Motion Modes: What Could Happen Next?

Karran Pandey[1] Matheus Gadelha[2] Yannick Hold-Geoffroy[2] Karan Singh[1]
Niloy J. Mitra[2,3] Paul Guerrero[2]
[1] University of Toronto [2] Adobe Research [3] UCL

https://motionmodes.github.io

Abstract

Predicting diverse object motions from a single static image remains challenging, as current video generation models often entangle object movement with camera motion and other scene changes. While recent methods can predict specific motions from motion arrow input, they rely on synthetic data and predefined motions, limiting their application to complex scenes. We introduce Motion Modes, a training-free approach that explores a pre-trained image-to-video generator's latent distribution to discover various distinct and plausible motions focused on selected objects in static images. We achieve this by employing a flow generator guided by energy functions designed to disentangle object and camera motion. Additionally, we use an energy inspired by particle guidance [6] to diversify the generated motions, without requiring explicit training data. Experimental results demonstrate that Motion Modes generates realistic and varied object animations, surpassing previous methods and even human predictions regarding plausibility and diversity. Code will be released upon acceptance.

1. Introduction

> Prediction is very difficult, especially if it's about the future.
>
> Niels Bohr

Consider Fig. 1. Can you imagine what could happen next in each case? Humans are good at imagining multiple ways the objects could move, even from single (image) snapshots. While we can train networks to predict videos starting from a conditioning text or image [3], most generated videos entangle camera motion, object motion, and other scene changes – predicting a diverse set of motions for a given object still remains an open challenge.

Authoring plausible animations for objects in a static image can be daunting. Researchers have recently been able to train networks to predict cyclic and small-scale motions [2, 16]. Another family of methods [15, 23] simplify this task by taking input motion arrows along with the starting image to predict videos with motions that follow the given arrows. However, such methods are trained on synthetic data and do not generalize to complex motions, such as the breaking ocean wave in Fig. 1. More importantly, they require motions to be given, rather than predicting them. In many scenarios, such as the roaring lion in Fig. 1, imagining a diverse set of motions and then conveying them with multiple corresponding motion arrows itself, can be very challenging. The ability to automatically discover diverse yet plausible object motions can thus assist users in cinematic exploration, motion illustration, and image/video editing.

Figure 1. Could you imagine how the scene evolves in each case? See Fig. 2 for plausible yet distinct motion videos predicted by our training-free approach Motion Modes.

The latest image-to-video generators provide this opportunity. Having been trained on a large variety of diverse data, such generators, conditioned on static images, encode distributions over plausible animations for scene objects and other scene properties. Our paper subsequently asks and affirmatively answers the research question: *is it possible to probe such a latent distribution to discover possible motions for a given object in a static image*?

Directly sampling these generators, conditioned on a starting image, produces random videos, some of which may include a motion of the selected object. Still, most will consist of motions pertaining to other random objects, camera

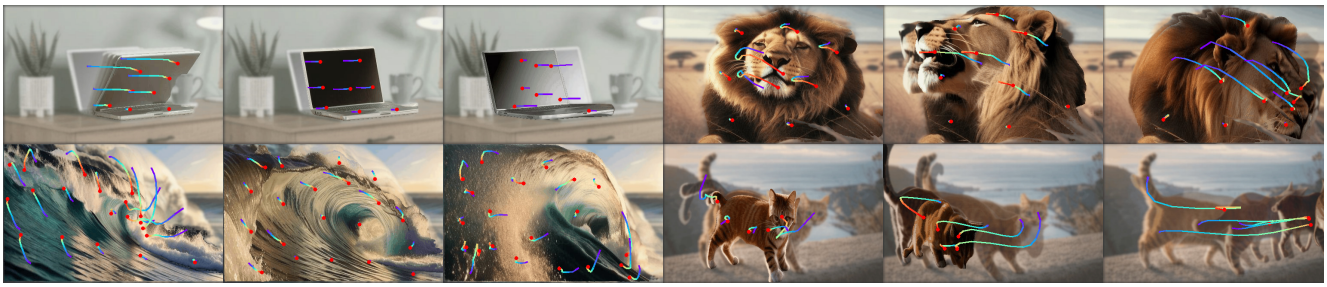

Figure 2. **Motion Modes** creates multiple distinct and plausible motions for a given object, disentangled from the motion of other objects, camera and other scene changes. We show three distinct object motions for each of Figure 1's images, representative of constrained rigid motion (latop), complex deformations (wave) and articulated characters (lion and cat). We visualize motions as flow trajectories from blue (first frame) to red (last frame). Ghosted intermediate frames further clarify complex motions. See supplemental for the result videos.

motion, lighting, and other changes to scene appearance. Hence, the main challenges are to discover the motions of an object in such a distribution that (i) disentangle motions of the selected object from other scene changes, and (ii) find multiple distinct object motions. We propose *Motion Modes* as a training-free method to find such object motions by exploring the prior of a pre-trained image-to-video generator.

We show that both of the above challenges can be addressed with a training-free approach that guides the denoising process of a flow generator [23] with carefully designed guidance energies. Using a flow generator naturally disentangles objects and camera motion from other scene changes. Our proposed guidance energies fulfill two purposes: (i) they further disentangle object and camera motion by encouraging non-zero object motion and zero camera motion, and (ii) encourage the generation of multiple distinct motions. We demonstrate that such guidance can be applied directly at inference time, without any fine-tuning of existing generators or access to suitable training data. Fig. 2 shows the output of Motion Modes on the images shown in Fig. 1.

We evaluated Motion Modes on a variety of input images and compared ours with possible baselines (e.g., random sampling, LLM-based) and ablated versions of our full method. We performed human evaluations to assess the quality of our generation, both in terms of plausibility and diversity of the predicted motions. The qualitative and quantitative evaluations show that we can reliably and accurately predict potential future outcomes, sometimes even surpassing human ability. We show that discovered motions can be used for motion exploration and to facilitate drag-based image editing. In summary, Motion Modes is the first training-free method to generate diverse and plausible videos of object motion from a single input image.

2. Related Work

Motion-aware video generators. Diffusion-based video generators have quickly advanced in the last years [4, 5, 12, 22], now producing realistic and temporally consistent videos. Adding extra control, Motion-I2V [23] introduced an image-guided video generation method as a two-stage process for consistent and controllable video generation. First, it uses a diffusion-based motion field predictor to determine pixel trajectories, followed by motion-augmented temporal attention that improves feature propagation across frames. We use this setup as our backbone and adapt it with our guidance energies during the denoising phase. AnimateAnything [7] presents an image animation method using a video diffusion generator's motion prior, enabling controlled animation by guiding motion areas and speed. They demonstrate fine-grained, text-aligned animations with intricate motion sequences, even on open-world settings. Such methods, however, require suitable text prompts to guide the generation, which may be non-trivial in more complex scenarios where mentally predicting future motions is challenging (see our LLM-based baseline in Sec. 4). Finally, towards train-free methods, similar to the analysis of image generators [10], Xiao et al. [28] identify (using PCA analysis) motion-aware features in video diffusion models and use them for interpretable and adaptable video motion control across different architectures.

