Schrödinger's Cat: Probabilistic Representation and Prediction of Potential Scene Kinematics

Timy Phan[*1,2], Jannik Wiese[*1,2], and Björn Ommer[1,2]

[1] CompVis @ LMU Munich, Germany
[2] Munich Center for Machine Learning (MCML)

Abstract. Predicting how a scene may evolve from partial observations requires reasoning about multiple possible futures rather than committing to a single trajectory. Existing approaches either generate appearance-dominated video predictions or sample a small number of trajectories without explicitly modeling the distribution of possible motion. We introduce Goal Aware Representations of Future kInEmatic Latent Distributions (GARFIELD), a probabilistic model of scene kinematics that learns a *structured spatio-temporal latent representation* of the distribution over possible futures given an image and optional *spatio-temporally sparse constraints*. The same latent representation enables both joint sampling of all trajectories and direct access to the underlying motion distribution through an efficient deterministic density decoder. As a result, uncertainty about future motion can be localized to specific scene elements and timesteps and progressively refined through additional constraints. Experiments demonstrate strong motion planning performance competitive with large video generation models while sampling trajectories 97× faster. Our method further estimates motion densities two orders of magnitude faster than Monte-Carlo sampling from motion generation models, enabling interactive exploration and uncertainty-aware planning.
Project page: `https://compvis.github.io/schroedingers_cat/`

Keywords: Motion Understanding & Planning · Generative Modeling

1 Introduction

Our world is constantly unfolding and every moment carries multiple possible futures: the future is fundamentally under-determined given the present due to stochasticity, free will, and other hidden factors. Acting in such environments requires anticipating how a scene might change. In particular, we need to predict scene kinematics – how its constituents may move – estimating the probability of *what could possibly be* rather than merely perceiving and tracking *what has already happened*. Video models used as world simulators [2, 3, 16] generate the future as sequences of images, producing individual frames rather than explicitly modeling scene changes between frames. Much of their capacity is spent modeling appearance details that do not affect scene kinematics, inducing unnecessary

[*] Equal contribution.

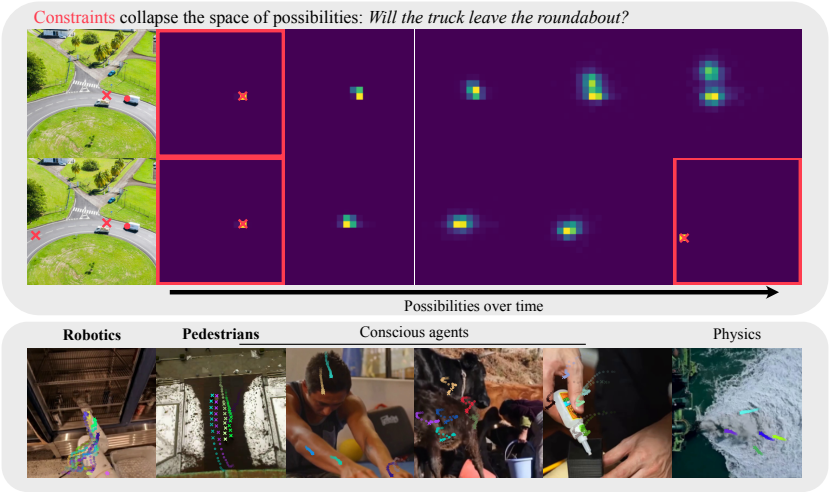


Fig. 1: Schrödinger's Cat: Given the initial scene context as an image and a set of *spatio-temporally sparse constraints or goals* (x), GARFIELD encodes the space of possible motions. By sampling from this space, our approach is able to plan realistic motion for conscious agents and inanimate objects, which we **evaluate** in applications such as robotics and trajectory prediction.

latency and limiting exploration of possibilities. Modeling motion directly [7,8,28] improves efficiency and focuses capacity on kinematics, but existing approaches predict dense motion fields, wasting computational effort on static background instead of targeting relevant scene elements. Moreover, generative motion models sample individual realizations rather than predicting the probability of various potential outcomes, their distribution, and uncertainty.

We take inspiration from Schrödinger's famous thought experiment: under limited observability, the state of a system is represented as a distribution of possible states, which collapses as new observations become available. Analogously, the key idea of GARFIELD is to represent the future as a *distribution over trajectories of scene constituents* that can be *progressively refined* by incorporating additional observations or constraints. Our encoder takes an image of a scene together with possible intermediate or goal positions to produce a latent representation that encodes the distribution over possible future movements of scene elements. Adding sparse constraints consistently sharpens this distribution, providing users with effective interactive control.

This kinematics distribution must be localized to specific scene elements and timesteps, so that uncertainty can be attributed correctly. We therefore learn a probabilistic embedding of scene kinematics using a joint motion encoder of the entire scene that produces *spatio-temporally localized latent variables*. Training with a generative decoder ensures that the representation captures the space of possible motions rather than a single trajectory. The structured latent space, which allows inspecting possible motion of specific elements at specific times is

enforced by training the encoder with a point-wise decoder model. We propose efficient decoders which – based on the same latent representation – enable both (i) joint trajectory sampling that captures interdependencies, as well as a (ii) *deterministic density decoding* that produces probability heatmaps in a single ViT forward pass. Compared to Monte-Carlo sampling, the density decoder provides estimates of densities and uncertainty orders of magnitude faster, which is crucial for real-time planning and fast exploration of uncertainty.

We consider the task of *motion planning*: generating trajectories that satisfy sparse goals while respecting scene constraints. Prior work typically relies on conditioning signals such as goal images [7], temporally dense trajectories [37], or text prompts [15, 37, 49], which are either ambiguous or difficult to obtain. In contrast, we enable *spatio-temporally sparse conditioning*, specifying constraints only for selected objects and timesteps while leaving the remaining motion for the model to infer. Coupled with fast density estimation, this enables efficient exploration of possible futures and uncertainty-aware motion planning with minimal user input.

Contributions: We learn a (i) *structured probabilistic latent representation of motion* (Sec. 3.1). The same latent representation enables (ii) jointly sampling *mutually consistent trajectories* that respect provided constraints (Sec. 3.4) and (iii) direct access to the underlying motion distribution via an efficient *density decoder* (Sec. 3.2, Fig. 4). Our model further supports (iv) *flexible spatio-temporally sparse conditioning*, allowing users to shape the distribution of possible futures through sparse constraints (Fig. 5). As a result, Goal Aware Representations of Future kInEmatic Latent Distributions (GARFIELD) enable uncertainty-aware motion planning and interactive exploration of future possibilities (Sec. 3.3, Fig. 3c). In experiments, our approach samples full trajectories 97× faster than video world models (Tab. 1) and estimates motion densities *two orders of magnitude faster* than Monte-Carlo sampling from motion prediction methods (Fig. 6b).

2 Related Work

Recently, video world models [2, 3, 10, 15, 16, 43, 49] based on diffusion transformers [30] have become the de facto standard approach to predict future scene evolution. These models are typically conditioned on a start image, text, and sometimes action classes, producing videos that depict future events in full pixel detail. They achieve impressive results in robotics [47, 52], autonomous driving [4, 44, 50], and physical simulation [3, 10, 25], but represent the future as a sequence of images instead of focusing on dynamics.

Low-level temporal dynamics can be represented as dense optical flow [18,41] or sparse point tracks [20, 51], which capture motion without modeling appearance. Recent methods such as Track2Act [7] and *What Happens Next?* [8] apply diffusion transformers to generate point tracks for motion-centric tasks (*e.g.* robotic control). Similarly, Motion-I2V [37] generates optical flow from a start image, text prompt, and spatially sparse conditioning tracks. The sampled flow then conditions a video generator to improve quality and control. However, these

approaches rely on ambiguous text [8, 38] or unavailable goal-state images [7]. In contrast, GARFIELD conditions on spatio-temporally sparse goal states. Moreover, these methods predict motion on dense grids, wasting compute on mostly static background instead of focusing on relevant scene elements.

