MULTI-SCALE FUTURE MODELING FOR LATENT WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Latent world models efficiently predict future world states in representation space, but most existing approaches model the future only at the timestep level, as a dense sequence of latent states. While this preserves fine-grained dynamics, it provides no explicit representation of future evolution over longer temporal intervals, potentially biasing forecasting toward local dynamics. We present MS-JEPA, a multi-scale latent forecasting framework that models future evolution at complementary temporal scales. The key idea is to represent the same future through both dense latent states and Future Abstractions: the former preserve fine-grained dynamics, while the latter summarize evolution over temporal intervals. A high-level predictor anticipates these abstractions from observed latents to guide dense forecasting. By explicitly modeling temporal structure across scales, our framework combines broader future context with fine-grained dynamics. Experiments across egocentric hand motion, whole-body motion, and physical-scene forecasting show consistent improvements over latent world modeling baselines, demonstrating the effectiveness and generality of our approach.

1 INTRODUCTION

World models (Ha & Schmidhuber, 2018; Hafner et al., 2019) capture how an environment evolves over time and enable prediction of future states. Recent latent world models (Zhou et al., 2025; Baldassarre et al., 2025) perform this prediction in a learned representation space, avoiding the need to explicitly reconstruct future observations at the pixel level. JEPA-style approaches (Assran et al., 2025; Mur-Labadia et al., 2026) have further demonstrated that such latent forecasting can effectively model motion and interaction dynamics from video. In most existing formulations (Assran et al., 2025; Zhang et al., 2026a; Xu et al., 2026), observed frames are encoded into latent representations, and a predictor directly forecasts a dense sequence of future latent states, describing the future at a detailed temporal granularity. These predicted representations can support a broad range of downstream tasks, including forecasting (Zhang et al., 2026a), planning (Zhang et al., 2026b; Maes et al., 2026), and control (Chen et al., 2026).

Despite its simplicity and effectiveness, dense latent forecasting represents the future only at the timestep level (Zhang et al., 2026a). While this preserves fine-grained dynamics, it does not explicitly represent future evolution over longer temporal intervals. As a result, forecasting is performed primarily at a single temporal scale, potentially biasing predictions toward local dynamics. This motivates reconsidering how the future is represented in latent space. Can the same future be represented at multiple temporal scales? If so, can such a representation improve dense forecasting?

Our hypothesis is that the same future should be represented at complementary temporal scales rather than solely as a sequence of individual future states. Dense timestep-level latents preserve fine-grained dynamics, while interval-level representations can summarize information shared across consecutive future states. We instantiate this idea using simple temporal averaging, obtaining interval-level summaries that we refer to as Future Abstractions. These abstractions do not replace dense prediction or serve as sparse future states; instead, they provide higher-level context for dense forecasting.

To test whether Future Abstractions provide useful information in addition to dense timestep-level states, we first conduct a controlled experiment. We construct Future Abstractions directly from ground-truth future latents and provide them to the dense predictor through cross-attention, while

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Controlled comparison of dense and oracle multi-scale forecasting. **Dense** uses only timestep-level future modeling, while **Multi-Scale (Oracle)** additionally conditions the dense predictor on ground-truth Future Abstractions.

Method	ADE ↓	FDE ↓	Acc ↑
Dense	0.049	0.067	0.656
Multi-Scale (Oracle)	**0.038**	**0.043**	**0.775**

keeping the remaining forecasting pipeline unchanged. As shown in Tab. 1, supplying ground-truth Future Abstractions leads to substantial improvements across all evaluation metrics, supporting our hypothesis that interval-level future representations provide complementary information for dense forecasting. For a more intuitive illustration, Fig. 1 shows one validation sample (wrist), where Multi-Scale (Oracle) more closely follows the ground-truth trajectory and yields consistently lower prediction error (more visualization examples are provided in Appendix A.3).

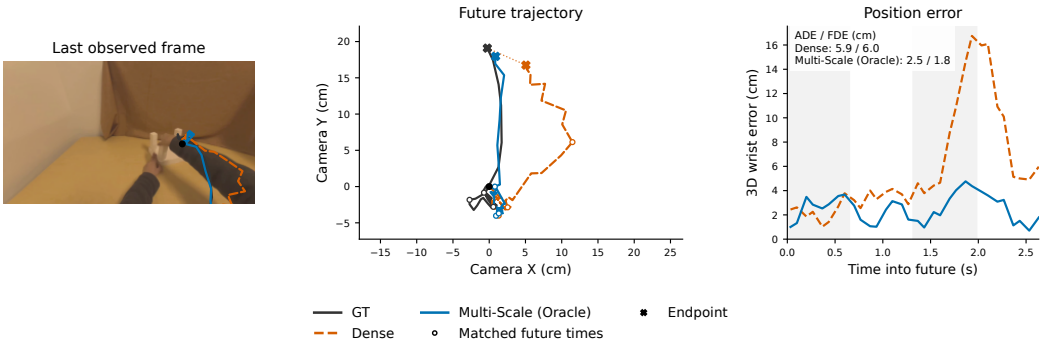

Figure 1: **Qualitative effect of oracle Future Abstractions**. Compared with dense forecasting, Multi-Scale (Oracle) follows the ground-truth future trajectory more closely and maintains lower prediction error over the forecast horizon.

However, ground-truth future representations are unavailable at inference time. We therefore introduce a high-level predictor that takes observed latent states as input. It predicts a compact set of Future Abstractions, each corresponding to a consecutive interval of the future horizon. These predicted abstractions are used to guide the dense latent predictor through cross-attention. The resulting framework, termed **MS-JEPA**, models the same future at two complementary temporal scales: the high-level pathway anticipates interval-level future evolution, while the low-level pathway resolves fine-grained timestep-level dynamics.

