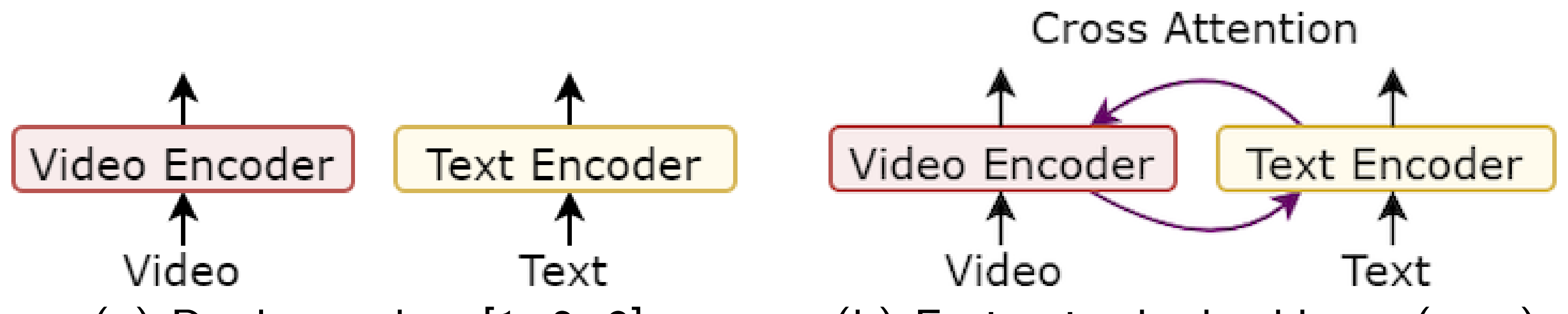
EgoVLPv2: Egocentric Video-Language Pre-training with Fusion in the Backbone

Shraman Pramanick[1,2] Yale Song[2] Sayan Nag[3] Kevin Qinghong Lin[4] Hardik Shah[2]
Mike Zheng Shou[4] Rama Chellappa[1] Pengchuan Zhang[2]
[1]Johns Hopkins University, [2]Meta AI, [3]University of Toronto, [4]National University of Singapore

Code and models are available.

Motivation

Existing egocentric VLP approaches [1, 2, 3] adopt dual encoders and perform late fusion. We aim to develop cross-modal fusion directly in the backbones while still flexibly supporting V/L/VL downstream tasks.