In comparatively narrow application domains such as autonomous driving and pedestrian prediction, methods like Social-LSTM [1], Social-GAN [14], and Trajectron++ [35] already focus on predicting motion only for relevant scene elements. This highlights the practical value of element-centric modeling. However, these models are tailored to specific domains and do not generalize to open-set motion prediction conditioned on arbitrary scene appearance.

Tracks and optical flow capture low-level motion but remain agnostic to semantics and constraints. Recent work therefore explores richer motion representations, including training with VAEs [11], normalizing flows [17], or self-supervised representation learning [32, 40, 45]. Motion Diffusion Autoencoders [33] further learn semantic motion representations in closed-domain settings. However, most approaches focus on reconstruction or semantics and do not model the stochasticity of future motion. Early works represented this stochasticity using probabilistic motion primitives [29], Bézier curve distributions [19], or Gaussian processes [21]. Recent diffusion models [6–8, 37, 38] enable sampling possible futures but do not provide direct access to the underlying distribution, requiring expensive Monte-Carlo sampling to estimate distributions. Motion Modes [28] improves exploration of this distribution through guided sampling but remains limited by high inference cost. In contrast, GARFIELD provides access to a non-parametric distribution over future positions in constant ($\mathcal{O}(1)$) time.

The Flow Poke Transformer (FPT) [5] made a significant step towards open-set, distribution-centric motion modeling by predicting the parameters of a Gaussian Mixture Model (GMM) over future positions given a set of spatially sparse known movements. However, the model is limited by the functional form of their output distribution, constraining the output space to those possibilities that can be represented by a GMM. More importantly, their approach only predicts distributions at one specific future timestep and does not account for scene evolution over multiple steps or allow specifying intermediate goals.

We propose GARFIELD to train an open-set motion foundation model that encodes possible future positions of scene elements given appearance cues and optionally sparsely known future goals. We further introduce decoder models that provide access to the underlying non-parametric distribution without repeated sampling and probabilistic sampling of future trajectories. Our general model can be applied to application domains either in a zero-shot setting or with minimal fine-tuning, highlighting the advantages of scalable open-set pretraining.

3 Goal Aware Representations of Future kInEmatic Latent Distributions (GARFIELD)

Predicting scene kinematics given a finite set of constraints or goals requires modeling the probability distribution over the many plausible futures. We represent

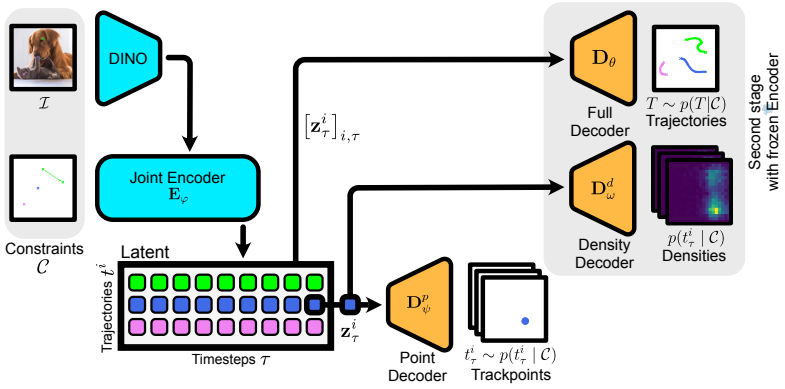


Fig. 2: GARFIELD Architecture Overview. Given a set of contstraints $\mathcal{C}$, namely an image $\mathcal{I}$ and kinematics $\widehat{T}$, our encoder produces conditional latents $\{\mathbf{z}^i_\tau\}$ representing distributions over possible future kinematics. These can be decoded into point estimates t^i_τ with $\mathbf{D}^{\mathrm{p}}_\psi$, densities $p(t^i_\tau)$ with $\mathbf{D}^{\mathrm{d}}_\omega$, and sets of kinematics T with $\mathbf{D}_\theta$.

one possible realization by the motion of scene elements i over time, described by trajectories $t^i = (t^i_1, t^i_2, \dots)$, where $t^i_\tau \in \mathbb{R}^2$ denotes the position of element i in the 2D image plane at timestep τ. The distribution $p(T)$ of possible futures, with $T = \{t^i_\tau\}_{i,\tau}$, can be narrowed and steered by conditioning on constraints or goals. Analogously to the *"Schrödinger's Cat"* thought experiment, additional constraints or observations collapse the space of possible futures. Providing scene context by means of a start frame I and a sparse set of positions $\widehat{T} \subset T$ through which certain trajectories must pass reduces uncertainty by partially controlling scene movement through intermediate goals. Goal Aware Representations of Future kInEmatic Latent Distributions (GARFIELD) therefore aim to model the conditional distribution $p(\bar{T} \mid I, \widehat{T})$ over the unknown trajectory points $\bar{T} = T \setminus \widehat{T}$. For brevity, we summarize the conditioning as $\mathcal{C} = (I, \widehat{T})$.

3.1 Learning the Distribution of Scene Kinematics

We learn a latent representation of the distribution of plausible scene kinematics given scene context and constraints $\mathcal{C}$. Our goal with GARFIELD is twofold: (i) to jointly sample trajectories of all scene elements, $\bar{T} \sim p(\bar{T} \mid \mathcal{C})$, and (ii) to estimate the probability density $p(t^i_\tau \mid \mathcal{C})$ for the position of a scene element i at time τ. Modeling such point-wise probability densities enables efficient parallel exploration of motion distributions across the spatio-temporal volume and reveals how changes to $\mathcal{C}$ affect individual scene elements at specific points in time.

By annotating video data with an optical tracker [20, 51], we can harness a rich and scalable source of data for learning and understanding scene kinematics. However, the videos and their track annotations are only one possible realization $T \sim p(T)$ of kinematics for the respective scene I. This leaves the distribution of *possible* kinematics out of reach and thus there is no direct groundtruth for

$p(T)$. Generative modeling is an established way to capture a data distribution without direct access to its underlying density function by learning from samples of that data distribution.

Current methods [6–8, 37, 38] can sample T with generative models which underlines the viability of learning scene kinematics from video data. However, the ability to sample T does not equate to easily accessible densities $p(t_\tau^i|\mathcal{C})$. In theory, the density function $p(t_\tau^i|\mathcal{C})$ can be approximated with Monte-Carlo estimation by drawing multiple realizations of T. In practice, reliable estimates require many samples and the expensive sampling procedure of generative models renders this avenue infeasible.

Because both $\bar{T} \sim p(\bar{T}|\mathcal{C})$ and $p(t_\tau^i|\mathcal{C})$ are conditioned on the same $\mathcal{C}$, we propose to learn a shared latent representation which represents $\mathcal{C}$ and informs both sampling and density estimation. Density estimation $p(t_\tau^i|\mathcal{C})$ needs to be spatio-temporally localized. Thus, we factorize the latent

$$\mathbf{z} = \left[\mathbf{z}_\tau^i\right]_{i,\tau} \tag{1}$$

into separate $\mathbf{z}_\tau^i$ components which each corresponds to a specific point-wise motion distribution $p(t_\tau^i|\mathcal{C})$. This representation is learned end-to-end with a joint encoder $\mathbf{E}_\varphi$, which produces $\mathbf{z}$, and a point-wise decoder $\mathbf{D}_\psi^{\mathrm{p}}$ which samples track positions from the point-wise distribution $t_\tau^i \sim \mathbf{D}_\psi^{\mathrm{p}}(\mathbf{z}_\tau^i) \approx p(t_\tau^i|\mathcal{C})$. We train $\mathbf{E}_\varphi$ and $\mathbf{D}_\psi^{\mathrm{p}}$ jointly to enforce this point-wise structural property in $\mathbf{z}$.

