DOPA: Can MLLMs Predict the Future Beyond Text?

Anonymous ACL submission

Abstract

Next-Event Prediction (NEP) trains multimodal large language models (MLLMs) to anticipate future events in videos, yet existing methods rely mainly on textual captions of future segments as supervision. Such caption-only supervision can under-specify fine-grained visual dynamics and limit future understanding to what is expressible in text. We propose **DOPA**, a **Dual-Objective Predictive Architecture** that complements future-caption supervision with visual-semantic embedding prediction. DOPA adds a lightweight prediction head to an MLLM, training it with both text-based supervision and future visual-semantic representation prediction to produce more accurate future-event answers. Experiments show that DOPA improves future prediction and temporal understanding by an average of +7.1 and +1.6 percentage points over the base model.

1 Introduction

The ability to predict the future is increasingly recognized as a core competency in recent research on intelligent systems, such as robotic planning, embodied decision-making, and general video understanding (Bruce et al., 2024; NVIDIA et al., 2025; Qu et al., 2025). It is becoming more and more important to reason about what will happen next, rather than to simply describe what is currently visible. Recent work on self-supervised world models shows that this anticipation ability can be learned from large-scale video, and can further enable zero-shot robot planning (Assran et al., 2025). A similar shift is happening for Multimodal Large Language Models (MLLMs), where recent efforts increasingly target higher-order capabilities such as reasoning, deliberation, and faithful explanation, going beyond simple next-token prediction (Guo et al., 2025; OpenAI, 2024; Zhang et al., 2024b; Xu et al., 2025; Rha et al., 2026). This trend is now accelerating in the video domain as well.

Next-Event Prediction (NEP) (Wang et al., 2025a) introduces a self-supervised learning task explicitly designed to foster temporal reasoning in MLLMs. It segments each video into past and future frames, and trains the model to predict a summary of events derived from the future frames based solely on the past frames. This aligns with the goal of modeling world dynamics, extending beyond static understanding of frames to reasoning about how scenes evolve over time. Despite this progress, existing NEP methods share one implicit design choice: to construct the supervision signal, they leverage automatically generated captions from future video segments as the prediction target. The model therefore does not learn from the future video itself, but only indirectly, through a textual representation of it.

As illustrated in Figure 1(a)–(c), text-based supervision (Ouyang et al., 2022; Wang et al., 2025c; Ho et al., 2023) is a natural choice for future prediction, but it can also become a bottleneck. While captions provide compact semantic summaries, they often under-specify visually distinct outcomes, fine-grained motion, spatial layouts, and subtle transitions. Because captions do not explicitly encode these visual details, caption-only supervision can restrict the model's future understanding to the aspects that are expressible in text. Moreover, relying solely on textual supervision risks encouraging the model to learn plausible-sounding future generation, rather than genuine visual reasoning about how the scene will evolve. Considering that the core objective of future prediction is to anticipate the future visual state, this motivates complementing caption supervision with visual signals that more directly capture scene evolution.

This view is also connected to cognitive accounts of human anticipation. Predictive coding theories suggest that the human brain shapes perception by predicting upcoming sensory states (Rao and Ballard, 1999; Clark, 2013). Anticipation is not

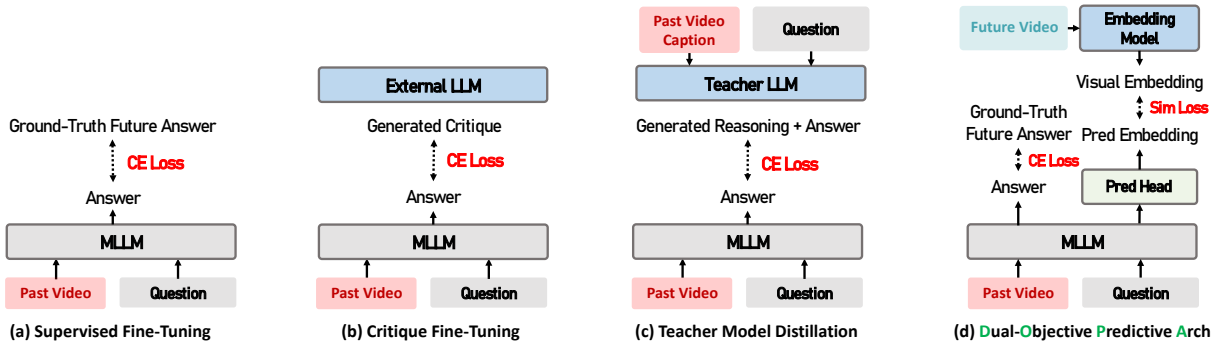


Figure 1: Comparison of training paradigms for video MLLMs in Next-Event Prediction. Existing text-based supervision methods train the model with (a) ground-truth future-caption targets, (b) external LLM-generated critiques, or (c) teacher-generated reasoning and answers. In contrast, (d) **DOPA** is a **Dual-Objective Predictive Architecture** that combines textual future-caption supervision with visual-semantic embedding prediction, enabling the model to learn from the future video beyond text-level supervision.

simply the generation of a verbal description, but it is closer to an internal simulation of a future perceptual state. Such a perspective supports the idea that learning to predict visual representations may better align with the nature of future prediction than only learning to predict captions. Based on this idea, we propose **DOPA**, a **Dual-Objective Predictive Architecture** that trains video MLLMs with both textual future-caption supervision and visual-semantic embedding prediction. This target complements caption supervision by preserving visual-semantic information that could be missed in captions.