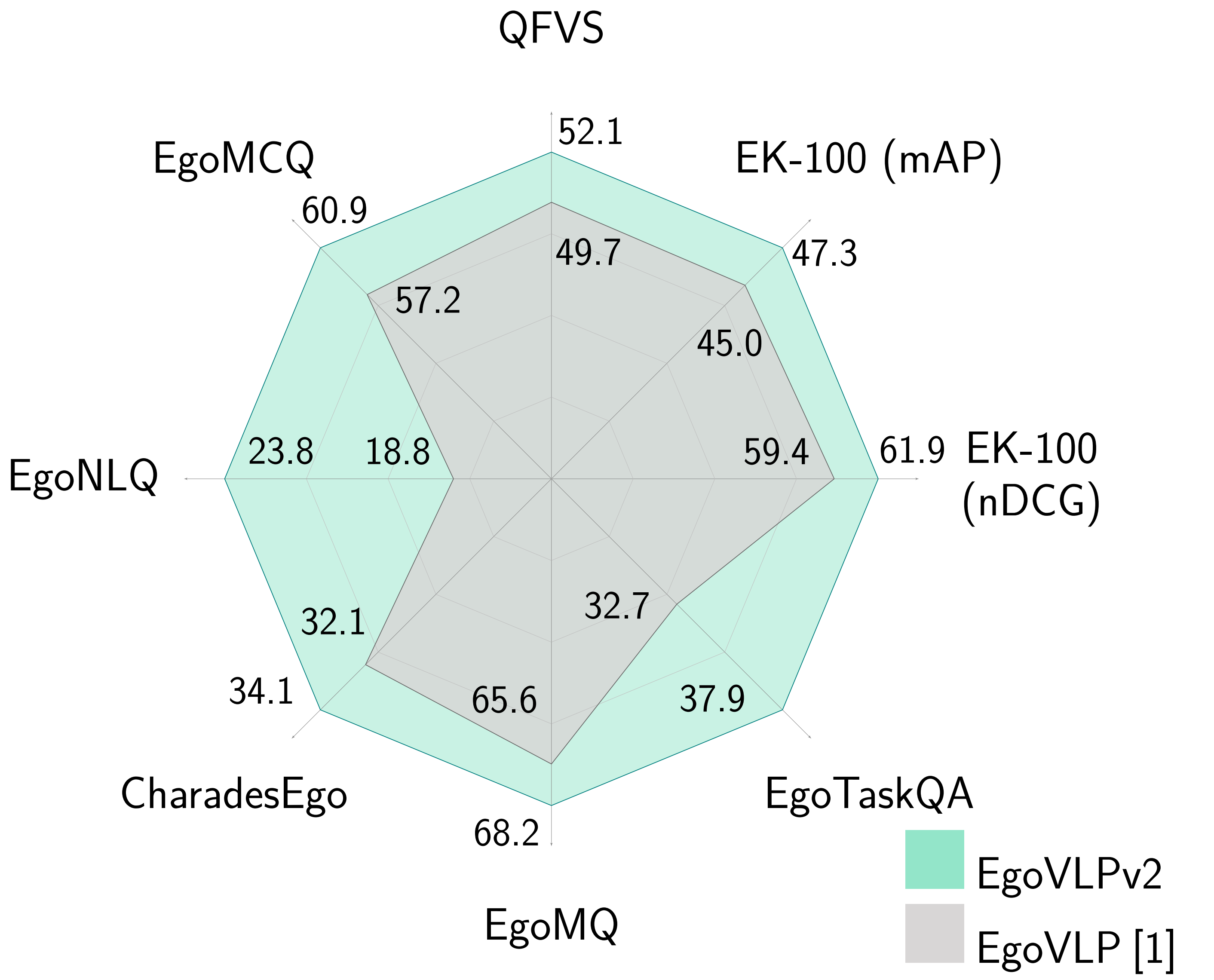


(a) Dual encoders [1, 2, 3]. (b) Fusion in the backbone (ours).

Primary Contributions

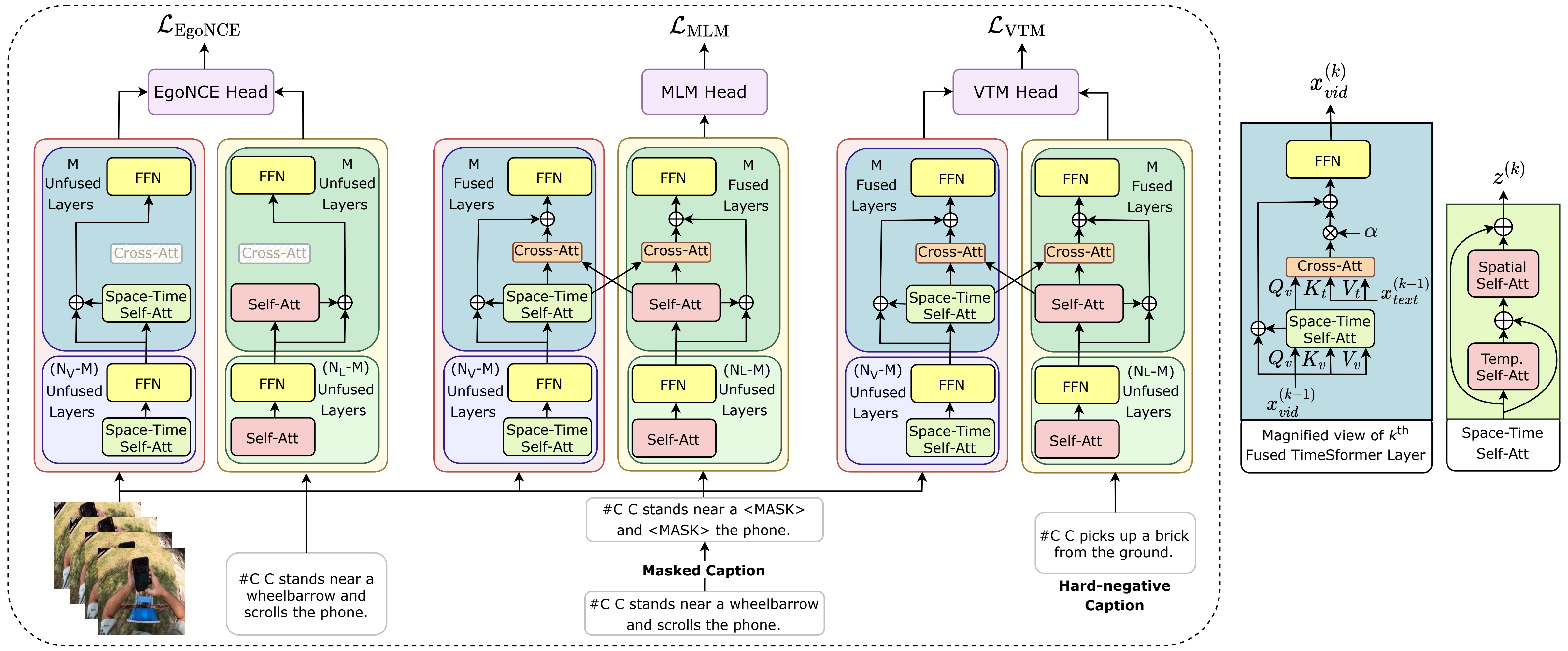
- **Flexible:** Gated cross-modal fusion in video and text backbones enables easy switch between dual and fusion encoders.
- **Efficient:** Requires 45% less compute (GMACs) than additional fusion-specific layers, and reduces fine-tuning cost compared to dual encoders.
- **Effective:** Strong results on various video- and video-text tasks.

