SYSU 2026
PRESENTATION SLIDE

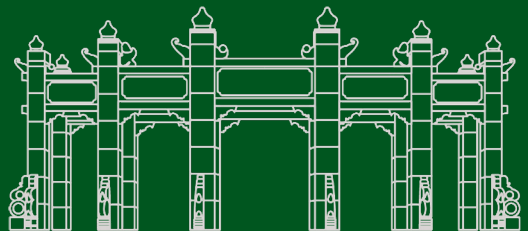
中山大學

SUN YAT-SEN UNIVERSITY

Thinking with Geometry: Active Geometry Integration for Spatial Reasoning

Shenzhen campus of Sun Yet-sen University

Haoyuan Li*, Qihang Cao*, Tao Tang, Kun Xiang, Zihan Guo,
Jianhua Han, JiaWang Bian, Hang Xu, Xiaodan Liang



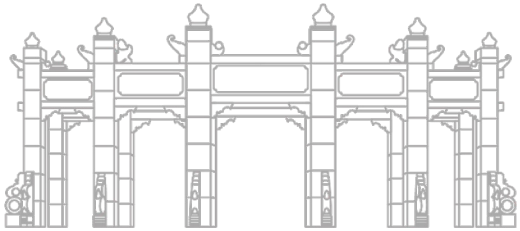
SYSU 2026
PRESENTATION SLIDE

中山大學
SUN YAT-SEN UNIVERSITY

01

Motivation

Why do we do this work?



From passive fusion to Active fusion

Differences between SOTA approaches

Geometry as Input

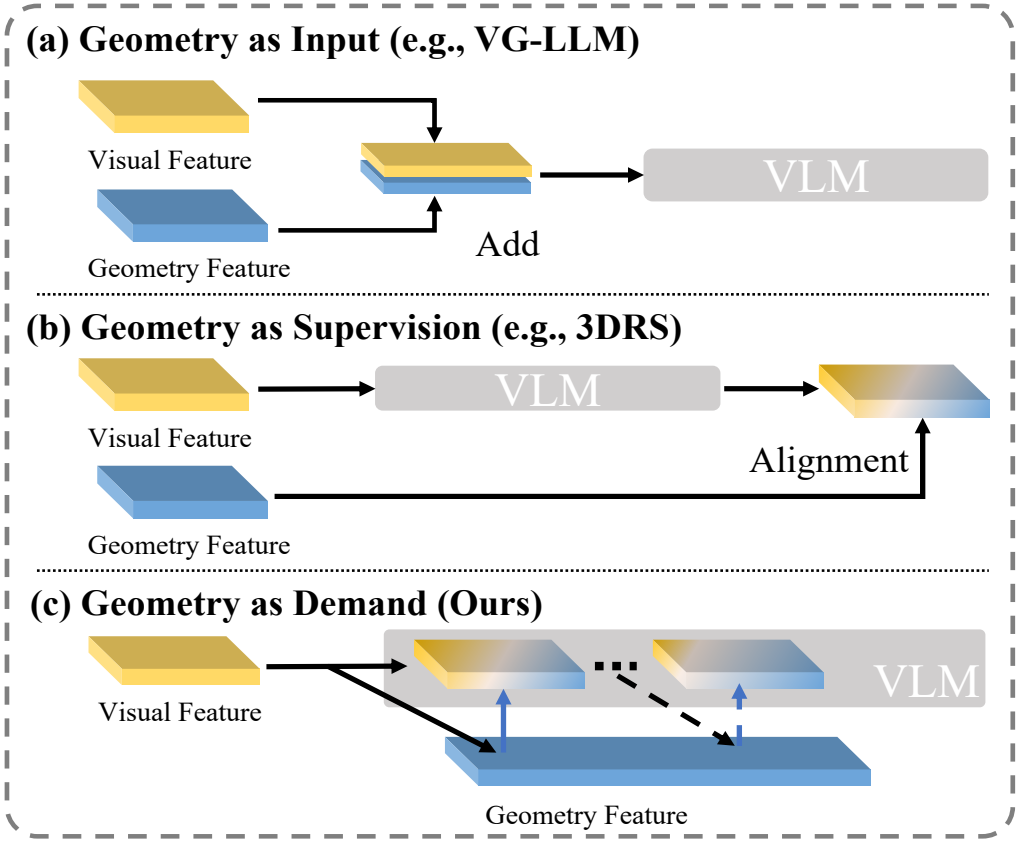
Passively fuses features at the input level, risking semantic misalignment and redundant noise injection.

Geometry as Supervision

Uses geometry uniformly for output alignment, overlooking task-dependent and spatially selective cues.

