ECCV 2026 | MALMÖ | SEPT 8–12

Paper Video

Token-level Response-visual Attention Guidance for Multimodal LLMs Knowledge Distillation

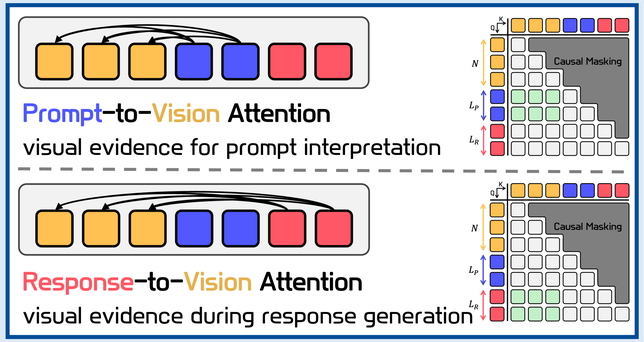
Jaehyun Jang[1], Eunseop Yoon[1], Hee Suk Yoon[1], SooHwan Eom[1], Mark A. Hasegawa-Johnson[2], and Chang D. Yoo[1]
[1]Korea Advanced Institute of Science and Technology (KAIST), [2]University of Illinois Urbana-Champaign (UIUC)

KAIST

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Takeaway) Distill visual grounding during response generation, token by token, with guidance adapted to how diffuse or focused the teacher's attention is.

Cross-modal Attention Extraction

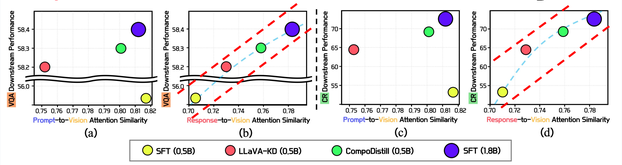


Prompt-to-Vision Attention
visual evidence for prompt interpretation

Response-to-Vision Attention
visual evidence during response generation

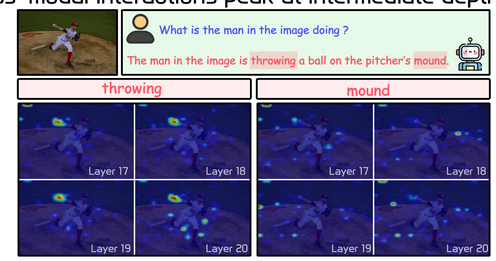
Observations

Response-to-Vision is the more informative signal



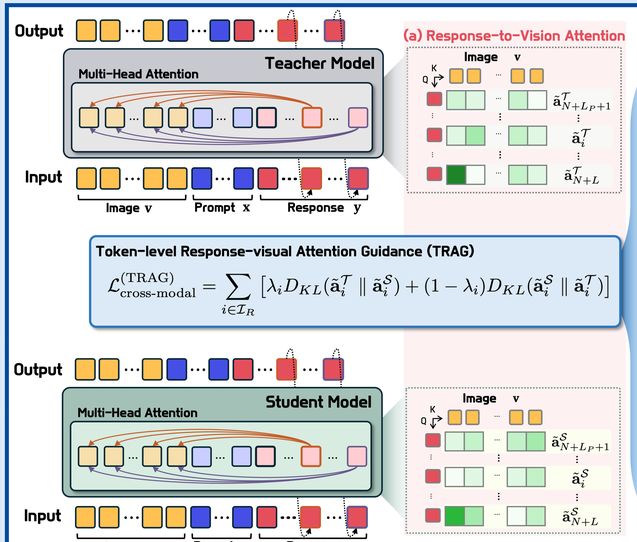
Prompt-to-Vision → Weak performance correlation
Response-to-Vision → Strong performance correlation
→ ① Supervise the response phase

Visual grounding varies across response tokens
Cross-modal interactions peak at intermediate depths



