Direct Visual Grounding by Directing Attention of Visual Tokens

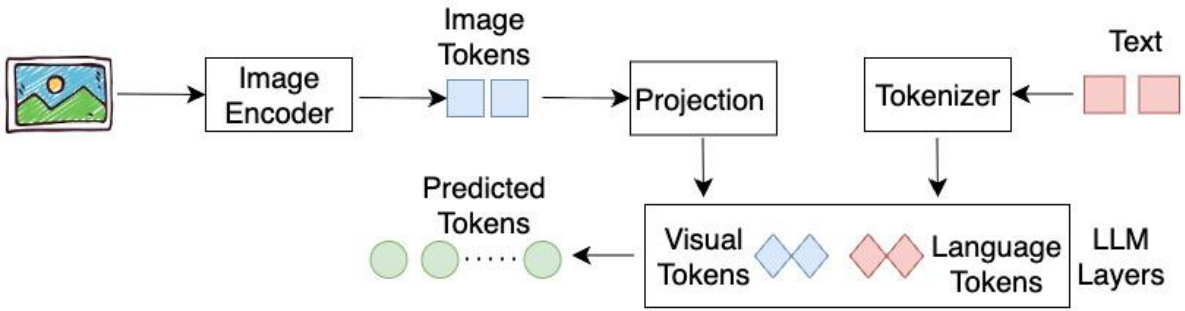
Parsa Esmaeilkhani, Longin Latecki

Temple University

WACV 2026

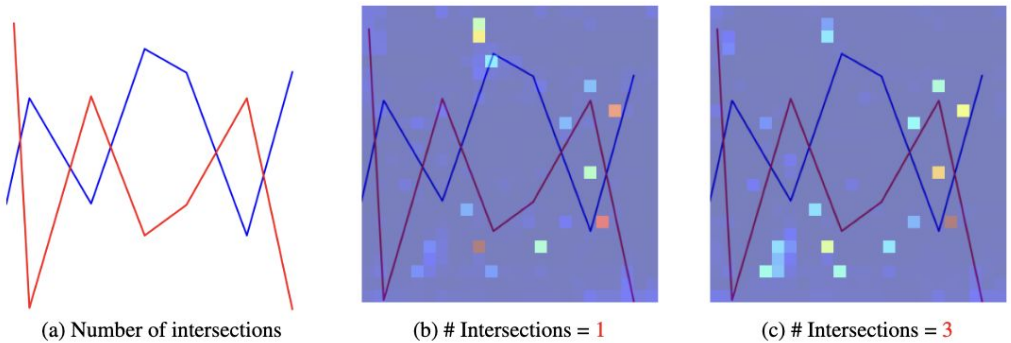
Problem Formulation

- Auto-regressive VLMs mix visual and language tokens in the LLM
- Attention layers treat all tokens uniformly
- Answer tokens allocate little attention to relevant visual tokens
- NTP loss provides weak cross-modal supervision



Empirical Observation

- **In Attention Maps: Attention to Relevant Patches is Low.**
- Target patches receive lower-than-average attention, even after NTP finetuning.



Method: KL Attention Loss

- **KLAL** aligns predicted and target visual-token attention distributions across layers

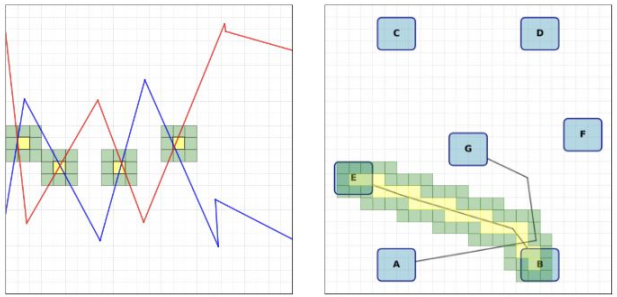
$$L_{\mathrm{KLAL}} = \frac{1}{L}\sum_{l=1}^{L} D_{\mathrm{KL}}\Big(P(S)\,\|\,Q^{(l)}(S)\Big) = \frac{1}{L}\sum_{l=1}^{L}\sum_{i\in I_V} P_i(S)\log\frac{P_i(S)}{Q_i^{(l)}(S)}$$

$$L_{\mathrm{total}} = L_{\mathrm{NTP}} + \lambda L_{\mathrm{KLAL}}$$

$P(S)$: **Ground-truth attention;** $Q^{(l)}(S)$: **Predicted visual-token attention.**

Ground Truth Maps

- **Target attention maps identify visual tokens relevant to the answer.**
- **No Labeling is required!**
- Synthetic tasks: derived from geometry (e.g., intersections, paths).
- Real images: derived from bounding boxes or point annotations.



(a) Line Intersection task (b) Is node E connected to node B?

Data & Tasks

We evaluate on synthetic, real-image, and benchmark grounding tasks.

- **Line Intersection:** count intersections between curves.
- **Line Tracing:** determine connectivity between nodes.
- **Grid Patch:** predict coordinates of target patch.
- **PixMo-Points:** predict object location in real images.
- **RefCOCO:** referring expression comprehension benchmark.