Main Results



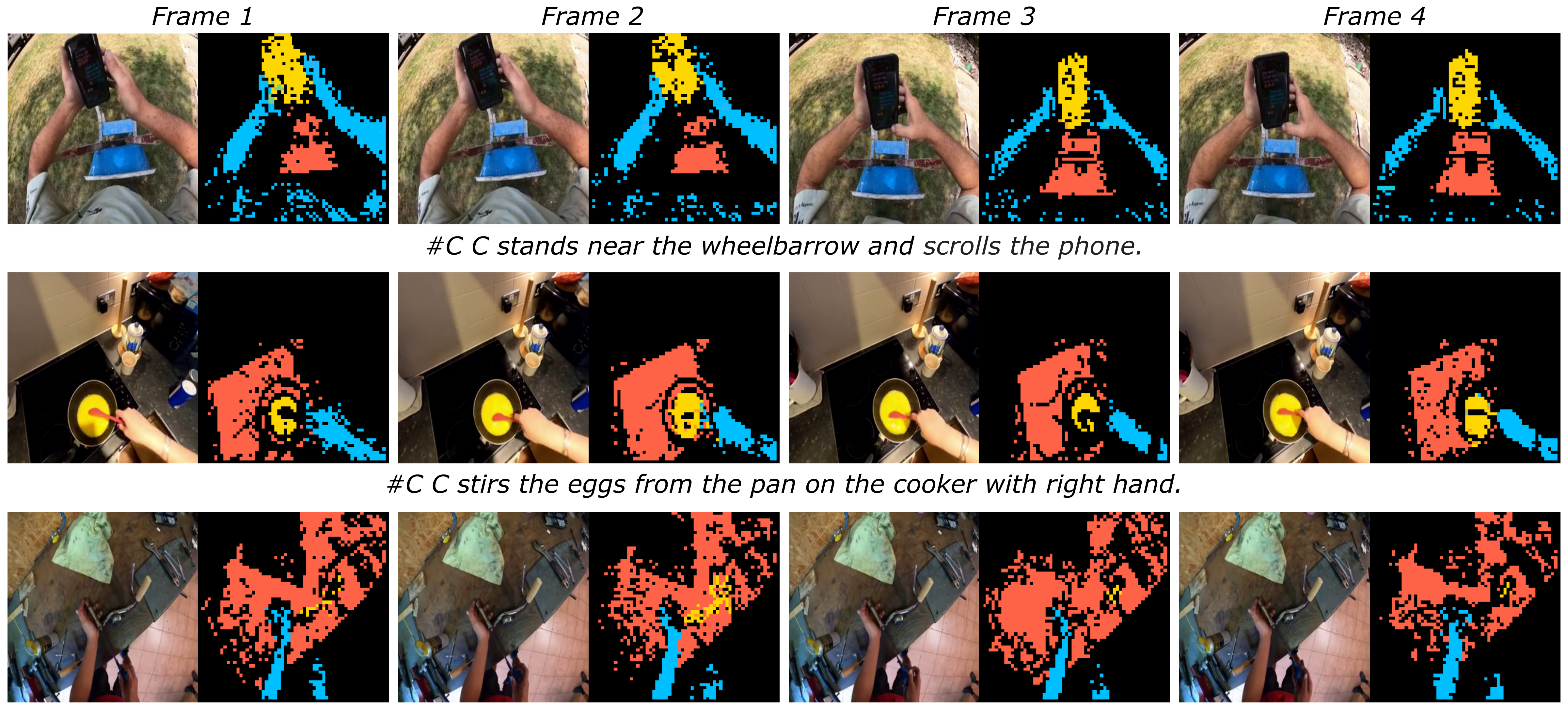
- EgoVLPv2 achieves the state-of-the-art performance across a broad range of egocentric video- and video-language tasks among similar-sized baselines [1, 2, 3] by incorporating cross-modal fusion in the backbones.