Specifically, we extract a visual-semantic representation from the future video segment and train the MLLM to predict this representation from the past video alone. This objective is optimized together with the standard caption generation loss. As a result, the model preserves its ability to describe future events in language, while also learning the visual semantics of the future through embedding-level supervision derived from the future video. During training, the model is encouraged to distinguish the true future state from alternative future candidates, rather than only generating plausible textual continuations. At inference, neither the future video nor the embedding model is required. The model generates future events autoregressively, using the future-oriented representations internalized during training.

Our main contributions are summarized as follows:

- We introduce DOPA, which incorporates embedding-space future prediction as an auxiliary signal within the instruction-tuning framework of MLLMs, jointly optimized with caption supervision.
- We instantiate this signal with two complementary objectives. The Future Embedding Alignment (FEA) loss aligns the predicted embedding with its paired future representation while contrasting it against other futures in the batch. The Past–Future Ranking (PFR) loss treats the observed past as a hard negative and encourages the predicted embedding to be closer to the actual future than to the past by an explicit margin.
- We demonstrate that DOPA improves both future prediction and temporal understanding across multiple benchmarks. Compared with caption-based training baselines, DOPA achieves stronger future-oriented reasoning while remaining competitive on general video understanding.

2 Related Works

2.1 Future Event Prediction

Future event prediction has been studied across both textual and multimodal settings. Early text-based benchmarks such as the Story Cloze Test (Mostafazadeh et al., 2016) formalized future prediction as selecting a plausible ending for a commonsense narrative, while subsequent forecasting benchmarks extended this direction to real-world events such as news articles, prediction markets, and dynamically updated forecasting scenarios (Jin et al., 2021; Zou et al., 2022; Karger et al., 2025; Zeng et al., 2026). In video-language settings, VLEP (Lei et al., 2020) introduced a task

that requires models to select the more plausible future event after observing a video, and VFE (Lai and Wang, 2025) further extended this direction to a generative setting by asking models to directly produce future-event descriptions from videos.

With the emergence of MLLMs (Qwen Team, 2025; Cheng et al., 2024; Zhang et al., 2024a; Wang et al., 2025b), future event prediction has recently been reformulated as an instruction-tuning problem over videos. NEP (Wang et al., 2025a) is a representative framework in this direction, where future video segments are converted into textual summaries and used as self-supervised targets for training MLLMs to predict what happens next. It further constructs the V1-33K dataset and FutureBench benchmark, and investigates multiple tuning strategies, including supervised fine-tuning (Ouyang et al., 2022), critique fine-tuning (Wang et al., 2025c), and teacher-model distillation (Ho et al., 2023).

2.2 Embedding-space Future Prediction

Predicting the future in an embedding space, rather than in token or pixel space, was established as a central idea in JEPA (LeCun, 2022). The key motivation is that prediction in the raw data space can force the model to reconstruct semantically irrelevant details, whereas prediction in an embedding space can provide a more semantic learning signal. I-JEPA (Assran et al., 2023) applies this principle to images, while V-JEPA (Bardes et al., 2024) and V-JEPA 2 (Assran et al., 2025) extend it to videos by learning representations that predict masked regions and capture temporal dynamics.

This paradigm has since been applied to other domains such as procedural video forecasting (Lin et al., 2025) and robotic control (Miao et al., 2026; Sun et al., 2026), where embedding prediction is typically used either as a standalone learning objective or as the core of a separate forecasting or world-model architecture. In contrast, we integrate embedding-space future prediction directly into the instruction-tuning framework of MLLMs, where it serves as an auxiliary supervision signal jointly optimized with caption-based objectives.

3 Methods

3.1 Preliminaries

Next-Event Prediction (NEP) (Wang et al., 2025a) partitions a video V along the temporal axis into an observed past segment V_p and an unobserved future segment V_f:

$$V = V_p \oplus V_f. \tag{1}$$

The goal of NEP is to infer the future event in V_f given only the observed past segment V_p. In caption-based NEP, each future segment V_f is paired with a textual caption $c_f = \{c_f^1, c_f^2, \ldots, c_f^T\}$, where T denotes the caption length.

In MLLM-based NEP, the model π_θ takes the past video V_p and a text query Q (e.g., *"Describe what happens next in the video"*) as a multimodal input prefix. Let this prefix be tokenized into a sequence of length k. During training, the ground-truth caption tokens c_f are appended to this prefix, and the full sequence is processed under a causal attention mask. This yields hidden states

$$H = \{ \underbrace{\mathbf{h}_1, \ldots, \mathbf{h}_k}_{V_p \,\&\, Q \text{ positions}}, \underbrace{\mathbf{h}_{k+1}, \ldots, \mathbf{h}_{k+T}}_{\text{caption positions}} \} \in \mathbb{R}^{(k+T)\times d_\pi}, \tag{2}$$

where d_π denotes the hidden dimension of the MLLM π_θ.

Under the standard next-token prediction objective, the hidden state at position $k+t-1$ is used to predict the t-th caption token c_f^t. The LM head $g_{\text{LM}} : \mathbb{R}^{d_\pi} \to \mathbb{R}^{|\mathcal{V}|}$ maps a hidden state to logits over the vocabulary $\mathcal{V}$, and the conditional probability of c_f^t is computed as

$$\pi_\theta(c_f^t \mid V_p, Q, c_f^{<t}) = \text{softmax}\big(g_{\text{LM}}(\mathbf{h}_{k+t-1})\big)[c_f^t]. \tag{3}$$

Text-based supervised fine-tuning minimizes the negative log-likelihood of the ground-truth future caption:

$$\mathcal{L}_{\text{SFT}} = -\sum_{t=1}^{T} \log \pi_\theta(c_f^t \mid V_p, Q, c_f^{<t}). \tag{4}$$

At inference time, neither the future segment V_f nor its caption c_f is available; the model generates a future-event description autoregressively conditioned only on V_p and Q.