Generating cyclic motions. Creating future animations from static scenes has received attention over the years. Davis et al. [8] create interactive elements in videos by analyzing subtle object vibrations to get motions, allowing manipulation of video elements as if they were physically interactive. The problem was recently revisited by Li et al. [16] to learn an image-space prior on scene motion from a collection of motion trajectories extracted from real video sequences depicting natural, oscillatory dynamics(e.g., leaves, trees, flowers, candles). Using a Fourier domain analysis, they learn a diffusion process to model the generation in the frequency space. Earlier, in the context of geometric objects, Mitra et al. [17] use symmetry analysis to infer plausible part movement in mechanical objects, focusing on gear assemblies and linkages. Hu et al. [13] present a model for predicting part mobility in 3D objects by learning how parts of an object can move based on their spatial configuration

in a single static snapshot by leveraging a linearity trait in typical object motions and creating a mapping that associates static snapshots with dynamic units. To model small and repetitive garment motion, Bertiche et al. [2] present an automatic method to generate human cinemagraphs from single RGB images to mimic garment dynamics arising from gentle winds. They introduce a cyclic neural network that produces looping cinemagraphs for the target loop duration. The network is trained with normal maps obtained from renderings of synthetic garment simulations. While they demonstrated that the learned dynamics can be applied to real RGB images, the reliance on training data does not allow these methods to be applied to the broader class of general motions.

Movements from generative priors. Priors learned by modern generators, trained on large datasets, have shown to be useful for handle-based image manipulation. DragGAN [20] presented an interactive tool for handle-based realistic editing of natural images that relied on a feature-based motion supervision that moves selected points toward target positions, leveraging GAN's internal features for precise localization. Similarly, image manipulators (e.g., point- or box-based) have exploited priors implicit in diffusion-based image [1, 18, 21, 25] or video [24, 26, 28] generators. Beyond zeroshot methods, Dragapart [15] presents a part-level editing system where they refine a pretrained image generator on a new synthetic dataset showing annotated part motion. The network, fine-tuned on synthetic data generalizes well across real-world images and diverse categories. However, the method fails on complex scenarios and object categories not seen in the training set (Sec. 4). Draganything [27] uses entity representation for drag-based plausible video generation in response to user arrows, but does not produce diverse results. There are also sampling strategies designed to increase the diversity of outputs in diffusion-based image generators [6]. They rely on concurrently denoising a batch of multiple samples guided by a repulsive energy. However, in the case of video generators, such strategies are limited by the memory cost of the number of samples that can be denoised together (≈10 GB per additional sample for Motion-I2V [23] with gradient checkpointing). On the other hand, we devise an iterative sampling strategy that is not capped by the number of samples that can fit in memory together (see Section 3.3).

3. Method

Our goal is to take an image $\mathbf{y} \in \mathbb{R}^{H\times W\times 3}$ and a mask $\mathbf{m} \in \mathbb{R}^{H\times W}$ marking an object in the image, and to find a set of likely *motions* $\mathcal{X} \coloneqq \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots\}$ of the object given its context in the image. In Figure 3, for example, the drawer could be opened or closed; however it could not plausibly be moved up- or downwards. We represent motions as time-dependent two-dimensional vector fields $\mathbf{x} \in \mathbb{R}^{F\times H\times W\times 2}$ for a motion that spans F frames. This vector field defines the trajectory of each pixel as per-frame 2D offsets from its initial position.

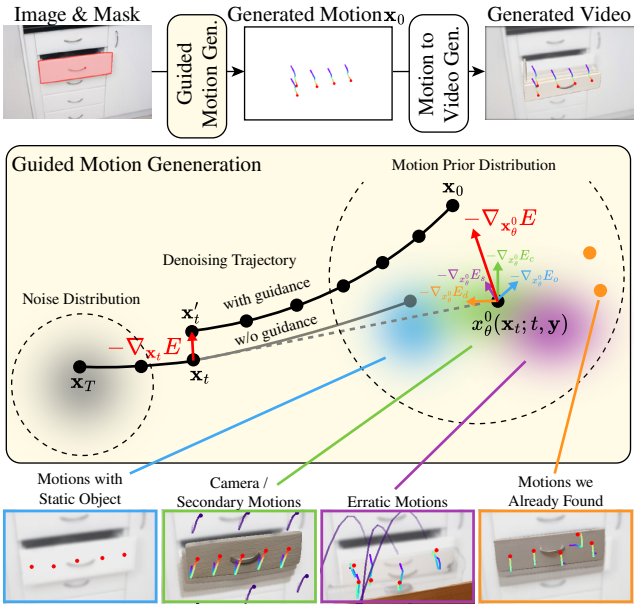

Figure 3. **Method Overview.** We generate a motion $\mathbf{x}$ using a guided denoising approach, where guidance energies encourage smooth object motions that are disentangled from camera motions and distinct from previously generated motions. Iterative sampling gives us a set of diverse motions $\mathcal{X}$.

We generate motions by sampling an existing image-to-video diffusion model that takes the image $\mathbf{y}$ as starting frame. The main challenges for generating motions of an object in an image are disentangling object motions from other types of scene changes and finding a diverse set of plausible object motions. To address these challenges, we (i) use a diffusion model that generates motion separately from appearance [23], effectively disentangling object / camera motions from other scene changes, such as lighting or shadows (Section 3.1), and (ii) define guidance energies that we minimize during the denoising process to further separate object motion from camera motion and to efficiently sample a diverse set of motions, rather than sampling motions randomly from the motion prior (Section 3.2). We build the motion set $\mathcal{X}$ by iteratively sampling the motion prior with our guidance energies, and define a simple stopping criterion to avoid implausible motions (Section 3.3).

3.1. Motion Generation

Our approach can be applied to any pre-trained diffusion-based image-to-video model which generates motion and appearance independently.

Training. Given an input image $\mathbf{y}$ and a motion $\mathbf{x}$ for this image, a noisy motion vector field $\mathbf{x}_t$ is first obtained by adding a random amount of noise to $\mathbf{x}$:

$$\mathbf{x}_t = \sqrt{\alpha(t)}\,\mathbf{x} + \sqrt{1-\alpha(t)}\,\epsilon, \tag{1}$$

where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is Gaussian noise, and $t \in [0, T]$ parameterizes a noise schedule α that determines the amount of noise in $\mathbf{x}_t$, with $\alpha(0) = 1$ (no noise) and $\alpha(T) = 0$ (pure noise). The denoiser ϵ_θ of the diffusion model is then trained to invert this noising process by minimizing the following loss through gradient-descent:

$$\mathcal{L}_{\text{diff}} := w(t)\|\epsilon_\theta(\mathbf{x}_t; t, \mathbf{y}) - \epsilon\|_2^2,$$

where $w(t)$ is a weighting scheme for different parameters t. In practice, we employ a latent-space diffusion model that operates on a lower-resolution latent representation of the motions and the input image, which is obtained through a VAE [14]. We omit this distinction in the notation, both for clarity and for generality, as the method is orthogonal to the choice of diffusion model. Specifically, our implementation utilizes Motion-I2V [23] as the backbone.