3.2 Sampling-free Density Estimation

Latent components $\mathbf{z}_\tau^i$ represent the distribution of possible positions of scene elements i at time τ. Using the point-wise decoder $\mathbf{D}_\psi^{\mathrm{p}}$ with N_{MC} different initial noise seeds allows sampling N_{MC} track points from the point-wise distribution and thus enable a Monte-Carlo (MC) estimation of the outcome probabilities. However, this presents a trade-off: accurate estimation requires many samples, leading to substantial computational overhead.

To circumvent this trade-off, we propose a deterministic density estimator $\mathbf{D}_\omega^{\mathrm{d}}$ for $\mathbf{z}_\tau^i$ with GARFIELD. We define the marginal probability density on a regular grid $\mathcal{G} = \{1, \ldots, g\} \times \{1, \ldots, g\}$ with $g \in \mathbb{N}$ as

$$p_\omega^{(u,v)}(\mathbf{z}_\tau^i) = \mathrm{P}\left(t_\tau^i \in \mathcal{B}_{u,v} \mid \mathbf{z}_\tau^i\right), \tag{2}$$

which gives the probability that t_τ^i lies within the spatial region $\mathcal{B}_{u,v}$ corresponding to grid location $(u, v) \in \mathcal{G}$.

The density estimator then directly predicts the point-wise non-parametric distribution over the possible future motion for scene element i at timestep τ given the known context $\mathcal{C}$ through $\mathbf{z}_\tau^i$:

$$\mathbf{D}_\omega^{\mathrm{d}}(\mathbf{z}_\tau^i) = \{p_\omega^{(u,v)}(\mathbf{z}_\tau^i)\}_{(u,v)\in\mathcal{G}} = p(t_\tau^i|\mathbf{z}_\tau^i) \approx p(t_\tau^i|\mathcal{C}) \tag{3}$$

We train the density estimator $\mathbf{D}_\omega^{\mathrm{d}}$ with a grid cross-entropy loss

$$\mathcal{L}_{\mathrm{grid}} = -\sum_{(u,v)\in\mathcal{G}} \mathbf{1}\left[t_\tau^i \in \mathcal{B}_{u,v}\right] \log p_\omega^{(u,v)}(t_\tau^i|\mathbf{z}_\tau^i), \tag{4}$$

where $\mathbf{1}\left[t_\tau^i \in \mathcal{B}_{u,v}\right]$ is 1 if t_τ^i falls into bin $\mathcal{B}_{u,v}$ and 0 otherwise, resulting in a one-hot target distribution.

Note that the density estimator also receives only the point-wise latent $\mathbf{z}_\tau^i$ as input. Thus, $\mathbf{D}_\omega^{\mathrm{d}}$ is a decoder of the distribution encoded in the latent representation, learned through generative pre-training. Instead of estimating densities through Monte Carlo sampling, we decode them directly with a single ViT forward pass in less than 0.8 ms (s. Table 6) from the latent representation $\mathbf{z}_\tau^i$, enabling fast uncertainty estimation, uncertainty-aware planning, and interactive exploration of these motion distributions.

3.3 Interactive Exploration of Possibilities

At inference, the density decoder $\mathbf{D}_\omega^{\mathrm{d}}$ provides access to the distribution $p(t_\tau^i \mid \mathcal{C})$ in a single forward pass. This enables uncertainty quantification with GARFIELD by measuring the Shannon entropy of the predicted distribution

$$H(t_\tau^i) = -\sum_{(u,v)\in\mathcal{G}} p_\omega^{(u,v)}(t_\tau^i|\mathbf{z}_\tau^i)\log p_\omega^{(u,v)}(t_\tau^i|\mathbf{z}_\tau^i). \tag{5}$$

Thus, we can efficiently evaluate where future motion remains under-determined: high entropy indicates multiple plausible futures, while low entropy corresponds to collapsed regions of the possibility space. This allows users to identify and explore regions where the range of possibilities remains diverse, instead of wasting effort on collapsed low-variance regions of the kinematics distribution.

The density estimator enables interactive visualization and exploration of possible futures in real-time with no need for sampling multiple realizations from a generative model. Users can therefore explore possible futures and iteratively modify the constraints $\mathcal{C}$ in a feedback loop, enabling interactive planning by guiding and narrowing the set of feasible trajectories. Once the user is satisfied with the result, the latents $\mathbf{z}_\tau^i$ can be decoded into a full, coherent trajectory realization with a generative model once.

3.4 Sampling Coherent Realizations

For many downstream applications in motion planning, concrete trajectory realizations are required. Naively, GARFIELD could infer trajectories from the point-wise decoder by factorizing the distribution of scene kinematics as

$$p(T \mid \mathcal{C}) \approx \prod_i \prod_\tau p(t_\tau^i \mid \mathcal{C}). \tag{6}$$

However, if the space of possibilities is not fully collapsed and multimodal futures are possible, the point-wise decoder cannot resolve mutual dependencies between points within and between trajectories. Thus, a sample for timestep τ might be drawn from one mode while the next timestep's sample $\tau + 1$ is drawn from another, resulting in temporally incoherent trajectories. The alternative

Table 1: Comparison on Motion Planning. GARFIELD achieves strong motion planning performance in open-set domains with reduced latency by modeling motion only for a subset of all scene elements and using spatio-temporally sparse goal conditioning.

	Method	Text Free	Num. Goals	EPE ↓	PCK ↑ @10%	PCK ↑ @1%	FDE ↓	Latency↓ [s]
video	CogVideoX [49]	✗	n/a	0.019	0.960	0.648	0.023	178
	LTX [15]	✗	n/a	0.025	0.936	0.601	0.046	39
explicit motion	Motion-I2V [37]	✗	4	0.033	0.927	0.412	0.044	13.2
			16	0.018	0.976	0.587	0.025	
	Track2Act [7]	✓	n/a	0.033	0.967	0.100	0.038	4.10
	FPT [5]	✓	4	0.112	0.669	0.109	0.169	0.60
			16	0.103	0.671	0.150	0.113	
	GARFIELD (Ours)	✓	4	0.014	0.969	0.795	0.018	**0.40**
			16	**0.012**	**0.977**	**0.821**	**0.015**	

of resolving these dependencies via auto-regressive updates of $\mathcal{C}$ would entail computational overhead and early greedy commitment to sampled points resulting in performance deterioration (Tab. S3 in supplementary materials).

Instead, we train a full decoder $\mathbf{D}_\theta$ as a joint generative model conditioned on the set of all latents, which learns to sample

$$T \sim \mathbf{D}_\theta\left(\left\{\mathbf{z}_\tau^i\right\}_{i,\tau}\right) = \mathbf{D}_\theta(\mathbf{z}) = \mathbf{D}_\theta(\mathbf{E}_\varphi(\mathcal{C})) \approx p(T \mid \mathcal{C}) \tag{7}$$

using latents produced by the frozen encoder (Section 3.1). By sampling trajectories jointly, GARFIELD's full decoder correctly handles inter-dependencies.

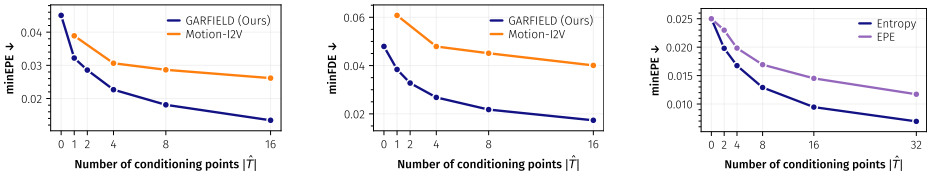
By learning structured latents (Eq. (1)) with a point-wise generative decoder described in Sec. 3.1, GARFIELD enables access to density estimates through Eq. (3) and can sample coherent trajectories from $p(T \mid \mathcal{C})$ using Eq. (7).

