

Forecasting Motion in the Wild

Neerja Thakkar^{1,2}, Shiry Ginosar^{* 3}, Jacob Walker², Jitendra Malik^{* 1}, Joao Carreira^{* 2}, and Carl Doersch^{* 2}

¹ UC Berkeley

² Google DeepMind

³ Toyota Technological Institute at Chicago

Abstract. Visual intelligence requires anticipating the future behavior of agents, yet vision systems lack a general representation for motion and behavior. We propose dense point trajectories as visual tokens for behavior, a structured mid-level representation that disentangles motion from appearance and generalizes across diverse non-rigid agents, such as animals in-the-wild. Building on this abstraction, we design a diffusion transformer that models unordered sets of trajectories and explicitly reasons about occlusion, enabling coherent forecasts of complex motion patterns. To evaluate at scale, we curate 300 hours of unconstrained animal motion from video through robust shot detection and camera-motion compensation. Experiments show that forecasting trajectory tokens achieves category-agnostic, data-efficient prediction, outperforms state-of-the-art baselines, and generalizes to rare species and morphologies, providing a foundation for predictive visual intelligence in the wild.

Keywords: Point Trajectories · Predictive Visual Intelligence · Behavior Forecasting · Non-Rigid Motion

1 Introduction

Predicting the future motion of objects and agents is a fundamental capability of visual intelligence. In dynamic environments, agents, from animals in the wild to humans in social settings, must anticipate the behavior of others in order to act effectively or survive. Despite major advances in visual recognition and generation, predicting behavior remains one of the least understood capabilities of modern vision systems.

A key reason for this gap is the lack of an appropriate representation for behavior. In language, prediction is enabled by discrete tokens that structure the modeling problem. Vision systems lack an analogous token for motion and behavior. In this work, we show that dense point trajectories can serve as such tokens, enabling scalable prediction of behavior across diverse agents. To understand why such a representation is needed, consider the limitations of existing approaches. Forecasting directly in pixel space is universal but poorly structured: while recent video diffusion models can generate realistic short clips,

* Equal advising contribution.

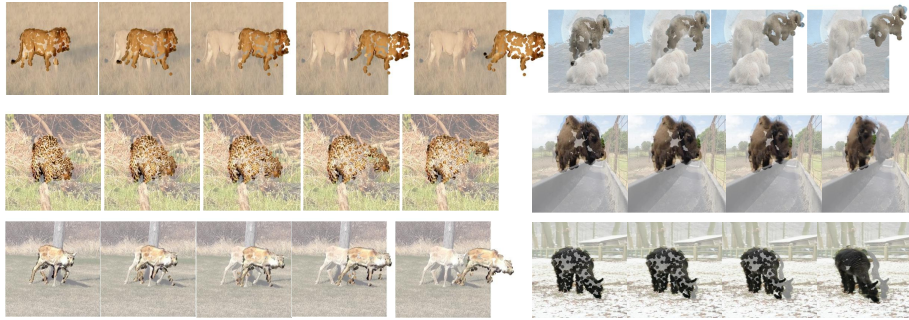


Fig. 1: Dense point trajectories act as visual tokens for behavior, enabling scalable forecasting of complex motion across diverse species. Our method takes as input a single RGB image, a history of motion, and an optional high-level motion vector, and forecasts future animal motion in the form of point trajectories. For each predicted point trajectory, we translate a small circular patch of the input image along the motion trajectory and superimpose it on the input image (**no pixels are generated!**). Leftmost shows the start locations on the input frame; the rest is forecast by our model. Our method is capable of forecasting many different animal species and behaviors, even long-tail ones—the polar bear on the top right is only present in 0.31% of the training data, the caribou on the bottom left in 0.025%, and the alpaca on the bottom right in 0.50%. See more results at <https://motion-forecasting.github.io/>.

forecasting behavior directly in pixel space entangles appearance, lighting, and camera motion with object dynamics, making the learning problem unnecessarily complex and data inefficient. At the opposite extreme, parameterized 3D models provide compact and physically valid representations for forecasting, but rely on strong object-specific priors and therefore apply only to a small number of carefully modeled categories, such as humans [43] and animals [62, 70, 80, 90–93]. Even in these settings they often miss fine-grained deformation and shape variation. Without an intermediate representation that captures motion structure while remaining general, scalable behavior prediction remains difficult. We therefore seek a representation that introduces structure without sacrificing generality.

We therefore propose dense point trajectories as visual tokens for behavior, providing a structured representation for forecasting motion across diverse agents. While sparse points carry little semantic meaning when static, their motion reveals rich information about 3D structure and intent, as demonstrated in Johansson’s classical biological motion studies [29] and subsequent work [2, 13, 17, 21, 35]. Representing behavior as evolving 2D point tracks focuses prediction directly on motion dynamics while remaining agnostic to appearance and scene variation. This formulation is significantly more data efficient than forecasting pixels directly [4, 5] and naturally applies to arbitrary non-rigid agents without requiring category-specific models. Point trajectories therefore occupy a principled middle ground between raw pixels and full 3D parameterizations: structured enough to constrain prediction, yet general enough to scale across species, morphologies, and environments.

Building on this abstraction, we introduce a diffusion transformer that forecasts behavior from short motion histories. Unlike prior trajectory-based approaches designed for robotics or rigid scenes [5, 11, 79], our formulation models motion for non-rigid agents in the wild. The model predicts future behavior as an unordered set of point trajectories (Fig. 1), treating each trajectory as a token augmented with local visual context from DINOv3 features. The architecture jointly models trajectories while explicitly reasoning about occlusion and visibility, enabling coherent predictions of complex non-rigid motion. Our model learns diverse motion patterns including gait, cyclical, and linear behaviors, and forecasts future motion across a wide range of species, outperforming state-of-the-art baselines. Training on the broad diversity of motion found in nature further enables generalization to previously unseen categories and morphologies of animate agents.

To study long-tailed motion at scale, we focus on unconstrained video of animals in the wild. Animals provide a particularly challenging testbed for motion prediction: they exhibit highly diverse morphologies and behavior patterns, and data for many species is inherently sparse. A representation that succeeds in this regime must generalize across categories without relying on category-specific models. We develop a large-scale pipeline for isolating animal motion from raw video, including robust shot detection and camera-motion compensation, and curate 300 hours of annotated footage for behavior forecasting which we release with this paper. Using this in-the-wild data, we demonstrate that our approach operates on tracks extracted from unconstrained video and is robust to the noise and partial observability inherent in real-world tracking. This dataset reveals previously unreported statistical structure in animal motion, and provides a foundation for studying predictive visual intelligence in natural environments.

Our results show that forecasting point trajectories enables structured, category-agnostic prediction of complex real-world behavior. These results suggest that trajectory tokens provide a scalable foundation for predictive visual intelligence.

Our contributions are:

1. **Point trajectories as visual tokens for behavior forecasting.** We introduce dense point tracks as a compact mid-level representation for modeling long-tailed natural-world behavior that disentangles motion from appearance and generalizes beyond category-specific 3D models.
2. **A diffusion transformer for trajectory forecasting.** We design a DiT-based architecture that treats trajectories as tokens and predicts diverse futures of non-rigid behavior from short histories while explicitly reasoning about occlusion in unordered track sets.
3. **MammalMotion, a large-scale dataset of animal motion.** We develop a robust pipeline for isolating animal motion in unconstrained video and release 300 hours of annotated footage for behavior forecasting in the wild.

2 Related Work

Pixel Forecasting. When it comes to forecasting visual information, pixels have been the natural choice for several years. Early approaches predicted future pixels deterministically, as a regression problem [54, 59, 69], which is exceedingly challenging, since the problem is ambiguous, and leads to blurry predictions.

While GANs [12, 73, 78] and variational models [37] were once promising, many modern approaches use diffusion models [26] which produce sharp videos [22, 25, 27, 81, 84, 86] – and have brought on a creative video revolution [19, 53]. However, training models directly on video is expensive and data-inefficient and models still struggle with hallucinations and basic physical interactions [3, 9, 31].