Inference. Given the trained denoiser ϵ_θ, a noise-free motion $\mathbf{x}_0$ for input image $\mathbf{y}$ is then generated by starting from pure noise $\mathbf{x}_T$ and iteratively denoising in small steps:

$$\begin{aligned} \mathbf{x}_T &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \\ \mathbf{x}_{t-1} &\sim \mathcal{N}(a_t \mathbf{x}_t - b_t \epsilon_\theta(\mathbf{x}_t; t, \mathbf{y}), \sigma_t^2 \mathbf{I}) \end{aligned} \tag{2}$$

where a_t, b_t, and the variance σ_t^2 are chosen according to a denoising schedule. This process creates a trajectory $\mathbf{x}_T, \mathbf{x}_{T-1}, ..., \mathbf{x}_0$ of motions with decreasing noise, where $\mathbf{x}_0$ is close to the natural motion manifold. Generated motions $\mathbf{x}_0$ are typically plausible, but they entangle camera motions with object motions. Additionally, exploring different motions by randomly sampling $\mathbf{x}_T$ is inefficient, as it requires a large number of samples to find multiple meaningful ways in which $\mathbf{y}$ can change in time.

3.2. Guidance Energies

Our key contribution is the guidance energies that we introduce into the inference process. The energies encourage the generation of motions that are different from any previously generated motions, where only the object in image $\mathbf{y}$ selected by the mask $\mathbf{m}$ moves and the camera is static. The goal is to significantly reduce the number of samples needed to get a diverse set of focused object motions.

(i) Static camera guidance. We encourage zero camera motion by penalizing the average magnitude of motion outside the object region defined by the object mask $\mathbf{m}$:

$$E_c(\mathbf{x}, \mathbf{m}) := \frac{\sum_{k,i,j} \|\mathbf{x}_{k,i,j}\| \, (1 - \mathbf{m}_{i,j})}{\sum_{k,i,j} (1 - \mathbf{m}_{i,j})},$$

where k, i, j are indices over frames, pixel rows, and pixel columns, respectively, so that $\mathbf{x}_{k,i,j}$ denotes a single offset vector of the motion $\mathbf{x}$. The mask $\mathbf{m}$ is 1 inside the object region and 0 everywhere else.

(ii) Object motion guidance. We encourage object motion by encouraging a difference between the average magnitude of motion inside the object mask $\mathbf{m}$ and outside:

$$E_o(\mathbf{x}, \mathbf{m}) := \phi\left(|E_c(\mathbf{x}, \mathbf{m}) - E_c(\mathbf{x}, 1 - \mathbf{m})|\right).$$

Here, ϕ is an activation function that gives higher energies for smaller differences, based on a soft inverse:

$$\phi(a) := \text{softplus}\left((a + e)^{-1} - \tau\right),$$

where e is a small epsilon to avoid division by zero, and τ is a threshold representing the point at which a satisfactory loss value is reached. τ is empirically set to 40 for the object motion guidance and 1 for the diversity guidance.

(iii) Diversity guidance. Given a set of previously generated motions $\mathcal{X}$, we encourage newly generated motions to be different by adding a repulsion energy from each of the motions in $\mathcal{X}$:

$$E_d(\mathbf{x}, \mathbf{m}, \mathcal{X}) := \sum_{\tilde{\mathbf{x}} \in \mathcal{X}} \frac{\sum_{k,i,j} \phi\left(d(\mathbf{x}_{k,i,j}, \tilde{\mathbf{x}}_{k,i,j})\right) \, \mathbf{m}_{i,j}}{\sum_{k,i,j} \mathbf{m}_{i,j}},$$

where d is a distance function between individual offset vectors in a motion based on angle and magnitude differences:

$$d(\mathbf{a}, \mathbf{b}) := w_{\text{mag}}(|\|\mathbf{a}\| - |\mathbf{b}\||) + w_{\text{angle}}\left(1 - \frac{\mathbf{a}^\top \mathbf{b}}{\|\mathbf{a}\| \, \|\mathbf{b}\|}\right),$$

with weights $w_{\text{mag}} = 0.25$ and $w_{\text{angle}} = 0.75$ to emphasize diverse motion directions.

(iv) Smoothness guidance. As a regularization, we also encourage smooth object motions by penalizing large changes in motion across consecutive frames within the object mask:

$$E_s(\mathbf{x}, \mathbf{m}) := \frac{\sum_{k,i,j} d\left(\mathbf{x}_{k,i,j}, \mathbf{x}_{k+1,i,j}\right) \mathbf{m}_{i,j}}{\sum_{k,i,j} \mathbf{m}_{i,j}}, \tag{3}$$

with $w_{\text{mag}} = 0.75$ and $w_{\text{angle}} = 0.25$ to minimize sudden changes in magnitude.

Guided Inference. We combine the four energies into a single (guidance) energy $E(\mathbf{x}, \mathbf{m}, \mathcal{X}) := \lambda_d E_d + \lambda_c E_c + \lambda_o E_o + \lambda_s E_s$, with weights $\lambda_d = 3.0$, $\lambda_c = 0.2$, $\lambda_o = 0.025$ and $\lambda_s = 0.1$. Similiar to classifier-free guidance and several image editing methods [9, 11, 21], we minimize these energies during the inference process, effectively changing the denoising trajectory, without requiring fine-tuning or retraining (which would be difficult as our tasks lacks suitable training data). Equation 2 takes the modified form:

$$\begin{aligned} \mathbf{x}_{t-1} &\sim \mathcal{N}\left(a_t \mathbf{x}_t - b_t \epsilon_\theta(\mathbf{x}'_t; t, \mathbf{y}), \sigma_t^2 \mathbf{I}\right), \text{with} \\ \mathbf{x}'_t &:= \mathbf{x}_t - \nabla_{\mathbf{x}_t} E\left(x_\theta^0(\mathbf{x}_t; t, \mathbf{y}), \mathbf{m}, \mathcal{X}\right). \end{aligned} \tag{4}$$

Here, $x_\theta^0(\mathbf{x}_t; t, \mathbf{y})$ is the non-noisy motion predicted at inference step t, derived from Eq. 1 as:

$$x_\theta^0(\mathbf{x}_t; t, \mathbf{y}) := \frac{1}{\sqrt{\alpha(t)}}\left(\mathbf{x}_t - \sqrt{1 - \alpha(t)} \, \epsilon_\theta(\mathbf{x}_t; t, \mathbf{y})\right).$$

3.3. Stopping Criterion

We build the set $\mathcal{X} \coloneqq \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots\}$ by iteratively sampling the motion prior as described above. We can obtain an arbitrary number of motions $\mathbf{x}$ using this strategy; for our experiments we sample up to 6 different motions. However, some objects and scenes may only admit a smaller number of distinct motions, after which motions either repeat or stop. We detect these cases using the guidance energy of the final denoised motion $E(\mathbf{x}_0, \mathbf{m}, \mathcal{X})$. We discard and re-sample motions with guidance energies above a threshold $\rho(= 5.0)$, and stop sampling after discarding two motions in a row.