We evaluate the proposed framework across different prediction horizons and diverse forecasting settings, including egocentric hand motion, whole-body motion, and physical-scene forecasting. The consistent improvements over dense latent forecasting baselines support the effectiveness of explicitly modeling future evolution at multiple temporal scales. Our contributions are threefold:

- We formulate latent future forecasting from a multi-scale temporal perspective, modeling the same prediction horizon through complementary timestep-level and interval-level representations. This augments fine-grained timestep-level forecasting with an explicit interval-level view of future evolution.
- We instantiate this formulation with Future Abstractions, interval-level summaries constructed from consecutive future latents. A high-level predictor anticipates these abstractions from observed latents and uses them to guide dense future prediction, complementing timestep-level dynamics with interval-level future context.
- We demonstrate the effectiveness and generality of our approach across different prediction horizons and diverse forecasting tasks, including egocentric hand motion, whole-body motion, and physical scene prediction.

2 Related Works

Latent World Models. Latent world models learn predictive dynamics in compact representation spaces, enabling forecasting, planning, and control without directly modeling high-dimensional observations. A major line of work learns latent dynamics together with observation modeling objectives, including PlaNet (Hafner et al., 2019) and the Dreamer family (Hafner et al., 2020; 2021; 2024). More recent approaches perform future prediction directly in learned visual feature spaces (Assran et al., 2025; Zhou et al., 2025; Baldassarre et al., 2025; Terver et al., 2026). V-JEPA 2 performs future representation prediction for video understanding and planning, while DINO-WM and DINO-World model future dynamics in pretrained DINO (Oquab et al., 2024) feature spaces. Most closely related to our work, ThinkJEPA augments dense latent forecasting with VLM-derived semantic guidance from observed video (Zhang et al., 2026a). Our work instead studies how the same latent future can be represented at complementary temporal scales, jointly modeling timestep-level states and interval-level summaries of future evolution.

Future Prediction from Video. Future prediction from video has been studied with a variety of prediction targets. One line of work formulates forecasting as future video prediction, directly synthesizing upcoming visual observations (Finn et al., 2016; Babaeizadeh et al., 2018; Denton & Fergus, 2018; Gupta et al., 2023; Voleti et al., 2022). Rather than synthesizing pixels, many methods focus on structured future states, covering semantic events and actions (Mittal et al., 2024; Chu et al., 2026; Girdhar & Grauman, 2021) as well as embodied motion and interactions (Ashutosh et al., 2025; Liu et al., 2022; Bao et al., 2023; Escobar et al., 2025). While effective for their respective tasks, these approaches formulate forecasting in predefined output spaces such as pixels, action labels, trajectories, or poses. They therefore do not directly address how future information should be represented across temporal scales in latent space. Our work focuses on this complementary problem by modeling the latent future at both timestep and interval levels.

3 Method

In this section, we present our multi-scale latent forecasting framework, MS-JEPA (Fig. 2). We first formulate the latent future forecasting problem and introduce a multi-scale representation that combines Future Abstractions with dense latent states. We then introduce a high-level predictor to anticipate Future Abstractions from observed latents, and a low-level predictor that uses them to guide dense future prediction.

3.1 Problem Formulation

Given a video sequence consisting of T_p observed frames and T_f future frames, we denote the observed and future segments as $\mathbf{X}^{\text{past}} = \{x_t\}_{t=1}^{T_p}$ and $\mathbf{X}^{\text{future}} = \{x_t\}_{t=T_p+1}^{T_p+T_f}$, respectively. Following JEPA-style latent forecasting, a frozen visual encoder $\mathcal{E}$ maps the video into a learned representation space, producing the observed latent states and future latent targets:

$$\mathbf{F}^{\text{past}} = \mathcal{E}(\mathbf{X}^{\text{past}}), \quad \mathbf{F}^{\text{future}} = \mathcal{E}(\mathbf{X}^{\text{future}}), \tag{1}$$

where $\mathbf{F}^{\text{future}} = \{\mathbf{F}_t\}_{t=1}^{T_f}$ denotes the dense latent sequence over the future horizon.

A conventional dense predictor directly infers the complete future latent sequence from the observed states. We refer to this dense forecasting module as the *low-level predictor*, denoted by $\mathcal{G}_{\text{low}}$. The predicted future representations are subsequently consumed by a task-specific head $\mathcal{H}$:

$$\hat{\mathbf{F}}^{\text{future}} = \mathcal{G}_{\text{low}}(\mathbf{F}^{\text{past}}), \quad \hat{\mathbf{Y}} = \mathcal{H}(\hat{\mathbf{F}}^{\text{future}}), \tag{2}$$

where $\mathbf{Y}$ denotes the downstream prediction target.

Instead of modeling the future solely as a dense sequence at a single temporal scale, we augment this formulation with a complementary interval-level representation of future evolution. Specifically, we represent the same future horizon using both dense latent states and a small set of temporally aggregated *Future Abstractions*. We introduce this multi-scale future representation in the following subsection.

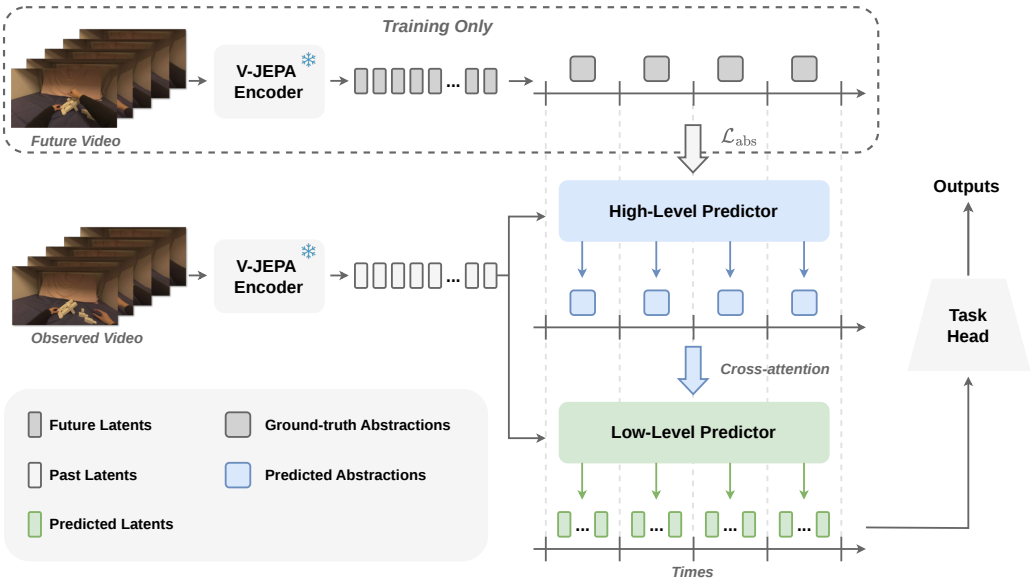

Figure 2: **Overview of the proposed multi-scale latent forecasting framework**. Observed video is encoded into past latents, from which the high-level predictor anticipates Future Abstractions. These abstractions guide the low-level predictor to forecast dense future latents, which are then used for downstream prediction. During training, future latents from the frozen encoder are used as dense forecasting targets and to construct ground-truth Future Abstractions for supervision.