Point Track Forecasting. Several works have pushed the frontier in high-quality point-tracking [14, 15, 32, 33, 89], with broad applications across different computer vision tasks. When it comes to forecasting point tracks, the most significant advancements have come from the robotics domain. Any-point Trajectory Modeling [79] introduced the paradigm of first training a regression model to predict point tracks from an image and language instruction, and learning a robot policy on top of the track prediction model. Several approaches have followed in this direction [5, 11, 18, 51, 82, 83, 87]. These works have explored different architectures for forecasting point tracks such as conditional diffusion transformers [5, 11] and latent diffusion models [82], all with the end-goal of learning good robotic manipulation policies. Similarly, [76] applies DiTs to forecast frozen video encodings along with future decoded point tracks. We draw inspiration from these conditional DiT architectures but focus on a different application, forecasting motion in the complex domain of in-the-wild animal data.

Most recently, [7] showed that point-track forecasting outperforms pixel generation for simple Kubric [20] object motions. Our work provides further evidence that point tracks can be a more data-efficient representation for motion, by expanding their scope to more challenging and non-rigid domain of in-the-wild animal data.

Behavioral Forecasting in Computer Vision.

Beyond pixels and tracks, there has also been work focusing on forecasting the behavior of intelligent entities as well as their interactions. For example, human trajectory prediction has a long history with a variety of approaches [34, 61]. For direct trajectory prediction, these range from RNN based approaches [64, 74] to VAEs [46] and GANS [24], leveraging generative modeling of future human trajectories. Many recent papers also focus on utilizing scene context [63, 71] for human trajectory forecasting. Behavioral forecasting has also been extensively explored in the context of autonomous driving [38, 39, 48, 66]. Relatively few vision papers have focused on forecasting animal motion. QuadForecaster [52] predicted the poses of animals in constrained contexts while [42] demonstrated a proof of concept of their approach on fish and mice. In contrast, our approach leverages large and diverse datasets and forecasts animal motion on a general level.

Animal Pose, Motion and Behavior. Ethology, the study of animal behavior, has a long history [44, 45, 72, 75]. Recent advances in computing and machine

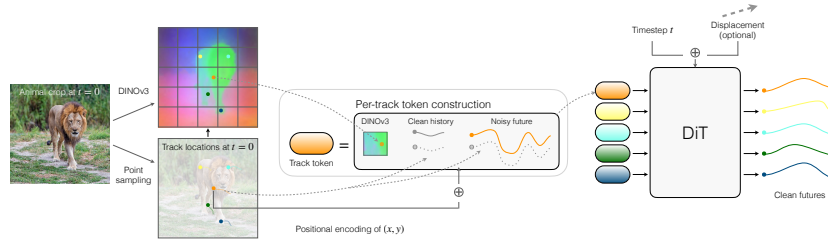


Fig. 2: Architecture. Given an input frame and (noisy) tracks, we construct a single token for every track, which includes a DINO feature at the start location, the motion history (gray), and the noisy track (orange) values, both with occlusion indicators (gray, dashed). After projection, we add a position encoding for the initial point location. Tokens are stacked and fed to a transformer (DiT) to predict clean tracks (right).

learning show promise in aiding discoveries – e.g. the emerging field of Computational Ethology [1] where computer vision and automated motion analysis plays a major role. For example, work such as DeepLabCut [36, 47, 49] and SLEAP [56] have accelerated annotating poses of animals in video. Video analysis has aided the ethology of a wide range of animals, from jumping plant lice [57], mice [85] and even zebrafish larvae [65]. However, as [1] notes, for most of these approaches, humans still need to manually annotate behaviors in training data – which can be subjective due to varied spatial and temporal scales, and limited by human perception and difficulties in discovering new behaviors. Our work leverages massive datasets of unlabeled videos and is a step towards automatic motion understanding of animals.

There has also been a line of work on reconstructing animal pose in 3D [91, 92], moving towards accurate reconstructions of individual species [62, 80, 90, 93], or creating species-specific models to generate 3D animal motion [70]. These works provide insights into individual animal species, but our work focuses on developing an approach that is data-efficient and can generalize to many species including long tail ones.

3 Forecasting Point Trajectories with a Diffusion Model

We present a diffusion-based approach for generating animal motion as a sequence of point tracks. Unlike video generation models that predict RGB pixels, our method operates directly on point trajectories. Given a single observation frame and optional conditioning information like motion history or desired velocity, our model generates plausible future trajectories.

3.1 Problem Formulation

We represent the motion of a single animal as a set of N point tracks, where each track describes the 2D trajectory of a single surface point over a time horizon of T timesteps. Formally, we aim to predict a set of tracks $\mathbf{X} \in \mathbb{R}^{T \times N \times 2}$. Each point

track $\mathbf{x}_n = [(x_n^1, y_n^1), (x_n^2, y_n^2), \dots, (x_n^T, y_n^T)]$, consists of a sequence of normalized coordinates (x_n^t, y_n^t) where t indexes time. Points may become occluded, in which case we assume the location is unknown: we represent the occlusion state as $\mathbf{O} \in \mathbb{R}^{T \times N}$, where $\mathbf{O}_n^t \in [0, 1]$ indicates that it the n 'th point is visible (1) or occluded (0) at time t .

Our forecasting model learns a conditional generative distribution:

$$p(\mathbf{X}_{T_c+1:T}, \mathbf{O}_{T_c+1:T} | \mathbf{I}, \mathbf{X}_{1:T_c}, \mathbf{O}_{1:T_c}, \mathbf{d})$$

Where $\mathbf{I}$ is the first frame, $\mathbf{X}_{1:T_c}$ and $\mathbf{O}_{1:T_c}$ are the observed conditioning motion history tracks and occlusion states over the first T_c timesteps, and a single optional 2D displacement vector $\mathbf{d} \in \mathbb{R}^2$ describing the average motion of tracks from the last frame: $\mathbf{d} = \sum_{n=1}^N \mathbf{O}_n^T [(x_n^T, y_n^T) - (x_n^1, y_n^1)] / \sum_{n=1}^N \mathbf{O}_n^T$. The model generates future trajectories $\mathbf{X}_{T_c+1:T}$ and occlusion states $\mathbf{O}_{T_c+1:T}$ conditioned on this observed history and the optional conditioning. Because the main challenge is to predict the track positions $\mathbf{X}_{T_c+1:T}$, which are a high-dimensional and continuous value, we draw inspiration from prior work [5] and model distribution with a diffusion process.

Parameterization of the diffusion target. Diffusion involves adding Gaussian noise to the inputs (tracks and occlusions) and training a network to denoise them. While we could directly denoise $\mathbf{X}$ and $\mathbf{O}$, there are two problems. First, $\mathbf{X}$ has missing values for occluded points (prior work, e.g. [79], assumes there are no missing points, which is untenable for longer horizons). Second, $\mathbf{X}$ values are extremely correlated, and most of the variance is due to the initial point that is tracked rather than due to the motion itself. We therefore reparameterize the tracks to improve training dynamics. Specifically, we construct the diffusion target $\mathbf{Z}_0^{\text{diff}} = \{\gamma \mathbf{V}, \beta \mathbf{O}\}$ where the n 'th row of $\mathbf{V} \in \mathbb{R}^{N \times T \times 2}$ is $[(\dot{x}_n^1, \dot{y}_n^1), (\dot{x}_n^2, \dot{y}_n^2), \dots, (\dot{x}_n^T, \dot{y}_n^T)]$, and γ and β are scaling parameters so the overall variance roughly matches the noise distribution. Here $\dot{x}_n^t = (x_n^{t+1} - x_n^t)$, and $\dot{y}_n^t = (y_n^{t+1} - y_n^t)$. We interpolate occluded values $\dot{x}_n^t = (x_n^i - x_n^j)/(i - j)$ where i and j are the next and previous visible points (for occluded points at the end of the sequence, which don't have any such j , we simply use 0). We don't do any special preprocessing for the occlusion indicator; even though it's discrete, we find that the model can still denoise to the discrete values provided that they are scaled appropriately.