4. Results

To evaluate a set of motions $\mathcal{X}$ generated by our method, we identify four desirable motion properties: (1) *Plausible*: motions appear natural and physically reasonable. (2) *Diverse*: motions are largely different from each other. (3) *Expected*: motions are plausible motions that match those imagined by a viewer for the selected object in the image. (4) *Focused*: motions avoid any scene motion (including camera motion) that does not pertain to the selected object or is directly caused by its motion (eg. the smoke from the selected train (top-left) in Figure 6).

We show with both quantitative metrics and a user study that our guided sampling strategy performs significantly better along these properties than alternatives, given the same sample budget and the same motion prior. We also provide several qualitative comparisons that demonstrate that Motion Modes can be used to explore object motions. As additional application, we also show how our motions can be used to assist users with drag-based image editing. More evaluation results are provided in the supplement.

Baselines. As far as we know, Motion Modes is the first training-free method to explore the problem of finding diverse motions for a given object in an image. However, there are several alternatives we can compare against. For a fair comparison, all baselines use the same Motion-I2V [23] backbone as our method. (1) *Prompt Generation*: We give GPT4-o an image with highlighted object and ask it to give us prompts for diverse object motions, which we then feed into Motion-I2V. Each prompt gives us one motion. (2) *ControlNet*: We use Motion-I2V's *MotionBrush* to restrict motions to the object region. This tool is a ControlNet trained to limit motions to originate in the given region. We obtain multiple motions by randomly sampling the starting noise $\mathbf{x}_T$. (3) *Random Arrows*: We use Motion-I2V's *MotionDrag* with random arrows to explore possible object motions. We sample an arrow by choosing a random starting position inside the object region, a random direction and a fixed length. Each arrow gives us a different motion. (4) *Random Noise*: We randomly sample the starting noise $\mathbf{x}_T$ of Motion-I2V. This is equivalent to our method without any guidance energies. (5) *Farthest Point Sampled (FPS) Noise*: We use farthest point sampling to sample distinct starting noise $\mathbf{x}_T$.

Table 1. **Quantitative comparison** of the *diverse* and *focused* property of our output motions to all baselines.

	diverse	focused		
	$\bar{E}_d \downarrow$	$\bar{E}_f \downarrow$	($\bar{E}_c \downarrow$	$\bar{E}_o \downarrow$)
Prompt Gen.	1.28	1.71	1.11	2.31
ControlNet	1.75	1.14	**0.07**	2.22
Random Arrows	1.77	1.17	**0.07**	2.27
Random Noise	1.27	2.20	1.36	3.05
FPS Noise	1.21	1.98	1.23	2.74
Motion Modes (ours)	**1.04**	**0.07**	**0.09**	**0.05**

Qualitative comparison. Figure 5 shows a qualitative comparison to all baselines on four scenes. (Please refer to the supplement for a comparison on a larger set of images.) We can see that the prompt generation baseline does tend to generate motions that are *diverse*, but the inaccurate nature of the prompt-based control results in less *focused* motions of the selected object. There is significant camera motion, and we can see motions of secondary objects in the basketball image, for example, where additional balls are hallucinated. Restricting the motion to the object region using the ControlNet baseline has the undesirable effect of significantly reducing the overall amount of motion, to the point of resulting in completely static scenes in many cases. Similar to the prompt generation baseline, sampling the motion prior randomly or with farthest point sampling without using our guidance energies entangles object motion with camera motion. Additionally, we can see that our approach produces more *plausible* and *expected* motions, compared to all baselines. For example, the opening and closing motion of the drawer is more natural without deforming parts, and the forward/backward motion of the tank generated by Motion Modes is probably closer to the motion we would expect from the tank than the more erratic motions generated by the baselines. We further confirm this trend on a larger set of scenes with the user study presented in the one of the following sections. We attribute the improved plausibility to our smoothness energy that avoids erratic motions.

Quantitative comparison. We measure two properties with explicit quantitative metrics: First, the *diversity* of motions in a set $\mathcal{X}$ can be measured with the average diversity guidance energy $\bar{E}_d(\mathbf{m}, \mathcal{X}) \coloneqq \sum_{\mathbf{x}\in\mathcal{X}} E_d(\mathbf{x}, \mathbf{m}, \mathcal{X})/|\mathcal{X}|$. Second, the *focus* of motions on only the selected object can be measured based on the average object motion and static camera guidance energies $\bar{E}_f \coloneqq 0.5(\bar{E}_o + \bar{E}_c)$, with $\bar{E}_o$ and $\bar{E}_c$ computed analogous to $\bar{E}_d$, but scaled by a factor of 0.01 and 0.1, respectively, to account for scaling differences.

We compare our method to all baselines on a dataset of 28 input images that were obtained either through a state-of-

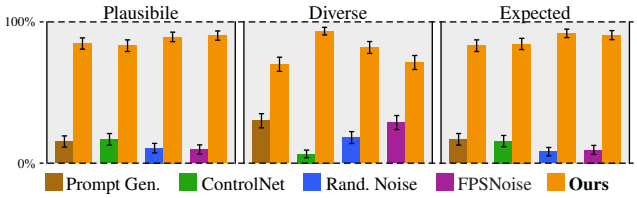

Figure 4. **User Study I.** We compare the *plausible*, *diverse*, and *expected* nature of our motions to four baselines. Each pair of bars shows the percentage of comparisons in which our method or a baseline was judged favorably with 95% confidence intervals.

the-art text-to-image generator, or from photographs. The images cover a wide range of scenes, including articulated objects, vehicles, animals, balls, and objects with and objects with complex motions, such as waves and flags. Please refer to the supplement for a full set of qualitative results.

Table 1 summarizes results. Due to our diversity guidance, we achieve significantly more diverse motions than any baseline. The ControlNet and Random Arrows baselines also achieve relatively good focus, but looking at Fig. 5 (as well as the camera and object guidance columns in Table 1), we can see that this is mostly caused by a lack of both camera *and* object motion. Our guidance energies fix the camera without fixing the object, giving us more focused motions.

User studies. We perform two user studies. The first study evaluates the *plausible*, *diverse* and *expected* nature of our output motions compared to baselines, while the second study examines the *expected* nature of our motions.