3.2 Dual-Objective Predictive Architecture

To ground future prediction in visual signals beyond text, we augment the standard caption-generation pipeline with a parallel visual representation prediction path (Figure 2). This path introduces two components: a prediction head g_ψ attached to the MLLM hidden representation, and a frozen multimodal embedding model ϕ that provides visual-semantic supervision targets.

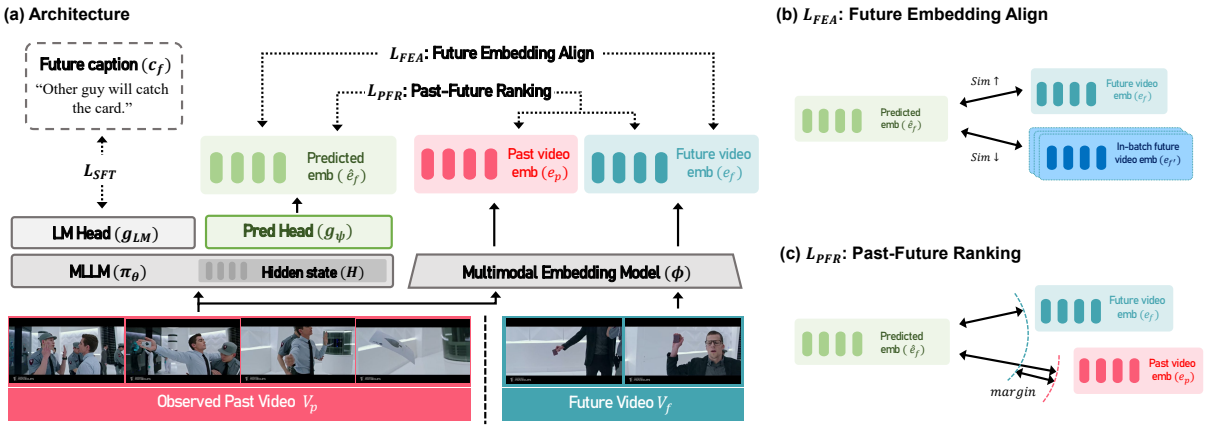


Figure 2: Overview of DOPA. (a) The model jointly learns future caption generation and future visual-semantic embedding prediction from the observed past video. (b) Future Embedding Alignment encourages the predicted embedding to match the target future representation while distinguishing it from other future embeddings in the batch. (c) Past–Future Ranking treats the observed past embedding as a hard negative and encourages the prediction to be closer to the actual future embedding.

Prediction Head g_ψ. While the language modeling head g_{LM} projects MLLM hidden states into vocabulary logits, the prediction head g_ψ maps the same hidden space into a visual embedding space:

$$g_\psi : \mathbb{R}^{d_\pi} \to \mathbb{R}^d, \tag{5}$$

where d denotes the output dimension of the frozen multimodal embedding model ϕ. We instantiate g_ψ as a lightweight MLP that maps the MLLM hidden representation to the visual embedding space, and train it jointly with the MLLM parameters θ.

Given the hidden states H from the MLLM, we extract a representation $\mathbf{h}_* \in \mathbb{R}^{d_\pi}$ for future visual prediction. In our default design, $\mathbf{h}_*$ is extracted from the input prefix so that visual prediction is conditioned only on the observed past video V_p and query Q. For example, one can use the hidden state of the final query token or that of a separately introduced future prediction token. The predicted future visual embedding is then computed as

$$\hat{\mathbf{e}}_f = g_\psi(\mathbf{h}_*) \in \mathbb{R}^d. \tag{6}$$

We compare different choices of $\mathbf{h}_*$ in Table 4.

Frozen Multimodal Embedding Model ϕ. The supervision target for $\hat{\mathbf{e}}_f$ is obtained from a frozen multimodal embedding model $\phi : \mathcal{X}_{\text{video}} \to \mathbb{R}^d$. The model ϕ maps a video segment into a vision-language aligned visual-semantic embedding space. Given a future segment V_f, the target future embedding is defined as

$$\mathbf{e}_f = \phi(V_f) \in \mathbb{R}^d. \tag{7}$$

Since $\mathbf{e}_f$ is extracted directly from the future video segment rather than from its caption, it provides a training signal grounded in the content of the future video beyond caption supervision.

We keep ϕ frozen throughout training. This design provides a fixed target distribution for stable supervision and improves training efficiency, since video embeddings can be pre-computed because both V_p and V_f are available during training. For the past–future ranking objective introduced in Section 3.3, we also extract the past video embedding using the same embedding model:

$$\mathbf{e}_p = \phi(V_p) \in \mathbb{R}^d. \tag{8}$$

3.3 Training Objective

The proposed architecture is trained with a multi-objective loss that combines standard caption supervision with visual future prediction objectives. Since the caption-based supervised fine-tuning loss $\mathcal{L}_{\text{SFT}}$ is already defined in Eq. 4, this section focuses on the two additional objectives used to train the visual prediction path: future embedding alignment and past–future ranking.

Future Embedding Alignment. Given the predicted future embedding $\hat{\mathbf{e}}_f$ and the target future embedding $\mathbf{e}_f$, we encourage the prediction to align with its paired future representation using a cosine alignment loss:

$$\mathcal{L}_{\cos} = 1 - \text{sim}(\hat{\mathbf{e}}_f, \mathbf{e}_f), \tag{9}$$

where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity.