4 Experiments

Our experiments evaluate GARFIELD's capabilities in goal-conditioned motion planning and its ability to predict distributions over future kinematics. We further show our method's usefulness as a foundation for downstream tasks.

4.1 General setup

Data For training and evaluation we annotate videos with off-the-shelf trackers [51] to obtain a pseudo-ground truth for supervision and metric calculation. We train on a dataset of 20M videos at 12 FPS that cover a wide range of subjects. Comparisons are performed on public OpenVid-1M [26] while ablation studies are performed on a held-out validation set of the training data. We further evaluate on domain-specific data [9,22,31,34,42] with either human annotated trajectories or CoTracker3 [20] annotations following standard practice.



(a) Average over time. **(b)** Final positional error. **(c)** Conditioning policy.

Fig. 3: Impact of Goal Conditioning. GARFIELD is able to infer accurate motion in terms of **(a)** minEPE and **(b)** minFDE with limited conditioning. **(c)** Selecting conditioning based on entropy yields faster collapse than an oracle.

Implementation details The encoder, full decoder, and density estimator of GARFIELD are transformer networks [13, 30]. We extract image features using DINOv2R-B [12, 27] and unfreeze DINO during encoder training. We use 3D Axial RoPE [39], yet mask the starting position for the full decoder to avoid information leakage. We use g =20 for the density grid and append an out of bounds token. To ensure smoothness of $\mathbf{z}_\tau^i$ we apply a tanh function and add Gaussian noise with σ =1e-5 to $\mathbf{z}_\tau^i$ during training [36]. The encoder adopts the static camera conditioning from FPT [5]. The pointwise decoder follows MAR [23] and is a 34M parameter MLP. More details are in Sec. A in the appendix.

Training details The GARFIELD encoder $\mathbf{E}_\varphi$ is trained for 460k steps with the point-wise generative decoder and a linearly decaying goal sparsity $\widehat{T}/T$ from 0.5 to 0.01 in the first 50k steps. The density estimator and full decoder are trained for ca. 100k steps while keeping the $\mathbf{E}_\varphi$ frozen. All training runs use a global batch size of 256, AdamW with a learning rate of 1e-4, and bfloat16 precision. We generally model $i \in \{1 \dots 64\}$ tracks across $\tau \in \{1 \dots 32\}$ timesteps.

Metrics For open-set evaluation, we measure motion accuracy using positional errors normalized to $[0, 1]$, where 1 corresponds to the image size. Our main metric is the *endpoint error* $\text{EPE} = \|\bar{t}_\tau^i - t_\tau^i\|_2$, averaged over all trajectories i and timesteps τ. We also report the *percentage of correct keypoints* $\text{PCK@}\alpha = \mathbb{E}[\text{EPE} < \alpha]$ with $\alpha \in \{10\%, 1\%\}$ relative to image size and the Final Distance Error (FDE), i.e., EPE at the final timestep. Calibration is evaluated using the Energy Score (ES), a proper scoring rule for probabilistic predictions. Domain-specific evaluations follow standard protocols from prior work. Appendix Sec. C provides more metric details.

4.2 Open-set Motion Modeling

Motion Planning We compare the GARFIELD full decoder to open-set baselines for motion planning on the public OpenVid-1M [26] dataset. Models predict future motion from a start frame and sparse future constraints (text, goal images, or positions depending on the method). Because multiple futures may satisfy the constraints, we generate five samples per model and report the best.

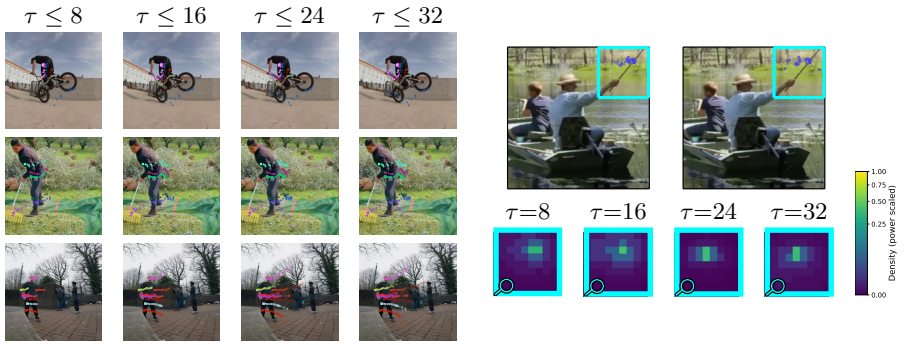


Fig. 4: Qualitative open-set motion and uncertainty. (**left**) With sparse conditioning ($\widehat{T}/T = 1\%$), GARFIELD samples realistic motion. Dots denote predictions and crosses conditioning points. (**right**) Complex motion gives non-collapsed distributions.

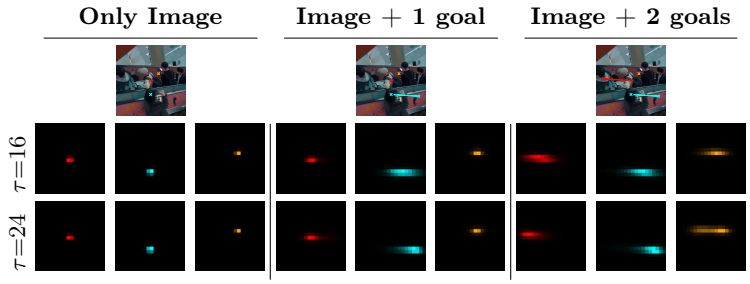
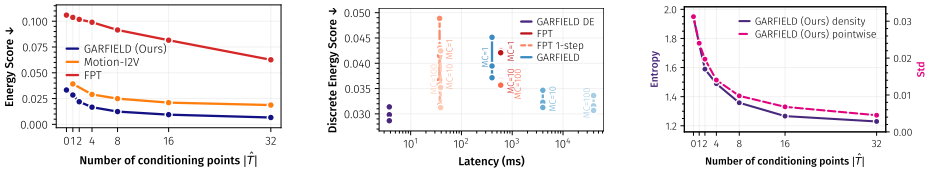


Fig. 5: Distributions with minimal conditioning. Given $|\widehat{T}|$=0, GARFIELD assumes a static scene. With $|\widehat{T}|$=1 the model captures motion but produces inaccurate object relations. With $|\widehat{T}|$=2, GARFIELD remains uncertain only about exact speed.

We evaluate both RGB video models and open-set trajectory models. For the former, we use TI2V models [15, 49] conditioned on the start frame and OpenVid prompt informing the model about temporal dynamics. Motion is extracted from generated videos using TapNext [51]. Motion-I2V [37] predicts optical flow from spatially sparse trajectories and interpolates temporally sparse signals. Track2Act [7] generates trajectories on a dense grid conditioned on start and goal RGB frames. FPT [5] predicts the distributions of changes between a start and final timestep. To obtain full trajectories, we rerun FPT for multiple timesteps using only the spatio-temporally sparse conditioning points GARFIELD observed.

Tab. 1 shows GARFIELD achieves the best EPE, PCK, and FDE. Therefore, we are competitive with much larger and orders-of-magnitude slower video generation models. We also outperform Track2Act, which relies on a dense final RGB frame that is typically unavailable in open-set motion planning. Rerunning FPT for each timestep yields subpar results, highlighting the importance of modeling temporal dependencies. Motion-I2V receives the same spatio-temporally sparse points as GARFIELD and additional text prompts, yet performs worse



(a) Energy Score. **(b)** Discretized energy scores. **(c)** Distribution tightening.

Fig. 6: Evaluation of Distributions. Our pointwise model achieves the best energy score. The density decoder attains superior discretized energy scores with orders-of-magnitude lower latency. Additional conditioning reduces entropy and variance.

and incurs higher latency, highlighting the advantage of modeling motion only for relevant scene elements. Overall, GARFIELD provides state-of-the-art motion planning in open-set domains, which we attribute to explicit motion modeling of relevant scene elements and spatio-temporally sparse conditioning that provides a more precise control signal than text or RGB-based conditioning used in prior video [15, 49] and motion generation [7, 37] methods.