In the first study, participants were asked to compare the top three motions of our method to top three motions of a baseline, and choose the best set of motions along each of the three metrics in three two-alternative forced choice questions. The methods were presented in a randomized order. We recruited 32 participants, each completed 10 comparisons per baseline (a total of 320 comparisons per baseline). For each comparison, a scene was chosen randomly from our dataset of 27 images. Results are shown in Figure 4: motions of our approach are judged to be more plausible, diverse and expected than motions found by baselines. Notedly, the prompt generation baseline also has a good amount of diversity, coming close to the diversity of our approach. We omitted the *Random Arrows* baseline due to its similarity (and worse performance) compared to ControlNet. It is included in an extended version of the study in the supplement.

In the second study, 12 new participants were first asked to describe all possible future motions of an object highlighted in an input image. We then revealed the first four of our motions to them, and asked them to make two independent sets of selections - (i) motions that align with their initial expectations and (ii) motions that are plausible. Each participant assessed 10 scenes, and we computed three metrics from their responses - *expected* (percentage of their expected motions predicted by our motions), *plausible* (percentage of our motions deemed plausible), and *inspirational* (percentage of our motions that were deemed plausible but outside the participant's expectation). Participants found on average, that 96% of motions were plausible, 92% of their expectations were produced by our approach, and 19% of motions were plausible but outside expectation. Overall, participants felt that our motions not only aligned well with their expectations, but also consistently provided inspiration for exploring unseen diverse motions in input scenes.

Table 2. **Ablation** of key components with metrics based on *diverse*, *focused* metrics and their tradeoff $\bar{E} := 0.5(\bar{E}_d + \bar{E}_f)$. Underlined values are closer to the best than to the worst value.

		div.	focused		
	$\bar{E}\downarrow$	$\bar{E}_d\downarrow$	$\bar{E}_f\downarrow$	($\bar{E}_c\downarrow$	$\bar{E}_o\downarrow$)
without E_c	0.83	1.02	0.64	1.29	**0.00**
without E_o	0.97	1.03	0.91	**0.06**	1.75
without E_d	0.72	1.36	0.08	0.13	0.04
FPS instead of E_d	0.79	1.49	0.10	0.11	0.08
ControlNet instead of E_c,E_o	0.88	**0.96**	0.80	0.15	1.45
Motion Modes	**0.55**	1.04	**0.07**	0.09	0.05

Ablation. We ablate several components of our method is shown in Table 2. We use the same metrics as in Table 1, but add another metric that illustrates the trade-off between diversity and focus: $\bar{E} := 0.5(\bar{E}_d + \bar{E}_f)$. We ablate the three main guidance energies, and show the effect of using farthest point sampling of the initial noise instead of the diversity guidance, and a ControlNet instead of the camera and object guidance. As expected, removing the camera or object guidance results in strong camera motions (high $\bar{E}_c$) or little object motion (high $\bar{E}_o$), and removing the diversity energy or using farthest point sampling instead results in less diversity (high $\bar{E}_d$). Swapping the object and camera guidance with a ControlNet tends to fix the object in place (high $\bar{E}_o$). We only achieve the best tradeoff between diversity and focus using all of our components.

Application. Motion Modes, as presented, can help artists efficiently explore a diverse set of motions for a selected object, without having to sieve through a large set of sampled videos containing disentangled object and camera motion.

Arrow-based motion prompting. We demonstrate a second application: completing a rough motion hint to be used as input to a drag-based image editor or a motion-to-video generator. Figure 6 shows examples on two recent drag-based image editors: DragonDiffusion [18] and Drag-A-Part [15], and one motion-to-video generator [23], comparing results with and without our motion completion. A single drag arrow given by the user (shown in red) is used to retrieve the closest one of our detailed motions (shown as multiple red arrows for the drag-based image editors). We define