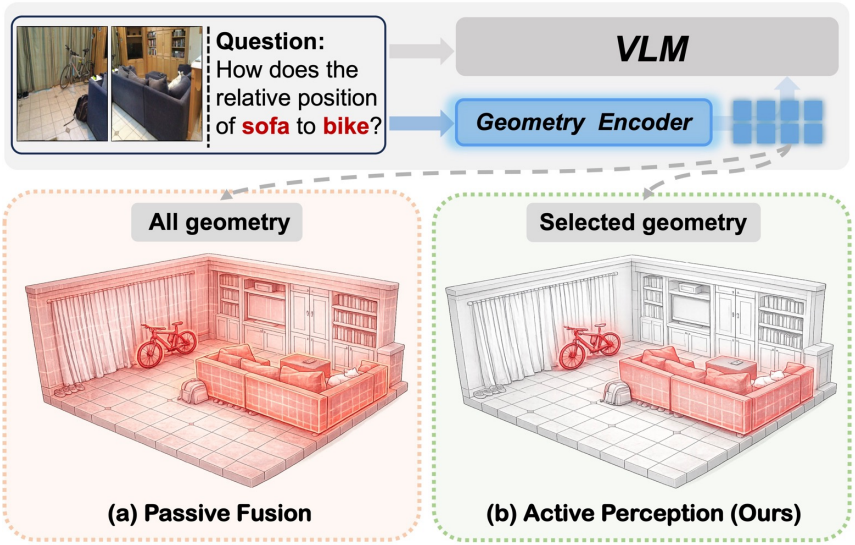
Geometry as Demand

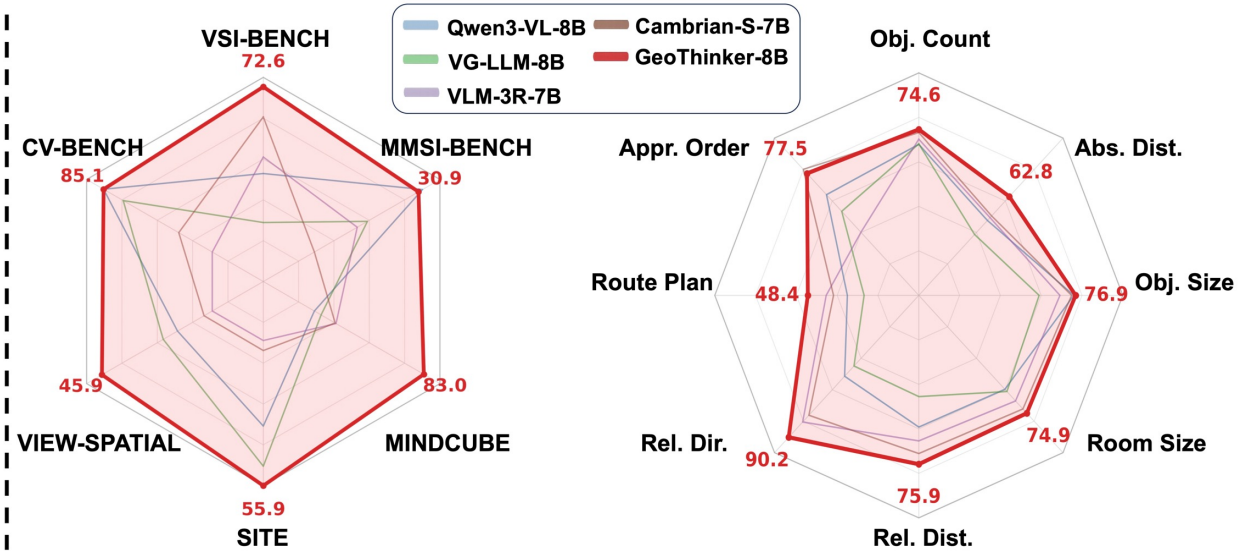
Dynamically integrates geometric features based on specific spatial demands, avoiding noise and enhancing reasoning.



From passive fusion to Active fusion

Towards active geometry integration for spatial reasoning





Key method： Active Geometry Integration

- **Core idea**: LLM should function as an active querier, autonomously **orchestrating when, where, and what type of information to integrate**. This demand-driven approach synthesizes a more robust and purposeful representation than passive fusion.