3.2 Multi-Scale Future Representation

The conventional JEPA formulation represents the future as a dense sequence $\mathbf{F}^{\text{future}} = \{\mathbf{F}_t\}_{t=1}^{T_f}$, where each latent state describes the visual dynamics at a particular future timestep. While this representation preserves detailed temporal evolution, the same future can also be characterized at a broader temporal scale. We therefore introduce a complementary representation that summarizes consecutive future states over longer temporal intervals.

Specifically, we partition the future horizon of T_f timesteps into N non-overlapping temporal chunks. Assuming T_f is divisible by N, each chunk contains $L = T_f/N$ consecutive future states, and the i-th chunk is defined as

$$\mathcal{C}_i = \{(i-1)L + 1, \ldots, iL\}. \tag{3}$$

For each temporal chunk, we construct a *Future Abstraction* by averaging the corresponding future representations in latent space:

$$\mathbf{A}_i^* = \frac{1}{|\mathcal{C}_i|} \sum_{t \in \mathcal{C}_i} \mathbf{F}_t^{\text{future}}. \tag{4}$$

This simple, parameter-free aggregation reduces timestep-specific variation and summarizes information shared across the interval. Collectively, these abstractions form $\mathbf{A}^* = \{\mathbf{A}_i^*\}_{i=1}^{N}$, which provides an interval-level representation of the future.

This construction represents the same future at two temporal scales: the dense sequence $\mathbf{F}^{\text{future}}$ preserves timestep-level dynamics, whereas $\mathbf{A}^*$ summarizes future evolution over consecutive temporal intervals. Importantly, Future Abstractions are not intended to replace dense future states or serve as sparse future predictions; instead, they provide predictive context for dense latent forecasting.

3.3 Predicting High-Level Future Abstractions

Given the observed latent states $\mathbf{F}^{\text{past}}$, we introduce a high-level predictor $\mathcal{G}_{\text{high}}$ to anticipate future evolution at the interval level. Rather than directly resolving every future latent state, the high-level

predictor outputs N Future Abstractions:

$$\hat{\mathbf{A}} = \mathcal{G}_{\text{high}}\left(\mathbf{F}^{\text{past}}\right) = \{\hat{\mathbf{A}}_i\}_{i=1}^{N}. \tag{5}$$

Each $\hat{\mathbf{A}}_i$ is associated with a fixed future interval $\mathcal{C}_i$ and is trained to approximate the corresponding target abstraction $\mathbf{A}_i^*$.

We directly supervise Future Abstraction prediction with a mean squared error objective:

$$\mathcal{L}_{\text{abs}} = \frac{1}{N}\sum_{i=1}^{N}\left\|\hat{\mathbf{A}}_i - \mathbf{A}_i^*\right\|_2^2. \tag{6}$$

The high-level predictor is trained solely with $\mathcal{L}_{abs}$ using targets constructed from frozen future representations, without additional semantic annotations or task-specific supervision. At inference time, $\mathcal{G}_{\text{high}}$ operates on the observed latent states and predicts all Future Abstractions in a single forward pass. The predicted abstractions are then passed to the low-level predictor.

3.4 Dense Forecasting with Future Abstractions

While the high-level predictor captures future evolution at the interval level, we retain dense timestep-level forecasting for downstream prediction. We therefore condition the *low-level predictor* on the predicted Future Abstractions. Given the observed latent states $\mathbf{F}^{\text{past}}$ and the predicted abstractions $\hat{\mathbf{A}}$, the low-level predictor forecasts the complete future latent sequence as

$$\hat{\mathbf{F}}^{\text{future}} = \mathcal{G}_{\text{low}}\left(\mathbf{F}^{\text{past}}; \hat{\mathbf{A}}\right). \tag{7}$$

Here, $\mathbf{F}^{\text{past}}$ remains the primary input for latent forecasting, while $\hat{\mathbf{A}}$ provides high-level context about the expected future evolution.

To incorporate this context, we introduce cross-attention adapters into the low-level predictor. Let $\mathbf{Z}$ denote the hidden predictor tokens at a low-level predictor layer. We use $\mathbf{Z}$ as queries and the predicted Future Abstractions as keys and values:

$$\text{CA}(\mathbf{Z}, \hat{\mathbf{A}}) = \text{softmax}\left(\frac{(\mathbf{Z}\mathbf{W}_Q)(\hat{\mathbf{A}}\mathbf{W}_K)^\top}{\sqrt{d}}\right)(\hat{\mathbf{A}}\mathbf{W}_V), \tag{8}$$

where $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V$ are learnable projections and d denotes the attention dimension. The cross-attention output is residually added to the predictor stream before being processed by the subsequent predictor block. This allows the low-level predictor to selectively exploit information from different future intervals.

The predicted dense future representations are finally passed to a task-specific head $\mathcal{H}$ to obtain the downstream prediction, $\hat{\mathbf{Y}} = \mathcal{H}(\hat{\mathbf{F}}^{\text{future}})$. In this way, the high-level pathway provides an explicit expectation of future evolution, while the low-level predictor resolves this guidance into dense latent states required by downstream forecasting tasks.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate our method on four benchmarks covering complementary forecasting scenarios. **EgoDex** (Hoque et al., 2026) serves as our primary benchmark for egocentric hand trajectory forecasting in dexterous manipulation. **EgoExo4D** (Grauman et al., 2024) extends the evaluation from hand motion to whole-body dynamics, allowing us to examine whether the proposed multi-scale forecasting framework generalizes to more complex human motion. For physical scene dynamics, we further evaluate on **MOVi-C** (Greff et al., 2022) and **MOVi-E** (Greff et al., 2022), two physical scene video datasets containing scenes with multiple interacting objects. These two benchmarks assess whether predicted future representations can preserve object-level structure throughout physical scene evolution.