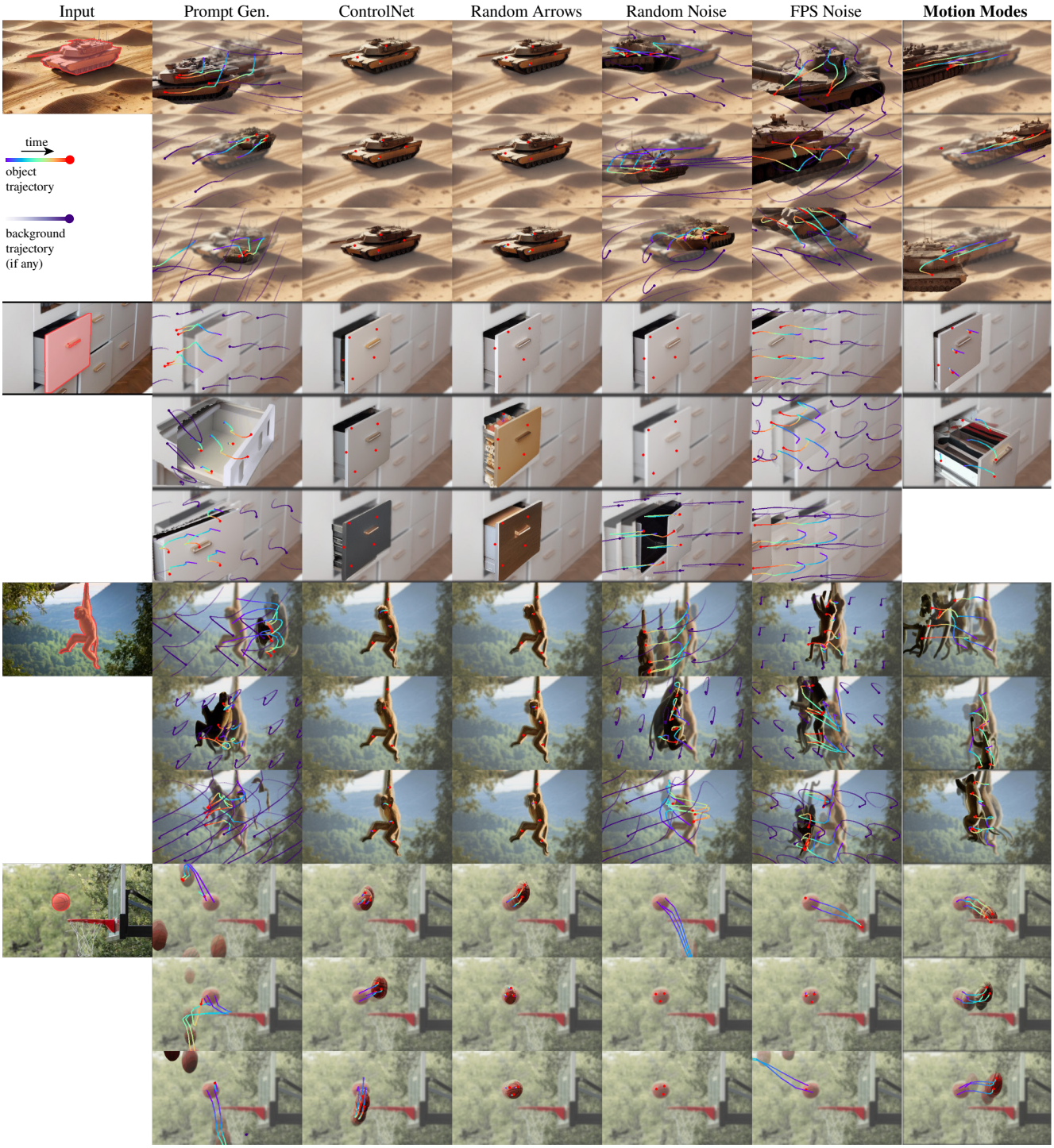

Figure 5. **Qualitative comparison.** Each column shows the first three motions for the masked object in the input (left). Object trajectories have red endpoints, background trajectories (usually due to camera motion) are purple. Motion is additionally visualized by overlaying ghosted intermediate frames. We can see that Motion Modes finds more plausible and diverse object motions disentangled from any other motions or scene changes, such as camera motions.

the closest motion as containing the 2D offset with lowest distance to the provided drag arrow, across all frames of the motion. We then use this motion, instead of the original drag arrow, as conditional input to image editor or video generator. Please refer to the supplement for details. This has two benefits: (i) Specifying complex image edits or video

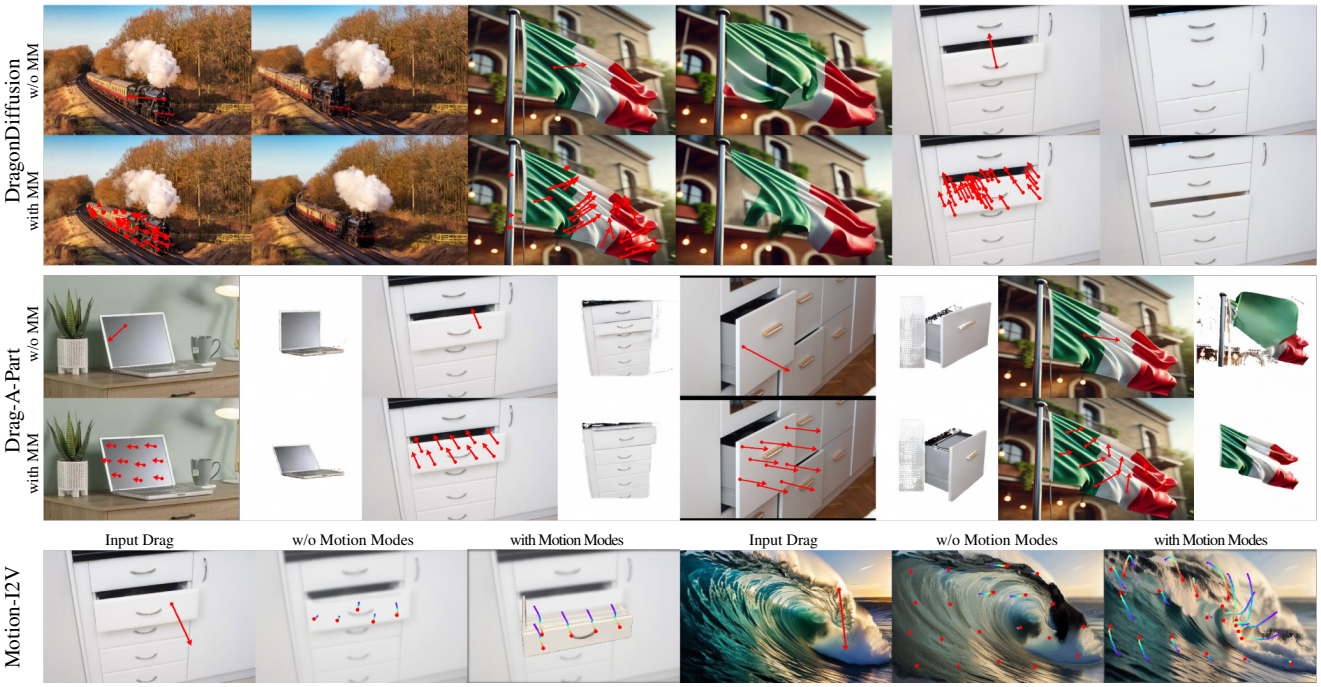

Figure 6. **Motion Completion.** We can use our set of motions $\mathcal{X}$ to complete rough motion hints (single red arrows) given by the user as conditional input to either drag-based image editors like DragonDiffusion or Drag-A-Part, or motion-to-video generators like Motion-I2V. Using the more detailed motions allows for complex motions that are hard to specify manually (like the flag or wave animations), and avoids ambiguities in the conditional input that can lead to implausible results, like the floating train, or the squashed drawer.

motions in detail is both difficult and time consuming, thus obtaining a complex edit/motion from a quick hint saves time and does not require artistic expertise. For example, it would be difficult to manually construct detailed drag arrows for the flag or the ocean wave. (ii) Rough motion hints are ambiguous and may be misinterpreted by the conditional generators, resulting in implausible motions. For example, dragging the train backwards with Dragon Diffusion results in a floating train, or dragging the drawer towards its closed position is misinterpreted as moving it upwards. Providing a more detailed motion removes this ambiguity and avoids implausible results.

5. Conclusion

We have presented Motion Modes as a training-free method to discover distinct motions for a selected object mask in a static image. Our primary contribution is a novel combination of guidance energies applied at inference time, to sample videos showing diverse object motions, from a pre-trained diffusion-based video generator. We evaluated our method on a range of complex images with both animate and inanimate objects to discover non-trivial motions, sometimes beyond those anticipated by viewers.

Limitations. Fig. 7 shows example limitations. Foremost, since Motion Modes is training-free, we inherit any data bias in our video generator (e.g., we will miss motions that cannot be expressed in our generator's sampling space). As we currently seek a discrete set of motions, we are only able to represent a continuous subspace of plausible motions a distinct set of discrete motions (eg. the laptop moving left-right, and front-back, instead of anywhere on the desk in Fig. 2). Further, since we progressively generate motions, we need a number of forward passes equal to the number of extracted motions. This can be slow and undesirable. Finally, very specific underlying modes can produce unrealistic motions.

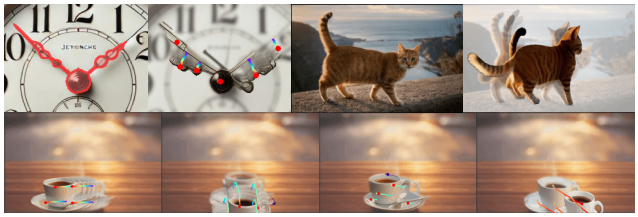

Figure 7. **Limitations.** (Top) The video prior can limit quality (bent clock handles, two cat tails). (Bottom) Continuous motion spaces can only be sampled discretely.

