DynaMind: Reasoning over Abstract Video Dynamics for Embodied Decision-Making

ICML 2025

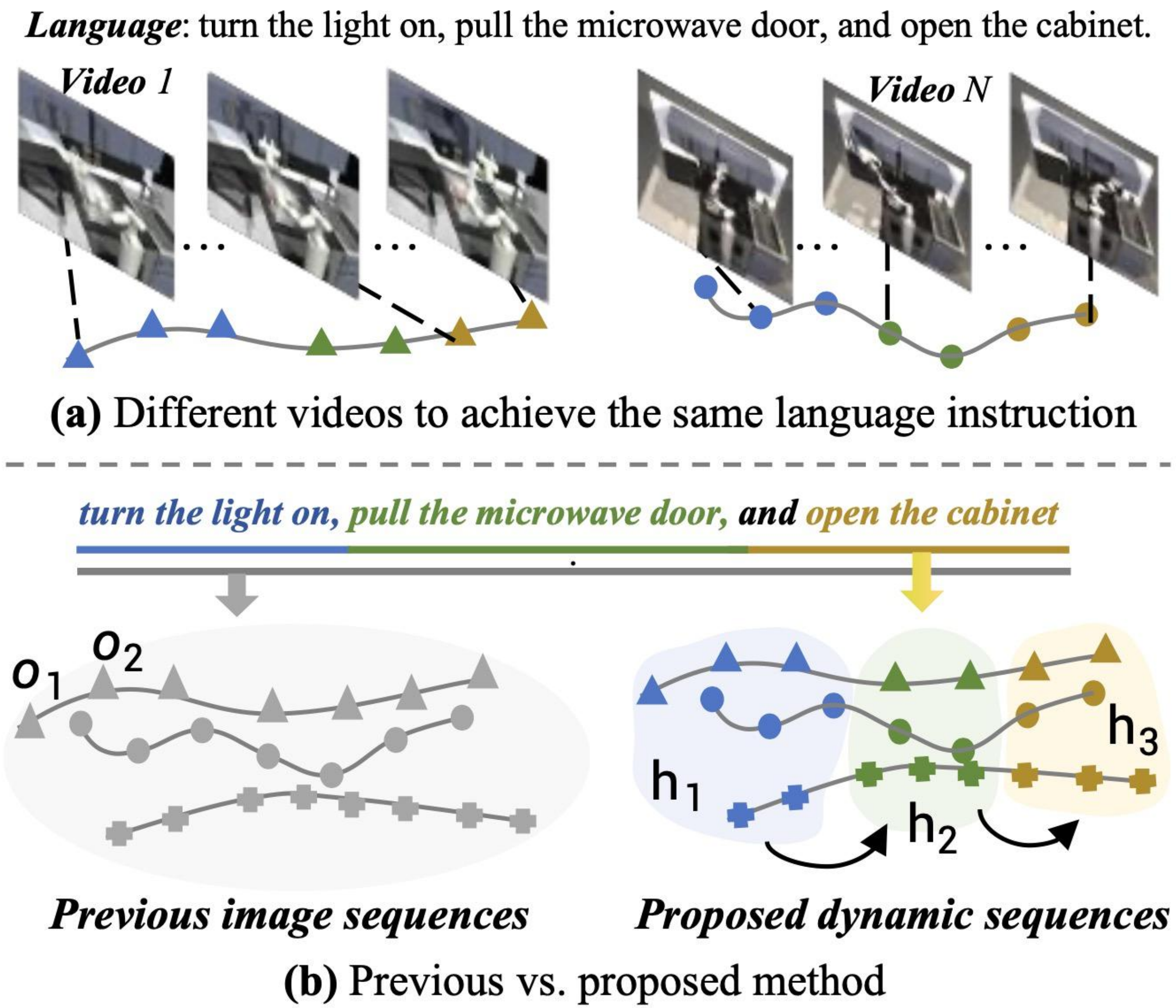
ICML
International Conference
On Machine Learning

Ziru Wang, Mengmeng Wang ✉, Jade Dai, Teli Ma, Guo-Jun Qi, Yong Liu, Guang Dai, Jingdong Wang