EgoVLPv2: Second Generation of Egocentric VLP


$\mathcal{L}_{\text{EgoNCE}}$
EgoNCE Head
$\mathcal{L}_{\text{MLM}}$
MLM Head
$\mathcal{L}_{\text{VTM}}$
VTM Head
M Unfused Layers
M Fused Layers
$(N_V$-M) Unfused Layers
$(N_L$-M) Unfused Layers
FFN
Cross-Att
Space-Time Self-Att
Self-Att
#C C stands near a wheelbarrow and scrolls the phone.
#C C stands near a <MASK> and <MASK> the phone.
Masked Caption
#C C stands near a wheelbarrow and scrolls the phone.
#C C picks up a brick from the ground.
Hard-negative Caption
$x_{vid}^{(k)}$
α
Q_v K_t V_t $x_{text}^{(k-1)}$
Q_v K_v V_v
$x_{vid}^{(k-1)}$
Magnified view of k^{th} Fused TimeSformer Layer
$z^{(k)}$
Spatial Self-Att
Temp. Self-Att
Space-Time Self-Att

- **EgoNCE** requires unfused video & text features. $\mathcal{L}_{\text{EgoNCE}}$ is computed with EgoVLPv2 acting as a dual encoder.
- **MLM & VTM** requires multi-modal representation. Cross-attention is switched on, $\mathcal{L}_{\text{MLM}}$ and $\mathcal{L}_{\text{VTM}}$ are computed with EgoVLPv2 acting as a fusion encoder.
- The three losses are added, $\mathcal{L}_{\text{total}} = (1-\gamma-\delta)\mathcal{L}_{\text{EgoNCE}} + \gamma\mathcal{L}_{\text{MLM}} + \delta\mathcal{L}_{\text{VTM}}$, and back-propagated end-to-end.