Experimental Results(1)

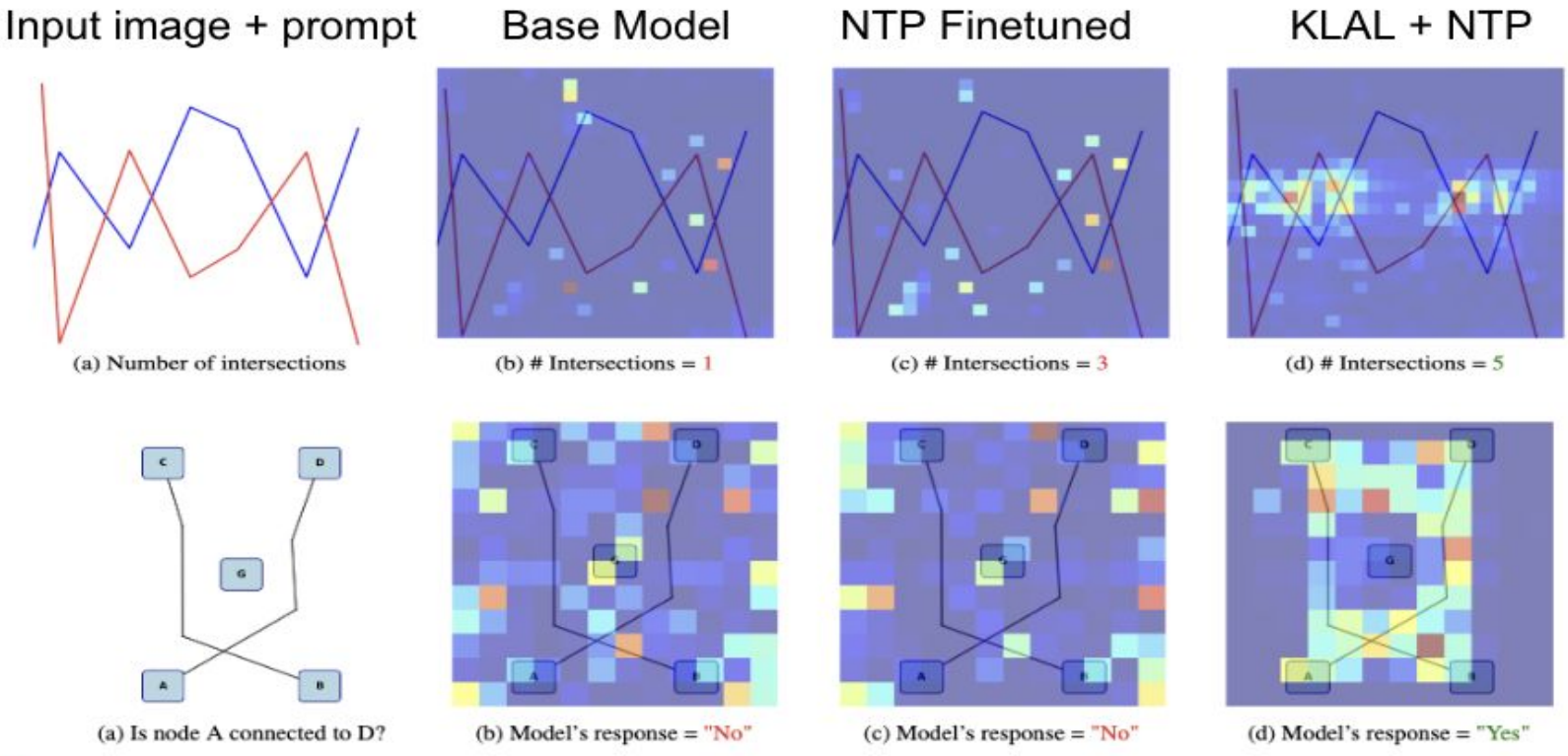
- Experiments Done using Qwen 2.5-VL.
- Grounding Accuracy in %.

Task	Base	NTP	KLAL + NTP (Ours)
Line Intersection	47.6	62.6	**70.2**
Line Tracing	49.6	53.8	**62.2**
Grid Patch	6.1	28.6	**44.9**
PixMo-Points	16.8	26.3	**35.8**
RefCOCO (val)	90.1	90.7	**91.5**

Experimental Results(2)

Model	Line Intersection	Line Tracing
NTP + KLAL (Ours)	**70.23**	**62.21**
Molmo-7B-D	41.07	49.51
GLaMM-FS-7B	27.50	39.03
InstructBLIP	36.67	46.23
GPT-4o	42.12	55.34
Gemini-2.0 Flash	56.41	59.25

KLAL Sharpens Attention Maps



Take Aways

- NTP does not sufficiently supervise visual Token.
- Direct supervision through KLAL improves grounding and attention maps.
- Model-agnostic and simple.
- Effective across synthetic and real tasks.