Goal Conditioning Figure 3 shows the impact of adding more sub-goals to the conditioning set $\mathcal{C}$. GARFIELD is able to perform more accurate trajectory planning when using very sparse information, *e.g.* no or limited future positional information. GARFIELD, given only $|\widehat{T}| = 4$, outperforms Motion-I2V using $|\widehat{T}| = 16$ in EPE and performs well with only image conditioning.

Qualitative Evaluation We present qualitative samples from GARFIELD with only $\widehat{T}/T = 1\%$ known trajectory points on the left side of Figure 4. GARFIELD is able to predict complex trajectories for a range of objects and activities given minimal future knowledge. Fig. S8 shows qualitative comparisons to baselines in the Supplementary Materials. The right side of Figure 4 shows that $\mathbf{z}_\tau^i$ can encode multiple possible futures given the same constraints $\mathcal{C}$. We show the zoomed-in densities obtained from the GARFIELD density estimator $\mathbf{D}_\omega^{\mathrm{d}}$ as well as multiple samples from $\mathbf{D}_\theta$ for the same latent representation, highlighting GARFIELD's ability to capture the distribution of future outcomes.

Figure 5 demonstrates qualitatively that the estimated densities reflect GARFIELD's improved motion understanding with more conditioning. While the model predicts a mostly static scene without future positions, and fails to account for object inter-dependencies with only one conditional position, it is able to infer complex motion distributions with as little as two conditioning points.

Distributions We aim to verify our encoded latent captures meaningful uncertainty about scene dynamics. Therefore, we evaluate the predicted distribution of the GARFIELD point-wise generative decoder $\mathbf{D}_\psi^{\mathrm{p}}$ to rely only on encoded information. Figure 6 shows the energy score of our point-wise decoder. Compared to Motion-I2V [37] and FPT [5] our approach achieves the best energy score. Evaluating the density estimator $\mathbf{D}_\omega^{\mathrm{d}}$, we further compute a discretized energy

Table 2: Robotic Planning. Given final goals GARFIELD can infer robot arm motion with similar quality to prior art without finetuning. We extend [7]'s table.

Method	BridgeData [42] ↑	RT1 [9]↑
Flow [48]	0.32	0.38
Track2Act [7]	**0.77**	0.75
GARFIELD (Ours)$_{\text{0-shot}}$	0.76	**0.77**

score and find that the density estimator achieves superior discrete energy scores while speeding up computation speed by two orders of magnitude. Additionally, providing more conditioning tightens the distributions. We provide in-depth evaluation of calibration in Sec. E in the supplementary materials.

Automated interactive conditioning We consider how we should perform interventions to collapse the distribution to the ground truth with minimal outside information (*e.g.* by prompting the user to provide further input). Using the GARFIELD density estimator we can quantify the model's uncertainty in each track position using the entropy (Eq. (5)). We propose using entropy as an indicator for where to provide conditioning by finding the most uncertain trackpoint given the current constraints. Fig. 3c compares our proposed entropy-based conditioning to an oracle-based variant that has access to the current EPE for each trackpoint. For automated evaluation, we insert ground-truth conditioning at either the most uncertain or highest-error trackpoint. We find that interactive entropy-based conditioning outperforms the oracle-based variant and achieves comparable performance with about half the conditioning density. Note that no oracle is available in downstream application while entropy-based conditioning selection does not rely on unavailable ground truth information. We evaluate the robustness of entropy-based conditioning in Sec. F in the Supplementary Materials.

4.3 Application Domains

Robotics Planning is commonly used to evaluate motion models [7, 28, 38, 46]. We compare GARFIELD following the protocol from Track2Act [7] and sample trajectories for a robotic arm given start and goal states. While Track2Act uses dense RGB for start and end states, we use positional information. We follow Track2Act and report *area under the curve* $\Delta = \frac{1}{A}\sum_{a=1}^{A} PCK@\frac{a}{S}$, where $A = 10$ is the highest PCK threshold in pixel units and $S = 256$ is the pixel resolution (more details: appendix Sec. C). Tab. 2 highlights GARFIELD is competitive with Track2Act without training on robotics data. In comparison, Track2Act trains on splits from RT1 [9] and BridgeData [42]. This highlights fundamental understanding of motion and kinematics constraints learned by our model.

Pedestrian Trajectory Prediction is important for autonomous driving and a long-standing benchmark. We compare zero-shot and finetuned performance of

Table 3: Pedestrian trajectory prediction. Errors are reported in pixels w.r.t. the original video resolution for Stanford Drone and meters for ETH/UCY. Baseline numbers are taken from IDM [24]. GARFIELD performs competitively.

Method	**ETH** [31]		**HOTEL** [31]		**SDD** [34]	
	ADE ↓	**FDE** ↓	**ADE** ↓	**FDE** ↓	**ADE** ↓	**FDE** ↓
Social-LSTM [1]	1.09	2.35	0.79	1.76	57.00	31.20
Social-GAN [14]	0.81	1.52	0.72	1.61	27.23	41.44
Trajectron++ [35]	0.39	0.83	0.12	0.21	8.98	19.02
IDM [24]	0.41	0.62	0.15	0.25	7.46	13.83
GARFIELD (Ours)$_{\text{0-shot}}$	1.09	1.79	0.41	0.71	36.77	49.86
GARFIELD (Ours)$_{\text{FT}}$	0.84	1.44	0.29	0.56	19.92	34.98

GARFIELD. As we require RGB input, we use only a subset of ETH/UCY. Further details are provided in Appendix Sec. E. Tab. 3 shows our zero-shot performance is competitive with established methods such as Social-LSTM [1] even though we *do not train for prediction*. With minimal finetuning (<15k iterations) GARFIELD achieves performance comparable to Social-GAN [14]. While we are outperformed by more recent domain-specific methods, our results underscore the utility of our representations.

4.4 Ablations

Entangled and Global Representations Training the point-wise decoder with a merged global token or without enforcing structure yields substantially degraded performance and does not allow decoding into accurate points. Intuitively, factorized $\mathbf{z}_\tau^i$ encode $p(t_\tau^i \mid \mathcal{C})$ and can be seamlessly decoded by the point-wise decoder and density estimator. In comparison, for non-localized embeddings the decoder has to identify relevant information by itself, resulting in a substantially harder task for $\mathbf{D}_\psi^{\text{p}}$ and $\mathbf{D}_\omega^{\text{d}}$. In comparison, our disentangled $\mathbf{z}_\tau^i$ enable high-quality track predictions and estimation of point-wise densities.

Latent Dimensionality Tab. 4 further shows the performance on in-distribution samples ($\widehat{T}/T = 1\%$ conditioning) is almost independent of latent dimensionality.

Table 4: Embeddings. Using non-disentangled representations yields performance degradation. Smaller embeddings achieve superior performance in OOD settings.

Latent structure	**EPE** ↓	**FDE** ↓
Global	0.479	0.482
Entangled	0.509	0.510
Ours	0.011	0.015

Latent size	**EPE** ↓ $\frac{\widehat{T}}{T} = 1\%$	**EPE (OOD)** ↓ $\frac{\widehat{T}}{T} = 90\%$
16	0.012	0.009
64	0.012	0.008
256	0.012	0.019

Table 5: Point-wise vs. Full Decoder. Given limited constraints the full decoder outperforms the point-wise model.

Decoder	$\lvert\widehat{T}\rvert$	EPE ↓	FDE ↓
Point-wise $\mathbf{D}_\psi^{\mathrm{p}}$	2	0.031	0.037
	4	0.025	0.031
	16	0.013	0.017
Full sample $\mathbf{D}_\theta$	2	0.028	0.032
	4	0.022	0.026
	16	0.013	0.017

Table 6: Latency. The density decoder predicts heatmaps faster than either generative decoder produces a sample.