Future work. Motion Modes produces videos with negligible camera motion. Extending our approach to generate object motion with moving cameras, as commonly observed in sport and action shots where the camera follows the trajectory of the moving object, is subject to future work. We would also like to extend our method beyond 2D motion fields, to produce 3D motions: this would allow us to directly output 4D dynamic shapes as animated mesh sequences, turning video generators into 4D asset generators.

References

[1] Omri Avrahami, Rinon Gal, Gal Chechik, Ohad Fried, Dani Lischinski, Arash Vahdat, and Weili Nie. Diffuhaul: A training-free method for object dragging in images. *arXiv preprint arXiv:2406.01594*, 2024. 3

[2] Hugo Bertiche, Niloy J. Mitra, Kuldeep Kulkarni, Chun-Hao Paul Huang, Tuanfeng Y. Wang, Meysam Madadi, Sergio Escalera, and Duygu Ceylan. Blowing in the wind: Cyclenet for human cinemagraphs from still images. In *CVPR*, 2023. 1, 3

[3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023. 1

[4] 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. 2

[5] 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. 2

[6] Gabriele Corso, Yilun Xu, Valentin De Bortoli, Regina Barzilay, and Tommi S. Jaakkola. Particle guidance: non-i.i.d. diverse sampling with diffusion models. In *ICLR*, 2024. 1, 3

[7] Zuozhuo Dai, Zhenghao Zhang, Yao Yao, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Animateanything: Fine-grained open domain image animation with motion guidance, 2023. 2

[8] Abe Davis, Michael Rubinstein, Neal Wadhwa, Gautham J Mysore, Fredo Durand, and William T Freeman. Interactive dynamic video. *ACM TOG (SIGGRAPH)*, 34(4):1–9, 2015. 2

[9] Dave Epstein, Allan Jabri, Ben Poole, Alexei A. Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation, 2023. 4

[10] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. In *NeurIPS*, pages 9841–9850. Curran Associates, Inc., 2020. 2

[11] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. 4

[12] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 2

[13] Ruizhen Hu, Wenchao Li, Oliver Van Kaick, Ariel Shamir, Hao Zhang, and Hui Huang. Learning to predict part mobility from a single static snapshot. *ACM TOG (SIGGRAPH)*, 36 (6), 2017. 2

[14] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 4

[15] Ruining Li, Chuanxia Zheng, Christian Rupprecht, and Andrea Vedaldi. Dragapart: Learning a part-level motion prior for articulated objects. In *ECCV*, 2024. 1, 3, 6, 11

[16] Zhengqi Li, Richard Tucker, Noah Snavely, and Aleksander Holynski. Generative image dynamics. In *CVPR*, 2024. 1, 2

[17] Niloy J. Mitra, Yong-Liang Yang, Dong-Ming Yan, Wilmot Li, and Maneesh Agrawala. Illustrating how mechanical assemblies work. *ACM TOG (SIGGRAPH)*, 29(3):58:1–58:12, 2010. 2

[18] Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. Dragondiffusion: Enabling drag-style manipulation on diffusion models. In *ICLR*, 2024. 3, 6, 11

[19] Muyao Niu, Xiaodong Cun, Xintao Wang, Yong Zhang, Ying Shan, and Yinqiang Zheng. Mofa-video: Controllable image animation via generative motion field adaptions in frozen image-to-video diffusion model. *ECCV*, 2024. 10, 11

[20] Xingang Pan, Ayush Tewari, Thomas Leimkühler, Lingjie Liu, Abhimitra Meka, and Christian Theobalt. Drag your gan: Interactive point-based manipulation on the generative image manifold. In *ACM TOG (SIGGRAPH)*, page 1–11, 2023. 3

[21] Karran Pandey, Paul Guerrero, Matheus Gadelha, Yannick Hold-Geoffroy, Karan Singh, and Niloy J. Mitra. Diffusion handles: Enabling 3d edits for diffusion models by lifting activations to 3d. 2024. 3, 4

[22] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024. 2

[23] Xiaoyu Shi, Zhaoyang Huang, Fu-Yun Wang, Weikang Bian, Dasong Li, Yi Zhang, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, et al. Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 1, 2, 3, 4, 5, 6, 10, 11

[24] Yujun Shi, Jun Hao Liew, Hanshu Yan, Vincent Y. F. Tan, and Jiashi Feng. Lightningdrag: Lightning fast and accurate drag-based image editing emerging from videos, 2024. 3

[25] Yujun Shi, Chuhui Xue, Jun Hao Liew, Jiachun Pan, Hanshu Yan, Wenqing Zhang, Vincent Y. F. Tan, and Song Bai. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. In *CVPR*, pages 8839–8849, 2024. 3

[26] Jiawei Wang, Yuchen Zhang, Jiaxin Zou, Yan Zeng, Guoqiang Wei, Liping Yuan, and Hang Li. Boximator: Generating rich and controllable motions for video synthesis, 2024. 3

[27] 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, 2024. 3

[28] Zeqi Xiao, Yifan Zhou, Shuai Yang, and Xingang Pan. Video diffusion models are training-free motion interpreter and controller. *arXiv preprint arXiv:2405.14864*, 2024. 2, 3

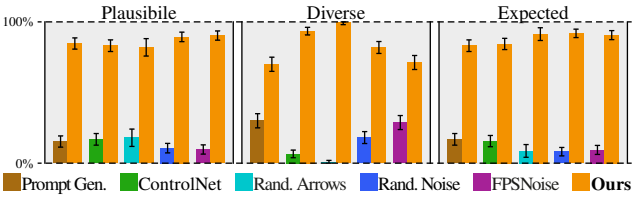

Figure 8. **Extended user study.** We compare the *plausible*, *diverse*, and *expected* nature of our motions to five baselines, including the Random Arrows baseline. Each pair of bars shows the percentage of comparisons in which our method or a baseline was judged favorably with 95% confidence intervals.

Table 3. **Extended ablation** of key components with metrics based on *diverse*, *focused* metrics and their tradeoff $\bar{E} := 0.5(\bar{E}_d + \bar{E}_f)$. Underlined values are closer to the best than to the worst value.

		div.	focused		
	$\bar{E}\downarrow$	$\bar{E}_d\downarrow$	$\bar{E}_f\downarrow$	($\bar{E}_c\downarrow$	$\bar{E}_o\downarrow$)
without E_c	0.83	1.02	0.64	1.29	**0.00**
without E_o	0.97	1.03	0.91	**0.06**	1.75
without E_d	0.72	1.36	0.08	0.13	0.04
without E_s	0.58	1.02	0.13	0.10	0.16
FPS instead of E_d	0.79	1.49	0.10	0.11	0.08
ControlNet instead of E_c,E_o	0.88	**0.96**	0.80	0.15	1.45
Motion Modes	**0.55**	1.04	**0.07**	0.09	0.05

A. Overview

In this appendix, we present extended versions of the user study (Section B) and the ablation study (Section C). Additionally, we examine how much a given motion constrains the video generator by showing different videos generated for the same motion (Section D) and provide additional implementation details as well as timing details (Section E). Finally, we provide a more detailed description for some of the baselines (Section F) and the arrow-based motion prompting application (Section G).