In addition, we use an in-batch contrastive objective to make the predicted future embedding

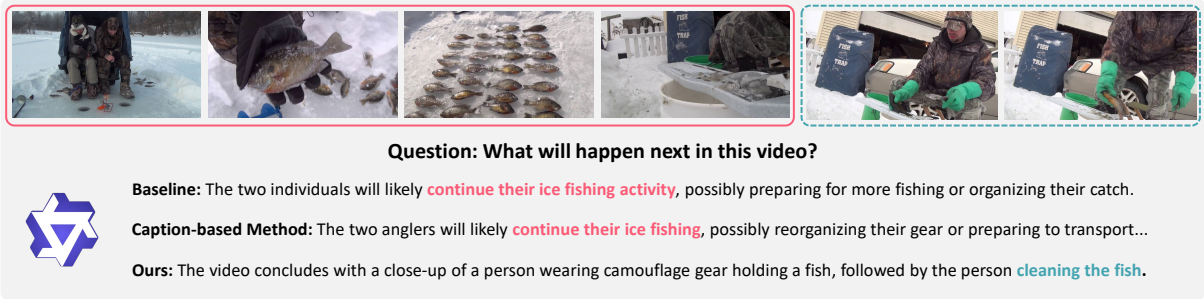


Figure 3: Illustration of the past–future ambiguity in future event prediction. Caption-based baselines often generate continuations close to the observed past, while DOPA better captures the actual future transition.

distinguishable from other future representations in the batch. For a mini-batch of size B, let $\hat{\mathbf{e}}_f^i$ and $\mathbf{e}_f^i$ denote the predicted and target future embeddings of the i-th sample, respectively. The InfoNCE loss is defined as

$$\mathcal{L}_{\text{NCE}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(\text{sim}(\hat{\mathbf{e}}_f^i, \mathbf{e}_f^i)/\tau)}{\sum_{j=1}^{B}\exp(\text{sim}(\hat{\mathbf{e}}_f^i, \mathbf{e}_f^j)/\tau)}, \quad (10)$$

where τ is a temperature hyperparameter. The final future embedding alignment loss is defined as

$$\mathcal{L}_{\text{FEA}} = \alpha\mathcal{L}_{\cos} + (1-\alpha)\mathcal{L}_{\text{NCE}}, \quad (11)$$

where $\alpha \in [0, 1]$ controls the trade-off between direct pairwise alignment and in-batch contrastive discrimination. The cosine term provides an absolute alignment signal by directly pulling $\hat{\mathbf{e}}_f$ toward its paired future embedding $\mathbf{e}_f$, while the InfoNCE term provides a relative discriminative signal by contrasting $\hat{\mathbf{e}}_f$ against other future embeddings in the batch.

Past–Future Ranking Loss. As shown in Figure 3, caption-based baselines can produce generic continuations that stay close to the observed context. For example, although the input video shows people sitting down after ice fishing to prepare the caught fish, the baseline predicts that they sit down to continue ice fishing. This motivates the need to distinguish the actual future transition from a plausible continuation of the past.

We find that this distinction is also non-trivial in the embedding space. To examine this, we conduct an analysis on 1,000 randomly sampled past–future video pairs. The cosine similarity between the past and future segments from the same video is 0.788, whereas the similarity between future segments from different videos in a random batch is only 0.327. This indicates that the observed past is a strong hard negative for future prediction, as it is close to the actual future but represents a different temporal state.

The key idea is to use the observed past as a hard negative and encourage the predicted embedding to be closer to the actual future than to the observed past. We formulate this as a ranking preference problem between two candidate visual states: the actual future embedding $\mathbf{e}_f$ and the observed past embedding $\mathbf{e}_p$. Given the predicted future embedding $\hat{\mathbf{e}}_f$, we aim to assign a higher preference score to $\mathbf{e}_f$ than to $\mathbf{e}_p$. Following the Bradley–Terry model, we model this preference probability using the difference between their scores:

$$P(\mathbf{e}_f \succ \mathbf{e}_p \mid \hat{\mathbf{e}}_f) = \sigma(s_f - s_p), \quad (12)$$

where $\sigma(\cdot)$ is the sigmoid function, and s_f and s_p denote the preference scores of the future and past candidates, respectively.

We instantiate each candidate score as its cosine similarity to the predicted future embedding:

$$s_f = \text{sim}(\hat{\mathbf{e}}_f, \mathbf{e}_f), \qquad s_p = \text{sim}(\hat{\mathbf{e}}_f, \mathbf{e}_p). \quad (13)$$

Minimizing the negative log-likelihood of this preference yields the basic past–future ranking objective:

$$\mathcal{L}_{\text{rank}} = -\log\sigma\left(\text{sim}(\hat{\mathbf{e}}_f, \mathbf{e}_f) - \text{sim}(\hat{\mathbf{e}}_f, \mathbf{e}_p)\right). \quad (14)$$

However, since the past and future segments are temporally adjacent segments from the same video, they can be highly similar in the visual embedding space. Thus, simply encouraging $\mathbf{e}_f$ to be preferred over $\mathbf{e}_p$ may not guarantee sufficient separation. To address this, we introduce a margin m, analogous to the margin used in triplet loss, and require the future score to exceed the past score by at least m:

$$s_f > s_p + m. \quad (15)$$

The final Past–Future Ranking loss is then defined as

$$\mathcal{L}_{\text{PFR}} = -\log\sigma\left(\text{sim}(\hat{\mathbf{e}}_f, \mathbf{e}_f) - \text{sim}(\hat{\mathbf{e}}_f, \mathbf{e}_p) - m\right). \quad (16)$$

This objective encourages the predicted future embedding not only to be closer to the actual future representation than to the observed past representation, but also to move toward the future representation with an explicit margin.

Overall Objective. The final training objective combines caption supervision, future embedding alignment, and past–future ranking:

$$\mathcal{L} = \mathcal{L}_{\text{SFT}} + \mathcal{L}_{\text{FEA}} + \mathcal{L}_{\text{PFR}}. \tag{17}$$

Through this multi-objective training, the model preserves its caption generation ability while learning to predict future visual representations grounded in the actual future video.