Component	Latency [ms]
Encoder $\mathbf{E}_\varphi$	3.0
Density $\mathbf{D}_\omega^{\mathrm{d}}$	0.8
Point-wise $\mathbf{D}_\psi^{\mathrm{p}}$	37.0
Full $\mathbf{D}_\theta$	330.4

However, in OOD settings, like when substantially more conditioning ($\widehat{T}/T = 90\%$) is given, smaller latents outperform larger variants, potentially due to over-adaptation. Unless noted otherwise, we use a latent dimensionality of 64 as a compromise with strong performance on both in-domain and OOD tasks.

Full vs. Point-wise Decoder Our point-wise decoder $\mathbf{D}_\psi^{\mathrm{p}}$ is able to infer the positions of scene elements at a specific timestep given a latent $\mathbf{z}_\tau^i$. Full trajectories can be reconstructed from independent point-wise estimates. However, in case of multi-modal uncertainty, independent point-wise estimation breaks down as samples per timestep can be assigned to different modes. The full decoder $\mathbf{D}_\theta$ resolves these inter-dependencies. Tab. 5 verifies this assumption: in uncertain cases with limited $\mathcal{C}$ the full decoder produces higher quality estimates, while the point-wise model performs competitively when the distribution is collapsed.

Grid Resolution Intuitively, higher g balance fine-granular density estimates with compute cost. Fig. 7 shows this trade-off empirically. As g influences the value range of the discrete energy score, we keep $g = 20$ fixed for all other evaluations and additionally report results with normalized grids ($g = 40$) for the ablation.

Alignment of Direct and MC Densities The density decoder $\mathbf{D}_\omega^{\mathrm{d}}$, point-wise decoder $\mathbf{D}_\psi^{\mathrm{p}}$, and full decoder $\mathbf{D}_\theta$ decode the same latents. Thus, heatmaps from

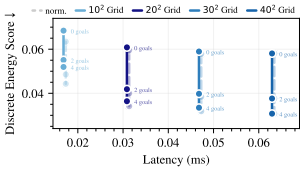


Fig. 7: Grid size ablation. Larger grid size g trades off improved density estimation for higher latency.

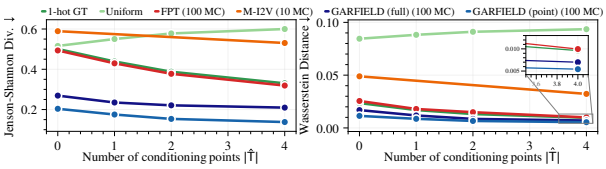


Fig. 8: Alignment of MC samples and densities. (**left**) MC samples from $\mathbf{D}_\theta$ and $\mathbf{D}_\psi^{\mathrm{p}}$ best match $\mathbf{D}_\omega^{\mathrm{d}}$ in Jensen-Shannon divergence. (**right**) Their MC heatmaps also achieve the lowest Wasserstein Distance to $\mathbf{D}_\omega^{\mathrm{d}}$.

$\mathbf{D}^{\mathrm{d}}_{\omega}$ should align with MC density estimates from $\mathbf{D}_{\theta}$ and $\mathbf{D}^{\mathrm{p}}_{\psi}$. Figure 8 confirms this: both decoders achieve lower Jensen–Shannon divergence and Wasserstein distance to $\mathbf{D}^{\mathrm{d}}_{\omega}$ than 1-hot ground-truth heatmaps or baseline samples.

Component Latency Tab. 6 shows runtimes for our individual components. The density estimator is able to decode a non-parametric distribution faster than either the point-wise or full decoder can produce a single example.

5 Conclusion

Reasoning about partially observable and uncertain environments requires modeling not only what will happen, but what *could happen*. We presented GARFIELD, a model that represents the space of possible scene kinematics through a structured probabilistic latent representation learned with a joint motion encoder and a point-wise diffusion decoder. Conditioned on sparse goal positions, this representation captures the distribution of plausible future motions and progressively sharpens as additional constraints become available. From the same latent representation, we obtain both coherent trajectory samples and probability densities over future positions through an efficient deterministic density decoder. These densities provide direct access to uncertainty about where scene elements may appear, thus letting users identify underdetermined future regions and refine them through sparse conditioning. This enables interactive exploration and motion planning with minimal user intervention. GARFIELD learns an open-set motion representation that performs competitively with large video models at substantially lower latency. The learned distributions collapse as additional information is given and transfer to downstream tasks such as robotic planning and trajectory prediction. More broadly, we view GARFIELD as a step toward models that reason about the space of possible futures rather than single outcomes, enabling interpretable, controllable, and efficient probabilistic modeling of scene evolution.

Acknowledgements

This project has been supported by the Horizon Europe project ELLIOT (GA No. 101214398), the project "GeniusRobot" (01IS24083) funded by the Federal Ministry of Research, Technology and Space (BMFTR), the BMWE ZIM-project (No. KK5785001LO4) "conIDitional LoRA", the German Federal Ministry for Economic Affairs and Energy within the project "NXT GEN AI METHODS - Generative Methoden für Perzeption, Prädiktion und Planung", and the bidt project KLIMA-MEMES. The authors gratefully acknowledge the Gauss Center for Supercomputing for providing compute through the NIC on JUWELS/JUPITER at JSC and the HPC resources supplied by the NHR@FAU Erlangen. We thank Stefan Baumann, Johannes Schusterbauer, Ming Gui, and Ulrich Prestel for helpful discussions and feedback, Nick Stracke, Frank Fundel, and Thomas Ressler-Antal for initial data work, and Owen Vincent for continuous technical support.

References

1. Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., Savarese, S.: Social lstm: Human trajectory prediction in crowded spaces. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 961–971 (2016). https://doi.org/10.1109/CVPR.2016.110, accessed 2025-11-18
2. Alonso, E., Jelley, A., Micheli, V., Kanervisto, A., Storkey, A., Pearce, T., Fleuret, F.: Diffusion for world modeling: Visual details matter in atari. In: Thirty-eighth Conference on Neural Information Processing Systems (2024), https://arxiv.org/abs/2405.12399, accessed 2026-06-26
3. Ball, P.J., Bauer, J., Belletti, F., Brownfield, B., Ephrat, A., Fruchter, S., Gupta, A., Holsheimer, K., Holynski, A., Hron, J., Kaplanis, C., Limont, M., McGill, M., Oliveira, Y., Parker-Holder, J., Perbet, F., Scully, G., Shar, J., Spencer, S., Tov, O., Villegas, R., Wang, E., Yung, J., Baetu, C., Berbel, J., Bridson, D., Bruce, J., Buttimore, G., Chakera, S., Chandra, B., Collins, P., Cullum, A., Damoc, B., Dasagi, V., Gazeau, M., Gbadamosi, C., Han, W., Hirst, E., Kachra, A., Kerley, L., Kjems, K., Knoepfel, E., Koriakin, V., Lo, J., Lu, C., Mehring, Z., Moufarek, A., Nandwani, H., Oliveira, V., Pardo, F., Park, J., Pierson, A., Poole, B., Ran, H., Salimans, T., Sanchez, M., Saprykin, I., Shen, A., Sidhwani, S., Smith, D., Stanton, J., Tomlinson, H., Vijaykumar, D., Wang, L., Wingfield, P., Wong, N., Xu, K., Yew, C., Young, N., Zubov, V., Eck, D., Erhan, D., Kavukcuoglu, K., Hassabis, D., Gharamani, Z., Hadsell, R., van den Oord, A., Mosseri, I., Bolton, A., Singh, S., Rocktäschel, T.: Genie 3: A new frontier for world models (2025), https://deepmind.google/blog/genie-3-a-new-frontier-for-world-models/, accessed 2026-06-26
4. Bartoccioni, F., Ramzi, E., Besnier, V., Venkataramanan, S., Vu, T.H., Xu, Y., Chambon, L., Gidaris, S., Odabas, S., Hurych, D., Marlet, R., Boulch, A., Chen, M., Zablocki, E., Bursuc, A., Valle, E., Cord, M.: VaViM and VaVAM: Autonomous Driving through Video Generative Modeling (Feb 2025), http://arxiv.org/abs/2502.15672, accessed 2025-07-10
5. Baumann, S.A., Stracke, N., Phan, T., Ommer, B.: What if: Understanding motion through sparse interactions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2025), https://openaccess.thecvf.com/content/ICCV2025/papers/Baumann_What_If_Understanding_Motion_Through_Sparse_Interactions_ICCV_2025_paper.pdf, accessed 2026-06-22
6. Baumann, S.A., Wiese, J., Martorella, T., Kalayeh, M.M., Ommer, B.: Envisioning the future, one step at a time. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6823–6836 (2026), https://openaccess.thecvf.com/content/CVPR2026/papers/Baumann_Envisioning_the_Future_One_Step_at_a_Time_CVPR_2026_paper.pdf, accessed 2026-06-22
7. 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). https://doi.org/10.1007/978-3-031-73116-7_18, accessed 2026-06-26
8. Boduljak, G., Karazija, L., Laina, I., Rupprecht, C., Vedaldi, A.: What happens next? anticipating future motion by generating point trajectories. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=t1vMYl1yhe, accessed 2026-06-26
9. Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., Ibarz, J., Ichter, B., Irpan, A., Jackson,