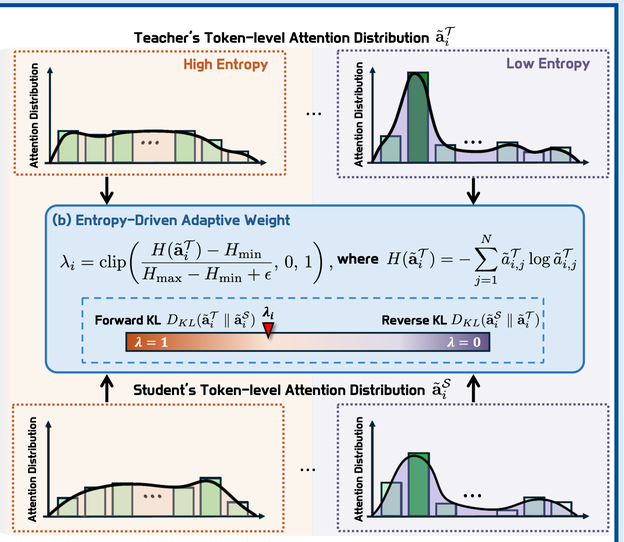
Across adjacent layers → Redundant attention
Across response tokens → Diverse visual grounding
→ ② Aggregate layers, ③ supervise each response token

TRAG: Token-level Response-visual Attention Guidance



Token-level Response-visual Attention Guidance (TRAG)

$$\mathcal{L}^{(\mathrm{TRAG})}_{\text{cross-modal}} = \sum_{i\in\mathcal{I}_R} \left[\lambda_i D_{KL}(\tilde{\mathbf{a}}_i^{\mathcal{T}} \,\|\, \tilde{\mathbf{a}}_i^{\mathcal{S}}) + (1-\lambda_i) D_{KL}(\tilde{\mathbf{a}}_i^{\mathcal{S}} \,\|\, \tilde{\mathbf{a}}_i^{\mathcal{T}})\right]$$



① Shifting Supervision to the Response Phase

$$\mathcal{L}^{(\mathrm{CompoDistill})}_{\text{cross-modal}} = \sum_{l\in\mathcal{M}_S} \mathcal{D}_{\cos}(\mathbf{A}^{\mathcal{T}}_{G_l}(I), \mathbf{A}^{\mathcal{S}}_l(I))$$

I = an index set of textual query

$I = I_P,\ I_P = \{N+1, \ldots, N+L_P\}$

$$\mathcal{L}^{(\mathrm{prior})}_{\text{cross-modal}} = \sum_{l\in\mathcal{M}_S} \mathcal{D}_{\cos}(\mathbf{A}^{\mathcal{T}}_{G_l}(I_P), \mathbf{A}^{\mathcal{S}}_l(I_P))$$

Query range shift! $I_P \Rightarrow I_R$

$I = I_R,\ I_R = \{N+L_P+1, \ldots, N+L\}$

$$\mathcal{L}^{(\mathrm{response})}_{\text{cross-modal}} = \sum_{l\in\mathcal{M}_S} \mathcal{D}_{\cos}(\mathbf{A}^{\mathcal{T}}_{G_l}(I_R), \mathbf{A}^{\mathcal{S}}_l(I_R))$$

$$\mathcal{L}^{(\mathrm{response})}_{\text{cross-modal}} = \sum_{l\in\mathcal{M}_S} \mathcal{D}_{\cos}(\mathbf{A}^{\mathcal{T}}_{G_l}(I_R), \mathbf{A}^{\mathcal{S}}_l(I_R))$$

