NEURAL INFORMATION PROCESSING SYSTEMS

EVIDENT: Routing MLLM Adaptation through Entity-Grounded Visual Evidence for Cross-Domain Video Temporal Grounding





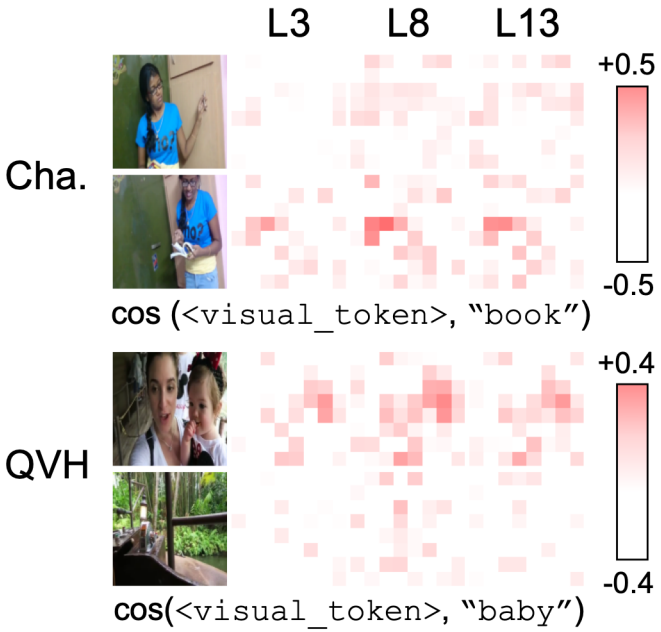


Geo Ahn[1*] Jiwook Han[1*] Youngrae Kim[2*] Joonseok Lee[3] Jinwoo Choi[1†]

[1] KHU VLLAB VISION & LEARNING LAB [2] USC [3] SEOUL NATIONAL UNIVERSITY

* Equal contribution † Corresponding author

The pretrained MLLM **already** knows what to look at.

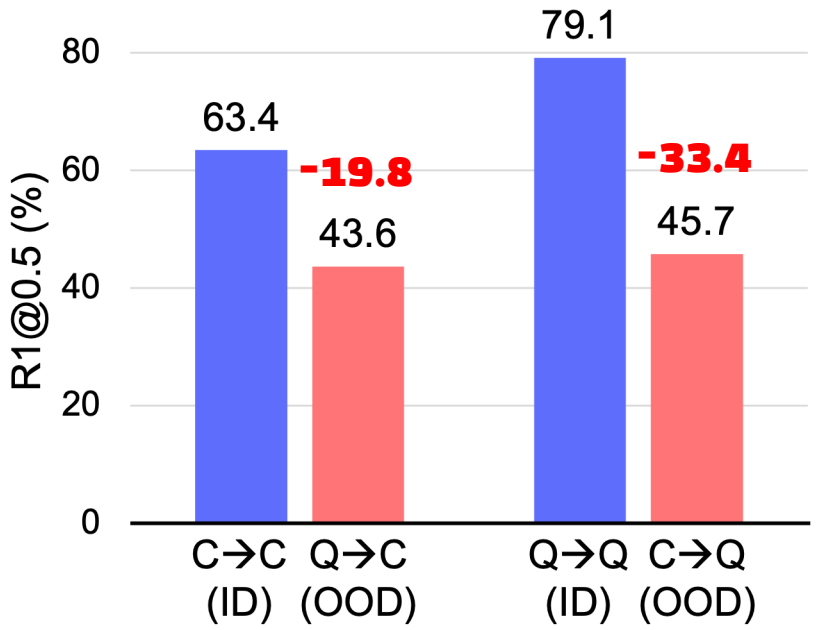


Entity alignment already generalizes.