3.2 Diffusion Process

Following DDPM [26], we define a forward diffusion process that gradually corrupts the diffusion targets $\mathbf{Z}_0^{\text{diff}}$ with Gaussian noise. The forward process over $\tau = 1, 2, \dots, S$ diffusion steps is:

$$q(\mathbf{Z}_\tau^{\text{diff}} | \mathbf{Z}_0^{\text{diff}}) = \mathcal{N}(\mathbf{Z}_\tau^{\text{diff}}; \sqrt{\bar{\alpha}_\tau} \mathbf{Z}_0^{\text{diff}}, (1 - \bar{\alpha}_\tau) \mathbf{I}), \quad (1)$$

where $\bar{\alpha}_\tau = \prod_{s=1}^\tau \alpha_s$ with $\alpha_s = 1 - \beta_s$ and $\{\beta_s\}_{s=1}^S$ is a linear noise schedule from $\beta_1 = 0.0001$ to $\beta_S = 0.02$.

Our diffusion model, f_θ , learns to reverse this process by predicting the clean diffusable data $\mathbf{Z}_0^{\text{diff}}$ directly. The training objective minimizes the L1 loss:

$$\mathcal{L} = \mathbb{E}_{\mathbf{Z}_0^{\text{diff}}, \tau, \epsilon} [\|\mathbf{Z}_0^{\text{diff}} - f_\theta(\mathbf{Z}_\tau^{\text{diff}}, \mathbf{Z}^{\text{cond}}, \tau)\|_1], \quad (2)$$

where $\mathbf{Z}_\tau^{\text{diff}} = \sqrt{\bar{\alpha}_\tau} \mathbf{Z}_0^{\text{diff}} + \sqrt{1 - \bar{\alpha}_\tau} \epsilon$ with $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$, and $\mathbf{Z}^{\text{cond}} = \{\mathbf{I}, \mathbf{X}_1, \mathbf{V}_{1:T_c}, \mathbf{O}_{1:T_c}, \mathbf{d}\}$ is conditioning information including the image $\mathbf{I}$, as well as motion and occlusion history and desired displacement, if available.

Diffusion Transformer Architecture. We now turn to the description of f_θ , which predicts the clean tracks given noisy tracks and conditioning information. We do not assume that tracks are given in any meaningful order or on any grid. However, similar to [5], we note that a transformer model, where each token corresponds to a track, can handle the permutation invariance, as long as we include relevant conditioning information within each token that encodes what the track corresponds to. This design means that the model can easily reason about the full motion forecast for a single point (since everything about a point is encoded within the same point), and yet it can also easily compare and contrast nearby points via attention. It also means that we can make our network is invariant to the input ordering of the tracks.

Figure 2 shows our overall architecture. Each input token corresponds to a full point trajectory; that is, we construct a token for each track before stacking them into a matrix to pass to the transformer. Each token contains all per-track conditioning information: image features, and clean history of conditioning velocities and occlusions $\{\mathbf{I}, \mathbf{X}_1, \mathbf{V}_{1:T_c}, \mathbf{O}_{1:T_c}\}$, as well as the noisy diffusion target for the track. We can then predict the clean data for each track via simple linear projection from the transformer’s output.

We construct a token for the n ’th point track in the following way (Fig. 2). We start with a visual feature derived from I , the image frame at time $t = 1$. We extract the full bounding box around the animal plus a 50% margin, and compute image features from a frozen DINOv3 [67], which should capture priors about animal parts. We then extract a feature for the track’s initial location (x_n^1, y_n^1) using bilinear interpolation. Next, we encode the velocity and occlusion history $(\dot{x}_n^{1:T_c-1}, \dot{y}_n^{1:T_c-1}, \mathbf{O}_n^{1:T_c})$; we embed the velocities $\dot{x}_n^{1:T_c-1}, \dot{y}_n^{1:T_c-1}$ using a sinusoidal embedding and scale by γ ; we keep the occlusions $\mathbf{O}_n^{1:T_c}$ as scalar and multiply by β . This component of the token is set to zero in the case where the conditioning is not provided. Finally, we add the noisy velocities and occlusion values $\mathbf{Z}_\tau^{\text{diff}} = \{\hat{\mathbf{V}}, \hat{\mathbf{O}}\}$. The full token construction is the concatenation of the clean conditioning DINOv3 features, the clean conditioning velocity history embedding and the occlusion history, the noisy velocities, and the noisy occlusions, along the channel dimension: $\mathbf{Z}_n = [\mathbf{Z}_n^{\text{diff}}, \mathbf{f}_n^{\text{DINO}}, \mathbf{V}_{1:T_c}, \mathbf{O}_{1:T_c}]$.

We project each token to the transformed dimension D_T and add a position encoding. Unlike sequence models, where the added position encoding is derived from the sequence index, we derive our position encoding from the initial location. (x_n^1, y_n^1) . We use a simple sinusoidal position encoding with length D_T and add it to the track token embedding. Finally we apply a standard DiT transformer [55], before linearly projecting the final layer to the dimension of each track in $\mathbf{Z}^{\text{diff}}$.

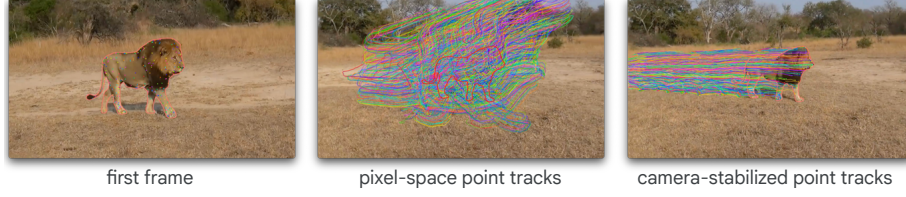


Fig. 3: Our processed data before and after camera stabilization. Given a first frame (left), the middle image shows the point tracks in pixel space, where the motion of the animals and the camera (panning, zooming out) are entangled. On the right are our point tracks in camera-stabilized space. We release all of our annotations, including the camera-stabilized point tracks.

The final conditioning information is global, rather than per-track: the diffusion timestep τ and optionally the desired total displacement d . We embed these values via a linear embedding, zeroing out the embedding for d in the cases where it is not given, and use adaptive layer norm [55] as input directly at each layer of the diffusion model, as is typical for encoding the diffusion timestep in a diffusion transformer. See appendix 3.1 for further details on the network architecture, embeddings, and other hyperparameters.

3.3 Sampling with DDIM

For efficient inference, we use the DDIM sampling algorithm [68], which enables deterministic sampling with fewer steps than the training diffusion process. DDIM defines a non-Markovian forward process that preserves the same marginals $q(\mathbf{Z}_\tau|\mathbf{Z}_0)$ but allows skipping diffusion timesteps during sampling.

Given the model’s prediction $\hat{\mathbf{Z}}_0 = f_\theta(\mathbf{Z}_\tau, \tau, \mathbf{d})$ at diffusion timestep τ , we compute the next state $\mathbf{Z}_{\tau-\Delta}$ as:

$$\boldsymbol{\epsilon}_\theta = \frac{\mathbf{Z}_\tau - \sqrt{\bar{\alpha}_\tau} \hat{\mathbf{Z}}_0}{\sqrt{1 - \bar{\alpha}_\tau}}, \quad (3)$$

$$\mathbf{Z}_{\tau-\Delta} = \sqrt{\bar{\alpha}_{\tau-\Delta}} \hat{\mathbf{Z}}_0 + \sqrt{1 - \bar{\alpha}_{\tau-\Delta} - \sigma_\tau^2} \boldsymbol{\epsilon}_\theta + \sigma_\tau \boldsymbol{\epsilon}, \quad (4)$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$ and $\sigma_\tau = \eta \sqrt{(1 - \bar{\alpha}_{\tau-\Delta}) / (1 - \bar{\alpha}_\tau)} \sqrt{1 - \bar{\alpha}_\tau / \bar{\alpha}_{\tau-\Delta}}$ controls stochasticity. We use deterministic sampling ($\eta = 0$) with 100 diffusion steps instead of the full 1000 training steps, yielding $10\times$ speedup with minimal quality degradation. After sampling in velocity space, we convert back to absolute coordinates via cumulative summation: $x_n^t = x_n^1 + \sum_{s=1}^{t-1} v_n^{x,s}$ and $y_n^t = y_n^1 + \sum_{s=1}^{t-1} v_n^{y,s}$ for each point n and trajectory time t .

