdXSrISM>.

Hyeonho Jeong, Geon Yeong Park, and Jong Chul Ye. Vmc: Video motion customization using temporal attention adaption for text-to-video diffusion models. In *CVPR*, pages 9212–9221, 2024.

Danah Yatim, Rafail Fridman, Omer Bar-Tal, Yoni Kasten, and Tali Dekel. Space-time diffusion features for zero-shot text-driven motion transfer. In *CVPR*, pages 8466–8476, 2024.

Pengyang Ling, Jiazi Bu, Pan Zhang, Xiaoyi Dong, Yuhang Zang, Tong Wu, Huaian Chen, Jiaqi Wang, and Yi Jin. Motionclone: Training-free motion cloning for controllable video generation. *arXiv preprint arXiv:2406.05338*, 2024.

Geon Yeong Park, Hyeonho Jeong, Sang Wan Lee, and Jong Chul Ye. Spectral motion alignment for video motion transfer using diffusion models. *arXiv preprint arXiv:2403.15249*, 2024.

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

Andrea Tacchetti, Leyla Isik, and Tomaso Poggio. Invariant Action Recognition Dataset, 2019. URL <https://doi.org/10.7910/DVN/DMTOPG>.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14549–14560, 2023.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video, 2024. URL <https://arxiv.org/abs/2404.08471>.

- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.
- BovikAC WangZhou, Hamid R Sheikh, et al. Image qualityassessment: From errorvisibilitytostructural similarity. *IEEE Transon ImageProcessing*, 13(4):600, 2004.
- Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.
- Yuecong Xu, Jianfei Yang, Haozhi Cao, Kezhi Mao, Jianxiong Yin, and Simon See. Arid: A new dataset for recognizing action in the dark. In *Deep Learning for Human Activity Recognition: Second International Workshop, DL-HAR 2020, Held in Conjunction with IJCAI-PRICAI 2020, Kyoto, Japan, January 8, 2021, Proceedings 2*, pages 70–84. Springer, 2021.
- Joanna Materzynska, Guillaume Berger, Ingo Bax, and Roland Memisevic. The jester dataset: A large-scale video dataset of human gestures. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019.
- Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.
- Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfruend, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE TPAMI*, pages 1–8, 2019b. ISSN 0162-8828. doi: 10.1109/TPAMI.2019.2901464.
- H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011.
- Noam Shazeer. Glu variants improve transformer, 2020. URL <https://arxiv.org/abs/2002.05202>.
- Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 69(6):066138, 2004.
- Brian C. Ross. Mutual information between discrete and continuous data sets. *PLOS ONE*, 9(2): 1–5, 02 2014. doi: 10.1371/journal.pone.0087357. URL <https://doi.org/10.1371/journal.pone.0087357>.
- Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Sparsectrl: Adding sparse controls to text-to-video diffusion models, 2023. URL <https://arxiv.org/abs/2311.16933>.
- Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer, 2025. URL <https://arxiv.org/abs/2408.06072>.
- Qiujiang Wang, Yukai Shi, Jiarong Ou, Rui Chen, Ke Lin, Jiahao Wang, Boyuan Jiang, Haotian Yang, Mingwu Zheng, Xin Tao, Fei Yang, Pengfei Wan, and Di Zhang. Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content, 2025. URL <https://arxiv.org/abs/2410.08260>.
- Christoph Feichtenhofer, Axel Pinz, and Richard P Wildes. Temporal residual networks for dynamic scene recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4728–4737, 2017.

## A Broader Societal Impact

While this work aims to advance realistic motion transfer with potential applications in digital storytelling, entertainment, and educational purposes, such progress also carries risks of misuse to generate harmful content without consent. We emphasize transparency in datasets and evaluation, and encourage ongoing efforts to address biases and promote responsible deployment.

## B Limitations

Our work is limited by the capabilities of the flow-matching-based decoder, which, while efficient, may struggle to generate high-fidelity or semantically rich images in complex scenarios. Additionally, the performance is constrained by the distribution and inherent biases of the training dataset, which can affect generalization, fairness, and robustness to out-of-distribution samples.

## C Implementation Details

### C.1 Pretraining details

**Architectures** We use DINOv2-B as our frame embedder. As our sequence embedder we use a 3D ViT-B/16 with RoPe positional embeddings. As our frame generator, we use a DiT-XL. The frame embedder takes as input a short video clip of 8 frames, sampled at 6 fps.

Table A: Pretraining hyper-parameters for our model.

Category	Hyper-parameter	Value
data	datasets	K-710/SSv2/MiT/OpenVid-1m
	resolution	256 × 256
	num_frames	8
	fps	6
geometric augmentations	scale	(0.75, 1.0)
	translate	(-0.3, 0.3)
	angle	(-30, 30)
	shear	(-15, 15)
	horizontal flip	✓
	padding mode	border
photometric augmentations	brightness	(0.5, 1.5)
	contrast	(0.5, 1.5)
	hue	(-0.3, 0.3)
	saturation	(0.5, 1.5)
optimization	batch_size	32
	total number of iterations	530k
	warmup iterations	5000
	lr scheduler	constant w/ warmup
	lr	$10^{-4}$
	AdamW $\beta$	(0.9, 0.95)
architecture	frame_embed_depth	12
	frame_embed_dim	768
	sequence_embed_depth	12
	sequence_embed_dim	768
	frame_generator_depth	28
	frame_generator_dim	1152
hardware	dtype	bfloat16
	accelerator	GH200 96G

**Training Data** To cover a wide variety of real-world videos, we train on a mixture of datasets covering over videos. This mixture is composed of Kinetics-710 [Kay et al., 2017], which combines Kinetics-400, 600 and 700, Something-Something-v2 (SSv2) [Goyal et al., 2017], Moments in Time

(MiT) [Monfort et al., 2019b], and OpenVid1M [Nan et al., 2025]. We show an overview of the size and domain of the various datasets in Table B. Overall, this dataset composition is similar to that used by V-JEPA [Bardes et al., 2024], which among others combines Kinetics-710 with Something-Something-v2. We further extend this mixture with the high-quality OpenVid-1M [Nan et al., 2025] dataset, which is more suitable for generative tasks compared to the other datasets.

Dataset	Domain	# Clips	Duration [h]
Kinetics-710 [Kay et al., 2017]	human-centered	0.65M	1.8k
Something-Something-v2 [Goyal et al., 2017]	open	108k	121
Moments in Time [Monfort et al., 2019b]	open	1M	833
OpenVid-1M [Nan et al., 2025]	open	1M	2.1k
<i>Combined</i>	open	2.8M	4.9k

Table B: Training Data Composition.

**Kinetics-710 Overlaps** There are significant overlaps between train and evaluation splits of the various subsets (K-400, K-600, K-700) that comprise the Kinetics-710 dataset. To enable fair comparisons, we remove any clips present in any of the three validation or test sets from our training dataset. We also de-duplicate the remaining training dataset by their respective YouTube video ids.

