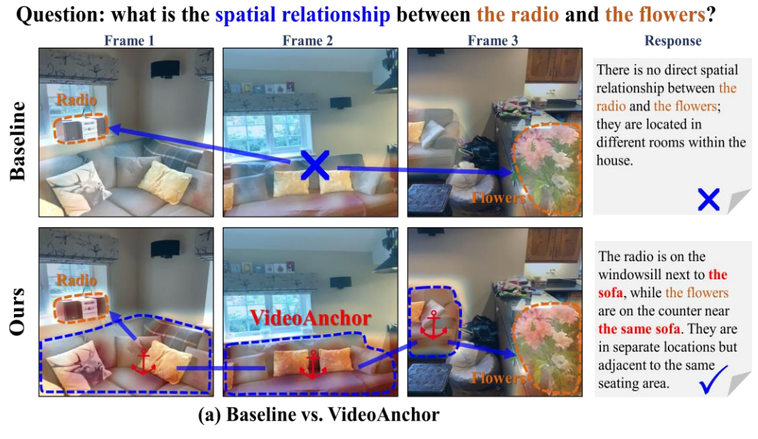
VideoAnchor: Reinforcing Subspace-Structured Visual Cues for Coherent Visual-Spatial Reasoning

Zhaozhi Wang, Tong Zhang, Mingyue Guo, Yaowei Wang, Qixiang Ye

ICLR
Rio de Janeiro, Brazil

Motivation

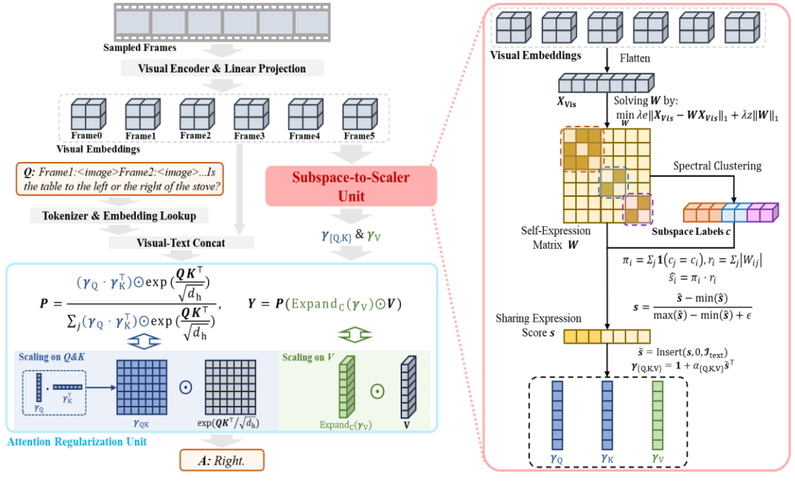
- MLLMs over-rely on textual priors and **underuse visual evidence**.
- Patch-level attention ignores **cross-frame structural consistency**.
- **Shared visual anchors** enable coherent spatial reasoning.



(a) Baseline vs. VideoAnchor

Methods

- **Subspace-to-Scaler Unit:** It discovers the geometric structure of tokens and compute sharing expression scores and attention scalers.
- **Attention Regularization Unit:** It applies the computed scalers at inference-time to modulate Q/K/V in attention operations.



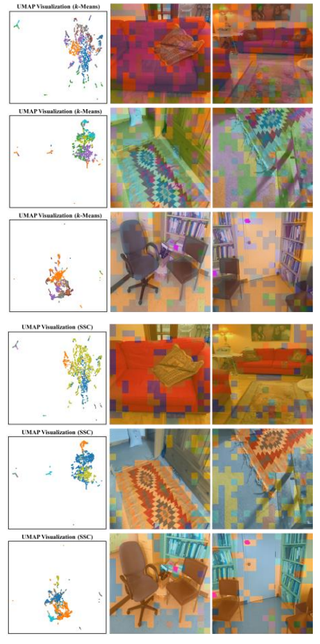
Results & Analysis

Table 1: Performance improvement by VideoAnchor with different MLLMs on VSI-Bench.