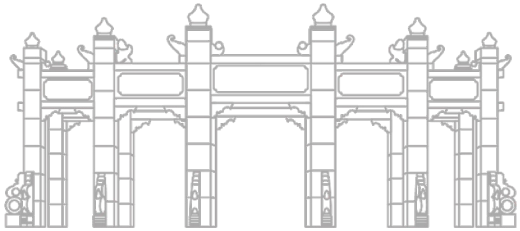
SYSU 2026
PRESENTATION SLIDE

中山大學
SUN YAT-SEN UNIVERSITY

02

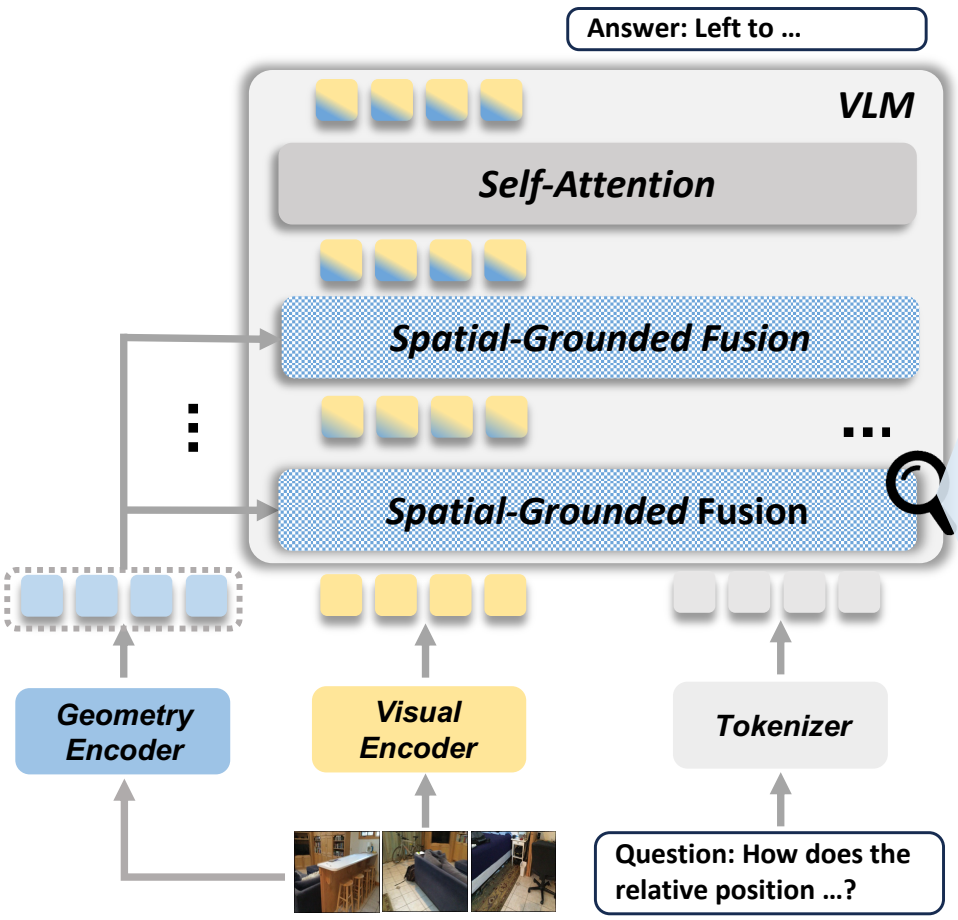
Method

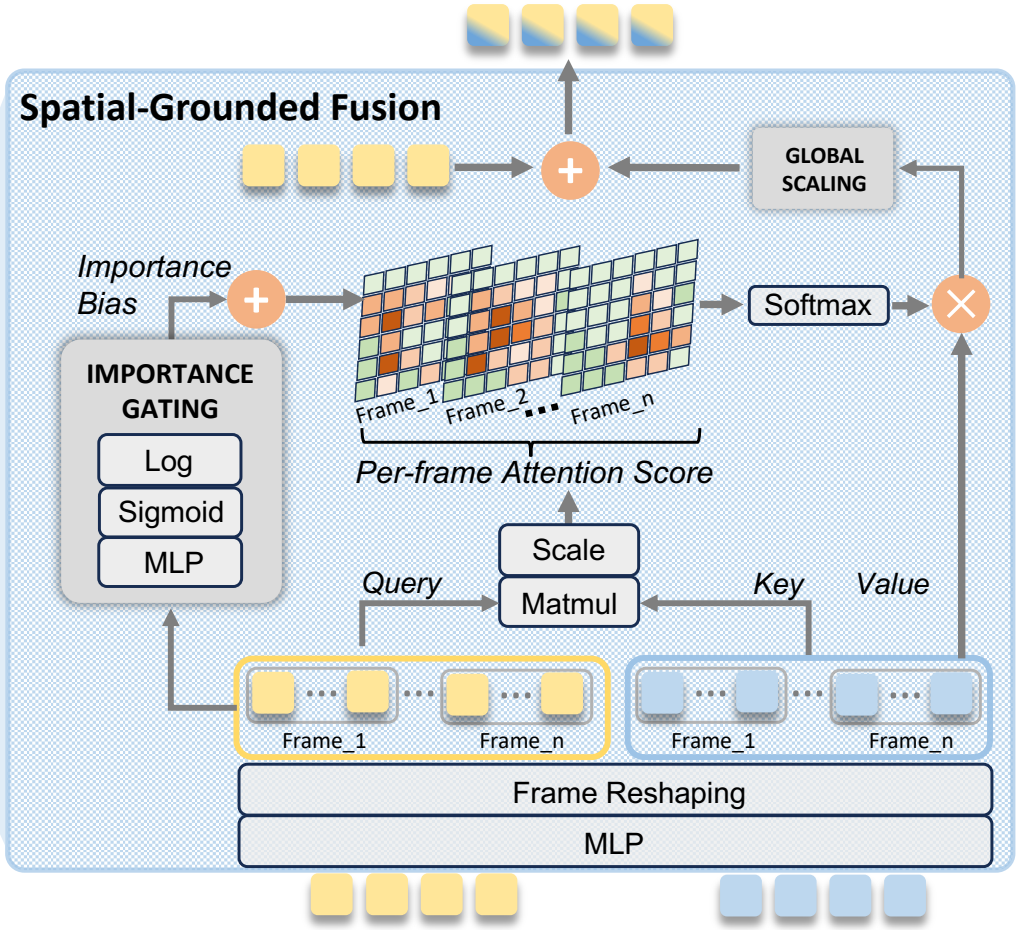
Model Architecture



Model Overview

Spatial-Grounded Fusion inside LLM





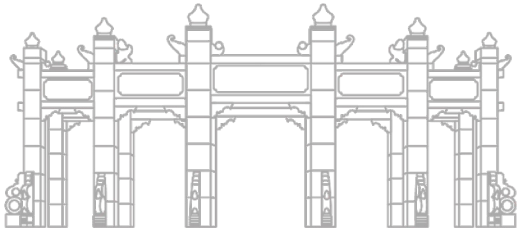
SYSU 2026
PRESENTATION SLIDE

中山大學
SUN YAT-SEN UNIVERSITY

03

Experiment

Performance across different setting



Vanilla Regime

Performance under out-of-domain model and data

Table 1. **Performance comparisons on VSI-Bench (Vanilla regime).** GeoThinker outperforms VG-LLM baseline across different model scales, revealing the effectiveness of Spatial-Grounded Fusion.

Methods	Active perception	Avg.	Numerical Answer				Multiple-Choice Answer			
			Obj.Count	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order
Baseline										
Chance Level (Random)		–	–	–	–	–	25.0	36.1	28.3	25.0
Chance Level (Frequency)		34.0	62.1	32.0	29.9	33.1	25.1	47.9	28.4	25.2
Proprietary Models (API)										
GPT-4o(Hurst et al., 2024)		34.0	46.2	5.3	43.8	38.2	37.0	41.3	31.5	28.5
Gemini-1.5 Flash(Gemini Team, 2024)		42.1	49.8	30.8	53.5	54.4	37.7	41.0	31.5	37.8
Gemini-1.5 Pro(Gemini Team, 2024)		45.4	56.2	30.9	64.1	43.6	51.3	46.3	36.0	34.6
Open-sourced Models										
LLaVA-OneVision-7B(Li et al., 2024a)		32.4	47.7	20.2	47.4	12.3	42.5	35.2	29.4	24.4
LLaVA-OneVision-72B(Li et al., 2024a)		40.2	43.5	23.9	57.6	37.5	42.5	39.9	32.5	44.6
LLaVA-NeXT-Video-7B(Liu et al., 2024a)		35.6	48.5	14.0	47.8	24.2	43.5	42.4	34.0	30.6