## C.2 Evaluation details

**Datasets** Something-Something v2 (SSv2) [Goyal et al., 2017] is a large dataset of human-object interactions designed to test fine-grained motion understanding while minimizing reliance on background or object identity. Kinetics-400 [Kay et al., 2017] contains 240,000 YouTube clips labeled with 400 action classes and serves as a general-purpose benchmark for appearance-driven video understanding. ARID [Xu et al., 2021] (Action Recognition in the Dark) includes 5,500 RGB-D videos captured in low-light settings to evaluate robustness to visibility degradation. Jester [Materzynska et al., 2019] contains over 148,000 short clips of hand gestures, providing a benchmark for evaluating fine-grained, local temporal motion sensitivity. IARD [Tacchetti et al., 2019] (Invariant Action Recognition Dataset) includes controlled recordings of five actors performing five actions from multiple viewpoints, enabling analysis of identity- and viewpoint-invariant motion recognition. HMDB51 [Kuehne et al., 2011] is a widely used benchmark for human action recognition, consisting of  $\sim 7,000$  video clips distributed across 51 action categories, collected from various sources including movies and YouTube, with diverse scene complexity and motion dynamics.

**LTX Finetuning details** We adapt both the attention and feedforward layers of LTX using conditional LoRA modules with a rank of 64. Each LoRA module receives a motion sequence as input. We use a batch size of 64 for training. For optimization, we use the AdamW optimizer with  $\beta_1 = 0.9$  and  $\beta_2 = 0.99$ , a learning rate of  $5 \times 10^{-5}$ , weight decay of 0.0001, and training is performed with 32-bit precision.

LTX operates on high frame rate videos and temporally downsamples them by a factor of 8 using its VAE. In contrast, our motion extractor processes videos at a lower frame rate, taking only 8 frames as input. To ensure alignment between the motion sequence and the LTX latent representations, we apply the following strategy:

- We sample a video at approximately 24 fps with a total of 29 frames, where the first frame serves as the source frame.
- These 29 frames are chosen such that we can extract 8 frames spaced 4 frames apart: the source frame (frame 0), followed by frames 4, 8, 12, 16, 20, 24, and 28. This satisfies the equation:

$$1 + 7 \times 4 = 29$$

- LTX downsamples this video by encoding the source frame and every 8<sup>th</sup> frame thereafter (frames 0, 8, 16, 24) into a latent representation.
- To align motion embeddings with these latents, we extract every 4<sup>th</sup> frame (frames 0, 4, 8, 12, 16, 20, 24, 28) and pass them through our motion extractor.

- For each pair of consecutive motion embeddings (except for the first one, which remains a singular embedding), we concatenate them to form a combined embedding. This results in a motion sequence of 4 embeddings aligned with the 4 LTX latents.
- This sequence is then transformed by a mapping network and used to condition the LoRA modules. The mapping network consists of two feedforward layers, each applying RMS normalization at the beginning and end, and using a Linear-GEGLU [Shazeer, 2020] activation. Each feedforward layer expands the dimensionality by a factor of 3 and subsequently reduces it back, operating at a width of 768.

**Identity Reconstruction from Motion Representations** We provide additional details on the identity classification experiment on the IARD dataset. We selected videos of 5 individuals and used only four actions (['eat', 'run', 'walk', 'jump']) for training, holding out the 'drink' action for testing. Motion representations were extracted using VideoMAE, VideoMAEv2, and Video JEPA. Videos were processed at 6 FPS with 16-frame windows, resized to  $224 \times 224$ , and uniformly preprocessed across models. Temporal features were averaged to produce a single embedding per video. A k-NN classifier ( $k = 20$ ) was trained to predict identity based on the training actions and evaluated on the held-out 'drink' action to assess the extent of identity information encoded in the representations.

**IARD identity-based split for action classification** In this experiment on the Invariant Action Recognition Dataset (IARD), we used an identity-based split to better test the generalization of motion representation models across individuals. The dataset includes 5 people performing 5 actions (walk, run, jump, drink, eat). We trained models on videos from four individuals (Andrea, Leyla, Gu, Georgios) and tested on videos from the fifth (Steve), ensuring no overlap in identity between train and test. This split contrasts with standard random splits and provides a more realistic measure of how well models capture action dynamics independently of personal appearance cues.

### C.3 Ablation Study

During training, we introduce several key strategies to enhance disentanglement between motion and appearance in our model. Additionally, we apply dropout to motion and frame conditioning, which serves a dual purpose: it stabilizes the training process and allows for subsequent evaluation of the frame-unconditional reconstructor. This evaluation is conducted by reconstructing frames solely from motion embeddings, without previous frame conditioning.

To comprehensively measure performance during ablation, we employ three complementary metrics: LPIPS [Zhang et al., 2018], CLIP [Radford et al., 2021] score, and Mutual Information Ratio (MIR). Ideally, if no appearance information leaks into the motion embeddings, we expect poorer reconstructions, indicated by higher LPIPS scores and lower CLIP similarities. We use a subset of the HMDB-51 dataset for LPIPS and CLIP, and a subset of the IARD dataset for MIR.

With LPIPS we quantify the perceptual similarity with respect to visual realism, whereas with CLIP we measure the overall semantic similarity between the reconstructed frame and the original one. MIR provides a formal quantification of disentanglement by comparing the mutual information between the motion embeddings and two distinct factors – actions and identities. Formally, we define MIR as:

$$\text{MIR} = \frac{I(z; c_{\text{action}})}{I(z; c_{\text{identity}})}, \quad (1)$$

where  $c_{\text{action}}$  and  $c_{\text{identity}}$  are the respective labels for embeddings  $z$  and the mutual information  $I(\cdot; \cdot)$  measures the shared information between variables, defined as:

$$I(z; y) = H(y) - H(y | z), \quad (2)$$

with entropy  $H(\cdot)$  quantifying uncertainty. We estimate entropy using  $k$ -NN ( $k = 5$ ) distances, as described in [Kraskov et al., 2004] and [Ross, 2014].

Table C summarizes the outcomes of our ablation study. Starting from a baseline model without conditioning or augmentations, adding previous frame conditioning significantly prevents appearance leaking and improves disentanglement. Integrating augmentations further boosts these improvements, especially regarding disentanglement. Collectively, these results underscore the efficacy and necessity of each component in achieving robust disentanglement between motion and appearance.

**Table C: Ablation Study.** Comparison of model variants with and without previous frame conditioning and augmentations.

Model	LPIPS $\uparrow$	CLIP $\downarrow$	MIR $\uparrow$
Baseline	0.47	0.81	0.47
+ Previous frame	0.72	0.65	3.07
+ Augmentations (ours)	0.76	0.65	5.56

**Table D: Disentanglement.** Comparison of appearance disentanglement performance between our model and V-JEPA.

Model	LPIPS $\uparrow$	CLIP $\downarrow$	MIR $\uparrow$
V-JEPA	0.70	0.70	3.72
Ours	<b>0.76</b>	<b>0.65</b>	<b>5.56</b>

To enable a direct performance comparison with V-JEPA, we trained an identical reconstructor using the same setup as our model, with the motion extractor replaced by a frozen V-JEPA. As shown in Table D, our model achieves superior disentanglement.