T., Jesmonth, S., Joshi, N.J., Julian, R., Kalashnikov, D., Kuang, Y., Leal, I., Lee, K.H., Levine, S., Lu, Y., Malla, U., Manjunath, D., Mordatch, I., Nachum, O., Parada, C., Peralta, J., Perez, E., Pertsch, K., Quiambao, J., Rao, K., Ryoo, M.S., Salazar, G., Sanketi, P.R., Sayed, K., Singh, J., Sontakke, S., Stone, A., Tan, C., Tran, H.T., Vanhoucke, V., Vega, S., Vuong, Q., Xia, F., Xiao, T., Xu, P., Xu, S., Yu, T., Zitkovich, B.: Rt-1: Robotics transformer for real-world control at scale. In: Bekris, K.E., Hauser, K., Herbert, S.L., Yu, J. (eds.) Robotics: Science and Systems (2023). https://doi.org/10.15607/RSS.2023.XIX.025, accessed 2026-06-26

10. Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., Ng, C., Wang, R., Ramesh, A.: Video generation models as world simulators (2024), https://openai.com/index/video-generation-models-as-world-simulators/, accessed 2026-06-22
11. Chen, G., Li, J., Zhou, N., Ren, L., Lu, J.: Personalized trajectory prediction via distribution discrimination. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15560–15569 (2021). https://doi.org/10.1109/ICCV48922.2021.01529, accessed 2025-07-21
12. Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers (2023)
13. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2020), https://openreview.net/forum?id=YicbFdNTTy, accessed 2024-12-13
14. 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), https://openaccess.thecvf.com/content_cvpr_2018/papers/Gupta_Social_GAN_Socially_CVPR_2018_paper.pdf, accessed 2026-06-22
15. HaCohen, Y., Chiprut, N., Brazowski, B., Shalem, D., Moshe, D., Richardson, E., Levin, E., Shiran, G., Zabari, N., Gordon, O., Panet, P., Weissbuch, S., Kulikov, V., Bitterman, Y., Melumian, Z., Bibi, O.: Ltx-video: Realtime video latent diffusion (2024), https://arxiv.org/abs/2501.00103, accessed 2026-06-26
16. Hafner, D., Yan, W., Lillicrap, T.: Training agents inside of scalable world models (2025), https://arxiv.org/abs/2509.24527, accessed 2026-06-26
17. Henter, G.E., Alexanderson, S., Beskow, J.: Moglow: Probabilistic and controllable motion synthesis using normalising flows. ACM Transactions on Graphics **39**(6), 1–14 (2020). https://doi.org/10.1145/3414685.3417836, accessed 2025-09-06
18. Horn, B.K.P., Schunck, B.G.: Determining optical flow. Artificial Intelligence **17**(1), 185–203 (1981). https://doi.org/10.1016/0004-3702(81)90024-2, accessed 2026-03-01
19. Hug, R., Hübner, W., Arens, M.: Introducing probabilistic bézier curves for n-step sequence prediction. Proceedings of the AAAI Conference on Artificial Intelligence **34**(06), 10162–10169 (2020). https://doi.org/10.1609/aaai.v34i06.6576, accessed 2025-09-06
20. Karaev, N., Makarov, I., Wang, J., Neverova, N., Vedaldi, A., Rupprecht, C.: CoTracker3: Simpler and Better Point Tracking by Pseudo-Labelling Real Videos (2024), http://arxiv.org/abs/2410.11831, accessed 2025-09-07
21. Kihwan Kim, Dongryeol Lee, Essa, I.: Gaussian process regression flow for analysis of motion trajectories. In: 2011 International Conference on Computer Vision. pp. 1164–1171. IEEE, Barcelona, Spain (2011). https://doi.org/10.1109/ICCV.2011.6126365, accessed 2025-09-06

22. Lerner, A., Chrysanthou, Y., Lischinski, D.: Crowds by example. In: Computer graphics forum. vol. 26, pp. 655–664. Wiley Online Library (2007). https://doi.org/10.1111/j.1467-8659.2007.01089.x, accessed 2026-06-26
23. Li, T., Tian, Y., Li, H., Deng, M., He, K.: Autoregressive image generation without vector quantization. In: Advances in Neural Information Processing Systems. vol. 37, pp. 56424–56445 (2024). https://doi.org/10.52202/079017-1797, accessed 2025-11-18
24. Liu, C., He, S., Liu, H., Chen, J.: Intention-aware denoising diffusion model for trajectory prediction. IEEE Transactions on Intelligent Transportation Systems (2025). https://doi.org/10.1109/TITS.2025.3553125, accessed 2026-06-26
25. Lu, H., Yang, G., Fei, N., Huo, Y., Lu, Z., Luo, P., Ding, M.: VDT: General-purpose video diffusion transformers via mask modeling. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=Un0rgm9f04, accessed 2026-06-26
26. Nan, K., Xie, R., Zhou, P., Fan, T., Yang, Z., Chen, Z., Li, X., Yang, J., Tai, Y.: Openvid-1m: A large-scale high-quality dataset for text-to-video generation. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=j7kdXSrISM, accessed 2026-06-26
27. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.Y., Xu, H., Sharma, V., Li, S.W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision (2023)
28. Pandey, K., Gadelha, M., Hold-Geoffroy, Y., Singh, K., Mitra, N.J., Guerrero, P.: Motion modes: What could happen next? (2024), https://arxiv.org/abs/2412.00148, accessed 2026-06-26
29. Paraschos, A., Daniel, C., Peters, J.R., Neumann, G.: Probabilistic movement primitives. In: Burges, C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K. (eds.) Advances in Neural Information Processing Systems. vol. 26. Curran Associates, Inc. (2013), https://proceedings.neurips.cc/paper_files/paper/2013/file/e53a0a2978c28872a4505bdb51db06dc-Paper.pdf, accessed 2026-06-26
30. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023), https://openaccess.thecvf.com/content/ICCV2023/papers/Peebles_Scalable_Diffusion_Models_with_Transformers_ICCV_2023_paper.pdf, accessed 2026-06-22
31. Pellegrini, S., Ess, A., Schindler, K., Van Gool, L.: You'll never walk alone: Modeling social behavior for multi-target tracking. In: 2009 IEEE 12th International Conference on Computer Vision. pp. 261–268. IEEE (2009). https://doi.org/10.1109/ICCV.2009.5459260, accessed 2026-06-26
32. Ressler-Antal, T., Fundel, F., Ben Alaya, M., Baumann, S.A., Krause, F., Gui, M., Ommer, B.: Dismo: Disentangled motion representations for open-world motion transfer. In: Belgrave, D., Zhang, C., Lin, H., Pascanu, R., Koniusz, P., Ghassemi, M., Chen, N. (eds.) Advances in Neural Information Processing Systems. vol. 38, pp. 158494–158534. Curran Associates, Inc. (2025), https://proceedings.neurips.cc/paper_files/paper/2025/file/e88870ec82f2469b0ddf32c817920c68-Paper-Conference.pdf, accessed 2026-06-26
33. Richardson, A., Putze, F.: Motion diffusion autoencoders: Enabling attribute manipulation in human motion demonstrated on karate techniques. In: Proceedings of the 27th International Conference on Multimodal Interaction. pp. 372–380 (2025). https://doi.org/10.1145/3716553.3750773, accessed 2026-06-26