Table 2: Main results on EgoDex. Future Modeling indicates how each method models the future: **Direct** performs direct future forecasting, **Dense** predicts timestep-level future latents, **Dense + VLM** augments dense forecasting with VLM-derived guidance, and **Multi-Scale** combines dense latent prediction with Future Abstractions.

Method	Future Modeling	ADE ↓	FDE ↓	Acc ↑
Direct Forecasting	Direct	0.065	0.072	0.423
V-JEPA Predictor	Dense	0.049	0.067	0.656
ThinkJEPA	Dense + VLM	0.050	0.071	0.639
MS-JEPA	Multi-Scale	**0.044**	**0.061**	**0.706**

Table 3: Results on EgoDex under different future prediction horizons. All methods observe 16 frames and predict the next 8, 16, 24, or 32 frames in a single forward pass.

Method	ADE ↓				FDE ↓				Acc ↑			
	8	16	24	32	8	16	24	32	8	16	24	32
Direct Forecasting	0.068	0.064	0.071	0.069	0.069	0.066	0.074	0.083	0.380	0.402	0.361	0.374
V-JEPA Predictor	0.041	0.042	0.044	0.051	0.043	0.049	0.060	0.071	0.759	0.735	0.703	0.614
ThinkJEPA	0.041	0.043	0.046	0.050	0.044	0.048	0.065	0.071	0.752	0.725	0.657	0.629
MS-JEPA	**0.037**	**0.038**	**0.040**	**0.047**	**0.039**	**0.042**	**0.050**	**0.069**	**0.815**	**0.792**	**0.768**	**0.695**

Metrics. For EgoDex, we report Average Displacement Error (ADE), Final Displacement Error (FDE), and trajectory Accuracy (Acc), following the trajectory forecasting protocol used in ThinkJEPA (Zhang et al., 2026a). For EgoExo4D, we evaluate whole-body pose forecasting using Mean Per Joint Position Error (MPJPE) and Mean Per Joint Velocity Error (MPJVE), which measure the accuracy of predicted joint positions and temporal motion, respectively. For MOVi-C and MOVi-E, we report foreground Adjusted Rand Index (FG-ARI), which evaluates object-centric grouping over the predicted future sequence. Higher FG-ARI indicates that the future representations better preserve object identity and temporal consistency throughout scene evolution.

Baselines. We primarily compare against three forecasting baselines. **Direct Forecasting** predicts the future task outputs directly from observed representations, without explicitly modeling future latent states. **V-JEPA Predictor** represents the standard dense JEPA-style forecasting paradigm, organizing the predicted future through timestep-level latent states without an explicit interval-level representation. **ThinkJEPA** augments this dense formulation with VLM-derived semantic guidance obtained from the observed video. It therefore provides a complementary comparison between observation-derived semantic guidance and our predicted interval-level Future Abstractions. All methods use the corresponding task-specific prediction setting for each benchmark.

Training Protocol. We adopt a two-stage training procedure to decouple learning how to *use* Future Abstractions from learning how to *predict* them. Joint training from scratch couples two immature components: the high-level predictor initially produces inaccurate abstractions, while the low-level predictor has not yet learned a reliable abstraction-conditioning mechanism. In Stage I, we train the low-level predictor using ground-truth Future Abstractions constructed from future latents, providing a clean and stable conditioning signal. In Stage II, we freeze the low-level forecasting pathway and train the high-level predictor to infer Future Abstractions from observed latents, replacing the oracle guidance used in Stage I. At inference time, only observed video is available, and all Future Abstractions are predicted by the high-level predictor. We compare this two-stage protocol with joint training from scratch in Appendix A.2.

Implementation Details. Unless otherwise specified, we use a V-JEPA2 encoder to extract video representations and a Transformer-based predictor for dense latent forecasting. The low-level predictor operates with an internal dimension of $D_p = 384$, using 12 Transformer blocks and 6 at-

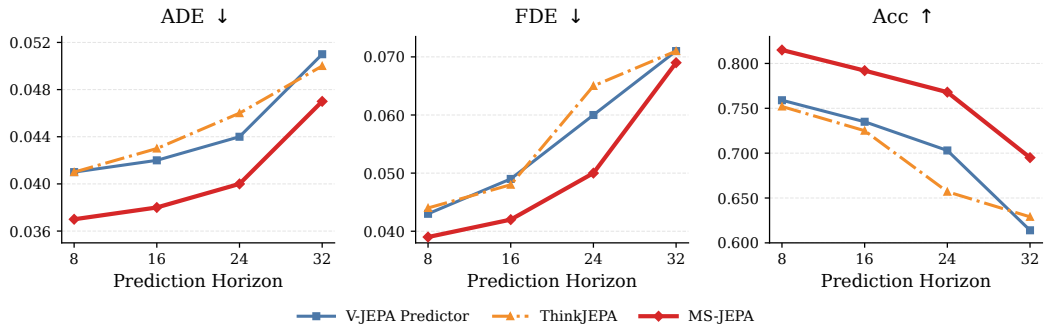

Figure 3: **Performance across different prediction horizons on EgoDex**. All methods observe 16 frames and predict 8, 16, 24, or 32 future frames in a single forward pass. MS-JEPA consistently improves over other latent forecasting baselines across different temporal extents.

Table 4: Whole-body motion forecasting results on EgoExo4D.

Method	Future Modeling	MPJPE ↓	MPJVE ↓
Direct Forecasting	Direct	16.70	83.30
V-JEPA Predictor	Dense	16.50	74.79
ThinkJEPA	Dense + VLM	16.49	73.49
MS-JEPA	Multi-Scale	**16.00**	**70.56**

tention heads. The high-level predictor uses a 12-layer Transformer with hidden dimension 512 and 8 attention heads, and predicts $N = 4$ Future Abstractions by default from the observed latent representations. These abstractions are injected into all layers of the low-level predictor through cross-attention adapters.