book, baby

KHU VLLAB VISION & LEARNING LAB

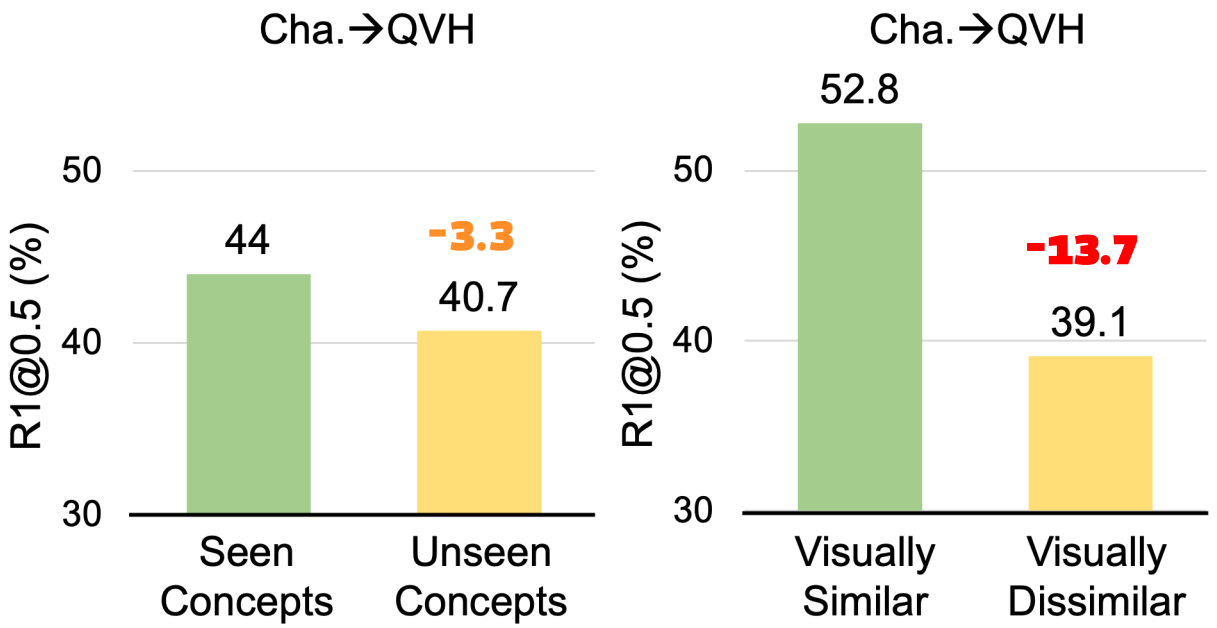
Fine-Tuning Struggles Across Video Datasets



C: Charades-STA
Q: QVHighlights

Performance drops when we test on a **different dataset**.

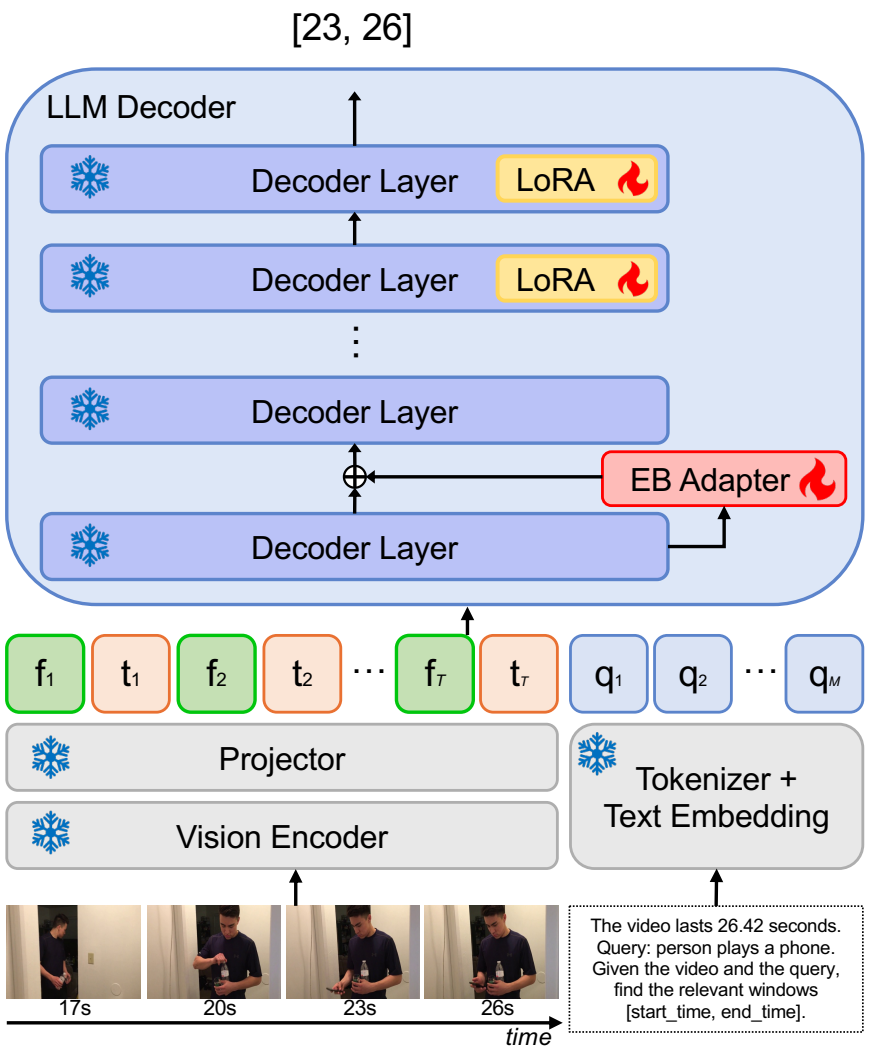
Visual Domain Shift Is the Bigger Challenge