LLaVA-NeXT-Video-72B(Liu et al., 2024a)		40.9	48.9	22.8	57.4	35.3	42.4	36.7	35.0	48.6
InternVL2-8B(Chen et al., 2024)		34.6	23.1	28.7	48.2	39.8	36.7	30.7	29.9	39.6
InternVL2-40B(Chen et al., 2024)		36.0	34.9	26.9	46.5	31.8	42.1	32.2	34.0	39.6
Qwen2.5VL-3B(Qwen Team, 2025a)		28.6	32.7	19.5	17.3	25.1	37.3	44.9	30.4	21.8
Qwen2.5VL-7B(Qwen Team, 2025a)		29.3	25.2	10.5	36.4	29.6	38.4	38.0	29.8	26.8
Open-source Spatial Intelligence Models										
SPAR-8B(Zhang et al., 2025)	–	44.1	–	–	–	–	–	–	–	–
SpatialLadder-3B(Li et al., 2025c)	–	44.8	–	–	–	–	–	–	–	–
Spatial-MLLM-4B(Wu et al., 2025)	✗	48.4	65.3	34.8	63.1	45.1	41.3	46.2	33.5	46.3
VG-LLM-4B(Zheng et al., 2025a)	✗	46.7	67.6	37.6	55.2	52.5	48.0	44.7	31.9	35.5
VG-LLM-8B (Zheng et al., 2025a)	✗	49.7	68.1	38.7	59.0	61.1	45.5	44.9	26.8	53.4
Ours										
GeoThinker Qwen2.5VL-3B	✓	**48.9**	68.5	36.1	57.3	62.5	43.7	47.9	34.5	40.9
GeoThinker Qwen2.5VL-7B	✓	**50.5**	69.5	38.5	57.9	62.2	45.2	46.2	31.4	52.6

Scaled Regime

Performance with in-domain model or data

Table 2. **Cross-benchmark comparison on spatial intelligence benchmarks (Scaled regime).** † indicates evaluation on reduced subsets. * indicates trained with S1+S2 dataset setting from VG-LLM. Benchmarks include VSI-Bench(Yang et al., 2025a), MMSI-Bench(Yang et al., 2025c), MindCube-Tiny(Yin et al., 2025), Viewspatial(Li et al., 2025a), SITE(Wang et al., 2025d), CV-Bench (Tong et al., 2024). ✗ denotes passive usage and – denotes methods without dedicated geometry encoders.