We additionally ablate *DisMo* with other high-quality video generation backbones, namely SparseCtrl [Guo et al., 2023] and CogVideoX-5B [Yang et al., 2025]. As shown in Table E, using a more capable generator like CogVideoX-5B leads to improved motion fidelity and temporal consistency, without any modification to the motion embeddings produced by *DisMo*. This demonstrates that *DisMo* is generally applicable to off-the-shelf video models without being constrained to a specific one. We further assessed the quality and realism of the generated videos using FID and FVD scores, observing consistent improvements with stronger video backbones. This demonstrates that *DisMo* benefits from more powerful architectures and remains compatible with future models. For this evaluation, we generate 2,000 videos per model, each 25 frames long, yielding 50,000 frames in total for FID computation. Motion is transferred from randomly sampled videos and prompts from the Koala-36M [Wang et al., 2025] dataset.

**Table E: Ablating Video Generation Backbones for Motion Transfer.** Effect of different video generation backbones on temporal consistency, motion fidelity, and visual quality metrics (FID, FVD) of generated motion transfers. Using a stronger generator like CogVideoX improves motion fidelity and temporal consistency without altering *DisMo*’s motion embeddings.

Model	Temporal Consistency $\uparrow$	Motion Fidelity $\uparrow$	FID $\downarrow$	FVD $\downarrow$
LTX [HaCohen et al., 2024]	0.95	0.75	88.54	298.4
SparseCtrl [Guo et al., 2023]	0.95	0.73	116.8	327.9
CogVideoX-5B [Yang et al., 2025]	<b>0.97</b>	<b>0.78</b>	<b>62.99</b>	<b>206.3</b>

## D Latent Space Analysis

### D.1 Cluster Visualization

We visualize embeddings from both *DisMo* and V-JEPA on the IARD dataset, which provides action and identity labels for each video sample. Using UMAP for dimensionality reduction and coloring the resulting sample points by either action or identity reveals notable differences between the two models, see Fig. A. When colored by action, embeddings from our model form well-defined clusters, while coloring by identity results in no clear grouping—indicating strong sensitivity to motion and minimal association with appearance. In contrast, V-JEPA’s embeddings show little to no clustering by either action or identity. Instead, its clusters primarily reflect viewpoint differences, consistent with the five distinct camera angles present in the dataset.

### D.2 Cycle Identification

Fig. B shows PCA visualizations of individual videos from three different classes in the IARD dataset: eat, walk, and run. The repetitive nature of these actions produces distinct cyclic patterns in the latent space. Slower movements, such as eating, yield smoother and more continuous trajectories, while faster actions like walking result in less smooth, more abrupt trajectories. None-repeating motions, such as someone diving into a pool, do not result in cycles. This highlights how differences in motion intensity and cycling movements are captured in our model’s latent representations.

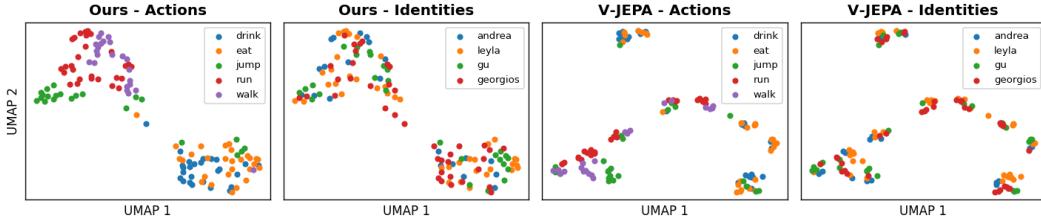


Figure A: **UMAP visualization of the IARD dataset.** Compared to V-JEPA, our model shows better grouping by action and almost no grouping by identity.

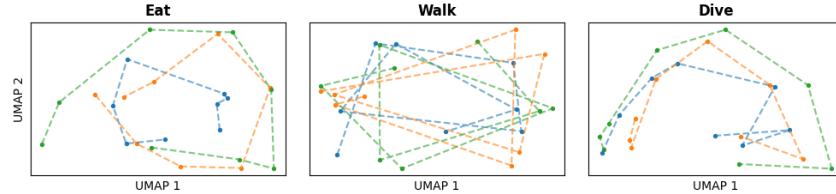


Figure B: **PCA cycle visualization of three different classes from the IARD dataset.** Three videos per class, indicated by different colors, showing clear cycles in our model’s latent representations.

### D.3 Reversibility

A task that is solely based on motion, where appearance provides no useful cues, is determining whether a video is played forward or backward. In such cases, temporal directionality must be inferred from the dynamics of the scene alone. To test this, we conducted an experiment where each video sample is paired with a temporally reversed version. We then visualize the latent representations using UMAP. As shown in Fig. C, our model separates forward and reversed sequences in the latent space, but only when the action is inherently irreversible, such as diving into a pool. In contrast, when the motion is reversible e.g. someone doing push-ups, no meaningful separation emerges. V-JEPA, on the other hand, fails to distinguish forward and reversed sequences in both reversible and irreversible cases. Its latent space shows no significant separation, suggesting its representations are largely appearance biased.

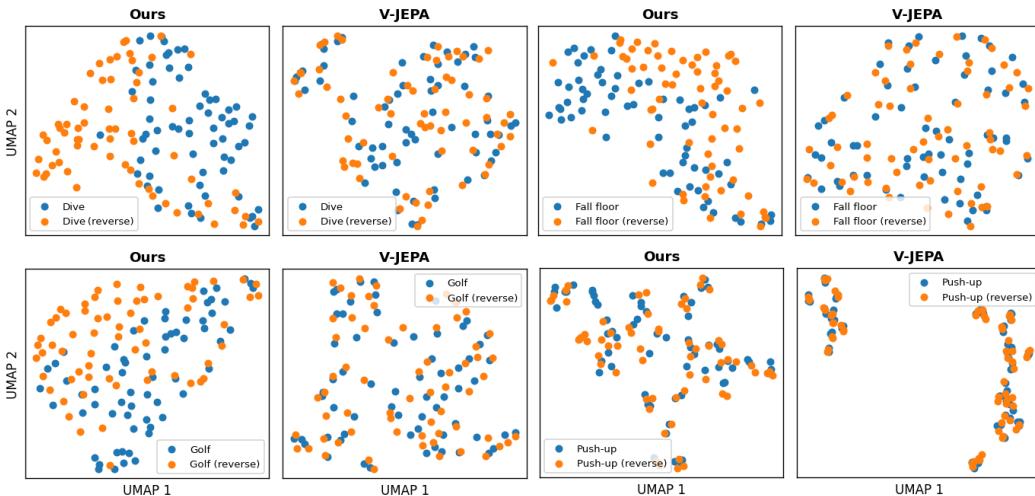


Figure C: **UMAP visualizations of reversible motion and irreversible motion on the HMDB-51 dataset.** Compared to V-JEPA, our model shows a better separation for irreversible motion.