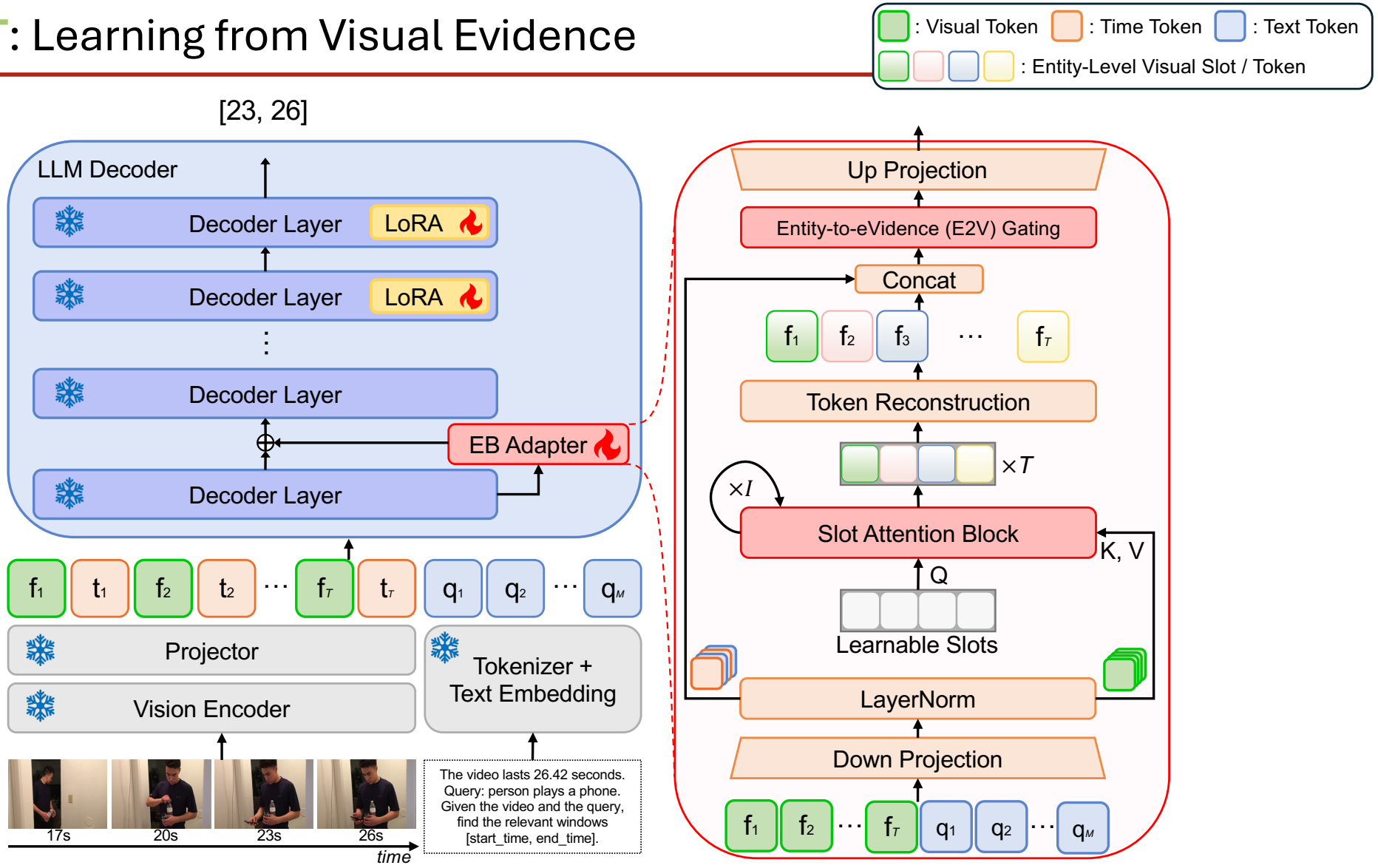
To improve OOD performance, VTG models should ground their predictions in **query-relevant visual evidence.**