Models	Active Perception	Avg.	VSI-Bench	MMSI-Bench	MindCube	ViewSpatial	SITE	CV-Bench
Human		–	79.2	97.2	94.5	–	67.5	–
Random Choice		–	34.0	25.0	33.0	26.3	0.0	–
Proprietary Models								
Seed-1.6(ByteDance Seed, 2025)		53.41	49.9	38.3	48.7	43.8	54.6	85.2
Gemini-2.5-Pro(Gemini Team, 2023)		56.33	53.5	38.0	57.6	46.0	57.0	85.9
GPT-5(OpenAI, 2025)		57.50	55.0	41.8	56.3	45.5	61.8	84.6
Gemini-3-Pro-Preview (Gemini, 2025)		62.16	52.5	45.2	70.8	50.3	62.2	92.0
Open-source General Models								
Bagel-7B-MoT(Deng et al., 2025)		41.90	31.4	31.0	34.7	41.3	37.0	76.0
Qwen2.5-VL-3B-Instruct(Qwen Team, 2025a)		38.60	28.6	28.6	37.6	31.9	33.1	71.8
Qwen2.5-VL-7B-Instruct(Qwen Team, 2025a)		40.31	29.3	26.8	36.0	36.8	37.6	75.4
Qwen3-VL-2B-Instruct(Qwen Team, 2025b)		42.40	49.4	11.9	31.4	34.2	35.6	78.4
Qwen3-VL-8B-Instruct(Qwen Team, 2025b)		47.70	57.7	28.8	29.8	39.0	45.8	85.1
InternVL3-2B(Zhu et al., 2025)		39.31	32.9	26.5	37.5	32.5	30.0	76.5
InternVL3-8B (Zhu et al., 2025)		45.38	42.1	28.0	41.5	38.6	41.1	81.0
Open-source Spatial Intelligence Models								
SpatialLadder-3B(Li et al., 2025c)	–	42.83	44.8	27.4	43.4	39.8	27.9	73.7
VST-3B-SFT (Yang et al., 2025b)	–	49.50	57.9†	30.2†	35.9	52.8	35.8	84.4
VST-7B-SFT(Yang et al., 2025b)	–	51.31	60.6†	32.0†	39.7	50.5	39.6	85.5
Cambrian-S-3B(Yang et al., 2025d)	–	42.91	57.3	25.2	32.5	39.0	28.3	75.2
Cambrian-S-7B (Yang et al., 2025d)	–	47.28	67.5	25.8	39.6	40.9	33.0	76.9
SenseNova-SI-1.1$_{\text{Qwen3VL-8B}}$ (Cai et al., 2025a)	–	-	64.8	38.1	73.8	51.2	49.6	-
SenseNova-SI-1.1$_{\text{InternVL3-8B}}$ (Cai et al., 2025a)	–	-	68.8	43.3	85.7	54.7	47.7	-
3DThinker*$_{\text{Qwen2.5-7B}}$(Chen et al., 2025b)	✗	-	63.7	43.3	76.0	68.6	-	81.1
VLM-3R-7B(Fan et al., 2025)	✗	45.40	60.9	27.9	40.0	40.5	31.3	71.8
VG-LLM-4B(Zheng et al., 2025a)	✗	47.46	46.6	28.0	36.9	42.5	49.8	81.0
VG-LLM-8B(Zheng et al., 2025a)	✗	48.15	49.6	28.4	32.7	42.9	52.6	82.7
~~VG-LLM-8B*(Zheng et al., 2025a)~~	✗	~~51.05~~	62.2	30.0	36.1	45.8	50.5	81.7
Ours								
GeoThinker $_{\text{Qwen2.5VL-7B}}$	✓	**60.43**	68.5	31.7	83.6	41.9	54.8	82.1
GeoThinker $_{\text{Qwen3VL-8B}}$	✓	**62.23**	72.6	30.9	83.0	45.9	55.9	85.1

Other evaluation

General video understanding and frame ablation

Table 3. **Performance comparison on general video data mixture.** (·) denotes the performance change compared to model trained without general video mixture. Notably, while the pure 2D-based Cambrian-S suffers from performance drops on VSI-Bench due to task interference, GeoThinker achieves consistent improvements across both specialized spatial tasks and general video benchmarks with much higher data efficiency.

Model	Video Mixture	VSI	VideoMME	MVBench
Cambrian-S-7B	✗	69.2	54.1	-
	3M	65.1(-4.1)	61.9(+7.8)	64.5
GeoThinker $_{\text{Qwen3vl-8B}}$	✗	72.0	53.7	42.8
	430k	72.6(+0.6)	59.4(+5.7)	69.1(+26.3)