4 Experiments

4.1 Experiment Settings

Training Data. We train our model on the V1-33K dataset introduced by NEP. Since our method requires both the observed past segment V_p and its corresponding future segment V_f to construct visual-semantic supervision targets, we use only samples with valid past–future video pairs. The original training split contains 31,809 QA pairs, and after filtering out samples without valid future-video pairs, we obtain 29,679 training samples.

Backbone MLLM. We use Qwen2.5-VL-7B (Qwen Team, 2025) as the backbone and follow its original video preprocessing pipeline. The backbone is fine-tuned with LoRA (Hu et al., 2022) using rank 64 and alpha 128, while the prediction head g_ψ is fully trained. The language modeling head is kept unchanged, and g_ψ is attached to the selected hidden representation for visual future prediction.

Multimodal Embedding Model. To construct visual supervision targets, we use Qwen3-VL-Embedding-2B (Li et al., 2026), which maps video segments into an embedding space, as the frozen multimodal embedding model ϕ. For each training sample, we pre-compute the future embedding $\mathbf{e}_f = \phi(V_f)$ and the past embedding $\mathbf{e}_p = \phi(V_p)$.

Prediction Head. The prediction head g_ψ is implemented as a lightweight MLP that maps the MLLM hidden dimension d_π to the visual embedding dimension d. By default, we use a three-layer MLP with a hidden dimension of 4096, GELU activation, and LayerNorm. The position used to extract the input representation $\mathbf{h}_*$ is compared in Section 4.4.4.

Training Objective. The model is trained with the combined objective in Eq. 17. This objective consists of the caption supervision loss $\mathcal{L}_{\text{SFT}}$, the future embedding alignment loss $\mathcal{L}_{\text{FEA}}$, and the Past–Future ranking loss $\mathcal{L}_{\text{PFR}}$. We set the temperature of the InfoNCE term to $\tau = 0.2$, the margin of the PFR loss to $m = 0.2$, and the embedding loss balancing coefficient to $\alpha = 0.5$. Unless otherwise specified, all hyperparameters are kept fixed across datasets and benchmarks.

Optimization. We use the AdamW optimizer with a learning rate of 5×10^{-5}, a weight decay of 0.01, and a batch size of 64. The model is trained for one epoch with a cosine learning-rate schedule. Training is performed with bfloat16 mixed precision. All trainings are conducted on four NVIDIA RTX A6000 GPUs.

4.2 Evaluation Benchmarks

Future Prediction Tasks. To evaluate future-event prediction ability, we use FutureBench and MMworld$_{fp}$. FutureBench, introduced by NEP (Wang et al., 2025a), measures whether models can generate coherent and causally consistent predictions of unseen future events. Because the V1-33K training set is drawn from the same data pool as FutureBench, we use FutureBench to evaluate in-domain future prediction performance. In contrast, MMworld$_{fp}$ is used to assess generalization to more diverse future prediction scenarios. MMworld$_{fp}$ is the future-prediction subset of MMworld (He et al., 2024), a multi-discipline and multi-faceted video understanding benchmark covering reasoning types such as explanation, counterfactual thinking, and future prediction.

Temporal Understanding Tasks. We further evaluate on TemporalBench (TB.) (Cai et al., 2024), TempCompass (TC.) (Liu et al., 2024), SEED-Bench-R1 (Chen et al., 2025), and TOMATO (Shangguan et al., 2025). These benchmarks assess general temporal understanding beyond future prediction, covering aspects such as temporal order, motion, direction, attribute change, and frame-order sensitivity. This evaluation helps verify that our method benefits broader temporal reasoning rather than only improving future-caption generation.

Table 1: Results on future prediction and temporal understanding tasks. DOPA achieves the best average performance on both future prediction benchmarks and temporal understanding benchmarks, demonstrating improved future-oriented reasoning and consistent gains in temporal understanding.

Model	Future Prediction Task			Temporal Understanding Task				
	FutureBench	MMworld$_{fp}$	FP-AVG.	TB.	TC.	Seed-R1	TOMATO	TU-AVG.
Qwen2.5-VL-7B	52.9	83.9	68.4	40.8	70.7	37.1	26.1	43.7
Training Methods								
SFT	60.0	79.6	69.8	40.9	70.0	**38.2**	26.2	43.8
CFT	56.8	78.1	67.4	34.7	72.5	37.2	27.0	42.9
Distill	58.4	79.6	69.0	39.1	73.6	35.9	27.4	44.0
Mix	60.5	83.2	71.8	39.1	70.6	37.9	**28.7**	44.1
DOPA (Ours)	**65.5**	**85.4**	**75.5**	**41.0**	**73.7**	37.9	**28.7**	**45.3**

4.3 Compared Methods

We compare our method with four representative NEP instruction-tuning baselines (Wang et al., 2025a): Supervised Fine-Tuning (Ouyang et al., 2022) (SFT), Critique Fine-Tuning (Wang et al., 2025c) (CFT), Distillation Tuning (Ho et al., 2023) (Distill), and Mix Tuning (Mix). SFT optimizes the model to generate ground-truth future captions with cross-entropy loss, CFT incorporates external LLM-generated critiques as additional textual feedback, Distill trains the model to imitate teacher-generated reasoning traces and final predictions, and Mix combines these supervision sources within a unified training set. While these baselines differ in their training signals, they remain text-based supervision methods. Our method instead augments caption-based training with future visual-semantic representation prediction, using the paired future video segment as an additional embedding-level supervision target.

4.4 Experiment Results

4.4.1 Results on Future Prediction and Temporal Understanding

Table 1 reports the results on future prediction and temporal understanding benchmarks. On the future prediction task, DOPA achieves a large improvement on FutureBench(+12.6%p) and is the only method that improves performance on MMworld$_{fp}$(+1.5%p) over the base model. This suggests that adding visual-semantic embedding supervision improves future prediction ability and provides better generalization beyond caption-only training.