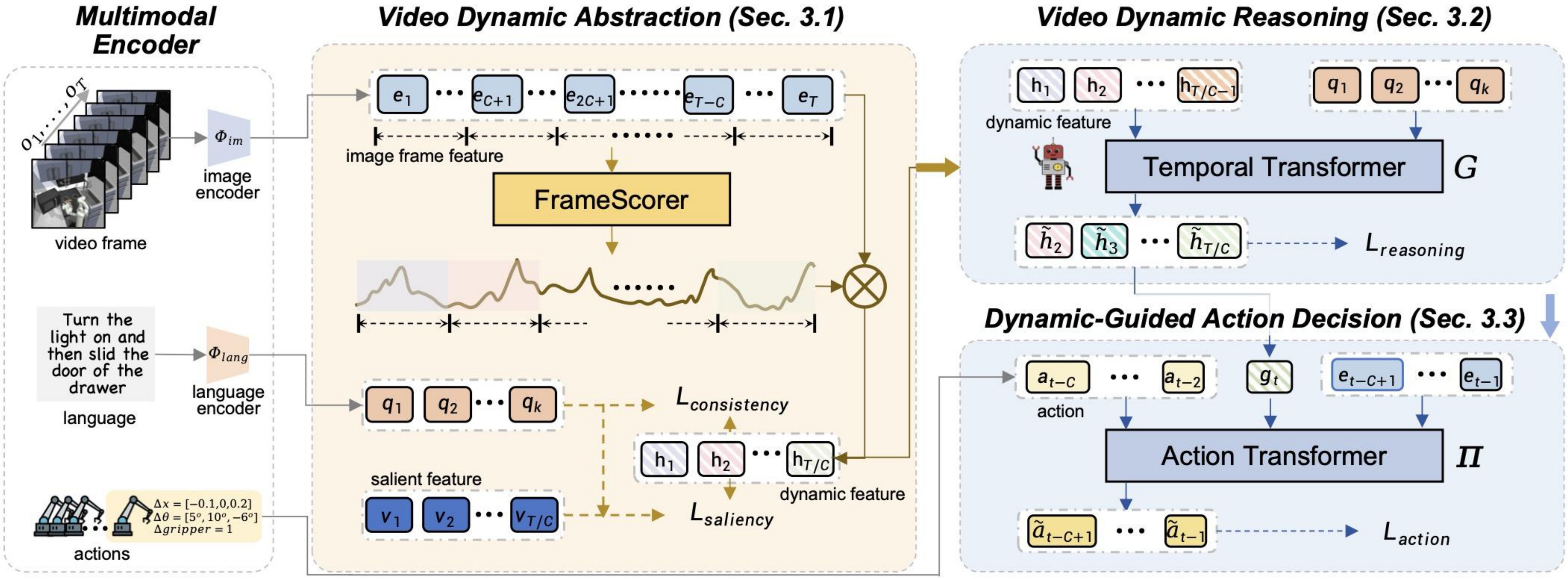
The mismatch between the simplicity and singularity of language and the diversity and complexity of videos

Goal of the work:

Bridge the gap between language instructions and video content for embodied agents, enabling more effective decision-making.



Overview framework of DynaMind



DynaMind consists of three core modules:

a) **Video Dynamic Abstraction** – transforms the input video into a compact dynamic representation.

b) **Video Dynamic Reasoning** – predicts the future evolution of the dynamics.

c) **Dynamic-Guided Action Decision** – uses the predicted dynamics to infer the corresponding action sequenceg.

To abstract a video into dynamic representations, we propose an adaptive **FrameScorer** that assigns importance scores based on semantic consistency and visual saliency.