Our project website, `motionmodes.github.io`, also contains, among other details, a full qualitative comparison on 28 images, results of our method on a total of 34 different input images, and our arrow-based motion prompting application using a different video generator [19].

B. Extended User Study

In Figure 8, we present an extended version of the user study that includes the random arrows baseline. Results for this baseline are collected from 16 instead of 32 participants, the other study details are the same as for all other baselines. Results confirm our findings for all other baselines: users find our motions significantly more plausible and diverse, and they also better agree with the motions users expected for the selected object.

Figure 9. **Multiple videos from one motion.** We generate multiple videos from the same motion $\mathbf{x}$. They differ in small details, but overall follow the motion accurately.

C. Extended Ablation

In Table 3, we provide an extended ablation study that includes an ablation of the smoothness guidance. Apart from its function as regularizer, surprisingly, this energy also improves object focus, i.e. it tends to better avoid static objects. Our interpretation is that object motions are suppressed by the motion generator's prior during the denoising process if they start out unrealistically jerky or jittery. Our smoothness energy guides the denoising trajectory away from these bad object motions early on, resulting in a less suppression from the prior.

D. Multiple Videos Generated for One Motion

All videos in our experiments are obtained by first generating a motion $\mathbf{x}$ and then generating a video conditioned on $\mathbf{x}$. To examine how closely the generated video follows $\mathbf{x}$, in Figure 9, we show multiple videos generated conditioned on the same motion $\mathbf{x}$ from different random noises. We can see that small details are different, but overall, the motions of the different videos are similar to each other and follow the generated motion $\mathbf{x}$ accurately.

E. Implementation Details

Guided Denoising As described in the paper, we use the flow generation module from Motion-I2V [23] as our backbone. We further disconnect the ControlNet module described in their paper, as we don't need the conditioning and we found that the constraints from ControlNet conditioning limits the diversity of our motions. The flow generator uses 25 total timesteps for denoising out of which the first 20 timesteps are guided in our approach.

Timing and Memory In our experiments, we further used gradient checkpointing on the U-Net to minimize the memory cost of backpropagating the guidance gradients in each denoising timestep. Given the time cost of gradient checkpointing and additional memory costs of backpropagation, our guided denoising approach has a peak memory usage of 21.7GB and requires on average 2 minutes 35 seconds to fully denoise a sample across 25 timesteps. Unguided vanilla denoising, on the other hand, has 12.3GB peak memory usage and requires 1 minute 18 seconds on average to fully denoise a sample.

F. Additional Baseline Details

Prompt Generation. Our backbone Motion-I2V [23] supports text-conditioning for image-to-video generation. In the Prompt Generation baseline, we aim to sample diverse and focused object motions using a set of distinct text prompts. To automate this process, we use GPT-4 to generate text prompts that correspond to distinct object motions for a given input image and object. The prompts are then used as text conditioning for Motion-I2V for video generation.

Specifically, we query GPT-4 for the prompts as follows. GPT-4 is first provided the following context: *"I am using a text-based video generator to discover all the different ways a specific object in an image can move, and I wish to generate a set of text prompts in order to achieve this. In particular, I will provide an image and specify an object. For each such specification, I would like to generate 6 text prompts that can be input to the video generator in order to explore the distinct motions the specified object can have in the scene. Remember that we want the motions to be focused only on the specified object and to each be distinct from the other."* We then provide the model with an image along with a text specification of the object in the context of the same conversation to retrieve the text prompts. Some examples of retrieved prompts follow. For a scene with a basketball near a net: *"video of a basketball swishing through the hoop after a jump shot"*, *"video of a basketball bouncing off the rim and falling away from the hoop"*, *"video of a basketball spinning around the rim before dropping in"*. For a scene with a cat on a ledge: *"video of a cat walking gracefully along a ledge with a scenic background"*, *"video of a cat jumping off the ledge gracefully"*, *"video of a cat stopping and looking around curiously"*.

Random Arrows. Our backbone Motion-I2V [23] can be conditioned on a drag arrow that describes the rough motion direction and motion magnitude of a point in the image, in an application the authors call *MotionDrag*. In the Random Arrows baseline, we use random drag arrows to explore a diverse set of motions for a selected object. Specifically, given an object mask $\mathbf{m}$, we set the starting point for the drag arrow to a random point inside the object mask, randomly sample a direction, and sample the length of the drag arrow uniformly from an interval of reasonable lengths (20 to 80 pixels in an image with 320p resolution). We found that arrow lengths outside this interval tended to either result in zero object motion or implausible motions.

G. Additional Arrow-based Prompting Details

Our arrow-based prompting application shows that Motion Modes can be used to facilitate user interaction with drag-controlled image editors and video generators. As image editors, we work with Drag-A-Part [15] and DragonDiffusion [18], and as video editors, we use MOFA [19] and the *MotionDrag* application of Motion-I2V [23]. We take as input a given drag arrow, defined by a start point $\mathbf{a} \in [1, H] \times [1, W]$ and end point $\mathbf{b} \in [1, H] \times [1, W]$, both given as pixel indices for resolution $W \times H$. We then use this drag arrow to retrieve the closest motion $\mathbf{x}$ from our motion set $\mathcal{X}$. Recall that in each frame, our motions describe the same offset of each image point from its starting position as a drag arrow. Thus we can simply compare the drag arrow to each frame of the motion $\mathbf{x}$ at the starting position $\mathbf{a}$ of the drag arrow:

$$\min_{\mathrm{k}} \left\| \mathbf{x}_{k,\mathbf{a}} - \overrightarrow{\mathbf{ab}} \right\|_2, \tag{5}$$

where $\mathbf{x}_{k,\mathbf{a}}$ is the offset vector of the motion $\mathbf{x}$ in frame k at the starting point $\mathbf{a}$ of the drag arrow. The motion $\mathbf{x}$ with closest distance to the drag arrow describes a motion similar to the drag arrow, but typically has good plausibility and much more detail than the drag arrow. We then convert the retrieved motion back into a representation that the image or video editors can use as input. Specifically, Drag-A-Part can take up to 10 drag arrows as input, for DragonDiffusion, we can fit up to 100 arrows into memory, for MOFA, we use up to 50 arrows (we found that more arrows result in non-static backgrounds), and for Motion-I2V, we can directly provide the retrieved motion $\mathbf{x}$ as conditional input. To convert a motion to n drag arrows, we cluster the offsets in the retrieved frame of the motion into n clusters using K-Means, and use the cluster means as drag arrows.