All compared methods are optimized with AdamW. For EgoDex and EgoExo4D, the task-specific prediction head is initialized with a learning rate of 1×10^{-3}, while MOVi-C/E use 1×10^{-4}. The low-level predictor uses a learning rate of 1×10^{-4} across all datasets. We adopt a stepwise learning-rate decay schedule during training. Models are trained for 500 epochs on EgoDex, 400 epochs on EgoExo4D, and 120 epochs on MOVi-C/E. For MS-JEPA, Stage I follows the same optimization settings as the corresponding dense forecasting baseline. In Stage II, the low-level forecasting pathway is kept fixed and only the high-level predictor is trained, using a learning rate of 1×10^{-4}.

4.2 Main Results

We first evaluate our method on EgoDex, where the model observes 32 frames and predicts the following 32 frames in a single forward pass. As shown in Table 2, MS-JEPA achieves the best performance across all three metrics, with an ADE of 0.044, FDE of 0.061, and Acc of 0.706. Compared with the V-JEPA Predictor, which obtains 0.049/0.067/0.656, MS-JEPA reduces both displacement errors while improving trajectory accuracy. MS-JEPA also outperforms ThinkJEPA (0.050/0.071/0.639), showing that predicted Future Abstractions provide effective guidance for dense latent forecasting.

4.3 Generalization Across Horizons and Tasks

Generalization across prediction horizons. We evaluate whether our method remains effective under different future horizons. Given a fixed observation window of 16 frames, we predict the next 8, 16, 24, or 32 frames in a single forward pass. Table 3 reports the detailed results, while Fig. 3 visualizes the performance trends across prediction horizons. MS-JEPA consistently achieves the best performance across all settings. In particular, its ADE increases from 0.037 at 8 frames to

Table 5: Physical-scene prediction results on MOVi-C and MOVi-E.

Method	Future Modeling	FG-ARI ↑	
		MOVi-C	MOVi-E
Direct Forecasting	Direct	71.82	85.91
V-JEPA Predictor	Dense	72.93	86.81
ThinkJEPA	Dense + VLM	73.30	86.58
MS-JEPA	Multi-Scale	**73.60**	**88.57**

Table 6: Ablation on the high-level prediction target. Future Abstractions aggregate latent states over each future interval, while Sparse Future States use a single latent state from each interval.

High-Level Target	ADE ↓	FDE ↓	Acc ↑
Sparse Future States	0.046	0.065	0.692
Future Abstractions	**0.044**	**0.061**	**0.706**

0.047 at 32 frames, compared with 0.041 to 0.051 for the V-JEPA Predictor. Similar advantages are observed in FDE and Acc across all four horizons, indicating that the benefit of Future Abstractions is maintained across different temporal extents.

Whole-body motion forecasting. On EgoExo4D, which evaluates egocentric whole-body motion forecasting, MS-JEPA achieves an MPJPE of 16.00 and an MPJVE of 70.56, compared with 16.50/74.79 for the V-JEPA Predictor and 16.49/73.49 for ThinkJEPA (Table 4). The improvement in both position and velocity errors shows that the proposed multi-scale formulation also benefits more complex human motion forecasting.

Physical scene prediction. As shown in Table 5, MS-JEPA achieves the highest FG-ARI on both MOVi-C and MOVi-E, reaching 73.60 and 88.57, respectively. The consistent improvements across human motion and physical scene prediction further support the general applicability of our approach across different forecasting tasks.

4.4 Ablation Studies and Analysis

Future Abstractions vs. Sparse Future States. We examine whether the benefit of the high-level pathway comes simply from providing sparse future anchors, or from summarizing information over temporal intervals. To isolate this difference, we replace each temporally aggregated Future Abstraction with the last future latent state within the corresponding interval, while keeping the remaining architecture unchanged. As shown in Table 6, Sparse Future States achieve an ADE/FDE/Acc of 0.046/0.065/0.692, whereas Future Abstractions improve the results to 0.044/0.061/0.706. Unlike sparse states, which represent isolated future instants, Future Abstractions summarize information shared across an interval. The improvement therefore suggests that interval-level aggregation provides more effective guidance than sparse future states alone.

Number of Future Abstractions. We study the effect of the number of Future Abstractions by varying it from 1 to 8. As shown in Table 7, using 2 or 4 abstractions gives the strongest overall performance. When $N = 1$, the entire future horizon is compressed into a single global summary, removing interval-level temporal organization. In contrast, intermediate values of N preserve multiple temporal intervals while remaining distinct from dense timestep-level prediction. When N is increased to 8, each abstraction covers a shorter interval, reducing the distinction between high-level temporal structure and low-level dynamics. Overall, an intermediate temporal scale is most effective in our setting.

Table 7: Ablation on the number of Future Abstractions. We vary the partition of the future horizon by using 1, 2, 4, or 8 abstractions.

Number	ADE ↓	FDE ↓	Acc ↑
1	0.046	0.062	0.696
2	0.045	**0.060**	**0.706**
4	**0.044**	0.061	**0.706**
8	0.046	0.063	0.688

Table 8: Results across different latent representation spaces. We compare dense forecasting and MS-JEPA using DINOv2 and V-JEPA2 encoders.

Latent Space	Method	ADE ↓	FDE ↓	Acc ↑
DINOv2	Dense Predictor	0.058	0.073	0.538
	MS-JEPA	**0.056**	**0.069**	**0.552**
V-JEPA2	Dense Predictor	0.049	0.067	0.656
	MS-JEPA	**0.044**	**0.061**	**0.706**

Effect Across Representation Spaces. We examine whether the benefit of multi-scale future modeling depends on the underlying latent representation. We replace the V-JEPA2 encoder with DINOv2 (Oquab et al., 2024) while keeping the forecasting architecture and evaluation protocol unchanged. As shown in Table 8, MS-JEPA improves over dense forecasting in both representation spaces. With DINOv2, FDE decreases from 0.073 to 0.069, while ADE and Acc improve from 0.058/0.538 to 0.056/0.552. The gains are larger with V-JEPA2, where MS-JEPA yields more substantial improvements across all three metrics. These results suggest that Future Abstraction guidance is not specific to the V-JEPA2 representation space, although its effectiveness varies with the underlying visual representation.