Cross Attention Visualization

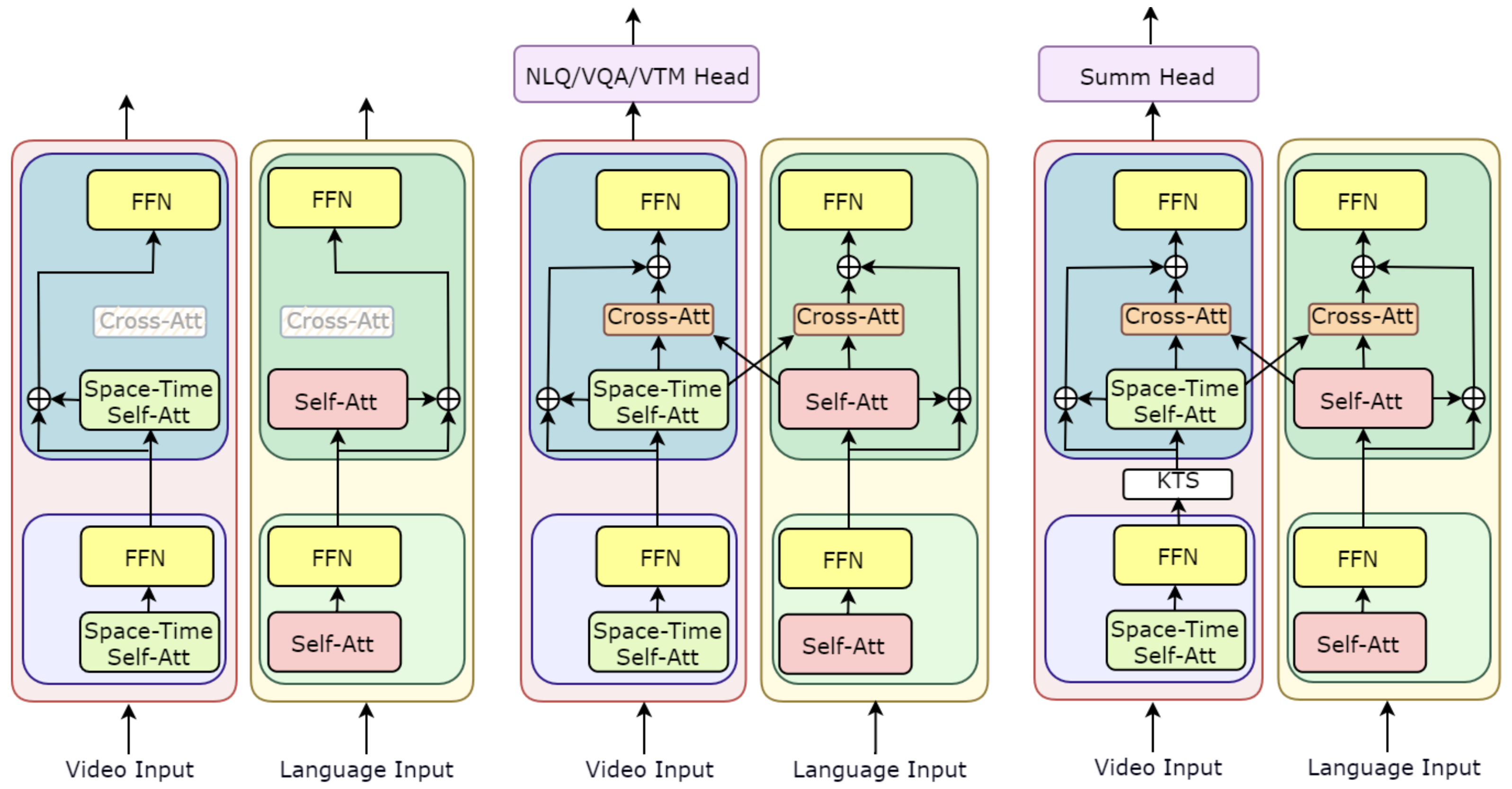


#C C stands near the wheelbarrow and scrolls the phone.

#C C stirs the eggs from the pan on the cooker with right hand.

#C C tightens the bolt on the bicycle handle on the table with the T-wrench in his right hand.

Adaptation to Downstream Tasks



(a) Retrieval using dual encoder. (b) VQA/NLQ/retrieval using fusion encoder. (c) QFVS using fusion encoder.

- EgoVLPv2 can be adapted to various dual- and fusion-encoder based video-language tasks, ranging from retrieval, grounding, video QA to query-focused video summarization.

Ablation on Pre-training Objectives

Pre-training Objectives				EgoMCQ (%)					
				Dual Enc.		Fusion Enc.		Ensemble	
EgoNCE	MLM	VTM	VTM-Hard	Inter	Intra	Inter	Intra	Inter	Intra
✓	–	–	–	89.5	52.6	–	–	–	–
✓	✓	–	–	89.6	52.4	–	–	–	–
✓	–	–	✓	89.6	53.4	**90.6**	59.1	**91.0**	60.0
✓	✓	✓	–	89.5	53.6	89.1	51.5	90.2	56.8
✓	✓	–	✓	**89.8**	**56.7**	**90.6**	**59.6**	**91.0**	**60.9**

- We evaluate on EgoMCQ using our model either as a dual encoder, as a fusion encoder, or an ensemble of both. Removing any objective leads to a performance drop.

References

1. Lin et al.; Egocentric Video-Language Pretraining. NeurIPS, 2022.
2. Zhao et al.; Learning Video Representations from Large Language Models. CVPR, 2023.
3. Ashutosh et al.; HierVL: Learning Hierarchical Video-Language Embeddings. CVPR, 2023.