② From Layer-wise Matching to Token-level Distribution

Aggregate intermediate-layer attention for each response token

$$\{a^{\mathcal{S}}_{l,i}\}_{l\in\mathcal{M}_S} \to \bar{\boldsymbol{a}}^{\mathcal{S}}_i = \frac{1}{|\mathcal{M}_S|}\sum_{l\in\mathcal{M}_S} \boldsymbol{a}^{\mathcal{S}}_{l,i} \qquad \{a^{\mathcal{T}}_{l,i}\}_{l\in\mathcal{M}_{\mathcal{T}}} \to \bar{\boldsymbol{a}}^{\mathcal{T}}_i = \frac{1}{|\mathcal{M}_{\mathcal{T}}|}\sum_{l\in\mathcal{M}_{\mathcal{T}}} \boldsymbol{a}^{\mathcal{T}}_{l,i}$$

Normalize each token's attention vector into a probability distribution

$$\tilde{\boldsymbol{a}}^{\mathcal{S}}_i \coloneqq \frac{\bar{\boldsymbol{a}}^{\mathcal{S}}_i}{\sum_{j=1}^{N} \bar{\boldsymbol{a}}^{\mathcal{S}}_{i,j} + \epsilon} \qquad \tilde{\boldsymbol{a}}^{\mathcal{T}}_i \coloneqq \frac{\bar{\boldsymbol{a}}^{\mathcal{T}}_i}{\sum_{j=1}^{N} \bar{\boldsymbol{a}}^{\mathcal{T}}_{i,j} + \epsilon}$$

Define token-level guidance over normalized attention distributions

$$\mathcal{L}^{(\mathrm{token-level})}_{\text{cross-modal}} = \sum_{i\in I_R} \mathcal{D}_{\text{guidance}}(\tilde{\boldsymbol{a}}^{\mathcal{T}}_i, \tilde{\boldsymbol{a}}^{\mathcal{S}}_i)$$

③ Entropy-Driven Adaptive Weighting

$$\mathcal{L}^{(\mathrm{token-level})}_{\text{cross-modal}} = \sum_{i\in I_R} \mathcal{D}_{\text{guidance}}(\tilde{\boldsymbol{a}}^{\mathcal{T}}_i, \tilde{\boldsymbol{a}}^{\mathcal{S}}_i)$$

$$\lambda_i = \mathrm{clip}\left(\frac{H(\tilde{\boldsymbol{a}}^{\mathcal{T}}_i) - H_{min}}{H_{\max} - H_{min} + \epsilon}, 0, 1\right), \text{ where } H(\tilde{\boldsymbol{a}}^{\mathcal{T}}_i) = -\sum_{j=1}^{N} \tilde{a}^{\mathcal{T}}_{i,j} \log \tilde{a}^{\mathcal{T}}_{i,j}$$

$\lambda_i \uparrow$ $\lambda_i \downarrow$

High Entropy
- Diffuse visual evidence
- Forward KL for coverage

$D_{\mathrm{KL}}(\tilde{\boldsymbol{a}}^{\mathcal{T}}_i \,\|\, \tilde{\boldsymbol{a}}^{\mathcal{S}}_i)$

Low Entropy
- Localized visual evidence
- Reverse KL for precision

$D_{\mathrm{KL}}(\tilde{\boldsymbol{a}}^{\mathcal{S}}_i \,\|\, \tilde{\boldsymbol{a}}^{\mathcal{T}}_i)$

$$\mathcal{L}^{(\mathrm{TRAG})}_{\text{cross-modal}} = \sum_{i\in I_R} \left[\lambda_i D_{\mathrm{KL}}(\tilde{\boldsymbol{a}}^{\mathcal{T}}_i \,\|\, \tilde{\boldsymbol{a}}^{\mathcal{S}}_i) + (1-\lambda_i) D_{\mathrm{KL}}(\tilde{\boldsymbol{a}}^{\mathcal{S}}_i \,\|\, \tilde{\boldsymbol{a}}^{\mathcal{T}}_i)\right]$$