## E Composite Motion Analysis

We designed a targeted experiment to evaluate the composability of DisMo’s motion embedding space. To assess which aspects of motion are encoded in the embeddings, we use the YUP++ dataset [Feichtenhofer et al., 2017], which contains 600 static videos without camera motion. We then construct three types of video clips:

1. **Object Motion Only.** A randomly sampled clip from the dataset, containing natural object motion but no camera movement.
2. **Camera Motion Only.** A different clip, augmented with temporally smooth geometric transformations simulating camera motion, applied only to the initial frame (thus eliminating object motion).
3. **Combined Motion.** The sampled clip with both object motion and the same simulated camera motion applied across all frames.

This setup allows us to isolate the two motion components, camera and object, and test whether embeddings extracted from the combined motion videos contain information from both individual motion types. We sample 3200 of such triplets, each containing the three types of clips mentioned before, and encode each of them using DisMo’s motion encoder. We then estimate the mutual information (MI) between the embeddings of the combined-motion clip ( $\mathbf{M}_{both}$ ) and those of the camera-only  $\mathbf{M}_{cam}$  and object-only  $\mathbf{M}_{obj}$  clips, respectively. High mutual information would suggest that  $\mathbf{M}_{both}$  shares information with  $\mathbf{M}_{cam}$  and  $\mathbf{M}_{obj}$ , which is a key indication of composability.

To quantify this, we use the Kraskov–Stögbauer–Grassberger (KSG) estimator [Kraskov et al., 2004] with  $k = 5$ . To compute statistical significance, we further estimate MI under 100 random permutations (i.e., computing MI between unpaired clips), yielding a null distribution with means of around 0.2 nats and standard deviations of around 0.4. The observed MI values between both ( $\mathbf{M}_{both}, \mathbf{M}_{cam}$ ) and ( $\mathbf{M}_{both}, \mathbf{M}_{obj}$ ) are depicted in Table F. They are significantly higher than the null baseline, indicating a strong statistical dependency and thus supporting the hypothesis that DisMo’s latent space is compositional with regard to simultaneous camera and object motion.

Table F: **Motion composability.** We compute the mutual information (MI) between motion embeddings extracted from videos with both camera and object motion, and embeddings extracted from videos consisting of only camera and object motion, respectively. We observe a high significance compared to a random baseline for both types of motion. Object motion MI is significant even with strong camera motion, albeit decreasing the heavier the added camera motion gets.

MI (nats)	Light Camera Motion	Medium Camera Motion	Heavy Camera Motion
Camera Motion	1.37	2.25	4.40
Object Motion	3.59	2.97	1.57

## F Motion Transfer Efficiency

We report inference time and number of parameters for DisMo and other baselines during motion transfer in Table G. DMT, VMC, and MotionDirector rely on per-sample fine-tuning or optimization, which significantly decreases inference efficiency. In contrast, our method is an off-the-shelf feed-forward generator. Across all configurations, we report DisMo to be the most efficient approach — except when paired with CogVideoX-5B as the video generation backbone, which increases inference time due to sheer model size. However, it is important to note that using CogVideoX-5B is not required to achieve strong motion transfer results: as evidenced by Table 1, DisMo combined with lighter backbones such as LTX-Video-2B already achieves better or comparable motion transfer fidelity.

## G Computational Resources for Experiments

This section gives an overview of the compute resources required to run the various individual experiments presented in this paper. We primarily use Nvidia GH200 96GB modules in an internal

Table G: **Inference Efficiency Comparison.** Inference time and parameter count of each model on an A100 GPU. Optimization-based baselines (VMC, DMT, MotionDirector) require per-sample fine-tuning, leading to slow inference, while our feed-forward *DisMo* variants enable substantially faster generation across all configurations.

Model	Inference Time (A100)	# Params
VMC	10 min / video	6B
DMT	7.5 min / video	1.7B
MotionClone	45 sec / video	0.983B
MotionDirector	~5 min / video	0.983B
DisMo-CogVideo	2 min / video	6.072B
DisMo-LTX	30 sec / video	2.172B
DisMo-SparseCtrl	10 sec / video	1.115B

<sup>†</sup> The reported number of parameters is approximate and refers to the largest component of each model, to the best of our knowledge. Exact values may vary depending on implementation details and auxiliary components.

cluster for pre-training and Nvidia H200 141GB in an internal cluster for evaluations. The project used approximately 600 H200-hrs across all experiments, including initial and other explorations that did not directly make it into the paper. In the following, we detail the required resources for each experiment.

- Pre-training: 192 H200 hours.
- Ablations:  $2 \times 96$  H200 hours.
- Action classification: 8 H200 hours
- Identity classification: 0.66 H200 hours
- Motion transfers: 17.2 H200 hours
- LTX fine-tuning: 200 H200 hours

## H Licenses of Used Assets

### H.1 Licences of video datasets

- **IARD**

License: CC0 1.0.

Terms: Community Norms in <https://dataverse.org/best-practices/dataverse-community-norms> as well as good scientific practices expect that proper credit is given via citation.

Source: <https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DMTOPG>

- **Something-Something v2 (SSv2)**

License: Research Use License (Qualcomm).

Terms: Available for research purposes only. Users must agree to Qualcomm's license terms before accessing the dataset.

Source: <https://www.qualcomm.com/developer/software/something-something-v-2-dataset>

- **Jester (Gesture Recognition)**

License: Research Use License (Qualcomm).

Terms: Available for research purposes only. Users must agree to Qualcomm's license terms before accessing the dataset.

Source: <https://www.qualcomm.com/developer/software/jester-dataset>

- **ARID (Action Recognition in the Dark)**

License: Creative Commons Attribution 4.0 International (CC BY 4.0).

Terms: Free to use, modify, and redistribute with proper attribution.

Source: <https://xuyu0010.github.io/arid.html>

- **OpenVid-1M**

License: Creative Commons Attribution 4.0 International (CC BY 4.0).

Terms: The video samples are collected from publicly available datasets. Users must follow the related licenses Panda, ChronoMagic, Open-Sora-plan, CelebVHQ(Unknow)) to use these video samples.

Source: <https://huggingface.co/datasets/nkp37/OpenVid-1M>

- **Kinetics-710**

License: Not explicitly specified.

Terms: Kinetics-710 is a video benchmark built upon Kinetics-400, -600, and -700. Its usage is subject to the licensing terms of these underlying datasets.

Source: <https://github.com/open-mmlab/mmaction2/blob/main/tools/data/kinetics710/README.md>

- **Moments in Time**

License: Not explicitly specified.

Terms: Intended for academic research.

Source: <http://moments.csail.mit.edu>

- **HMDB51**

License: Creative Commons Attribution 4.0 International (CC BY 4.0).

Terms: Free to use, modify, and redistribute with proper attribution.

Source: <https://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-database/>

## I Qualitative Comparisons on Text-based Motion Transfer

We present additional qualitative results for motion transfer and compare them with the previously introduced baseline methods. As most existing approaches do not support motion transfer from a source image, we restrict our comparison to examples generated solely from text prompts. All experiments were conducted using our adapted LTX model on videos that were not seen during training.