Predicted vs. Oracle Future Abstractions. We analyze the gap between predicted and oracle Future Abstractions. On EgoDex, dense forecasting achieves an ADE/FDE/Acc of 0.049/0.067/0.656, while MS-JEPA improves the results to 0.044/0.061/0.706. Using oracle abstractions further improves performance to 0.038/0.043/0.775. This progression shows the current high-level predictor already captures useful interval-level future structure, while a substantial portion of the potential benefit remains unrealized. In particular, the larger remaining gap in FDE suggests that accurately anticipating future evolution toward the end of the prediction horizon remains challenging. At the same time, the strong oracle performance indicates the low-level predictor can effectively exploit high-quality abstraction guidance. These results therefore suggest that the main opportunity for further improvement lies in strengthening the prediction of Future Abstractions rather than changing how they are consumed by the dense forecasting pathway.

5 Conclusion

We presented MS-JEPA, a multi-scale latent forecasting framework that represents the same future through both dense timestep-level states and interval-level Future Abstractions. This formulation makes longer-range temporal structure explicit while preserving fine-grained dynamics, enabling complementary modeling across temporal scales. Experiments across egocentric hand motion, whole-body motion, and physical-scene prediction consistently show improvements over latent forecasting baselines under different prediction horizons and task settings. Additional ablations further support the design of the proposed Future Abstractions and the robustness of the framework across representation spaces. While our framework achieves promising results, the remaining gap between predicted and oracle Future Abstractions indicates substantial room for improving high-level future anticipation, which may further strengthen latent world modeling.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on methodology or experiments. Other required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for code assistance and for improving the organization and readability of the manuscript. All AI-assisted content was reviewed and verified by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve new data collection or human-subject studies. All experiments are conducted on existing research datasets following their respective usage terms and protocols. We do not identify any significant ethical risks specific to the proposed methodology other than those associated with the underlying datasets and general-purpose predictive models.

REPRODUCIBILITY STATEMENT

We provide the key implementation and optimization details in Sec. 4.1 (*Implementation Details* and *Training Protocol*). Additional dataset-specific configurations and task-head settings are described in Appendix A.1. We plan to release the source code upon acceptance to facilitate reproduction of our results.

REFERENCES

Kumar Ashutosh, Georgios Pavlakos, and Kristen Grauman. Fiction: 4d future interaction prediction from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew J. Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H. Campbell, and Sergey Levine. Stochastic variational video prediction. In *International Conference on Learning Representations*, 2018.

Federico Baldassarre, Marc Szafraniec, Basile Terver, Vasil Khalidov, Francisco Massa, Yann LeCun, Patrick Labatut, Maximilian Seitzer, and Piotr Bojanowski. Back to the features: Dino as a foundation for video world models. *arXiv preprint arXiv:2507.19468*, 2025.

Wentao Bao, Lele Chen, Libing Zeng, Zhong Li, Yi Xu, Junsong Yuan, and Yu Kong. Uncertainty-aware state space transformer for egocentric 3d hand trajectory forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Jialei Chen, Kai Wang, Kang Chen, Shuaihang Chen, Feng Gao, Wenhao Tang, Zhiyuan Li, Weilin Liu, Zhuyu Yao, Boxun Li, Yuanbo Xu, and Chao Yu. Lawam: Latent world action models for efficient dynamics-aware robot policies. *arXiv preprint arXiv:2606.15768*, 2026.

Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Qiaohui Chu, Haoyu Zhang, Meng Liu, Yisen Feng, Haoxiang Shi, and Liqiang Nie. Intention-guided cognitive reasoning for egocentric long-term action anticipation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Emily Denton and Rob Fergus. Stochastic video generation with a learned prior. In *International Conference on Machine Learning*, 2018.

Maria Escobar, Juanita Puentes, Cristhian Forigua, Jordi Pont-Tuset, Kevis-Kokitsi Maninis, and Pablo Arbelaez. Egocast: Forecasting egocentric human pose in the wild. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025.

Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. *Advances in Neural Information Processing Systems*, 2016.

Rohit Girdhar and Kristen Grauman. Anticipative video transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zach Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, Maria Escobar, Cristhian Forigua, Abrham Gebreselasie, Sanjay Haresh, Jing Huang, Md Mohaiminul Islam, Suyog Jain, Rawal Khirodkar, Devansh Kukreja, Kevin J Liang, Jia-Wei Liu, Sagnik Majumder, Yongsen Mao, Miguel Martin, Effrosyni Mavroudi, Tushar Nagarajan, Francesco Ragusa, Santhosh Kumar Ramakrishnan, Luigi Seminara, Arjun Somayazulu, Yale Song, Shan Su, Zihui Xue, Edward Zhang, Jinxu Zhang, Angela Castillo, Changan Chen, Xinzhu Fu, Ryosuke Furuta, Cristina Gonzalez, Prince Gupta, Jiabo Hu, Yifei Huang, Yiming Huang, Weslie Khoo, Anush Kumar, Robert Kuo, Sach Lakhavani, Miao Liu, Mi Luo, Zhengyi Luo, Brighid Meredith, Austin Miller, Oluwatumininu Oguntola, Xiaqing Pan, Penny Peng, Shraman Pramanick, Merey Ramazanova, Fiona Ryan, Wei Shan, Kiran Somasundaram, Chenan Song, Audrey Southerland, Masatoshi Tateno, Huiyu Wang, Yuchen Wang, Takuma Yagi, Mingfei Yan, Xitong Yang, Zecheng Yu, Shengxin Cindy Zha, Chen Zhao, Ziwei Zhao, Zhifan Zhu, Jeff Zhuo, Pablo Arbelaez, Gedas Bertasius, Dima Damen, Jakob Engel, Giovanni Maria Farinella, Antonino Furnari, Bernard Ghanem, Judy Hoffman, C.V. Jawahar, Richard Newcombe, Hyun Soo Park, James M. Rehg, Yoichi Sato, Manolis Savva, Jianbo Shi, Mike Zheng Shou, and Michael Wray. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J. Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, Thomas Kipf, Abhijit Kundu, Dmitry Lagun, Issam Laradji, Hsueh-Ti (Derek) Liu, Henning Meyer, Yishu Miao, Derek Nowrouzezahrai, Cengiz Oztireli, Etienne Pot, Noha Radwan, Daniel Rebain, Sara Sabour, Mehdi S. M. Sajjadi, Matan Sela, Vincent Sitzmann, Austin Stone, Deqing Sun, Suhani Vora, Ziyu Wang, Tianhao Wu, Kwang Moo Yi, Fangcheng Zhong, and Andrea Tagliasacchi. Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Agrim Gupta, Stephen Tian, Yunzhi Zhang, Jiajun Wu, Roberto Martín-Martín, and Li Fei-Fei. Maskvit: Masked visual pre-training for video prediction. In *International Conference on Learning Representations*, 2023.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International Conference on Machine Learning*, 2019.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2020.