KHU VLLAB VISION & LEARNING LAB

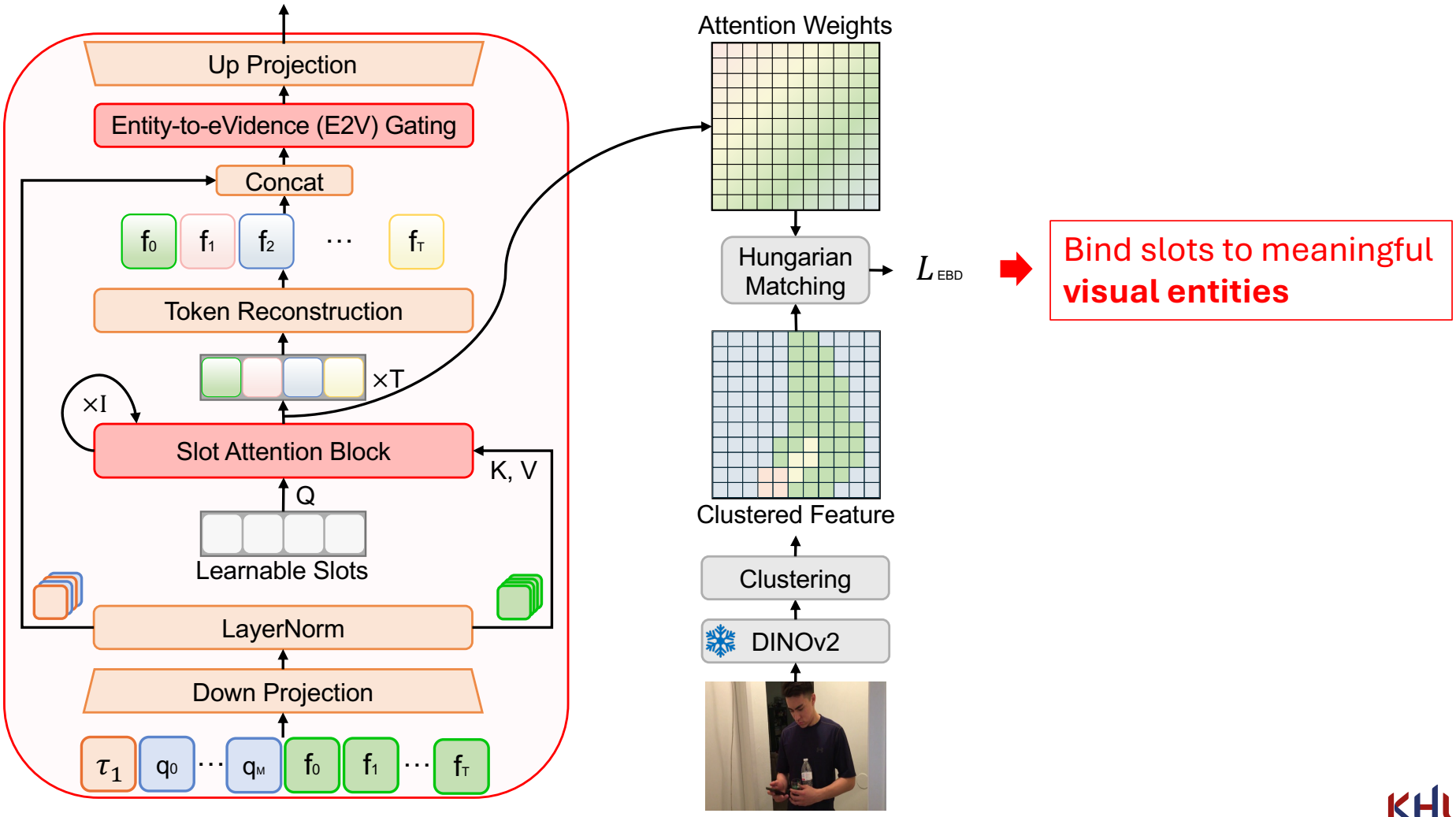
EVIDENT: Learning from Visual Evidence



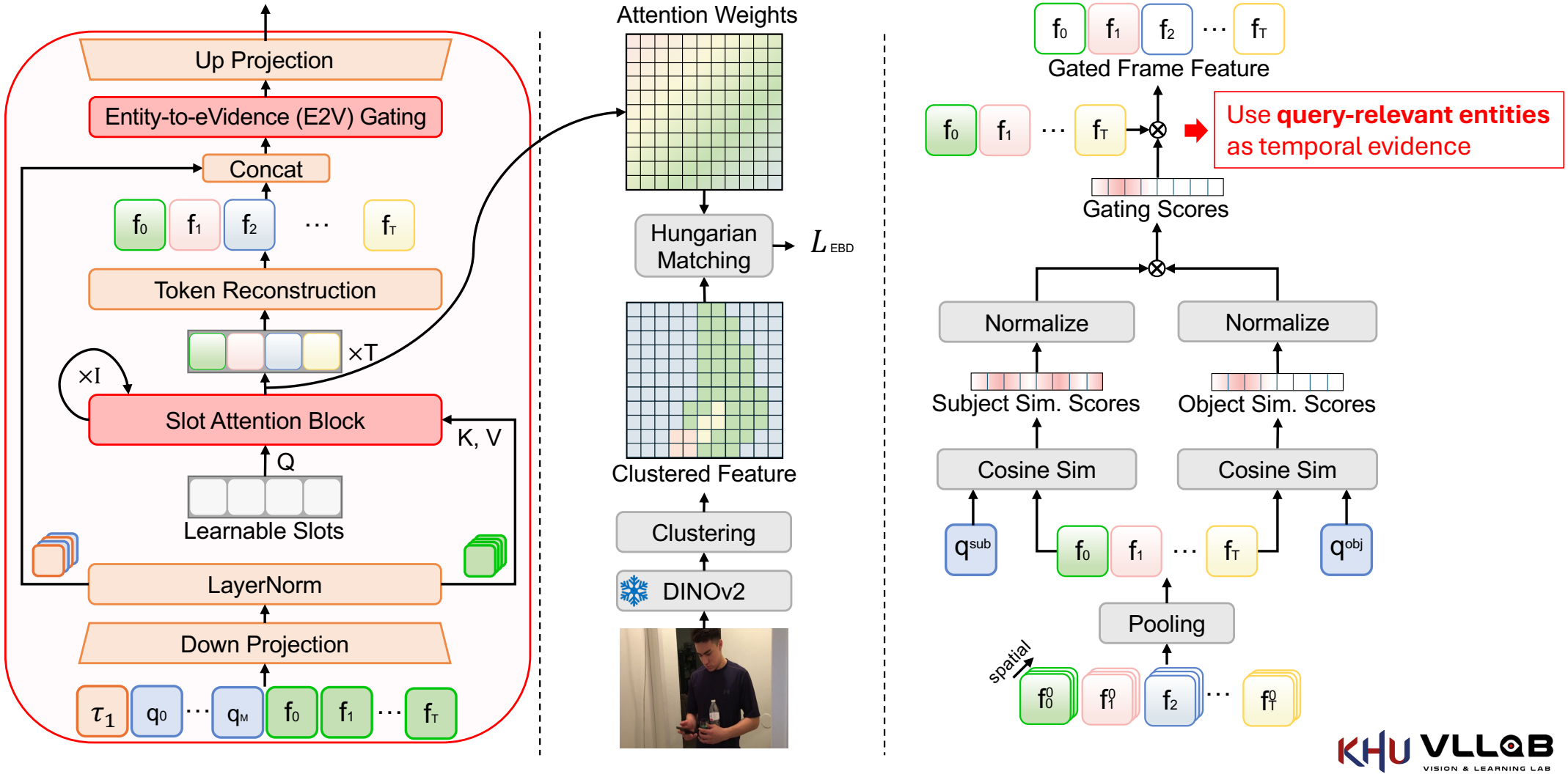
EVIDENT: Learning from Visual Evidence



EVIDENT: Learning from Visual Evidence



EVIDENT: Learning from Visual Evidence



EVIDENT: Generalizes across Video Datasets