These results further demonstrate the effectiveness of our abstract motion representation in enabling motion transfer across diverse scenarios. Our approach consistently preserves both the high-level motion semantics and the content specified in the text prompt without relying on the structure of the driving video, a trade-off that competing methods often struggle to achieve.

In Fig. D, Fig. E, Fig. F, and Fig. G, we show that our approach outperforms existing methods in same-category motion transfer, especially when structural differences exist between the reference and generated videos. Additionally, our abstract motion representations enable effective cross-category transfer, allowing motion to be applied across semantically different domains. We demonstrate this capability in Fig. H, Fig. I and Fig. J, where motion is successfully transferred across categories. In contrast, existing methods often struggle in such scenarios, as they typically rely heavily on the structure of the input video and face a trade-off between accurately transferring motion and remaining faithful to the content described in the text prompt. Our method is decoupled from the reference structure and instead describes motion as the change between states. This becomes apparent in Fig. I, where DisMo applies the change of the pose, rather than aligning the structure. We showcase that our approach performs extraordinarily well for camera motion transfer while being completely independent of the reference structure. This capability is demonstrated in Fig. L, Fig. M and Fig. N, which highlight examples of successful camera motion transfer.



Figure D: **Same-category Transfer.** The reference video displays a man leaning towards the camera, rotating his torso and raising his arm. Our method is able to transfer these movements to a new identity with differing appearance, pose, and structure. Other methods either fail to capture the motion completely (**VMC**, **MD**) or rely heavily on the source structure (**DMT**, **MC**).

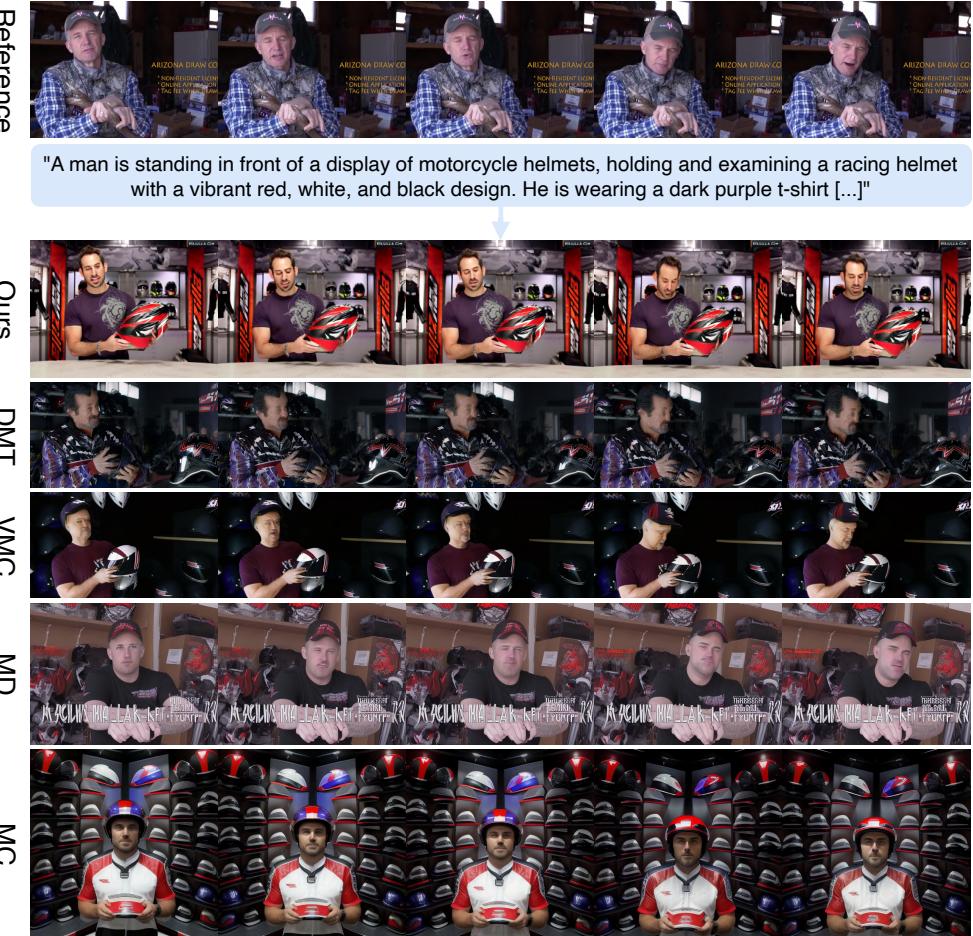
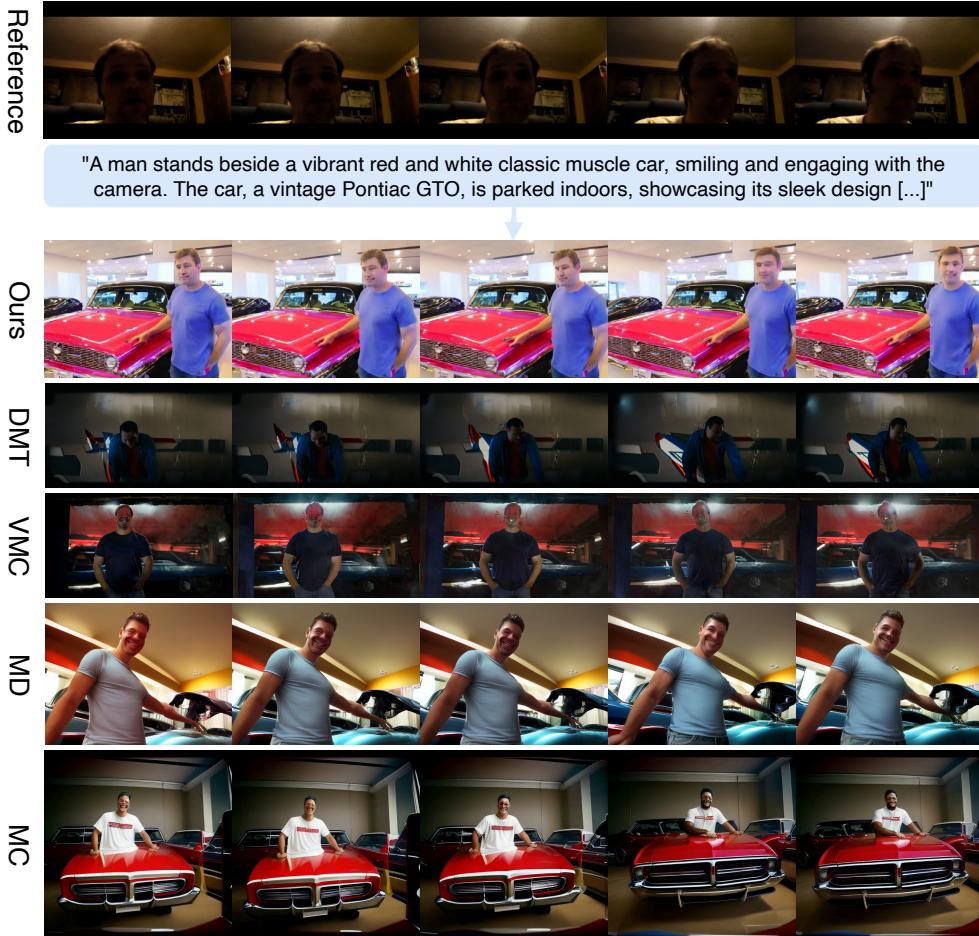