On temporal understanding benchmarks, DOPA achieves a consistent though modest improvement over the base model and most caption-based baselines (TU-AVG +1.6%p). While the gain is smaller than on future prediction, it is positive across all four benchmarks, suggesting that visual-semantic embedding supervision provides a complementary signal that also benefits general temporal understanding, rather than improving future-caption generation alone. Overall, DOPA improves future prediction and yields a smaller but consistent gain in broader temporal understanding .

We further evaluate general video QA in Appendix A.1 and find that DOPA remains competitive on general video understanding. Qualitative comparisons on future prediction benchmarks in Appendix A.2 further corroborate these results: caption-based baselines often remain anchored to actions already completed in the observed video, whereas DOPA more accurately anticipates the actual future transition.

4.4.2 Ablation Study on Training Objectives

Table 2 shows the ablation study on the training objectives. Starting from the caption-based supervised fine-tuning loss $\mathcal{L}_{\text{SFT}}$, we progressively add the Future Embedding Alignment loss $\mathcal{L}_{\text{FEA}}$ and the Past–Future Ranking loss $\mathcal{L}_{\text{PFR}}$ to evaluate the contribution of each objective.

Adding $\mathcal{L}_{\text{FEA}}$ leads to clear improvements over text-only supervised fine-tuning on both future prediction and temporal understanding tasks. In particular, the FP-AVG score improves from 69.8 to 74.0, and the TU-AVG score also increases from 43.8 to 45.2. This indicates that predicting visual-semantic embeddings provides complementary supervision to caption-based training, helping the model capture visual and temporal cues that may be underspecified in text.

Further adding $\mathcal{L}_{\text{PFR}}$ improves the future prediction performance from 74.0 to 75.5 in FP-AVG. This suggests that explicitly distinguishing the actual future from the observed past is beneficial

Table 2: Ablation on loss components. We add each objective step by step to evaluate its contribution to future prediction and temporal understanding. Using all three losses yields the best overall performance.

Loss Type			Future Prediction Task			Temporal Understanding Task				
L_{SFT}	L_{FEA}	L_{PFR}	FutureBench	MMworld$_{fp}$	FP-AVG.	TB.	TC.	Seed-R1	TOMATO	TU-AVG.
✓			60.0	79.6	69.8	40.9	70.0	**38.2**	26.2	43.8
✓	✓		64.1	83.9	74.0	**41.7**	73.0	37.3	**28.7**	45.2
✓	✓	✓	**65.5**	**85.4**	**75.5**	41.0	**73.7**	37.9	**28.7**	**45.3**

Table 3: Ablation on the PFR margin m. We vary the margin value to evaluate its effect on future prediction performance.

margin m	Future Prediction Task		
	FutureBench	MMworld$_{fp}$	FP-AVG.
0	64.1	**85.4**	74.7
0.2	**65.5**	**85.4**	**75.5**
0.4	65.1	83.9	74.5

Table 4: Ablation on the target hidden representation for future embedding prediction.

Target Position	Future Prediction Task		
	FutureBench	MMworld$_{fp}$	FP-AVG.
Create New Token	64.1	84.6	74.3
Use Last Token	64.5	83.9	74.2
Use Generation Prompt Token	**65.5**	**85.4**	**75.5**

for future-oriented reasoning. By treating the observed past as a hard negative, $\mathcal{L}_{\text{PFR}}$ encourages the model to learn more discriminative representations between visually similar but temporally different states. Using all three losses yields the best overall performance.

4.4.3 Sensitivity Analysis on the Ranking Margin

Table 3 shows the effect of the margin m in the PFR loss. The model achieves the best average performance when $m = 0.2$, indicating that a moderate margin is beneficial for distinguishing the actual future from the observed past.

Without a margin ($m = 0$), the model is only encouraged to rank the future above the past, but not to separate them sufficiently. In contrast, an overly large margin ($m = 0.4$) slightly reduces performance, likely because the separation constraint becomes too strict for visually similar past–future pairs. Based on this result, we set $m = 0.2$ as the default value.

4.4.4 Ablation on Hidden State Selection

Table 4 compares different target positions for extracting the hidden representation $\mathbf{h}_*$ used by the prediction head. We consider three choices: (i) a newly introduced prediction token (‘<pred_token>’), placed at the beginning of the target response sequence; (ii) the last input token, corresponding to the final position of the input sequence; and (iii) the generation prompt token (‘<|im_start|>assistant\n’), located immediately before the assistant response begins.

As shown in Table 4, the generation prompt token achieves the highest FP-AVG (75.5), with a modest margin over the newly introduced prediction token and the last input token, which perform comparably (74.3 vs. 74.2). This may suggest that the hidden state at the beginning of response generation tends to provide a more useful summary of the video-question context for visual-semantic prediction. Based on this observation, we use the generation prompt token as the default target position.

5 Conclusion

We presented DOPA, a Dual-Objective Predictive Architecture that reframes next-event prediction in MLLMs as both a caption-generation and a future visual-state prediction problem. By jointly training caption supervision with two embedding-prediction objectives, Future Embedding Alignment and Past–Future Ranking, DOPA learns future-oriented visual dynamics beyond what is expressible in text. Across multiple benchmarks, it consistently improves future prediction and temporal understanding while remaining competitive on general video understanding.

More broadly, our findings suggest that future prediction in MLLMs can benefit from moving beyond purely language-based supervision toward predictive representation learning grounded in future visual dynamics. We hope DOPA offers a practical step toward integrating world-model-inspired predictive learning into multimodal language models for more robust anticipation and temporal reasoning.