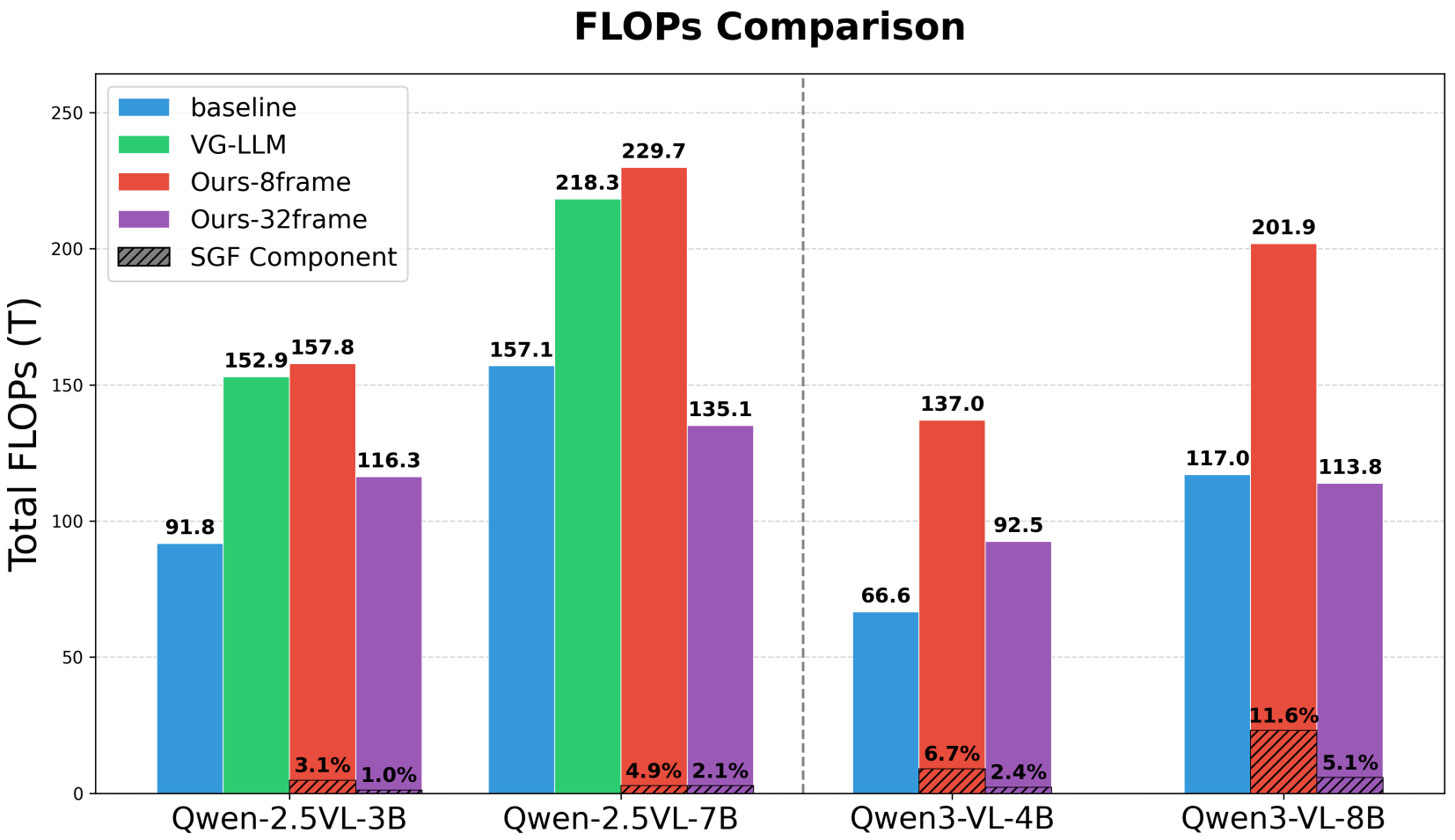
Table 4. **Performance comparison and frame ablation on VSI and VSI-Debiased.** While Cambrian-S-7B (Yang et al., 2025d) is trained on 64/128 frames and VG-LLM-8B*(Zheng et al., 2025a) is trained on 8 frames with S1+S2 setting, GeoThinker is trained on a maximum of 8/32 frames. We evaluate the zero-shot extrapolation capability of all models by scaling inference frames to 128.

Model	Benchmark	# Frames			
		16	32	64	128
Cambrian-S-7B	VSI	58.6	63.6	66.4	**67.5**
	VSI-Debiased	49.7	55.6	59.1	**59.9**
VG-LLM-8B*	VSI	60.5	62.2	**63.7**	63.1
	VSI-Debiased	51.6	52.4	**55.2**	55.1
GeoThinker $_{\text{Qwen3vl-8B-8frame}}$	VSI	67.1	69.8	70.3	**71.2**
	VSI-Debiased	60.7	64.8	64.3	**65.3**
GeoThinker $_{\text{Qwen3vl-8B-32frame}}$	VSI	69.2	72.6	73.4	**73.4**
	VSI-Debiased	64.3	66.3	67.7	**68.1**

GeoThinker achieve better performance with general video mixture and frame ablation study.