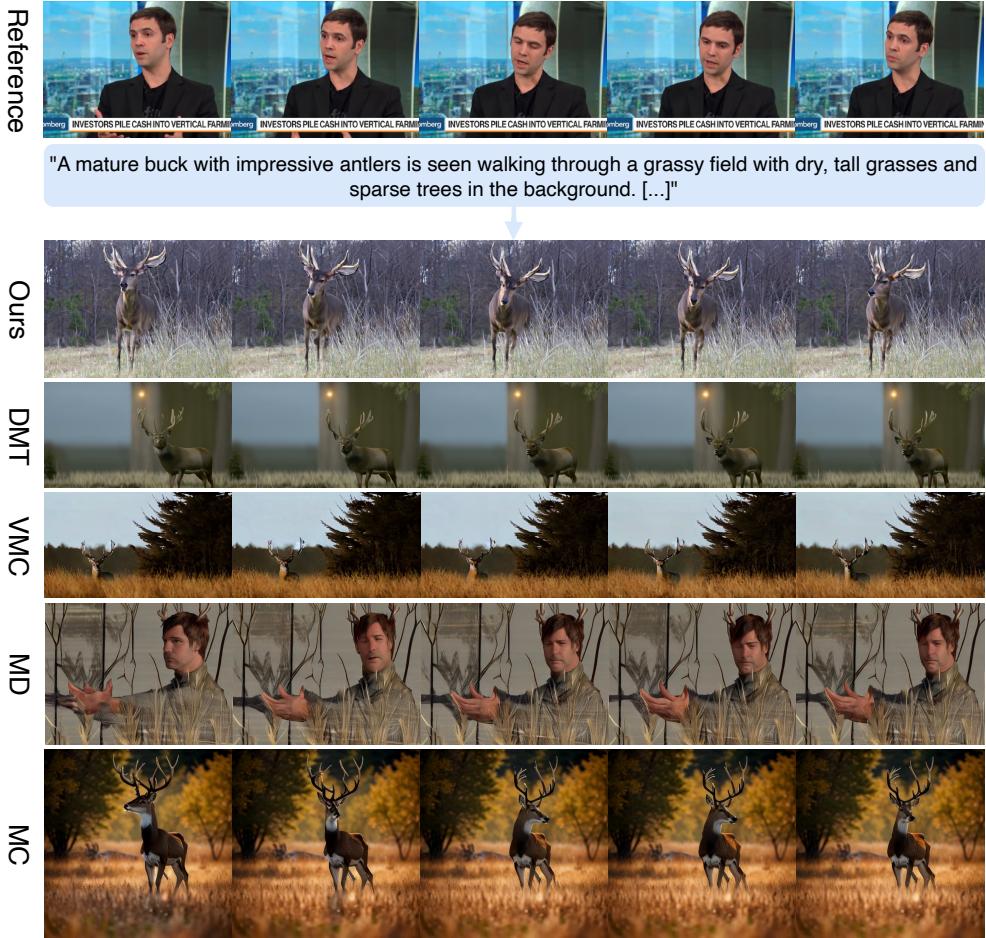
Figure E: **Same-category Transfer.** The reference video shows the upper torso of a person, moving their head and upper body. Our method successfully captures this motion and transfers it to the generated scene, closely following the instructions of the text prompt. Other methods fail to interpret the motion and the input prompt in a meaningful manner, the only exception being **VMC**.



**Figure F: Same-category Transfer.** The reference video depicts a talking head in a dark setting, slowly leaning to the left-hand side. Our approach effectively captures the details of the text prompt and accurately attaches the reference head movement to the semantically appropriate part of the target video, despite significant structural differences. In contrast, other methods fail to transfer the reference motion correctly.



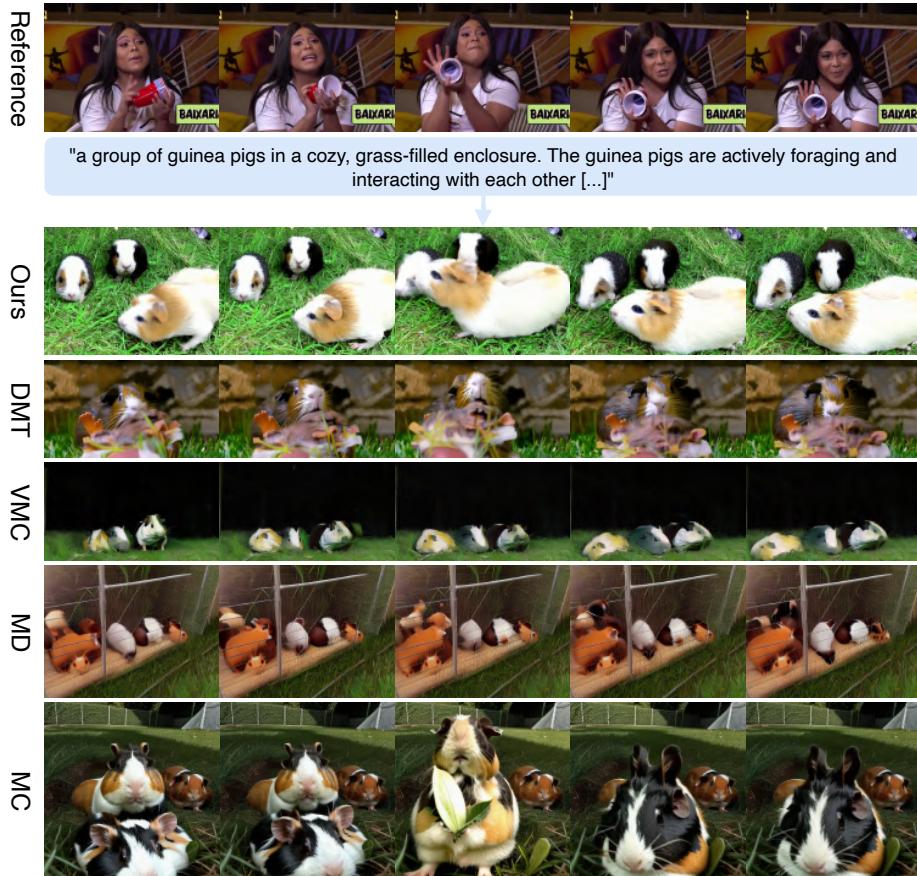
**Figure G: Same-category Transfer.** The reference video depicts a close-up shot of a person looking at the camera while performing head and arm movements. Our approach successfully transfers the nuances of this motion to the two target characters and even reinterprets it to represent the action described in the text prompt (eating). In contrast, **DMT** overly relies on the reference structure, limiting its ability to capture the intended semantics. Furthermore, **VMC**, **MD**, and **MC** fail to capture key nuances of the reference motion, such as the arm movement.