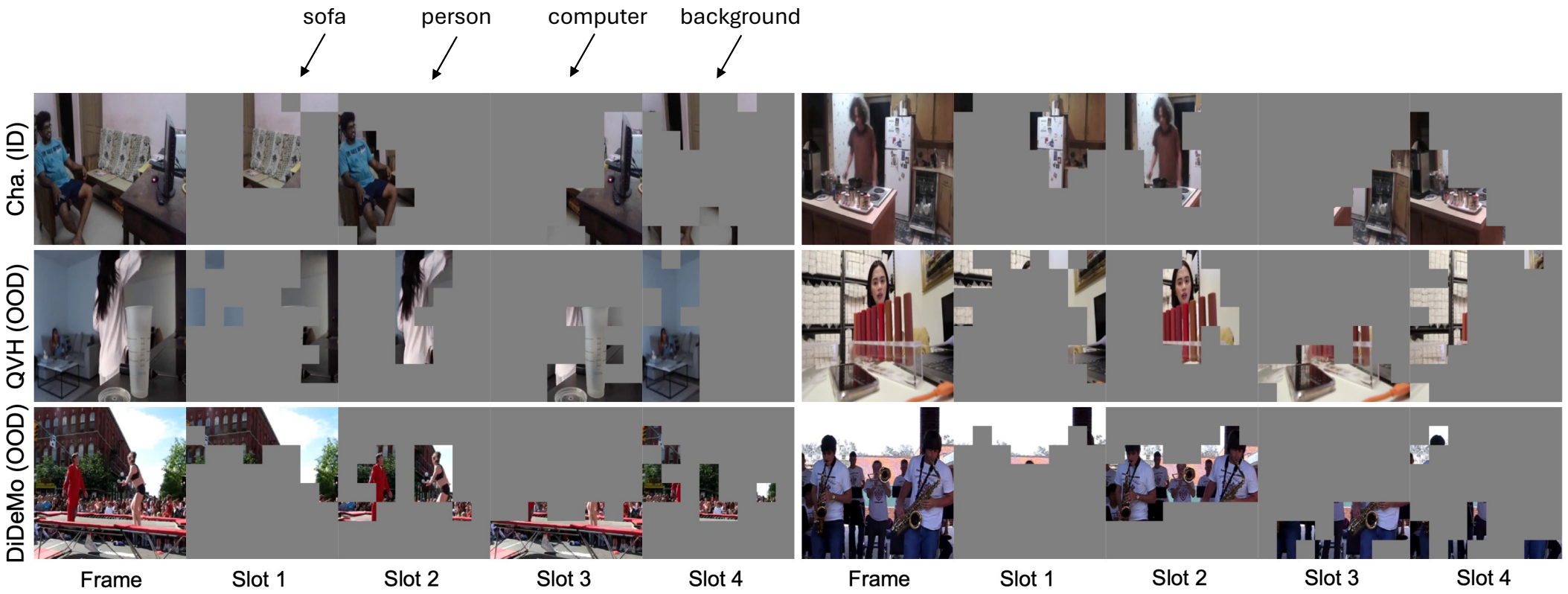
Source Dataset	Method	Target Dataset					
		Charades-STA		QVHighlights		DiDeMo	
		R1@0.5	R1@0.7	R1@0.5	R1@0.7	R1@0.5	R1@0.7
Cha.	EaTR	55.2	33.1	17.0	6.4	20.1	7.8
	CG-DETR	57.6	35.1	22.8	10.5	21.8	11.0
	Chrono-BLIP	68.9	48.7	44.5	23.7	43.2	23.1
	Qwen2.5-VL	67.7	43.9	54.3	29.7	53.8	30.3
	EVIDENT	**69.1**	**49.2**	**59.8**	**36.9**	**55.4**	**32.3**
QVH	EaTR	27.2	13.0	59.6	40.3	17.8	6.5
	CG-DETR	25.5	12.2	65.6	52.1	21.2	8.9
	Chrono-BLIP	39.3	20.4	76.6	62.7	24.4	11.8
	Qwen2.5-VL	48.1	26.2	78.2	64.8	34.1	16.9
	EVIDENT	**50.2**	**28.0**	**82.2**	**69.1**	**36.0**	**17.4**