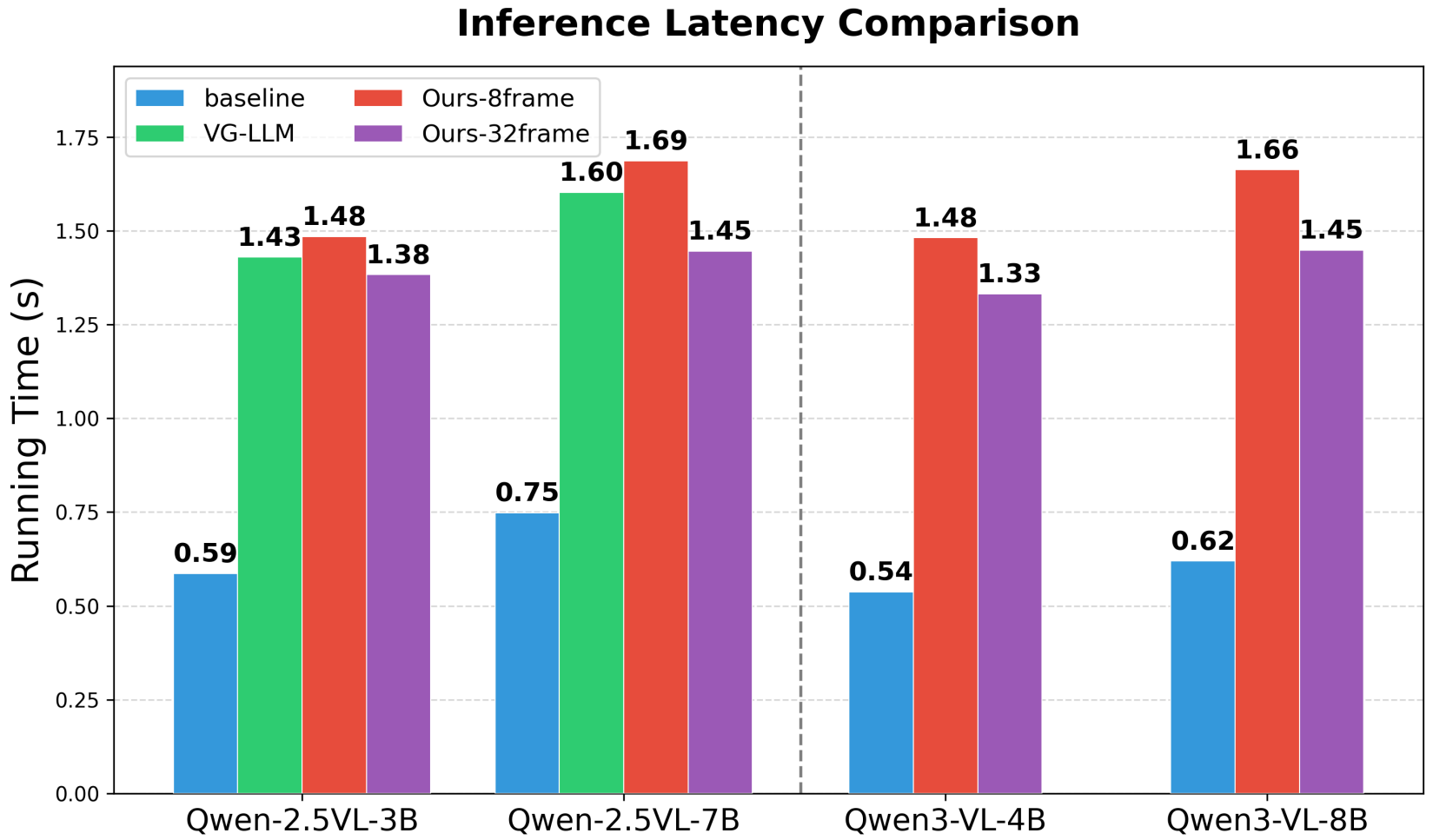
Time cost study

FLOPs and inference latency comparsion



Time cost study

FLOPs and inference latency comparsion



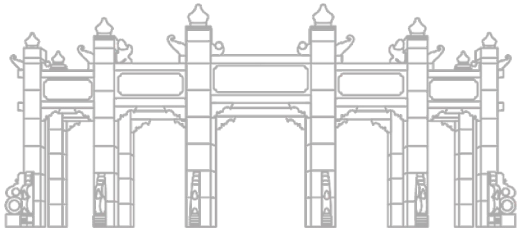
SYSU 2026
PRESENTATION SLIDE

中山大學
SUN YAT-SEN UNIVERSITY

04

Visualization

Visualization of importance score



MindCube

Visualization of importance score on MindCube

Question: Based on these four images (image 1, 2, 3, and 4) showing the pink bottle from different viewpoints (front, left, back, and right), with each camera aligned with room walls and partially capturing the surroundings: If I am standing at the same spot and facing the same direction as shown in image 1, then I turn right and move forward, will I get closer to the pink plush toy and headboard? A. No B. Yes