Models	Frames	Avg.	Obj. Count	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order
			Numerical Answer				Multiple-Choice Answer			
InternVL2-2B (Chen et al., 2024b)	8	27.4	21.8	24.9	22.0	35.0	33.8	44.2	30.5	7.1
* + VideoAnchor	8	**28.6 (+1.2)**	29.0	25.4	22.0	32.5	34.6	44.6	34.0	6.5
InternVL2-4B (Chen et al., 2024b)	8	31.7	26.3	26.9	37.5	25.4	34.5	42.9	35.0	25.4
* + VideoAnchor	8	**34.5 (+2.8)**	43.3	29.7	33.0	26.2	35.2	44.3	40.2	23.8
InternVL2-8B (Chen et al., 2024b)	8	34.6	23.1	28.7	48.2	39.8	36.7	30.7	29.9	39.6
* + VideoAnchor	8	**37.8 (+3.2)**	39.3	30.3	50.5	38.8	38.6	31.0	36.1	37.8
Qwen2.5-VL-3B (Bai et al., 2025)	16	26.9	18.6	16.9	16.0	26.5	39.1	45.9	29.2	23.7
* + VideoAnchor	16	**27.8 (+0.9)**	20.6	17.2	16.0	26.8	40.4	46.7	30.5	24.0
Qwen2.5-VL-7B (Bai et al., 2025)	16	31.8	28.9	15.2	45.8	30.3	37.6	40.5	25.8	30.2
* + VideoAnchor	16	**33.3 (+1.5)**	28.3	16.4	46.9	31.6	38.6	43.2	28.9	32.5
LLaVA-Video-7B (Zhang et al., 2025b)	16	34.6	44.3	14.6	45.5	25.9	41.3	40.4	36.5	29.5
* + VideoAnchor	16	**35.7 (+1.1)**	47.0	15.1	48.3	25.2	43.1	41.1	34.6	31.2
LLaVA-Video-72B (Zhang et al., 2025b)	16	39.4	43.1	23.7	54.6	38.0	41.1	35.4	30.4	49.2
* + VideoAnchor	16	**40.9 (+1.5)**	48.5	24.9	56.4	35.5	41.6	36.2	34.5	49.4

Table 2: Performance improvement by VideoAnchor with different MLLMs on All-Angles-Bench.

Models	Avg.	Attribute	Cam. Pose	Counting	Manipul.	Rel. Dir.	Rel. Dist.
		Multiple-Choice Answer					
InternVL2.5-2B (Chen et al., 2024a)	44.2	58.7	26.7	38.6	47.8	31.8	47.5
* + VideoAnchor	**45.5 (+1.3)**	61.1	27.9	41.1	47.5	33.3	48.6
InternVL2.5-8B (Chen et al., 2024a)	48.5	77.5	26.7	48.6	39.9	34.6	51.8
* + VideoAnchor	**50.2 (+1.7)**	76.3	24.3	52.4	44.0	38.2	52.5
InternVL2.5-38B (Chen et al., 2024a)	53.1	80.5	32.3	54.9	45.7	40.9	54.0
* + VideoAnchor	**55.0 (+1.9)**	82.3	33.0	57.0	48.6	43.8	54.7
LLaVA-OneVision-0.5B (Li et al., 2025a)	42.4	51.4	35.2	25.1	49.4	34.0	45.9
* + VideoAnchor	**43.9 (+1.5)**	55.4	36.4	26.0	49.6	37.3	46.2
LLaVA-OneVision-7B (Li et al., 2025a)	44.1	63.7	16.5	37.0	42.4	35.8	50.0
* + VideoAnchor	**46.7 (+2.6)**	63.3	23.9	40.6	46.0	37.7	52.0
LLaVA-Video-7B (Zhang et al., 2025b)	43.1	65.5	14.8	37.4	42.6	31.5	47.3
* + VideoAnchor	**44.3 (+1.2)**	65.8	16.5	41.0	40.1	34.1	50.4
LLaVA-Video-72B (Zhang et al., 2025b)	49.9	75.7	29.0	40.6	44.5	42.3	52.8
* + VideoAnchor	**53.0 (+3.1)**	77.1	33.0	41.9	45.8	44.1	60.2



Key insights: VideoAnchor leverages subspace self-expressiveness to transform independent visual tokens into structures that guide attention across frames, enabling coherent visual-spatial reasoning.

https://www.arxiv.org/abs/2509.25151

https://github.com/feufhd/VideoAnchor

wangzhaozhi22 AT mails.ucas.ac.cn