4 Data Processing for In-the-Wild Animal Videos

Isolating high-fidelity animal motion trajectories from in-the-wild video requires extensive data processing. To this end, we propose a comprehensive pipeline that

leverages state-of-the-art perception models to filter data, detect animals, track points on them, and disentangle camera motion from animal motion, ultimately yielding a dataset of stabilized point trajectories suitable for training our animal motion generation model.

Predicting motion is simpler in world coordinates, as much of the scene remains static and accelerations are predictable from forces [50]. While there has been much recent work on extracting full 3D coordinates of objects and cameras [77, 88], these algorithms remain unreliable for in-the-wild video of uncommon categories. However, we note that for naturalist videos, most of the time the cameras are either stationary or shot from a long distance so parallax is minimal. Therefore, we approximate a world coordinates via simple camera stabilization and normalizing the coordinates with respect to the animal’s initial position, as shown in Figure 3.

We start by densely tracking points on our videos with BootsTAPIR [14], and detect individual shots by thresholding the number of points that move smoothly from one frame to the next. For each shot, we run Grounding-DINO [41] to get initial animal segments and track them with VideoSAM [60]. To perform camera stabilization, we remove tracks that fit mostly inside animal segments, and then perform RANSAC on the rest to estimate a homography H_t for every frame t .

We then extract point tracks within the animal segment, with a bias towards thin structures to ensure they are represented in the resulting tracks. We then construct training examples by selecting an animal and a particular starting frame t . We extract the input image by taking a bounding box around the segment and expanding it by 50% on each side. We transform all other points with respect to this bounding box using the homographies (i.e. multiply each point on frame t' by $H_t H_{t'}^{-1}$). We then normalize all coordinates with respect to the first bounding box, so that $(0, 0)$ corresponds to the upper-left corner and $(1, 1)$ the bottom right. See appendix Sec. 1 for more details. We release the resulting MammalMotion dataset.

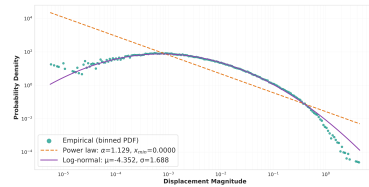


Fig. 4: Animal motion follows a log normal distribution: We plot a histogram of animal displacement. Horizontal axis is a binned log displacement, while vertical axis is log frequency. We find that log-normal (purple) fits much better than both a power law (orange).

Motion distribution. To validate our motion processing, we compute a histogram of the average displacement (i.e. average distance between start and end for points that remain visible) in Figure 4. While one might expect animal motion to follow a power law, we instead find that a log-normal distribution fits far better (i.e., the log displacements are normally distributed). Such distributions have been found in other datasets of animal motions, e.g. Lévy flights [8, 23, 28], foraging decisions in rats [30], and general spontaneous behavior in animals [58], and are suggested to imply that motion magnitude the result of a *multiplicative* interaction of independent factors. We believe this is the first time such a result

has arisen from such a diverse dataset of many different species, and without any painstaking manual annotation or use of tracking devices.

5 Experimental Setup

5.1 Experimental Dataset

Before processing the data to create MammalMotion, we filter the full 539-hour MammalNet dataset [10], cutting it down to just under 300 hours. Videos were excluded if they failed to meet minimum requirements for temporal and spatial resolution or displayed a low dynamic range, see appendix Sec. 1 for details.

We evaluate our approach on our filtered *all-species* dataset spanning the entire MammalNet taxonomy, as well as a *Panthera*-only subset comprising lions, tigers, and leopards. For each configuration, we construct evaluation sets by randomly sampling from the validation split with different levels of motion. In the all-species setting, random samples are also drawn using stratified sampling across species $\times$ behavior classes to ensure balanced representation of rare categories. In contrast, the *Panthera*-only setting uses uniform random sampling due to its more homogeneous taxonomy. In both cases, we draw even amounts of samples where the animal averages the following amounts of frame-to-frame absolute motion: less than half a pixel, half to 1.5 pixels, and greater than 1.5 pixels.

5.2 Metrics

We evaluate our model’s performance using a suite of metrics that assess both example-level trajectory accuracy and distribution-level motion. All metrics are computed on predicted trajectories compared against our ground truth.

Distribution-Level Motion Statistics: we apply several metrics to the overall distributions of predicted trajectories.

Fréchet Distance (FD). To assess whether our model captures the statistical properties of animal motion, we compute the Fréchet distance [16] between predicted and ground truth trajectory distributions. It fits multivariate Gaussian distributions to a set of vectors and compares them. We compute FD on two representations: first-order differences (velocities), and second-order differences (accelerations), capturing motion dynamics, and motion smoothness, respectively. Following prior work [76], we restrict this analysis to individual tracks visible in all predicted frames to ensure complete motion sequences.

Trajectory Variance. We measure the temporal variance of predicted trajectories $\text{Var}_{\text{pred}} = \text{Var}(\text{flat}(\mathbf{P}^{\text{pred}}))$ where $\mathbf{P}^{\text{pred}} \in \mathbb{R}^{N_{\text{samples}} \times T \times 2}$ is the matrix of all predicted track samples. This captures the diversity and magnitude of motion in generated trajectories. We report this alongside ground truth variance Var_{gt} to assess whether the model reproduces natural motion magnitudes.

Fréchet Video Motion Distance (FVMD). To evaluate temporal coherence, we use the Fréchet Video Motion Distance (FVMD) [40]. FVMD quantifies

the discrepancy between the distributions of motion feature vectors, where the features are local histograms of motion orientation and magnitude.

Example-Level Metrics: Since diffusion models are stochastic, we follow common practice and report best-of- K metrics by sampling $K = 5$ predictions with different random seeds for each test example. For metrics where lower is better (ADE, FDE, VMD), we compute $\min_k \text{metric}_k$ for each example, and average across examples. When higher is better (PWT) we use max.

Displacement Error (ADE and FDE). Following standard protocols, we evaluate trajectory accuracy using Average Displacement Error (ADE) and Final Displacement Error (FDE). ADE is the mean squared Euclidean distance between predicted and ground truth trajectories for all visible points across the predicted timesteps. FDE measures endpoint accuracy at the terminal timestep T .

Points Within Threshold (PWT). As established in point tracking literature [15], we report the fraction of predicted points within pixel-wise distance thresholds of the ground truth $\delta \in \{1, 2, 4, 8, 16\}$, in pixel space, where the input bounding boxes are all resized to (256,256).

Video Motion Distance (VMD). This is a straightforward extension of FVMD to an example-level metric: we compute the same feature vector used for FVMD for both the sample and ground-truth, and report the average Euclidean distance.

5.3 Baselines

We first compare our approach against three non-learned baselines. We then also compare our approach with learned baselines ATM and Track2Act. All baselines and our model use $N_{\text{cond}} = 4$ and predict 28 timesteps at 15 FPS.

No-Motion Baseline. The simplest prediction strategy, repeating the last conditioning position for all future timesteps: $\hat{\mathbf{p}}_t = \mathbf{p}_{N_{\text{cond}}-1}$ for $t \geq N_{\text{cond}}$.

Constant Velocity Baseline. We estimate a per-point-track velocity from conditioning frames as $\mathbf{v} = (\mathbf{p}_{N_{\text{cond}}-1} - \mathbf{p}_0) / (N_{\text{cond}} - 1)$ and linearly extrapolate future positions: $\hat{\mathbf{p}}_t = \mathbf{p}_0 + t \cdot \mathbf{v}$. This provides a simple physics-based predictor assuming constant motion dynamics.