Prediction: B

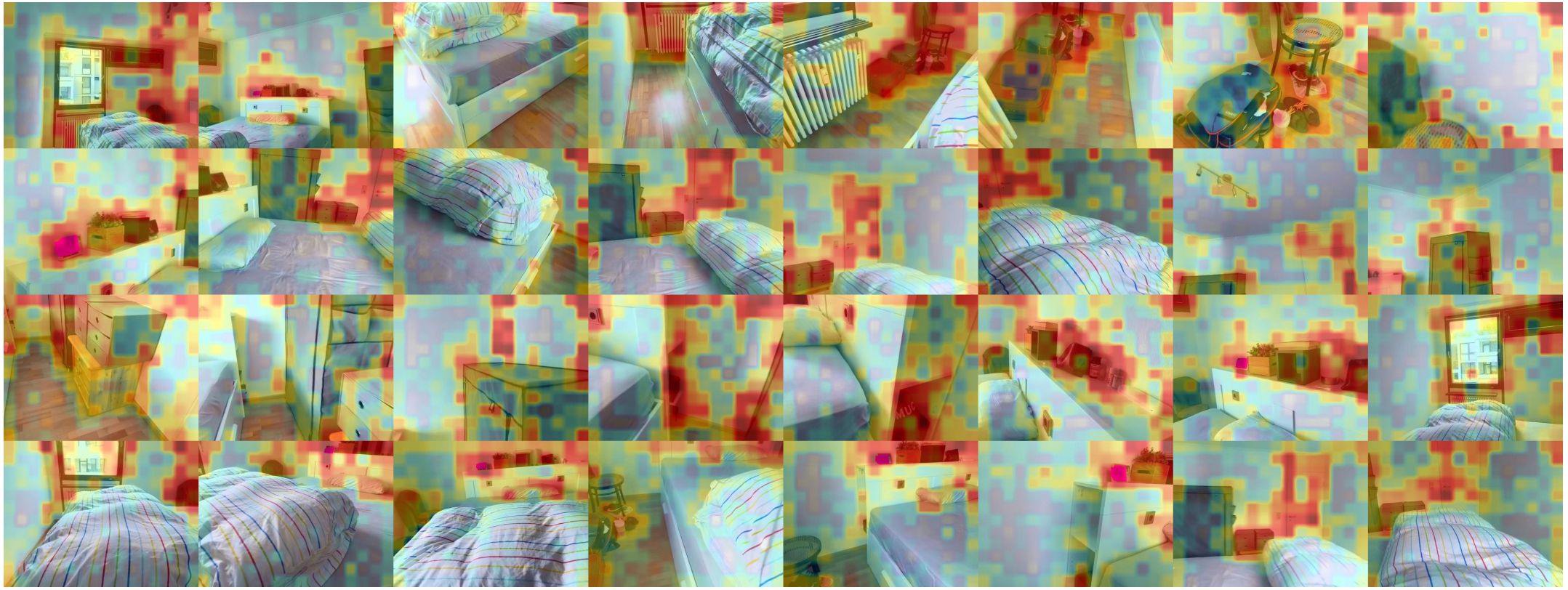
Question: Based on these four images (image 1, 2, 3, and 4) showing the red bottle from different viewpoints (front, left, back, and right), with each camera aligned with room walls and partially capturing the surroundings: If I am standing at the same spot and facing the same direction as shown in image 2, then I turn left and move forward, will I get closer to the TV and electric fan? A. No B. Yes

Prediction: B

Without object mask supervision, GeoThinker can autonomously identify edges and objects that require geometry

VSI-Bench

Visualization of importance score on VSI-Bench



Without object mask supervision, GeoThinker can autonomously identify edges and objects that require geometry

Robustness

Visualization of robustness to image resolution

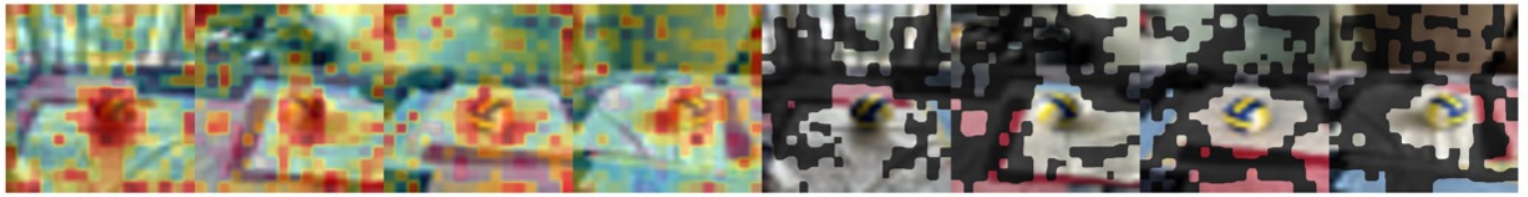
Question: Based on these four images (image 1, 2, 3, and 4) showing the color ball from different viewpoints (front, left, back, and right), with each camera aligned with room walls and partially capturing the surroundings: If I am standing at the same spot and facing the same direction as shown in image 4, then I turn left and move forward, will I get closer to the glass wall? A. Yes B. No

Prediction: A

(a) Input question and images with resolution of [448,488]

(b) Images with 100% original resolution

(c) Images with 25% original resolution



(d) Images with 6.25% original resolution

中山大學
SUN YAT-SEN UNIVERSITY

Thanks for listening

SUN YAT-SEN UNIVERSITY 2026