Danijar Hafner, Timothy P Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. In *International Conference on Learning Representations*, 2021.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2024.

Ryan Hoque, Peide Huang, David J. Yoon, Mouli sivapurapu, and Jian Zhang. Egodex: Learning dexterous manipulation from large-scale egocentric video. In *International Conference on Learning Representations*, 2026.

Shaowei Liu, Subarna Tripathi, Somdeb Majumdar, and Xiaolong Wang. Joint hand motion and interaction hotspots prediction from egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. Leworld-model: Stable end-to-end joint-embedding predictive architecture from pixels. *arXiv preprint arXiv:2603.19312*, 2026.

Himangi Mittal, Nakul Agarwal, Shao-Yuan Lo, and Kwonjoon Lee. Can't make an omelette without breaking some eggs: Plausible action anticipation using large video-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mido Assran, Koustuv Sinha, Mike Rabbat, Yann LeCun, Nicolas Ballas, and Adrien Bardes. V-jepa 2.1: Unlocking dense features in video self-supervised learning. *arXiv preprint arXiv:2603.14482*, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Basile Terver, Tsung-Yen Yang, Jean Ponce, Adrien Bardes, and Yann LeCun. What drives success in physical planning with joint-embedding predictive world models? *arXiv preprint arXiv:2512.24497*, 2026.

Vikram Voleti, Alexia Jolicoeur-Martineau, and Chris Pal. Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. *Advances in Neural Information Processing Systems*, 2022.

Kai Xu, Hang Zhao, Ruizhen Hu, Yuhang Huang, Ziqiao Zhou, Wancheng Feng, Li Yi, Sida Peng, Xing Liu, Zihao Liu, Jin Zhang, Chenyang Zhu, Renjiao Yi, Qin Zou, Bo Du, and Haibin Yu. From specialist to generalist: A comprehensive survey on world models. *TechRxiv*, 2026.

Haichao Zhang, Yijiang Li, Shwai He, Tushar Nagarajan, Mingfei Chen, Jianglin Lu, Ang Li, and Yun Fu. Thinkjepa: Empowering latent world models with large vision-language reasoning model. *arXiv preprint arXiv:2603.22281*, 2026a.

Wancong Zhang, Basile Terver, Artem Zholus, Soham Chitnis, Harsh Sutaria, Mido Assran, Randall Balestriero, Amir Bar, Adrien Bardes, Yann LeCun, and Nicolas Ballas. Hierarchical planning with latent world models. *arXiv preprint arXiv:2604.03208*, 2026b.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. In *International Conference on Machine Learning*, 2025.

A Appendix

A.1 Additional Training Details

A.1.1 Dataset-specific Forecasting Settings

For **EgoDex**, we use a 64-frame clip with a 32/32 observation-prediction split for the main experiments. For the horizon generalization experiments, the observation window is fixed to 16 frames, while the prediction horizon is varied among 8, 16, 24, and 32 frames.

For **EgoExo4D**, we uniformly sample the 30-FPS videos at 10 FPS and use a 20/10 observation-prediction split, corresponding to two seconds of observed video followed by one second of future prediction. The model receives only egocentric RGB frames from the Aria camera and predicts the 3D coordinates of 17 body joints over the ten future frames. Training is performed in the camera coordinate system, while world-coordinate metrics are computed at evaluation time using the annotated camera extrinsics.

For **MOVi-C/E**, each 24-frame video is divided into 12 observed and 12 future frames. The model predicts future latent representations from the observed clip, which are decoded into class-agnostic instance segmentation masks for all 12 future frames. Observed and future windows are encoded separately to avoid information leakage from bidirectional attention. Future frames and future latent representations are used only for constructing training targets and are unavailable during inference.

A.1.2 Task-specific Prediction Heads

For **EgoDex**, we use a lightweight trajectory head on top of the predicted future latents. Spatial tokens at each latent timestep are first aggregated through multi-head attention with a learnable query, followed by a residual MLP for feature refinement. A final linear layer predicts the 3D coordinates of 52 hand and arm joints for the future frames.

For **EgoExo4D**, we use a lightweight pose prediction head on top of the predicted future latents. Spatial tokens at each latent timestep are aggregated through multi-head attention with a learnable query, followed by an MLP-based feature refinement module. A final linear layer predicts the 3D coordinates of 17 body joints for the future frames. The head is trained with a visibility-masked mean squared error over valid joints.

For **MOVi-C/E**, we use a lightweight Mask2Former-style (Cheng et al., 2022) instance segmentation head. Predicted future latents are projected to frame-level features and processed into a multi-scale spatial feature pyramid. Learned object queries are then decoded through masked cross-attention to predict class-agnostic instance-mask tubes over the future frames. Training uses bipartite matching with classification, focal, and Dice losses.

A.2 Additional Experiments and Analysis

Two-Stage Training vs. Joint Training. To examine whether decoupling abstraction conditioning from abstraction prediction is beneficial, we compare the proposed two-stage protocol with joint training from scratch, where the high-level and low-level predictors are optimized simultaneously. As shown in Table 9, Joint Training already improves over dense forecasting, achieving an ADE/FDE/Acc of 0.047/0.062/0.673, showing that Future Abstractions provide useful guidance even without staged optimization. Two-Stage Training further improves performance to 0.044/0.061/0.706, suggesting that separating the two learning objectives enables more effective use of the predicted abstractions.