Oracle Velocity Baseline. Uses ground truth average velocity computed from all of the points on the animal, giving a fair lower bound for the setting of our model that takes ground-truth displacement.

What Happens Next (WHN). WHN [7] aims for general-purpose point track forecasting, but the model architecture has a grid constraint that makes it difficult to train on our non-constrained data. Therefore we apply it zero-shot.

Any Trajectory Modeling (ATM). ATM’s [79] Track Transformer is a regression-based method. Similarly to our method, it treats each point track over time as a token. It masks out future timesteps and learns to regress these coordinates. ATM does not handle visibility, regresses on absolute xy-coordinates, and can only predict one plausible future. We train this baseline using our Panthera subset, using $N_{\text{cond}} = 4$.



Fig. 5: Samples from our model Sampling from our model with different random seeds (each row) and no displacement conditioning. The frame on the left is the input state after the motion history.

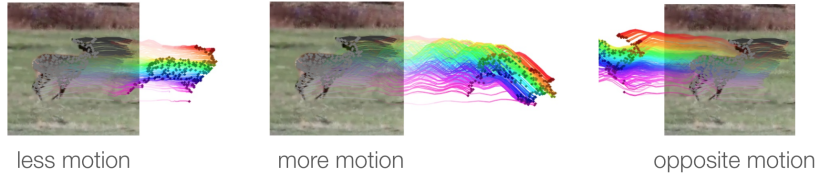


Fig. 6: Prompting our model with different levels of motion. We ask the model for less motion (left), more motion (middle), or motion in an opposite direction (right). Grey represents motion history, colors are our predictions.

Track2Act [5]. Most similar to our model, using a diffusion backbone, point track as tokens, and point conditioning setup. We use public Track2Act code and diffuse directly on absolute XY-coordinates, without any positional encoding following the original implementation. We use the [5]’s learned ResNet visual features integrated through AdaLN, and use a standard L2 loss. We omit the goal image (unavailable in our setting), condition the model solely on the initial image, and train the model using $N_{\text{cond}} = 4$.

6 Results

Samples from our model. Figure 1 shows qualitative results which exhibit convincing forecasts across a variety of behaviors: e.g. the lion’s legs follows natural articulation, the jaguar scouts out prey, the alpaca grazes naturally. This also works for rare animal categories. Figure 5 shows the diversity that our model produces with only different seeds: we see different frequencies of the grooming behavior for the jaguar and the dog’s head moves different directions. Fig. 6 demonstrates prompting our model for more motion, less motion, or motion in a different direction, and the model produces plausible behaviors consistent with these motions. Static visualizations do not do justice to motion accuracy; we urge our readers to check the results webpage for better visualizations of these effects.

Out-of-distribution examples. Fig. 7 displays qualitative results of our model on out-of-distribution data. We note that the MammalNet dataset does have videos that contain humans and other types of animals; the ostrich on the bottom right was found in our validation set, so this type of generalization is not



Fig. 7: Out-of-distribution: Our model *generalizes* to humans and non-mammals.



Fig. 8: Our Model vs. Stable Diffusion. Our approach can model the behavior of less common animals in our dataset such as hares, while conventional video models struggle with these animals.

surprising. However, the Lego robot (bottom left) and butterfly (top right) are unlike the expected MammalNet data distribution, but still observe physically plausible motion.

Comparison with Video Generation Models. Figure 8 displays a side-by-side comparison with Stable Video Diffusion [6]. As video models are forced to simulate information beyond motion—textures, lighting, and more—they can struggle to produce physically realistic motions. In contrast, because we model trajectory tokens directly, we can produce more realistic biological behavior with less compute and data. These extend the findings of [7] on synthetic data of rigid objects to the in-the-wild nonrigid setting. E.g., our model is able to forecast the foraging behavior of a hare even though hares only constitute 0.39% of the training data. The video model not only struggles to model this behavior but even fails to model the animal itself, morphing the ears of the hare into wings.

6.1 Quantitative Comparison with Baselines

Results comparing our method with baselines are presented in table 1 and table 2 for our full MammalMotion dataset and table 3 and table 4 for the *Panthera* genus subset. Each table shows models trained (for learned baselines) and evaluated either on the full dataset or restricted to the *Panthera* subset.

Simple baselines, such as the no-motion and constant-velocity models, can achieve competitive performance on the combined dataset due to the prevalence of low-motion examples. Tables 1-4 in the appendix show results for all motion buckets. We see that on the low and medium motion data subsets (where an animal might be eating, sleeping, or just not moving), an unlearned baseline of just predicting no motion outperforms the other baselines on several metrics. These simple baselines, however, fail for higher motion, and particularly for FVMD which accurately scores motion statistics. Interestingly, WHN gives an accurate acceleration distribution despite not being trained on this data, yet

Table 1: Distribution-level quantitative results on **All Data**. FD values are multiplied by 10^3 ; Variance values are multiplied by 10^5 ; FVMD values are divided by 10^3 . V and A refer to velocity and acceleration respectively. Best results in **bold**, second best underlined. $\uparrow$ indicates higher is better; $\downarrow$ indicates lower is better.

	Method	FD(V) $\downarrow$	FD(A) $\downarrow$	Var(V)	Var(A)	FVMD $\downarrow$
High motion	GT	-	-	31.5	8.94	-
	No motion	27.1	7.51	0	0	481.99
	Constant vel	13.7	7.51	23.8	0	210.47
	WHN	25.2	3.19	1.1	3.34	280.77
	Track2Act	<u>11.8</u>	4.81	9.89	1.63	<u>76.28</u>
	Ours (uncond)	8.96	<u>3.74</u>	13.1	1.68	49.30
	Ours (cond)	4.86	3.33	28.3	2.14	40.24
Combined	GT	-	-	5.41	1.82	-
	No motion	4.66	1.53	0	0	204.14
	Constant vel	<u>2.59</u>	1.53	4.11	0	89.77
	WHN	5.34	0.691	1.23	3.7	94.7
	Track2Act	<u>2.52</u>	1.09	2.37	0.516	<u>26.17</u>
	Ours (uncond)	1.96	<u>0.877</u>	2.94	0.453	17.0
	Ours (cond)	1.07	0.778	6.26	0.57	14.38

Table 2: Example-level quantitative results on **All Data**. Best results in **bold**, second best underlined. For non-learned baselines we report single-sample metrics; for WHN, Track2Act, and Ours we report best of $K = 5$.

	Method	ADE $\downarrow$	FDE $\downarrow$	VMD $\downarrow$	PWT $\uparrow$
High motion	No motion	0.325	0.596	6.50	12.44%
	Constant vel	0.286	0.591	5.02	11.94%
	WHN	0.262	0.538	5.74	11.62%
	Track2Act	0.136	<u>0.294</u>	4.50	21.79%
	Ours (uncond)	0.119	0.275	4.33	26.01%
	Ours (cond)	0.068	0.103	4.25	31.50%
	Ours (cond)	0.068	0.103	4.25	31.50%
Combined	No motion	0.099	0.180	4.82	53.94%
	Constant vel	0.104	0.215	5.02	41.15%
	WHN	0.105	0.200	4.85	29.92%
	Track2Act	<u>0.053</u>	<u>0.110</u>	<u>3.48</u>	<u>56.59%</u>
	Ours (uncond)	0.046	0.102	3.31	60.01%
	Ours (cond)	0.028	0.042	3.26	63.48%
	Ours (cond)	0.028	0.042	3.26	63.48%

fails to estimate overall velocity and other statistics well (qualitatively it gives low-motion, jittery predictions that don’t match animal skeletons).

More advanced baselines, including ATM and Track2Act—both retrained on the *Panthera* subset, with Track2Act additionally trained on the full dataset—give predictions that are closer to the ground truth in terms of endpoint error and velocity statistics. These methods actually perform worse in terms of acceleration and point-level accuracy, suggesting they learn overall motion but miss motion details.