34. Robicquet, A., Sadeghian, A., Alahi, A., Savarese, S.: Learning social etiquette: Human trajectory prediction in crowded scenes. In: Proceedings of the European Conference on Computer Vision (ECCV) (2016), https://svl.stanford.edu/assets/publications/pdfs/ECCV16social.pdf, accessed 2026-06-26
35. Salzmann, T., Ivanovic, B., Chakravarty, P., Pavone, M.: Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data (2021). https://doi.org/10.48550/arXiv.2001.03093, accessed 2025-11-18
36. Schusterbauer, J., Wiese, J., Stracke, N., Phan, T., Ommer, B.: Probabilistic precipitation nowcasting with rectified flow transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25742–25756 (2026), https://openaccess.thecvf.com/content/CVPR2026/papers/Schusterbauer_Probabilistic_Precipitation_Nowcasting_with_Rectified_Flow_Transformers_CVPR_2026_paper.pdf, accessed 2026-06-26
37. Shi, X., Huang, Z., Wang, F.Y., Bian, W., Li, D., Zhang, Y., Zhang, M., Cheung, K.C., See, S., Qin, H., et al.: Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling. SIGGRAPH 2024 (2024). https://doi.org/10.1145/3641519.3657497, accessed 2026-06-26
38. Stracke, N., Bauer, K., Baumann, S.A., Bautista, M.A., Susskind, J., Ommer, B.: Learning long-term motion embeddings for efficient kinematics generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 42581–42591 (2026), https://openaccess.thecvf.com/content/CVPR2026/papers/Stracke_Learning_Long-term_Motion_Embeddings_for_Efficient_Kinematics_Generation_CVPR_2026_paper.pdf, accessed 2026-06-22
39. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024). https://doi.org/10.1016/j.neucom.2023.127063, accessed 2026-06-26
40. Suris Coll-Vinent, D., Vondrick, C.: Representing spatial trajectories as distributions. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35, pp. 13731–13744. Curran Associates, Inc. (2022), https://proceedings.neurips.cc/paper_files/paper/2022/file/58f4a9ca9031ff197cb4a61b456574bf-Paper-Conference.pdf, accessed 2026-06-26
41. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: European conference on computer vision. pp. 402–419. Springer (2020), https://www.ecva.net/papers/eccv_2020/papers_ECCV/papers/123470392.pdf, accessed 2026-06-22
42. Walke, H.R., Black, K., Zhao, T.Z., Vuong, Q., Zheng, C., Hansen-Estruch, P., He, A.W., Myers, V., Kim, M.J., Du, M., Lee, A., Fang, K., Finn, C., Levine, S.: Bridgedata v2: A dataset for robot learning at scale. In: 7th Annual Conference on Robot Learning. PMLR (2023), https://openreview.net/forum?id=f55MlAT1Lu, accessed 2026-06-26
43. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., Wang, J., Zhang, J., Zhou, J., Wang, J., Chen, J., Zhu, K., Zhao, K., Yan, K., Huang, L., Feng, M., Zhang, N., Li, P., Wu, P., Chu, R., Feng, R., Zhang, S., Sun, S., Fang, T., Wang, T., Gui, T., Weng, T., Shen, T., Lin, W., Wang, W., Wang, W., Zhou, W., Wang, W., Shen, W., Yu, W., Shi, X., Huang, X., Xu, X., Kou, Y., Lv, Y., Li, Y., Liu, Y., Wang, Y., Zhang, Y., Huang, Y., Li, Y., Wu, Y., Liu, Y., Pan, Y., Zheng, Y., Hong, Y., Shi, Y., Feng, Y., Jiang, Z., Han,

Z., Wu, Z.F., Liu, Z.: Wan: Open and advanced large-scale video generative models (2025). https://doi.org/10.48550/arXiv.2503.20314, accessed 2025-11-18
44. Wang, X., Zhu, Z., Huang, G., Chen, X., Zhu, J., Lu, J.: Drivedreamer: Towards real-world-drive world models for autonomous driving. In: European conference on computer vision. pp. 55–72. Springer (2024), https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/06416.pdf, accessed 2026-06-22
45. Wang, Y., Zhang, P., Bai, L., Xue, J.: Fend: A future enhanced distribution-aware contrastive learning framework for long-tail trajectory prediction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1400–1409 (2023), https://openaccess.thecvf.com/content/CVPR2023/papers/Wang_FEND_A_Future_Enhanced_Distribution-Aware_Contrastive_Learning_Framework_for_Long-Tail_CVPR_2023_paper.pdf, accessed 2026-06-22
46. Wen, C., Lin, X., So, J., Chen, K., Dou, Q., Gao, Y., Abbeel, P.: Any-point trajectory modeling for policy learning (2024). https://doi.org/10.48550/arXiv.2401.00025, accessed 2025-07-02
47. Wen, Y., Lin, J., Zhu, Y., Han, J., Xu, H., Zhao, S., Liang, X.: Vidman: Exploiting implicit dynamics from video diffusion model for effective robot manipulation. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) Advances in Neural Information Processing Systems. vol. 37, pp. 41051–41075. Curran Associates, Inc. (2024). https://doi.org/10.52202/079017-1298, accessed 2026-06-26
48. Xu, H., Zhang, J., Cai, J., Rezatofighi, H., Yu, F., Tao, D., Geiger, A.: Unifying flow, stereo and depth estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(11), 13941–13958 (2023). https://doi.org/10.1109/TPAMI.2023.3298645, accessed 2026-06-26
49. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., Yin, D., Yuxuan.Zhang, Wang, W., Cheng, Y., Xu, B., Gu, X., Dong, Y., Tang, J.: Cogvideox: Text-to-video diffusion models with an expert transformer. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=LQzN6TRFg9, accessed 2026-06-26
50. Zhang, K., Tang, Z., Hu, X., Pan, X., Guo, X., Liu, Y., Huang, J., Yuan, L., Zhang, Q., Long, X.X., et al.: Epona: Autoregressive diffusion world model for autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 27220–27230 (2025), https://openaccess.thecvf.com/content/ICCV2025/papers/Zhang_Epona_Autoregressive_Diffusion_World_Model_for_Autonomous_Driving_ICCV_2025_paper.pdf, accessed 2026-06-22
51. 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), https://openaccess.thecvf.com/content/ICCV2025/papers/Zholus_TAPNext_Tracking_Any_Point_TAP_as_Next_Token_Prediction_ICCV_2025_paper.pdf, accessed 2026-06-22
52. Zhu, C., Yu, R., Feng, S., Burchfiel, B., Shah, P., Gupta, A.: Unified World Models: Coupling Video and Action Diffusion for Pretraining on Large Robotic Datasets. In: Proceedings of Robotics: Science and Systems. LosAngeles, CA, USA (2025). https://doi.org/10.15607/RSS.2025.XXI.015, accessed 2026-06-26