6 Limitations

DOPA relies on a frozen multimodal embedding model ϕ to provide visual supervision targets. Because the predicted future embedding is trained to match $\phi(V_f)$, the supervision signal is bounded by the representational quality of ϕ. Visual details or temporal dynamics that are not encoded by ϕ cannot be fully transferred to the MLLM, and the achievable improvement may therefore depend on the choice of embedding model. Future work could explore stronger or more specialized embedding models, as well as supervision from multiple complementary embedding spaces.

References

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. 2023. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew 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, and 10 others. 2025. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. 2024. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*.

Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal Behbahani, Stephanie Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, and 6 others. 2024. Genie: Generative interactive environments. *Preprint*, arXiv:2402.15391.

Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, Yao Dou, Jaden Park, Jianfeng Gao, Yong Jae Lee, and Jianwei Yang. 2024. Temporalbench: Benchmarking fine-grained temporal understanding for multimodal video models. *arXiv preprint arXiv:2410.10818*.

Yuxuan Chen, Zhang Ge, Yixuan Liu, Yixiao Zhang, Weiyun Wang, Zhaoyang Liu, Yichen Liu, Bo Li, Junnan Li, and Wenhai Wang. 2025. Exploring the effect of reinforcement learning on video understanding: Insights from seed-bench-r1. *arXiv preprint arXiv:2503.24376*.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. 2024. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*.

Andy Clark. 2013. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3):181–204.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Rongrong Ji, and Xing Sun. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Daya Guo, Dejian Yang, Haowei Zhang, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Xuehai He, Yongqi Zong, Xinyuan Li, Jing Ma, Yujie Zhang, Guohai Li, Jing Xue, Jingjing Pan, Yali Zhao, Ziwei Liu, Yu-Gang Wang, Yu Qiao, Kai Zhang, and Limin Wang. 2024. Mmworld: Towards multi-discipline multi-faceted world model evaluation in videos. *arXiv preprint arXiv:2406.08407*.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2023. Large language models are reasoning teachers. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 14852–14882.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3.

Woojeong Jin, Rahul Khanna, Suji Kim, Dong-Ho Lee, Fred Morstatter, Aram Galstyan, and Xiang Ren. 2021. Forecastqa: A question answering challenge for event forecasting with temporal text data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

Ezra Karger, Houtan Bastani, Yueh-Han Chen, Zachary Jacobs, Danny Halawi, Fred Zhang, and Philip Tetlock. 2025. Forecastbench: A dynamic benchmark of ai forecasting capabilities. In *The Thirteenth International Conference on Learning Representations*.

Ching-Hao Lai and Hong-Shuo Wang. 2025. Vfe: A large-scale video future event description dataset for

evaluating video temporal prediction. *Applied Intelligence*.

Yann LeCun. 2022. A path towards autonomous machine intelligence. *OpenReview*.

Jie Lei, Licheng Yu, Tamara L. Berg, and Mohit Bansal. 2020. What is more likely to happen next? video-and-language future event prediction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. 2024. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, and 1 others. 2026. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*.

Han Lin, Tushar Nagarajan, Nicolas Ballas, Mido Assran, Mojtaba Komeili, Mohit Bansal, and Koustuv Sinha. 2025. Vedit: Latent prediction architecture for procedural video representation learning. In *The Thirteenth International Conference on Learning Representations*.

Yuanxin Liu, Shicheng Li, Yi Chen, Yuxuan Wang, Shuhuai Ren, Yu Li, Yang Chen, Mingjie Zhang, Xiaoxue Zhang, Ke Li, Dit-Yan Yang, and Limin Wang. 2024. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*.

Shangchen Miao, Ningya Feng, Jialong Wu, Ye Lin, Xu He, Dong Li, and Mingsheng Long. 2026. Jepa-vla: Video predictive embedding is needed for vla models. *arXiv preprint arXiv:2602.11832*.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

NVIDIA, :, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, Jiaojiao Fan, Michele Fenzi, Francesco Ferroni, Sanja Fidler, Dieter Fox, Songwei Ge, and 60 others. 2025. Cosmos world foundation model platform for physical ai. *Preprint*, arXiv:2501.03575.

OpenAI. 2024. Learning to reason with llms. https://openai.com/index/learning-to-reason-with-llms/.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Yan Ding, Zhigang Wang, JiaYuan Gu, Bin Zhao, Dong Wang, and Xuelong Li. 2025. Spatialvla: Exploring spatial representations for visual-language-action model. *Preprint*, arXiv:2501.15830.

Qwen Team. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Rajesh P. N. Rao and Dana H. Ballard. 1999. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2:79–87.

Hyeongseop Rha, Jeong Hun Yeo, Yeonju Kim, and Yong Man Ro. 2026. Emotion-coherent reasoning for multimodal llms via emotional rationale verifier. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 2029–2037.

Ziyao Shangguan, Chuhan Li, Yuxuan Ding, Yanan Zheng, Yilun Zhao, Tesca Fitzgerald, and Arman Cohan. 2025. Tomato: Assessing visual temporal reasoning capabilities in multimodal foundation models. In *International Conference on Learning Representations*, volume 2025, pages 7593–7734.

Jingwen Sun, Wenyao Zhang, Zekun Qi, Shaojie Ren, Zezhi Liu, Hanxin Zhu, Guangzhong Sun, Xin Jin, and Zhibo Chen. 2026. Vla-jepa: Enhancing vision-language-action model with latent world model. *arXiv preprint arXiv:2602.10098*.

Haonan Wang, Hongfu Liu, Xiangyan Liu, Chao Du, Kenji Kawaguchi, Ye Wang, and Tianyu Pang. 2025a. Fostering video reasoning via next-event prediction. *arXiv preprint arXiv:2505.22457*.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, and 56 others. 2025b. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *Preprint*, arXiv:2508.18265.