In contrast, our method, trained from scratch on MammalMotion data (or the *Panthera* subset), substantially outperforms others in prediction accuracy on every metric. Furthermore, our method can take the true velocity as conditioning to improve results even further, even though for many metrics simply using the oracle velocity provides little boost.

Comparing our model trained on the *Panthera* data subset vs the full dataset, our model trained on the full data is substantially better, e.g. FVMD for high motion examples falls from 84.8 to 49.3 and PWT rises from 20.6 to 26.0. This isn’t because the full dataset is easier than the *Panthera* subset; other baselines actually perform similarly or worse on these metrics. Instead, this suggests that training on the full dataset improves performance due to transfer between species.

7 Conclusion

Our paper demonstrates a data-efficient and yet highly-general framework for motion forecasting. The critical ingredient is a token representation based on

Table 3: Distribution-level quantitative results on **Panthera**. FD values are multiplied by 10^3 ; Variance values are multiplied by 10^5 ; FVMD values are divided by 10^3 . V and A refer to velocity and acceleration respectively. Best results in **bold**, second best underlined. $\uparrow$ indicates higher is better; $\downarrow$ indicates lower is better.

	Method	FD(V) $\downarrow$	FD(A) $\downarrow$	Var(V)	Var(A)	FVMD $\downarrow$
High motion	GT	-	-	29.5	10.8	-
	No motion	16.6	5.61	0	0	335.41
	Constant vel	7.49	5.61	37.3	0	149.52
	WHN	15.2	3.27	1.37	4.11	247.56
	ATM	6.52	6.18	10	6.99	112.71
	Track2Act	<u>6.32</u>	5.06	8.01	0.45	<u>104.85</u>
	Ours (uncond)	3.71	<u>4.3</u>	12.8	1.02	84.79
	Oracle vel	5.7	5.61	20.9	0	218.73
	Ours (cond)	2.82	4.19	16.6	1.18	79.38
Combined	GT	-	-	6.93	2.66	-
	No motion	3.77	1.38	0	0	149.53
	Constant vel	1.86	1.38	8.58	0	62.51
	WHN	3.37	<u>1.12</u>	1.43	4.29	86.89
	ATM	1.49	1.4	2.42	1.75	<u>35.50</u>
	Track2Act	<u>1.43</u>	1.21	1.89	.121	38.44
	Ours (uncond)	.874	1.05	2.94	.226	24.82
	Oracle vel	1.43	1.38	4.73	0	118.61
	Ours (cond)	.679	1.02	3.73	.262	24.90

Table 4: Example-level quantitative results on **Panthera**. Best results in **bold**, second best underlined. For non-learned baselines and ATM (single output), we report single-sample metrics; for WHN, Track2Act, and Ours we report best of $K = 5$.

	Method	ADE $\downarrow$	FDE $\downarrow$	VMD $\downarrow$	PWT $\uparrow$
High motion	No motion	.211	.393	5.51	13.22%
	Constant vel	.193	.413	<u>4.91</u>	<u>16.73%</u>
	WHN	.215	.393	5.82	10.24%
	ATM	.143	.262	5.95	16.31%
	Track2Act	<u>.135</u>	<u>.245</u>	5.04	16.72%
	Ours (uncond)	.107	.209	4.77	20.68%
	Oracle vel	.082	.095	6.03	17.05%
	Ours (cond)	.067	.097	4.61	27.31%
	No motion	.076	.138	4.02	<u>54.55%</u>
Combined	Constant vel	.079	.164	4.33	46.16%
	WHN	.086	.146	5.05	30.43%
	ATM	.057	.101	4.48	48.39%
	Track2Act	<u>.053</u>	<u>.092</u>	3.81	51.13%
	Ours (uncond)	.042	.078	3.62	58.11%
	Oracle vel	.035	.042	4.51	53.14%
	Ours (cond)	.028	.039	3.55	61.67%
	No motion	.076	.138	4.02	<u>54.55%</u>
	Constant vel	.079	.164	4.33	46.16%

individual trajectories, using DINO features and track velocities as input, and a transformer-based diffusion network to model them. Our results show strong performance for animal forecasting across many species, producing diverse samples and allowing control of output velocity. Results even transfer out of distribution. We believe this model can serve as a foundation for more general modeling of arbitrary object motion, and extensions may extract knowledge about animal species that are useful to ecologists, potentially saving the incredible labor that’s currently expended for behavior modeling [1].

8 Acknowledgments

We thank Noah Snavely for challenging us with general motion forecasting over a lovely Parisian lunch. We thank Andrew Zisserman, Drew Purves, Aleksander Holynski, Linyi Jin, Sander Dieleman, Mark Hamilton, Jathushan Rajasegeran, and Amanda Seed for helpful discussions and feedback. This work originated [in part] while the authors were visiting the Simons Institute for the Theory of Computing. This work was supported by ONR MURI N00014-21-1-280, and a NSF Graduate Fellowship to NT.

References

1. Anderson, D.J., Perona, P.: Toward a science of computational ethology. *Neuron* **84**(1), 18–31 (2014)

2. Atkinson, A.P., Dittrich, W.H., Gemmell, A.J., Young, A.W.: Emotion perception from dynamic and static body expressions in point-light and full-light displays. *Perception* **33**(6), 717–746 (2004)
3. Bansal, H., Lin, Z., Xie, T., Zong, Z., Yarom, M., Bitton, Y., Jiang, C., Sun, Y., Chang, K.W., Grover, A.: Videophy: Evaluating physical commonsense for video generation. In: The Thirteenth International Conference on Learning Representations (2025)
4. Bharadhwaj, H., Dwibedi, D., Gupta, A., Tulsiani, S., Doersch, C., Xiao, T., Shah, D., Xia, F., Sadigh, D., Kirmani, S.: Gen2act: Human video generation in novel scenarios enables generalizable robot manipulation. In: Conference on Robot Learning. pp. 3936–3951. PMLR (2025)
5. Bharadhwaj, H., Mottaghi, R., Gupta, A., Tulsiani, S.: Track2act: Predicting point tracks from internet videos enables generalizable robot manipulation. In: European Conference on Computer Vision (ECCV) (2024)
6. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127* (2023)
7. Boduljak, G., Karazija, L., Laina, I., Rupprecht, C., Vedaldi, A.: What happens next? anticipating future motion by generating point trajectories. *arXiv preprint arXiv:2509.21592* (2025)
8. Breed, G.A., Severns, P.M., Edwards, A.M.: Apparent power-law distributions in animal movements can arise from intraspecific interactions. *Journal of the Royal Society Interface* **12**(103) (2015)
9. Chefer, H., Singer, U., Zohar, A., Kirstain, Y., Polyak, A., Taigman, Y., Wolf, L., Sheynin, S.: Videojam: Joint appearance-motion representations for enhanced motion generation in video models. In: Forty-second International Conference on Machine Learning
10. Chen, J., Hu, M., Coker, D.J., Berumen, M.L., Costelloe, B., Beery, S., Rohrbach, A., Elhoseiny, M.: Mammalnet: A large-scale video benchmark for mammal recognition and behavior understanding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13052–13061 (2023)
11. Chen, Y., Li, P., Huang, Y., Yang, J., Chen, K., Wang, L.: Ec-flow: Enabling versatile robotic manipulation from action-unlabeled videos via embodiment-centric flow. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11958–11968 (October 2025)
12. Clark, A., Donahue, J., Simonyan, K.: Adversarial video generation on complex datasets. *arXiv preprint arXiv:1907.06571* (2019)
13. Cutting, J.E., Kozlowski, L.T.: Recognizing friends by their walk: Gait perception without familiarity cues. *Bulletin of the psychonomic society* **9**(5), 353–356 (1977)
14. Doersch, C., Luc, P., Yang, Y., Gokay, D., Koppula, S., Gupta, A., Heyward, J., Rocco, I., Goroshin, R., Carreira, J., et al.: Bootstrap: Bootstrapped training for tracking-any-point. In: Proceedings of the Asian Conference on Computer Vision. pp. 3257–3274 (2024)
15. Doersch, C., Yang, Y., Vecerik, M., Gokay, D., Gupta, A., Aytar, Y., Carreira, J., Zisserman, A.: Tapir: Tracking any point with per-frame initialization and temporal refinement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10061–10072 (2023)
16. Dowson, D.C., Landau, B.: The fréchet distance between multivariate normal distributions. *Journal of multivariate analysis* **12**(3), 450–455 (1982)