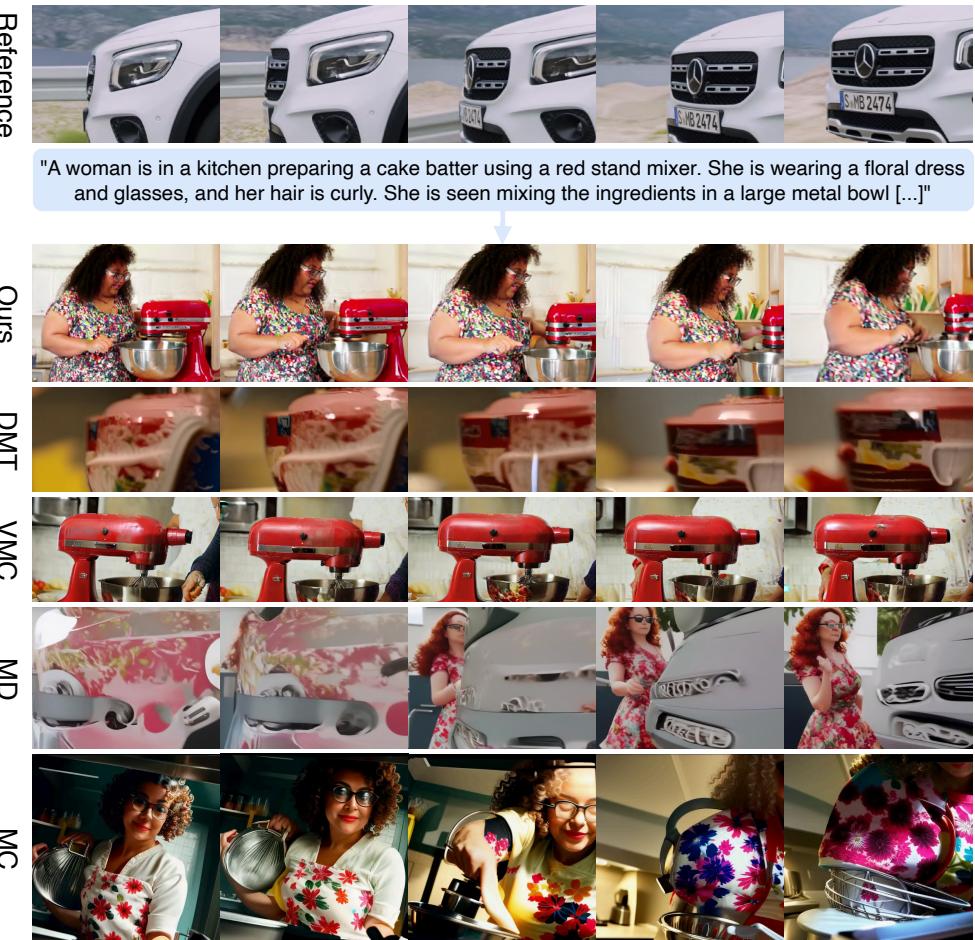
**Figure H: Cross-category Transfer.** The reference video shows a person rotating their head. Among all methods, our method and **MC** are the only approaches capable of performing this cross-category transfer effectively. Other methods either leak the appearance from the source video or fail to transfer the motion accurately.



**Figure I: Cross-category Transfer.** The reference video shows the upper body of a person performing rotating head movements. Our method successfully identifies the relative head pose changes as motion and applies them to the target character, starting from a different initial head pose. In contrast, most other methods fail to capture this motion. The only exception is **MC**, which, however, remains tied to the reference pose.



**Figure J: Cross-category Transfer.** The source video depicts a woman holding a cup and moving it with both hands to the left side. Our method is the only approach that effectively captures this motion transfer across categories. It interprets the rotation and translation of the reference objects (cups and arms) as the movements of the guinea pigs, that serve as the general directional guidance. **MD** is the only other method that performs reasonably well, while most others fail significantly.



**Figure K: Camera Motion Transfer.** The reference video depicts a rotating camera movement around a car. Our method successfully transfers this motion while remaining faithful to the text prompt. In contrast, other methods either overly rely on the input structure (e.g., **DMT**) or even leak visual appearance (e.g., **MD**), failing to reflect the prompt accurately. While **VMC** and **MC** capture aspects of the prompt, they largely disregard the reference motion.