Yubo Wang, Xiang Yue, and Wenhu Chen. 2025c. Critique fine-tuning: Learning to critique is more effective than learning to imitate. *arXiv preprint arXiv:2501.17703*.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. 2024. Longvideobench: A benchmark for long-context interleaved video-language understanding. In *Advances in Neural Information Processing Systems*.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. 2025. Llava-cot: Let vision language models reason step-by-step. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2087–2098.

Zhiyuan Zeng, Jiashuo Liu, Siyuan Chen, Tianci He, Yali Liao, Jinpeng Wang, Zaiyuan Wang, Yang Yang, Lingyue Yin, Mingren Yin, Zhenwei Zhu, Tianle Cai, Zehui Chen, Jiecao Chen, Yantao Du, Xiang Gao, Jiacheng Guo, Liang Hu, Jianpeng Jiao, and 11 others. 2026. Futurex: An advanced live benchmark for llm agents in future prediction. In *The Fourteenth International Conference on Learning Representations*.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2024a. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2024b. Multimodal chain-of-thought reasoning in language models. *Preprint*, arXiv:2302.00923.

Andy Zou, Tristan Xiao, Ryan Jia, Joe Kwon, Mantas Mazeika, Richard Li, Dawn Song, Jacob Steinhardt, Owain Evans, and Dan Hendrycks. 2022. Forecasting future world events with neural networks. In *Advances in Neural Information Processing Systems*.

A Appendix

A.1 Results on General Video Question Answering

We evaluate on VideoMME (Fu et al., 2025), MVBench (Li et al., 2024), and LongVideoBench (Wu et al., 2024) to examine whether DOPA preserves general video reasoning ability. These benchmarks assess general video understanding across diverse video QA settings, rather than focusing on future prediction or temporal reasoning.

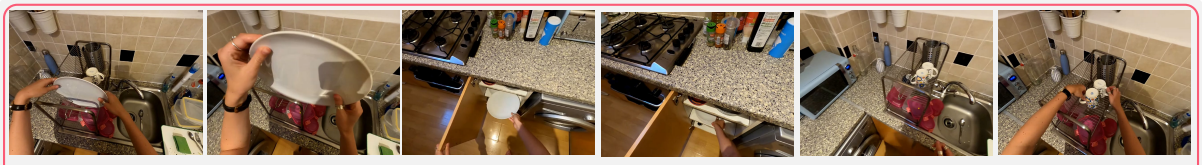
As shown in Table 5, most training methods show a slight performance drop compared to the base model. This suggests that training on a task-specific future prediction dataset can mildly affect general video QA performance. Nevertheless, DOPA maintains competitive results, achieving a G-AVG of 60.0% compared to 60.5% for the base model. This indicates that adding visual-semantic embedding supervision does not substantially compromise general video understanding, while improving future prediction and temporal understanding in the main evaluations.

A.2 Qualitative Results

We present qualitative examples in Figure 4 and Figure 5, showing that DOPA is not merely improved at generating plausible future guesses, but better captures visually grounded future events.

Table 5: Comparison of MLLMs on general video question answering benchmarks. All results are reported in accuracy (%).

Model	General VQA Task			
	VideoMME	MVBench	LongVid.	G-AVG.
Qwen2.5-VL-7B	**60.5**	65.2	**55.9**	**60.5**
Training Methods				
SFT	57.9	64.0	49.7	57.2
CFT	58.4	63.9	52.8	58.4
Distill	59.9	66.2	54.6	60.2
Mix	58.3	65.3	53.3	59.0
DOPA (Ours)	60.0	**66.4**	53.7	60.0



Video caption: A person is taking a white plate from the rack and putting it inside the counter cabinet.

Question: What might probably happen then?
A: The person might try to **take a white mug from the rack.**
B: The person might try to **take a white plate from the rack.**
C: The person might try to take a white mug from the counter cabinet.
D: The person might try to **put a white plate inside the counter cabinet.**

Baseline: D
SFT: B
Ours: A

Figure 4: Qualitative result on $\text{MMworld}_{\text{fp}}$. Although the plate-placement action has already been completed in the observed video, the baseline and SFT remain focused on the previously observed plate-related action. In contrast, DOPA captures the hand reaching toward the mug rack and correctly predicts the next plausible action.

Video caption: The chef is explaining the dish while demonstrating the cooking process.

Question: Based on the given video, complete the future sequence by selecting the pair of intermediate events that best bridges the observed scenes to the final scene:

1. [?]

2. [?]

3. In the final scene, two white bowls of pasta are garnished by chefs with a blue finishing tool and a drizzle of green sauce, and the plated pasta is placed on a counter ready for serving.

A: 1. A chef rapidly rolls out dough and cuts it into shapes using a knife;
2. Chefs exchange ingredients across the kitchen countertop before switching to dough feeding.
B: 1. Chefs feed the dough directly into a pasta machine without further preparation;
2. The pasta is immediately tossed with herbs in a bowl.
C: 1. A chef demonstrates the **dough-cutting technique by first slicing the dough into precise shapes**;
2. The cut dough is run through the pasta machine and its output is briefly inspected by other chefs.
D: 1. A chef sautés fresh ingredients including green herbs and vegetables in a metal pan on **the stove while explaining the cooking process**;
2. A chef uses a metal pan to transfer the sautéed mixture into a white bowl and then applies final plating techniques, **setting the stage for the dish to be garnished in the final scene**.

Baseline: C
SFT: C
Ours: D

Figure 5: Qualitative result on FutureBench. Although the pasta has already been cut in the observed video, the baseline and SFT remain focused on past dough-cutting steps, whereas DOPA predicts the later cooking and plating process leading to the final garnish.