17. Fox, R., McDaniel, C.: The perception of biological motion by human infants. *Science* **218**(4571), 486–487 (1982)
18. Gao, C., Zhang, H., Xu, Z., Cai, Z., Shao, L.: Flip: Flow-centric generative planning for general-purpose manipulation tasks. arXiv preprint arXiv:2412.08261 (2024), <https://arxiv.org/abs/2412.08261>
19. Google DeepMind: Veo 3 technical report. Tech. rep., Google DeepMind (2025)
20. Greff, K., Belletti, F., Beyer, L., Doersch, C., Du, Y., Duckworth, D., Fleet, D.J., Gnanapragasam, D., Golemo, F., Herrmann, C., et al.: Kubric: A scalable dataset generator. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3749–3761 (2022)
21. Grossman, E., Donnelly, M., Price, R., Pickens, D., Morgan, V., Neighbor, G., Blake, R.: Brain areas involved in perception of biological motion. *Journal of cognitive neuroscience* **12**(5), 711–720 (2000)
22. Gu, X., Wen, C., Ye, W., Song, J., Gao, Y.: Seer: Language instructed video prediction with latent diffusion models. arXiv preprint arXiv:2303.14897 (2023)
23. Gunner, R., Wilson, R., Lurgi, M., Borger, L., Redcliffe, J., Shepard, E., Holton, M., Crofoot, M., Alagaili, A., Andrzejaczek, S., et al.: High resolution data reveal fundamental steps and turning points in animal movements (2024)
24. Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., Alahi, A.: Social gan: Socially acceptable trajectories with generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2255–2264 (2018)
25. Gupta, A., Yu, L., Sohn, K., Gu, X., Hahn, M., Fei-Fei, L., Essa, I., Jiang, L., Lezama, J.: Photorealistic video generation with diffusion models. In: ECCV (2024)
26. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *Advances in neural information processing systems* **33**, 6840–6851 (2020)
27. Höppe, T., Mehrjou, A., Bauer, S., Nielsen, D., Dittadi, A.: Diffusion models for video prediction and infilling. arXiv preprint arXiv:2206.07696 (2022)
28. Humphries, N.E., Queiroz, N., Dyer, J.R., Pade, N.G., Musyl, M.K., Schaefer, K.M., Fuller, D.W., Brunnenschweiler, J.M., Doyle, T.K., Houghton, J.D., et al.: Environmental context explains lévy and brownian movement patterns of marine predators. *Nature* **465**(7301), 1066–1069 (2010)
29. Johansson, G.: Visual perception of biological motion and a model for its analysis. *Perception & psychophysics* **14**(2), 201–211 (1973)
30. Jung, K., Jang, H., Kralik, J.D., Jeong, J.: Bursts and heavy tails in temporal and sequential dynamics of foraging decisions. *PLoS computational biology* **10**(8), e1003759 (2014)
31. Kang, B., Yue, Y., Lu, R., Lin, Z., Zhao, Y., Wang, K., Huang, G., Feng, J.: How far is video generation from world model: A physical law perspective. In: International Conference on Machine Learning. pp. 28991–29017. PMLR (2025)
32. Karaev, N., Makarov, Y., Wang, J., Neverova, N., Vedaldi, A., Rupperecht, C.: Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6013–6022 (2025)
33. Karaev, N., Rocco, I., Graham, B., Neverova, N., Vedaldi, A., Rupperecht, C.: Cotracker: It is better to track together. In: European conference on computer vision. pp. 18–35. Springer (2024)
34. Kitani, K.M., Ziebart, B.D., Bagnell, J.A., Hebert, M.: Activity forecasting. In: European conference on computer vision. pp. 201–214. Springer (2012)
35. Kozlowski, L.T., Cutting, J.E.: Recognizing the sex of a walker from a dynamic point-light display. *Perception & psychophysics* **21**(6), 575–580 (1977)

36. Lauer, J., Zhou, M., Ye, S., Menegas, W., Schneider, S., Nath, T., Rahman, M.M., Di Santo, V., Soberanes, D., Feng, G., et al.: Multi-animal pose estimation, identification and tracking with deeplabcut. *Nature Methods* **19**(4), 496–504 (2022)
37. Lee, A.X., Zhang, R., Ebert, F., Abbeel, P., Finn, C., Levine, S.: Stochastic adversarial video prediction. *arXiv preprint arXiv:1804.01523* (2018)
38. Lee, N., Choi, W., Vernaza, P., Choy, C.B., Torr, P.H., Chandraker, M.: Desire: Distant future prediction in dynamic scenes with interacting agents. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 336–345 (2017)
39. Li, S., Liu, C., Xu, X., Yeo, S.Y., Yang, X.: Future-aware interaction network for motion forecasting. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 7505–7515 (October 2025)
40. Liu, J., Qu, Y., Yan, Q., Zeng, X., Wang, L., Liao, R.: Fr\`echet video motion distance: A metric for evaluating motion consistency in videos. *arXiv preprint arXiv:2407.16124* (2024)
41. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499* (2023)
42. Liu, Z., Su, P., Wu, S., Shen, X., Chen, H., Hao, Y., Wang, M.: Motion prediction using trajectory cues. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 13299–13308 (October 2021)
43. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. In: *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pp. 851–866 (2023)
44. Lorenz, K.: Der kumpan in der umwelt des vogels. der artgenosse als auslösendes moment sozialer verhaltensweisen. *Journal für Ornithologie. Beiblatt.*(Leipzig) (1935)
45. Lorenz, K., Tinbergen, N.: Taxis und instinkthandlung in der eirollbewegung der graugans. *Zeitschrift für Tierpsychologie* (1938)
46. Mangalam, K., Girase, H., Agarwal, S., Lee, K.H., Adeli, E., Malik, J., Gaidon, A.: It is not the journey but the destination: Endpoint conditioned trajectory prediction. In: *European conference on computer vision*. pp. 759–776. Springer (2020)
47. Mathis, A., Mamidanna, P., Cury, K.M., Abe, T., Murthy, V.N., Mathis, M.W., Bethge, M.: Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nature neuroscience* **21**(9), 1281–1289 (2018)
48. Moon, S., Woo, H., Park, H., Jung, H., Mahjourian, R., Chi, H.g., Lim, H., Kim, S., Kim, J.: Visiontrap: Vision-augmented trajectory prediction guided by textual descriptions. In: *European Conference on Computer Vision*. pp. 361–379. Springer (2024)
49. Nath, T., Mathis, A., Chen, A.C., Patel, A., Bethge, M., Mathis, M.W.: Using deeplabcut for 3d markerless pose estimation across species and behaviors. *Nature protocols* **14**(7), 2152–2176 (2019)
50. Newton, I.: *Philosophiae Naturalis Principia Mathematica* (1687)
51. Niu, D., Sharma, Y., Xue, H., Biamby, G., Zhang, J., Ji, Z., Darrell, T., Herzig, R.: Pre-training auto-regressive robotic models with 4d representations. *arXiv preprint arXiv:2502.13142* (2025)
52. Noronha, I., Chowdhury, A., Bharti, S., Kaur, U.: Quadforecaster: Diffusion-based quadruped pose prediction for animal communication analysis. In: *The Thirty-Ninth Annual Conference on Neural Information Processing Systems workshop: AI for non-human animal communication*