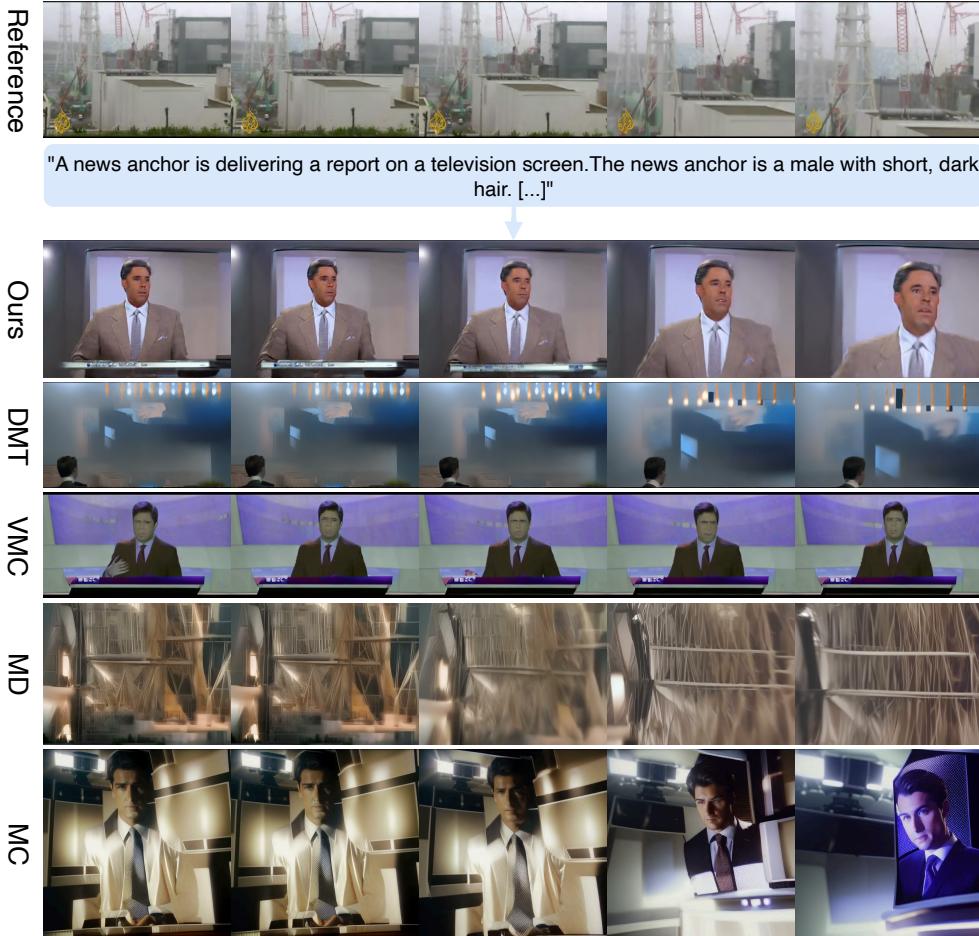
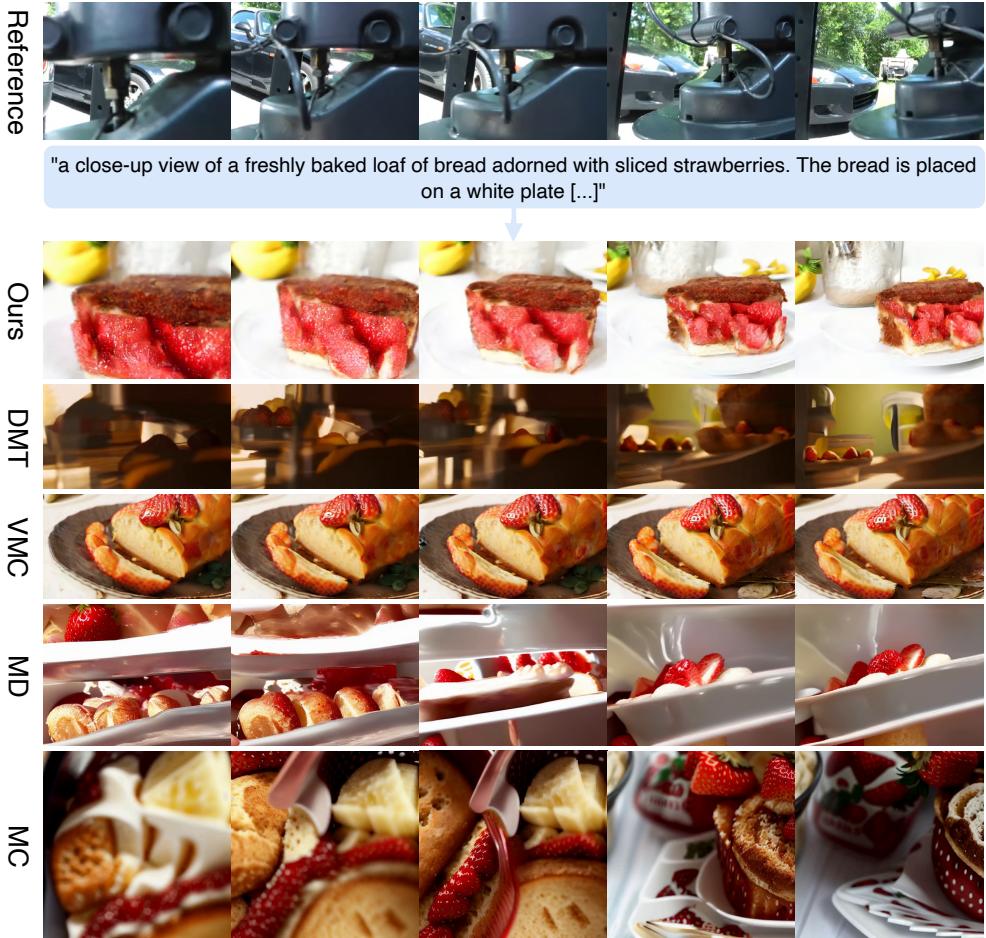


Figure L: **Camera Motion Transfer.** The reference video solely shows a zoom-in to the top left of the frame. Our method captures this motion even while transferring to a substantially different target category, while the others fail to capture it or even fail to produce realistic videos due to the strong change in category.



**Figure M: Camera Motion Transfer.** The reference video depicts a camera zoom-out combined with a rotation. Our method is the only approach capable of capturing both the intended motion while maintaining a coherent and realistic generated video.

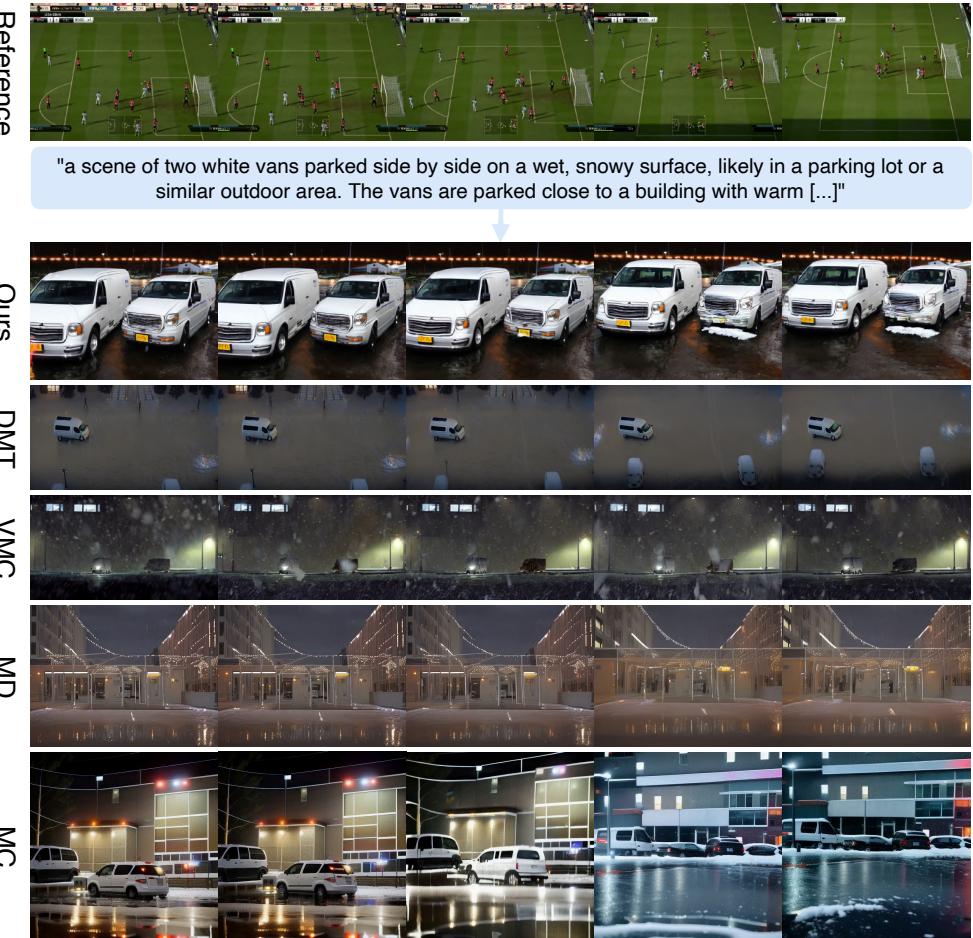


Figure N: **Camera Motion Transfer.** The reference video depicts a soccer match with camera moving backward and then to the right. Our approach is the only method that faithfully captures this motion, while adhering to the text prompt, notably while filtering out unnecessary noise from the individual soccer players. Other methods either fail to replicate the motion accurately or do not align with the prompt.