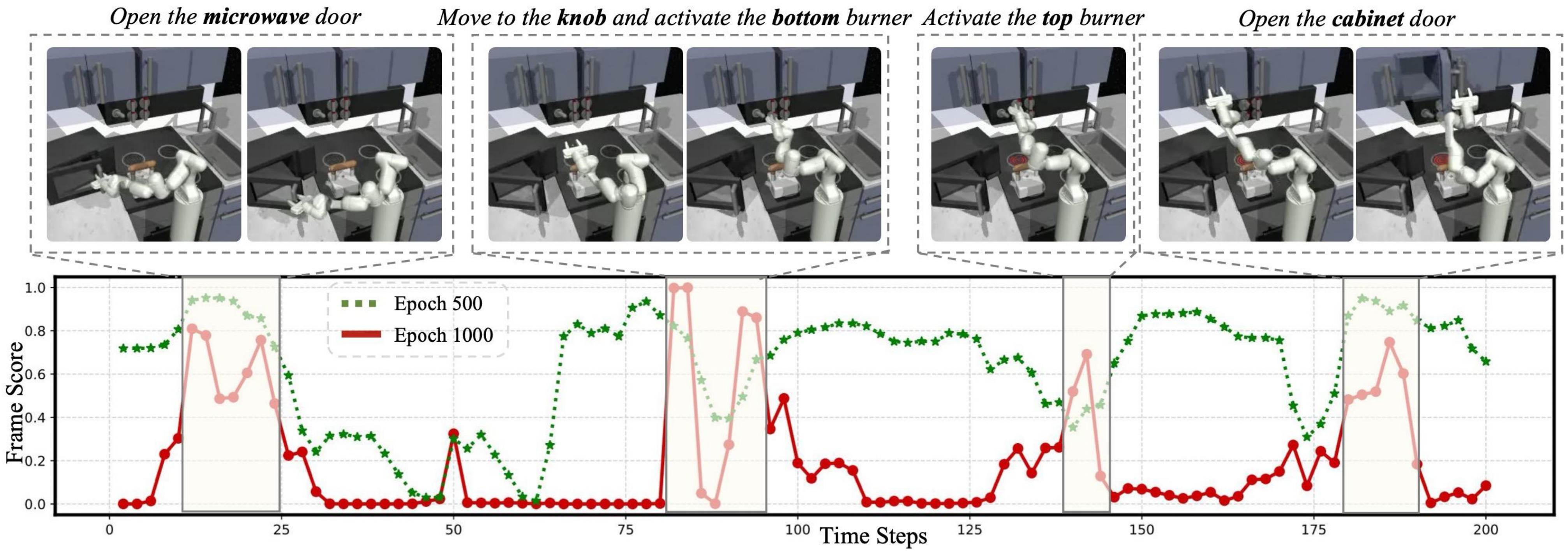


Figure 5. **Visualization of our method in adaptive scoring image frames.** The top row displays critical frames within an episode. The bottom row shows the importance score of the frame at each time step. This allows DynaMind to extract relevant information from the video while filtering out redundant content, effectively bridging the gap between complex video and concise language instructions.

Visualization: Abstracted dynamic representations convey key video information.

Performance comparison

Table 1. **Task-wise success rates on LOReL Sawyer.** DynaMind outperforms all other methods in terms of average performance. The results are calculated over 3 seeds. Best methods and those within 10% of the best are highlighted in bold.

Task	Random	Vanilla BC	RL	DT	LISA	SkillDiffuser	**DynaMind** (ours)
closer drawer	52%	50%	58%	10%	**100%**	**95%**	**100%**
open drawer	14%	0%	8%	60%	20%	55%	**80%**
turn faucet left	24%	12%	13%	0%	0%	**55%**	**57%**
turn faucet right	15%	**31%**	0%	0%	**30%**	**25%**	**26%**
move black mug right	12%	**73%**	0%	20%	60%	18%	39%
move while mug down	5%	6%	0%	0%	**30%**	10%	**20%**
Average over tasks	20%	29%	13%	15%	40%	43%	**53.67%**

Table 3. Performance on BabyAI.

Task	Vanilla BC	DT	LISA	DynaMind
GoToSeq	33.3%	49.3%	59.4%	**72.7%**
SynthSeq	12.9%	42.3%	46.3%	**50.7%**
BossLevel	20.7%	44.5%	49.1%	**52.3%**

Method	Success Rate
DT	28.63%
LISA	28.69%
GR-1	32.94%
MT-R3M	30.50%
DynaMind	**39.81%**

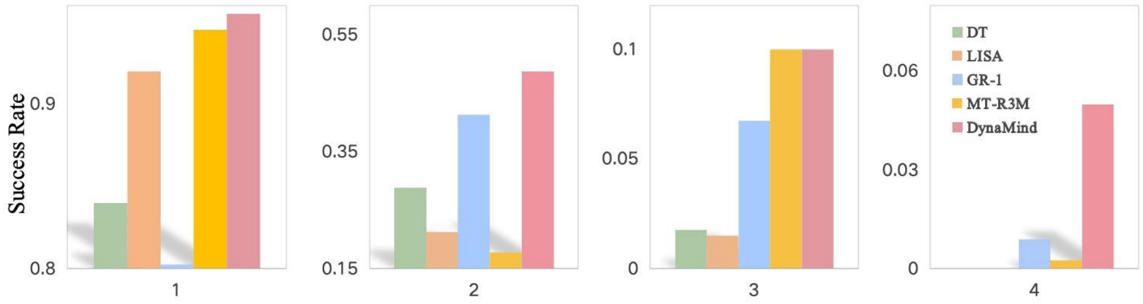


Figure 3. **Success rates on Franka Kitchen.** The four plots on the right illustrate the success rates of completing 1 to 4 subtasks within a single episode, while the left plot shows the average success rate across all tasks. The evaluation is repeated 100 times.