53. OpenAI: Sora (12 2024), <https://openai.com/sora/>
54. Oprea, S., Martinez-Gonzalez, P., Garcia-Garcia, A., Castro-Vargas, J.A., Orts-Escolano, S., Garcia-Rodriguez, J., Argyros, A.: A review on deep learning techniques for video prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**(6), 2806–2826 (2022). <https://doi.org/10.1109/TPAMI.2020.3045007>
55. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 4195–4205 (2023)
56. Pereira, T.D., Tabris, N., Matsliah, A., Turner, D.M., Li, J., Ravindranath, S., Papadoyannis, E.S., Normand, E., Deutsch, D.S., Wang, Z.Y., et al.: Sleep: A deep learning system for multi-animal pose tracking. *Nature methods* **19**(4), 486–495 (2022)
57. Polajnar, J., Kvinikadze, E., Harley, A.W., Malenovsky, I.: Wing buzzing as a mechanism for generating vibrational signals in psyllids (hemiptera: Psylloidea). *Insect science* **31**(5), 1466–1476 (2024)
58. Proekt, A., Banavar, J.R., Maritan, A., Pfaff, D.W.: Scale invariance in the dynamics of spontaneous behavior. *Proceedings of the National Academy of Sciences* **109**(26), 10564–10569 (2012)
59. Ranzato, M., Szlam, A., Bruna, J., Mathieu, M., Collobert, R., Chopra, S.: Video (language) modeling: a baseline for generative models of natural videos. *arXiv preprint arXiv:1412.6604* (2014)
60. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714* (2024)
61. Rudenko, A., Palmieri, L., Herman, M., Kitani, K.M., Gavrilu, D.M., Arras, K.O.: Human motion trajectory prediction: A survey. *The International Journal of Robotics Research* **39**(8), 895–935 (2020)
62. Rüegg, N., Tripathi, S., Schindler, K., Black, M.J., Zuffi, S.: Bite: Beyond priors for improved three-d dog pose estimation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 8867–8876 (2023)
63. Salzmann, T., Chiang, H.T.L., Ryll, M., Sadigh, D., Parada, C., Bewley, A.: Robots that can see: Leveraging human pose for trajectory prediction. *IEEE Robotics and Automation Letters* **8**(11), 7090–7097 (2023)
64. Salzmann, T., Ivanovic, B., Chakravarty, P., Pavone, M.: Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In: *European conference on computer vision*. pp. 683–700. Springer (2020)
65. Scholz, L.A., Mancienne, T., Stednitz, S.J., Scott, E.K., Lee, C.C.: Plug-and-play automated behavioral tracking of zebrafish larvae with deeplabcut and sleap: pre-trained networks and datasets of annotated poses. *bioRxiv* (2025)
66. Seff, A., Cera, B., Chen, D., Ng, M., Zhou, A., Nayakanti, N., Refaat, K.S., Al-Rfou, R., Sapp, B.: Motionlm: Multi-agent motion forecasting as language modeling. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 8579–8590 (2023)
67. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khali-dov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. *arXiv preprint arXiv:2508.10104* (2025)
68. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: *International Conference on Learning Representations* (2021)
69. Srivastava, N., Mansimov, E., Salakhudinov, R.: Unsupervised learning of video representations using lstms. In: *International conference on machine learning*. pp. 843–852. PMLR (2015)

70. Sun, K., Litvak, D., Zhang, Y., Li, H., Wu, J., Wu, S.: Ponymation: Learning articulated 3d animal motions from unlabeled online videos. In: European Conference on Computer Vision. pp. 100–119. Springer (2024)
71. Thakkar, N., Mangalam, K., Bajcsy, A., Malik, J.: Adaptive human trajectory prediction via latent corridors. In: European Conference on Computer Vision. pp. 297–314. Springer (2024)
72. Tinbergen, N.: On aims and methods of ethology. *Zeitschrift für tierpsychologie* **20**(4), 410–433 (1963)
73. Tulyakov, S., Liu, M.Y., Yang, X., Kautz, J.: Mocogan: Decomposing motion and content for video generation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1526–1535 (2018)
74. Vemula, A., Mueller, K., Oh, J.: Social attention: Modeling attention in human crowds. In: 2018 IEEE international Conference on Robotics and Automation (ICRA). pp. 4601–4607. IEEE (2018)
75. Von Frisch, K.: The dancing bees, vol. 354. A Harvest (1953)
76. Walker, J.C., Vélez, P., Cabrera, L.P., Zhou, G., Kabra, R., Doersch, C., Ovsjanikov, M., Carreira, J., Ginosar, S.: Generalist forecasting with frozen video models via latent diffusion (2025), <https://arxiv.org/abs/2507.13942>
77. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
78. Wang, Y., Bilinski, P., Bremond, F., Dantcheva, A.: Imaginator: Conditional spatio-temporal gan for video generation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 1160–1169 (2020)
79. Wen, C., Lin, X., So, J., Chen, K., Dou, Q., Gao, Y., Abbeel, P.: Any-point trajectory modeling for policy learning (2023)
80. Wu, S., Li, R., Jakab, T., Rupprecht, C., Vedaldi, A.: Magicpony: Learning articulated 3d animals in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8792–8802 (2023)
81. Xing, Z., Dai, Q., Weng, Z., Wu, Z., Jiang, Y.G.: Aid: Adapting image2video diffusion models for instruction-guided video prediction. arXiv preprint arXiv:2406.06465 (2024)
82. Xu, M., Xu, Z., Xu, Y., Chi, C., Wetzstein, G., Veloso, M., Song, S.: Flow as the cross-domain manipulation interface. arXiv preprint arXiv:2407.15208 (2024)
83. Yang, J., Zhu, H., Wang, Y., Wu, G., He, T., Wang, L.: Tra-moe: Learning trajectory prediction model from multiple domains for adaptive policy conditioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6960–6970 (2025)
84. Yang, S., Zhang, L., Liu, Y., Jiang, Z., He, Y.: Video diffusion models with local-global context guidance. arXiv preprint arXiv:2306.02562 (2023)
85. Ye, S., Filippova, A., Lauer, J., Schneider, S., Vidal, M., Qiu, T., Mathis, A., Mathis, M.W.: Superanimal pretrained pose estimation models for behavioral analysis. *Nature communications* **15**(1), 5165 (2024)
86. Ye, X., Bilodeau, G.A.: Stdif: Spatio-temporal diffusion for continuous stochastic video prediction. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 6666–6674 (2024)
87. Yuan, C., Wen, C., Zhang, T., Gao, Y.: General flow as foundation affordance for scalable robot learning. arXiv preprint arXiv:2401.11439 (2024)
88. Zhang, C., Le Moing, G., Koppula, S., Rocco, I., Momeni, L., Xie, J., Sun, S., Sukthankar, R., Barral, J.K., Hadsell, R., Ghahramani, Z., Zisserman, A., Zhang,

- J., Sajjadi, M.S.M.: Efficiently reconstructing dynamic scenes one d4rt at a time. arXiv preprint (2025)
89. Zholus, A., Doersch, C., Yang, Y., Koppula, S., Patraucean, V., He, X.O., Rocco, I., Sajjadi, M.S., Chandar, S., Goroshin, R.: Tapnext: Tracking any point (tap) as next token prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9693–9703 (2025)
 90. Zuffi, S., Kanazawa, A., Berger-Wolf, T., Black, M.J.: Three-d safari: Learning to estimate zebra pose, shape, and texture from images" in the wild". In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5359–5368 (2019)
 91. Zuffi, S., Kanazawa, A., Black, M.J.: Lions and tigers and bears: Capturing non-rigid, 3d, articulated shape from images. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 3955–3963 (2018)
 92. Zuffi, S., Kanazawa, A., Jacobs, D.W., Black, M.J.: 3d menagerie: Modeling the 3d shape and pose of animals. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6365–6373 (2017)
 93. Zuffi, S., Mellbin, Y., Li, C., Hoeschle, M., Kjellström, H., Polikovsky, S., Hernlund, E., Black, M.J.: Varen: Very accurate and realistic equine network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5374–5383 (June 2024)