Main Results

General Visual Question Answering Benchmarks

Size	Method	LLM	# Samples	Visual Question Answering Benchmarks								Avg_6	Avg_8
				GQA	ScienceQA	TextVQA	MME	MMB	MMB^{CN}	POPE	MMMU		
≤ 4B	TinyLLaVA†	Qwen1.5-4B	1.2 M	62.7	70.2	60.0	70.5	67.5	67.2	86.9	38.6	69.6	65.4
≤ 2B	TinyLLaVA†	Qwen1.5-1.8B	1.2 M	60.9	65.2	47.7	61.3	57.8	56.2	83.4	34.8	62.7	58.4
	KD†	Qwen1.5-1.8B	1.2 M	60.9	64.6	53.3	66.3	65.8	63.1	86.5	32.7	66.2	61.6
	CompoDistill†	Qwen1.5-1.8B	1.2 M	61.2	66.5	53.5	67.0	64.5	63.0	85.5	34.1	66.4	61.9
	LLaVA-KD	Qwen1.5-1.8B	1.2 M	**62.3**	64.7	53.4	**69.1**	64.0	**63.7**	86.3	33.6	66.6	62.1
	TRAG	Qwen1.5-1.8B	1.2 M	61.7	**66.9**	**55.6**	67.0	**66.6**	63.3	**86.8**	**35.2**	**67.4**	**62.9**
≤ 0.5B	TinyLLaVA†	Qwen1.5-0.5B	1.2 M	56.7	60.9	46.6	59.4	55.2	51.8	83.9	31.9	60.4	55.8
	KD†	Qwen1.5-0.5B	1.2 M	56.8	60.4	46.1	58.9	54.7	49.7	84.9	32.3	60.3	55.5
	CompoDistill†	Qwen1.5-0.5B	1.2 M	59.1	61.1	48.2	63.8	59.6	54.9	85.7	**34.0**	62.9	58.3
	LLaVA-KD	Qwen1.5-0.5B	1.2 M	**59.6**	60.6	**49.9**	64.5	60.1	**55.5**	85.9	30.2	63.4	58.2
	TRAG	Qwen1.5-0.5B	1.2 M	59.5	**62.1**	48.9	**65.2**	**60.9**	54.6	**86.5**	30.9	**63.8**	**58.5**

Across backbones & scales → Consistent gains

Compositional Reasoning Benchmarks

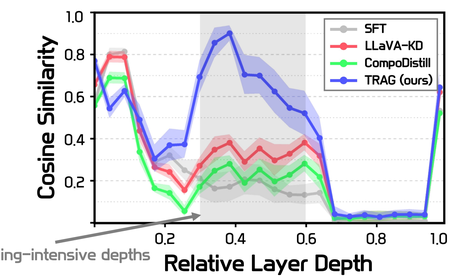
↳ Directly probe fine-grained visual grounding

LLM	Method	Size	Compositional Reasoning Benchmarks			Avg
			SugarCrepe	BiVLC	Winoground	
Qwen1.5	TinyLLaVA	4B	87.3	93.2	70.1	83.5
	TinyLLaVA	1.8B	74.8	83.8	62.6	73.7
	KD		76.1	85.3	61.5	74.3
	LLaVA-KD		76.6	85.5	60.9	74.3
	CompoDistill		81.7	86.5	62.2	76.8
	TRAG		**83.1**	**89.1**	**67.2**	**79.8**
	TinyLLaVA	0.5B	52.8	56.5	50.2	53.2
	KD		51.0	50.1	52.1	51.0
	LLaVA-KD		66.9	74.5	51.7	64.4
	CompoDistill		72.1	78.6	56.6	69.1
	TRAG		**73.4**	**80.2**	**57.3**	**70.3**

TRAG → Up to +3.0 Avg points over prior baselines

Response-to-Vision Attention Fidelity

Teacher-Student R→V Attention Cosine Similarity



30–60% Depth → TRAG ≈ 0.72
SFT ≈ 0.18 | LLaVA-KD ≈ 0.32 | CompoDistill ≈ 0.41
→ Stronger response-time grounding fidelity

Check out the paper for more experiments and analyses!