Open Vocabulary Long-Term Action Anticipation

Syed Talal Wasim, Jinhui Yi, Hamid Suleman, Ahmad Javed, Yanan Luo, Muzammal Naseer, Juergen Gall

LAMARR INSTITUTE FOR MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE

UNIVERSITÄT BONN

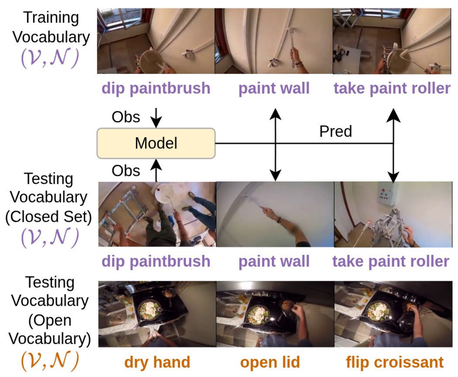
Khalifa University جامعة خليفة

SEEK WISDOM

ECCV 2026 | MALMÖ | SEPT 8–12

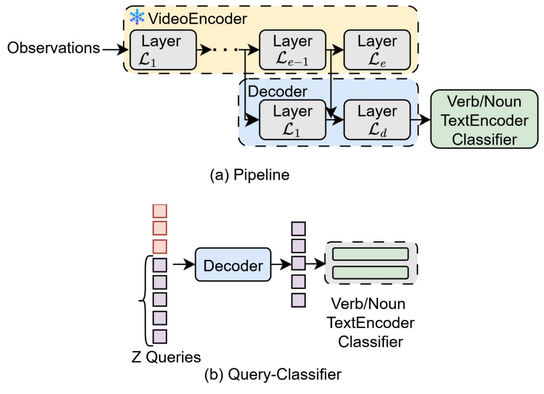
Motivation

Closed-set anticipation reuses the training vocabulary at test time; we test on entirely new actions from new datasets.

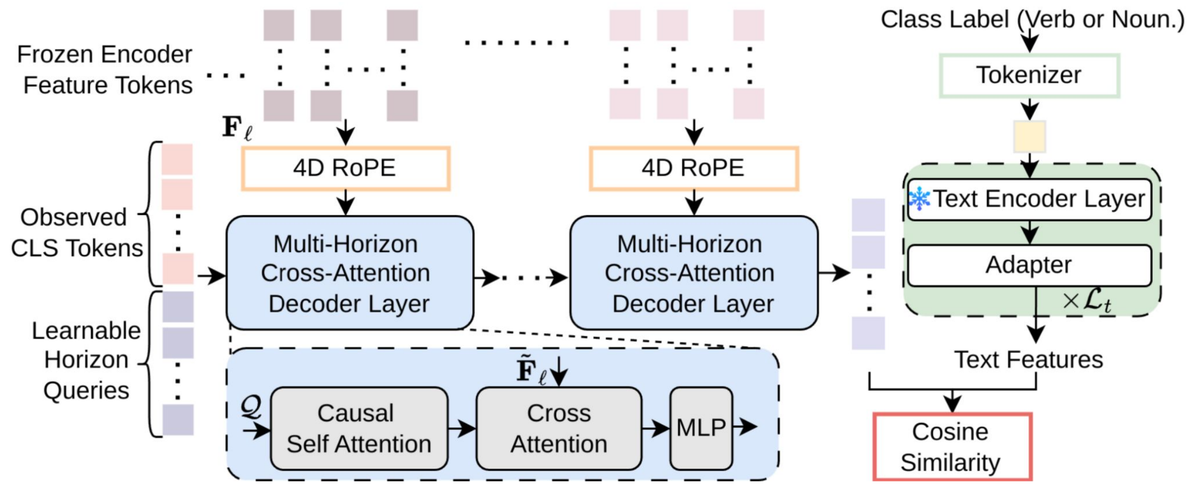


Contributions

1. First open-vocabulary anticipation protocol: train on Ego4D, test on EK100 and EGTEA+.
2. We adapt AntGPT, Video-LLaVA, XCLIP, and LaViLa to the task; all either overfit to Ego4D actions or predict incoherent sequences.
3. **MHCAD:** horizon-specific queries matched against adapted text encoders.
4. A single architecture handles both settings, with no separate heads or retraining per benchmark.
5. **Closed-set:** on par with models 12× larger.
6. **Open-vocabulary:** 0.728 verb error on EK100 vs. 0.849 for the best adapted baseline.



Main Architecture



1. **Frozen video encoder:** LaViLa-Large stays frozen; we keep dense spatial tokens from the last L_d layers, not just the CLS token, since fine-grained detail matters most for unseen actions.
2. **Horizon-specific queries:** Z learnable queries, one per future action, so each horizon aggregates its own visual evidence instead of sharing one pooled vector.
3. **Causal self-attention:** each query attends only to earlier ones, matching the fact that future actions unfold in order.
4. **Layer-matched cross-attention:** decoder layer m attends to encoder layer L_e - L_d + m, so shallow queries see low-level features and deep queries see semantic ones.
5. **4D RoPE:** positions are factored into segment, frame, height, and width, letting queries distinguish where and when a token came from.
6. **Adapted text encoders:** frozen verb and noun encoders with bottleneck adapters, scored against query embeddings by cosine similarity, so the vocabulary is swappable at test time.