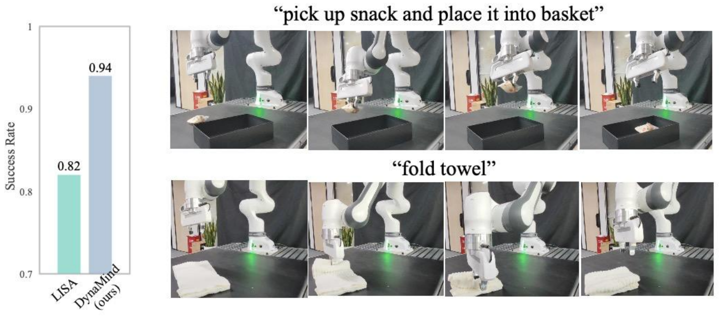


Figure 9. Left: **Success rate** averaged over 5 tasks. Right: **Qualitative results** of DynaMind for 2 tasks in real-world experiments. More results and details can be found in §F.

Ablation and efficiency experiments

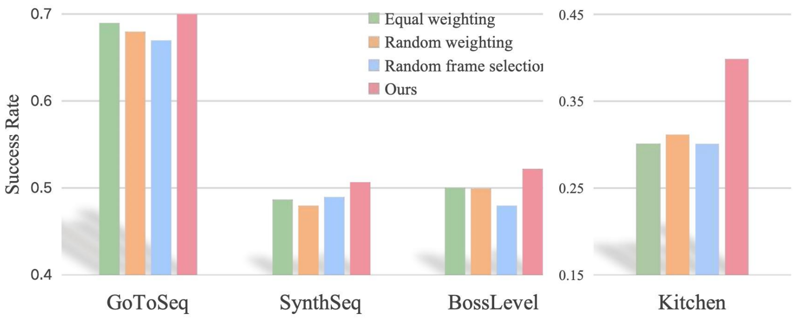


Figure 6. Ablation on dynamic abstraction.

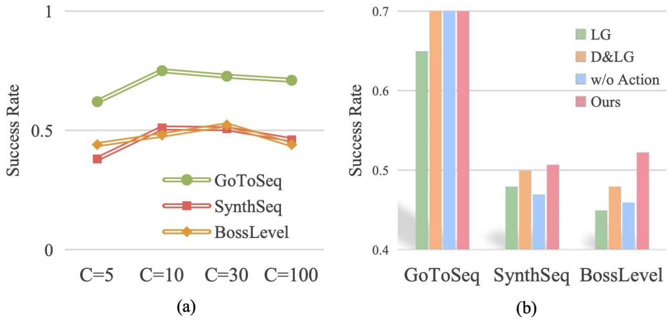


Figure 7. (a) Ablation on dynamic reasoning. (b) Ablation on dynamic-guided action decision.

Table 4. Comparison of training efficiency.

Method	Params(M)	GPU Memory(MiB)	Success Rate
LISA	7.52	690	40.0%
SkillDiffuser	60.29	1136	43.0%
DynaMind	7.84	854	53.7%

DynaMind capture the correlation between dynamics and language.

Method	LISA	DynaMind+LISA	DynaMind
Success Rate	28.69%	33.19%	**39.81%**

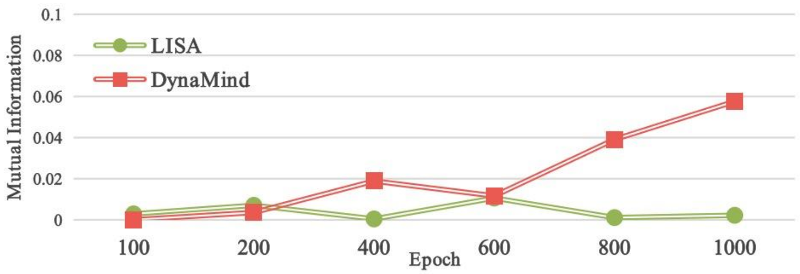


Figure 8. Top: Results of the combined method. Bottom: Mutual information over training.

The learned dynamic representations can be used to perform new tasks

Table 5. Performance on unseen tasks.

Unseen Task	DT	LISA	DynaMind
SynthSeq	31.0%	33.1%	**40.0%**
BossLevel	31.2%	32.4%	**35.7%**

Table 6. Performance on unseen compositional tasks on LOReL Sawyer.

Method	DT	LISA	SkillDiffuser	DynaMind
Success Rate	13.33%	20.89%	25.21%	**36.67%**

Conclusions

- We introduce the **DynaMind framework**, which abstracts video content into dynamic representations and aids decision-making through dynamic reasoning, thus reducing the mismatch between language and video.
- We design a **dynamic abstraction** module with an adaptive FrameScorer to convert video into compact, expressive dynamic sequences, followed by a **generation** module to generate future dynamics and a **decision** module to predicts appropriate actions.
- We empirically demonstrate DynaMind’s effectiveness and generalization capabilities across various simulation experiments, provide visualizations of abstract video dynamics, and confirm its effectiveness in real-world tasks.

Feel free to contact with alexw_rob@163.com.