Table 9: Ablation on different training strategies. Joint Training learns both predictors simultaneously from scratch, while Two-Stage Training decouples abstraction conditioning from abstraction prediction.

Training Strategy	ADE ↓	FDE ↓	Acc ↑
Joint Training	0.047	0.062	0.673
Two-Stage Training	**0.044**	**0.061**	**0.706**

Model Complexity and Computational Overhead. We measure computational cost under the standard EgoDex setting, where the model observes 32 frames and predicts the subsequent 32 frames in a single forward pass. As summarized in Table 10, the V-JEPA Predictor requires 5.745 TFLOPs and 101.26 ms per sample, whereas MS-JEPA requires 6.530 TFLOPs and 114.06 ms on a single NVIDIA RTX 4090, corresponding to increases of approximately 13.7% and 12.6%, respectively. Despite moderate increases in computation and latency, MS-JEPA improves Acc from 0.656 to 0.706. This additional overhead mainly comes from the high-level predictor, which predicts interval-level Future Abstractions to condition dense forecasting.

Table 10: Computational cost under the standard EgoDex 32/32 observation-prediction setting. FLOPs are measured for the complete pipeline including the visual encoder, and inference latency is measured on a single NVIDIA RTX 4090.

Method	TFLOPs ↓	Latency (ms) ↓	Acc ↑
V-JEPA Predictor	**5.745**	**101.26**	0.656
MS-JEPA	6.530	114.06	**0.706**

A.3 Additional Visualization Examples

Figure 4 provides additional qualitative examples on EgoDex, comparing dense forecasting with Multi-Scale (Oracle) predictions against the ground-truth future trajectories.

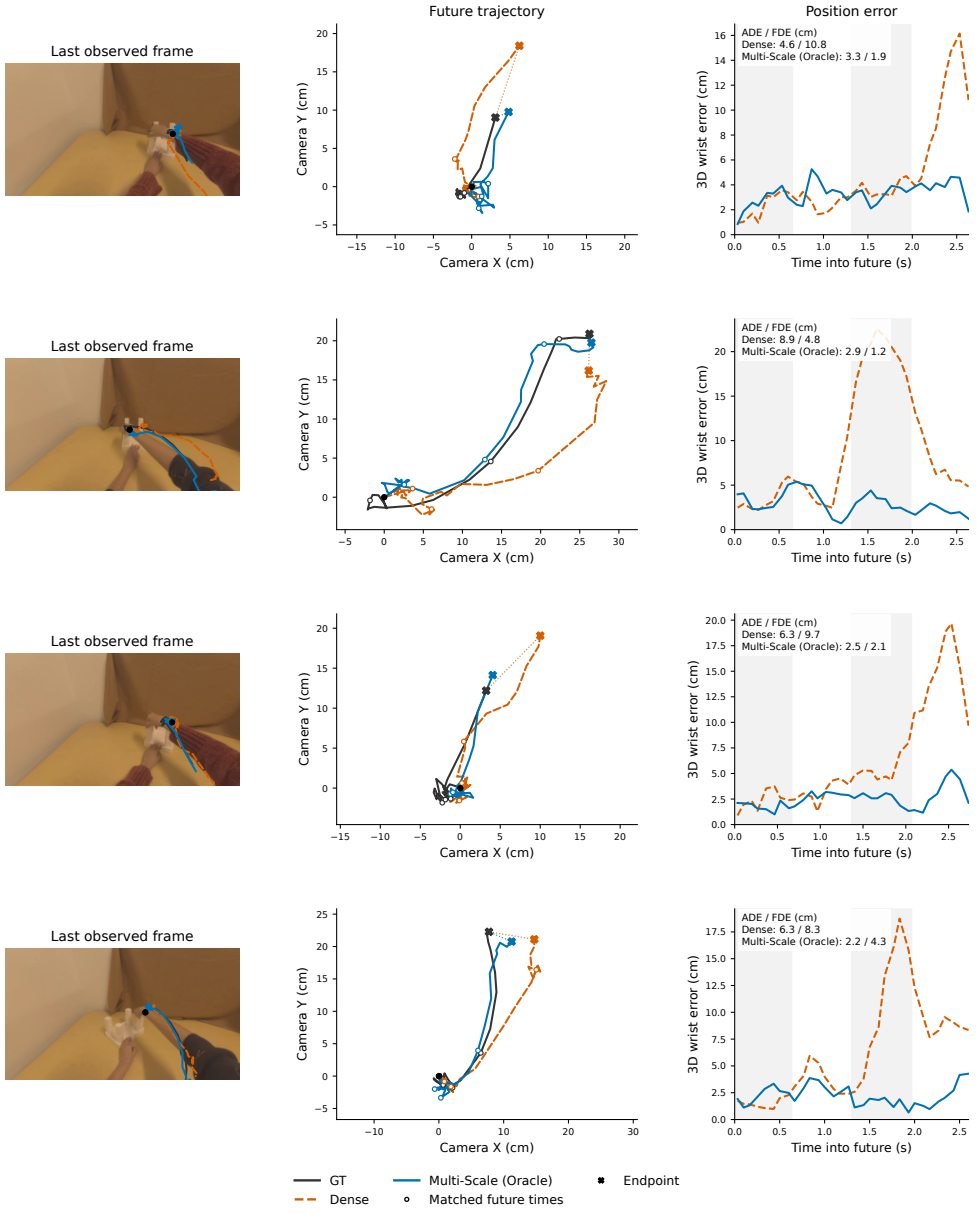

Figure 4: Additional Visualization Examples.

A.4 Limitations and Future Directions

Several aspects of the current formulation remain underexplored and offer promising directions for future work. First, Future Abstractions are constructed using fixed temporal partitions and mean aggregation. While this design is simple and parameter-free, it may not fully capture videos with non-uniform temporal dynamics. Future work could explore adaptive interval partitioning and learned aggregation mechanisms that better adjust to changes occurring at different temporal scales.

Second, a noticeable gap remains between predicted and oracle Future Abstractions, suggesting substantial room for improving high-level future prediction. Future work could investigate more expressive abstraction predictors to improve the accuracy of interval-level future anticipation.