Results

Method	Frozen Params	Trained Params	Ego4D v1 Val Verb ↓	Ego4D v1 Val Noun ↓	Ego4D v1 Test Verb ↓	Ego4D v1 Test Noun ↓	Ego4D v2 Test Verb ↓	Ego4D v2 Test Noun ↓
LLM/VLM Foundation Models with >7B parameters								
VideoLLM [5]	7.0B*	67M*	-	-	-	-	0.721	0.725
VideoLLaMA [45]	8.2B*	42M*	0.703	0.721	-	-	-	-
PlausiVL [26]	7.74B*	188M*	0.679	0.681	-	-	-	-
AntGPT [47]	13.3B*	246M*	0.700	0.717	0.658	0.655	0.650	0.650
PALM [19]	11.4B*	219M*	-	-	0.656	0.640	0.647	0.612
Video-LLaVA	303M	7.2B	0.697	0.711	-	-	-	-
Methods with <1B parameters								
RepLAI [27]	0	22M*	0.755	0.834	-	-	-	-
SlowFast [14]	63.4M	181M	0.745	0.779	-	-	0.717	0.736
VCLIP [7]	149M*	181M*	0.715	0.748	0.739	0.769	-	-
ICVAE [25]	63.4M*	55M*	-	-	0.741	0.740	-	-
HierVL [3]	0	192M	-	-	0.724	0.735	-	-
XCLIP-MultiHead [30]	575M	91M	0.736	0.761	-	-	-	-
LaViLa-MultiHead [48]	489M	91M	0.724	0.749	-	-	0.720	0.727
LaViLa-MHCAD (Ours)	489M	157M	**0.699**	**0.717**	**0.719**	**0.722**	**0.689**	**0.695**

Closed-set on Ego4D. Best among sub-1B methods, and on par with models using 7B or more parameters (greyed).

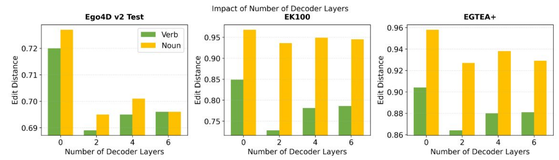
Method	Frozen Params	Trained Params	EK100 Verb ↓	EK100 Noun ↓	EK100 Action ↓	EGTEA+ Verb ↓	EGTEA+ Noun ↓	EGTEA+ Action ↓
LLM/VLM Foundation Models with >7B parameters								
AntGPT [47]	13.3B*	246M*	0.933	0.958	0.993	0.961	0.974	0.998
Video-LLaVA [23]	303M	7.2B	0.925	0.945	0.991	0.959	0.970	0.994
Methods with Base Transformer Backbone								
XCLIP-MultiHead [30]	195M	66M	0.915	0.983	-	0.940	0.978	-
LaViLa-MultiHead [48]	152M	66M	0.889	0.981	-	0.931	0.970	-
LaViLa-MHCAD (Ours)	152M	76M	**0.779**	**0.960**	-	**0.903**	**0.950**	-
Methods with Large Transformer Backbone								
XCLIP-MultiHead [30]	575M	91M	0.868	0.971	0.985	0.912	0.966	0.981
LaViLa-MultiHead [48]	489M	91M	0.849	0.968	0.982	0.904	0.958	0.979
LaViLa-MHCAD (Ours)	489M	157M	**0.728**	**0.936**	**0.969**	**0.864**	**0.927**	**0.968**

Open-vocabulary on EK100 and EGTEA+. Trained on Ego4D only, no fine-tuning. We beat every baseline at both backbone sizes.

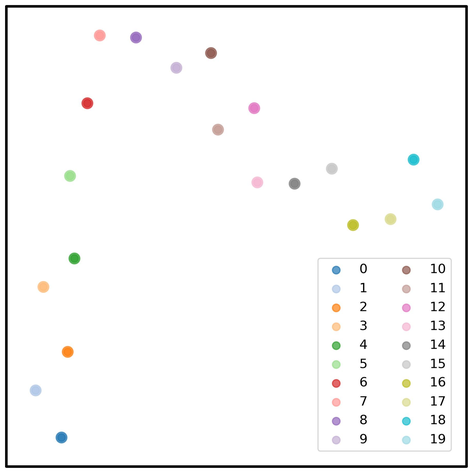
Ablations and Analysis

Num Queries	Head Style	Ego4D v2 Test Verb ↓	Ego4D v2 Test Noun ↓	EK100 Verb ↓	EK100 Noun ↓	EGTEA+ Verb ↓	EGTEA+ Noun ↓
Single	Per Horizon	0.719	0.727	0.759	0.992	0.901	0.970
20	Per Horizon	0.701	0.705	0.837	0.947	0.879	0.946
20	Shared	0.689	0.695	0.728	0.936	0.864	0.927

Horizon-specific queries with shared heads win. Separate heads hurt open-vocabulary transfer.



Two decoder layers are optimal. No decoder is much worse, and deeper degrades.



Cross-attention of the 20 horizon queries, t-SNE projected. Queries trace a path from near to distant future. This ordering is never supervised.

Qualitative Results