+5.3

+7.2

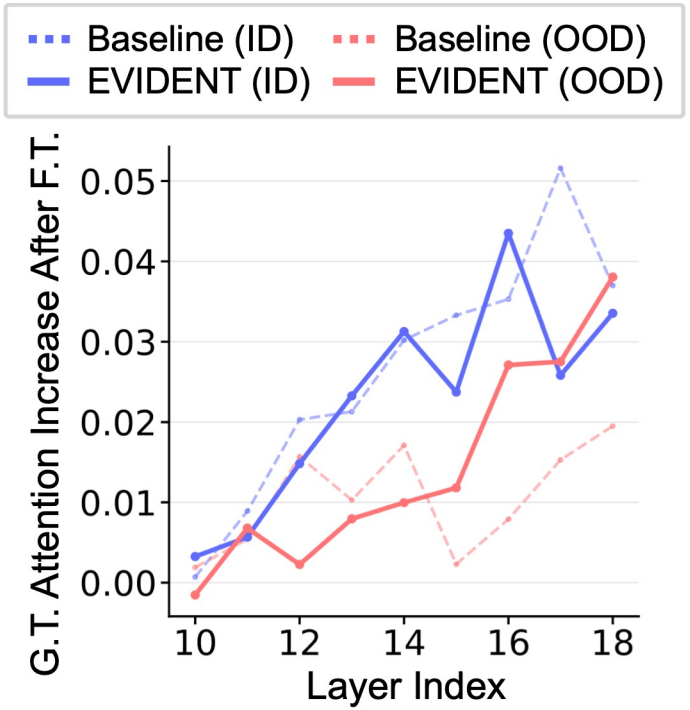
KHU VLLAB VISION & LEARNING LAB

EVIDENT: Generalizes across Video Datasets



KHU VLLAB VISION & LEARNING LAB

Did We Fix **the Failure** We Started With?



Model	Δ ID (Cha.)	Δ OOD (QVH)	Transfer Rate
Qwen2.5-VL†	0.0516	0.0195	37.8%
EVIDENT	0.0435	**0.0381**	**87.6%**

➡ EVIDENT carries more attention gain into the new domain.

KHU VLLAB VISION & LEARNING LAB

Key Takeaways

Problem

- On a new visual domain, naïvely fine-tuned MLLMs **don't look at the query's entities anymore**.

Idea

- Route adaptation through **entity-level slots** instead of dense visual tokens.

Result

- EVIDENT demonstrates **robust VTG performance** across multiple cross-domain settings.

KHU VLLAB
VISION & LEARNING LAB