

Perception Chains: Stabilizing Multimodal Reasoning with Explicit Visual Intermediates

Dongqi Zuo
dongqizuo@gmail.com
University of California, Berkeley
Berkeley, California, USA

Weixing Guo
The Chinese University of Hong Kong
Hong Kong

Xinyu Wang
Johns Hopkins University
Baltimore, Maryland, USA

Abstract

Multimodal chain-of-thought methods produce textual reasoning steps, but the visual grounding underlying each step is never made explicit. When a model attends to the wrong region, nothing in the chain records this: the error is unverifiable during inference, unlocalizable after the fact, and uncorrectable without regenerating the entire chain. We propose **PERCEPTION CHAINS**, which treats each reasoning step as a grounding commitment: before drawing any conclusion, the model must name a region, extract an attribute from that region, and assign it to a reasoning slot. With this structure in place, every grounding decision is independently checkable and a wrong answer is traceable to the step where the binding failed. Correct output format, however, is not sufficient. A model satisfying the format can still leave its hidden states at anchor positions in a non-attenuating drift regime. Across our measured settings, task-loss gradients are nearly orthogonal to the stability objective at these positions, and standard fine-tuning does not eliminate the drift. A stability loss directly constrains this geometry and enforces transversal contraction. High-quality reference trajectories are constructed without human annotation by retaining only chains where independent samples from a frozen 72B VLM converge, filtering at the anchor level rather than by answer correctness alone. At inference time, a consistency check identifies anchors whose evidence conflicts with the provisional answer; only those anchors are re-localized before the final answer is committed. On eight benchmarks against six baselines, **PERCEPTION CHAINS** achieves the highest average accuracy, with the largest gains on tasks requiring precise region-attribute binding. On DIOR-RSVG, grounding accuracy rises from 61.7% to 72.6% over the strongest spatial baseline. The stability loss, not the anchor format, is the transferable component: gains carry to seven held-out benchmarks unseen during training.

CCS Concepts

• **Computing methodologies** → **Scene understanding**; *Image representations*.

Keywords

multimodal reasoning, visual grounding, chain-of-thought, representation stability, knowledge distillation

ACM Reference Format:

Dongqi Zuo, Weixing Guo, and Xinyu Wang. 2026. Perception Chains: Stabilizing Multimodal Reasoning with Explicit Visual Intermediates. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26)*, November 10–14, 2026, Rio de Janeiro, Brazil. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3767308.3836512>

1 Introduction

Multimodal large language models have made rapid progress on visual question answering by coupling language reasoning with visual perception [2]. A prominent line of work improves multi-step accuracy by having the model produce a chain of reasoning steps before committing to a final answer [17, 19, 25]. Structuring the reasoning process in this way reduces answer hallucination, but it leaves a fundamental problem untouched: every intermediate step makes a claim about the image without naming any spatial region. When a model attends to the wrong part of the image, nothing in the chain records this. The claim is accepted, propagated, and only revealed as wrong by the final answer. At that point there is no basis for identifying which step failed, and no region to re-examine. Grounding errors are unverifiable during inference, unlocalizable after the fact, and uncorrectable without regenerating the entire chain from scratch.

A complementary line of work produces explicit spatial outputs. Referring models [21] and visual grounding approaches [3] localize objects accurately, but treat localization as an endpoint rather than an intermediate step in a reasoning chain. Post-hoc correction [20] and preference-based alignment [8] reduce surface hallucination but do not localize where in the reasoning chain the visual error arose. The gap these two lines leave open is the same: no method requires a reasoning step to spatially commit before drawing a conclusion, so the source of a grounding failure stays opaque by construction.

We propose **PERCEPTION CHAINS**, which treats grounding as a two-level problem.

At the output level, *Anchored Visual Evidence* (AVE, §3.1) requires each reasoning step to name a region, extract an attribute from that region, and assign that attribute to a reasoning slot before drawing any conclusion. With this structure in place, every grounding decision is checkable against the image, and a wrong answer is traceable to the specific step where the binding failed.

At the representation level, satisfying AVE at the output does not mean the model's internal states are well-grounded. A model can produce correctly structured triples while its hidden states remain in a non-attenuating drift regime. Standard task loss does not explicitly constrain hidden-state geometry at intermediate anchor positions. Empirically, its gradient is nearly orthogonal to the stability objective at these positions. The *Stability Loss* (§3.2)



This work is licensed under a Creative Commons Attribution 4.0 International License. *MM '26, Rio de Janeiro, Brazil.*

© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2213-4/2026/11
<https://doi.org/10.1145/3767308.3836512>

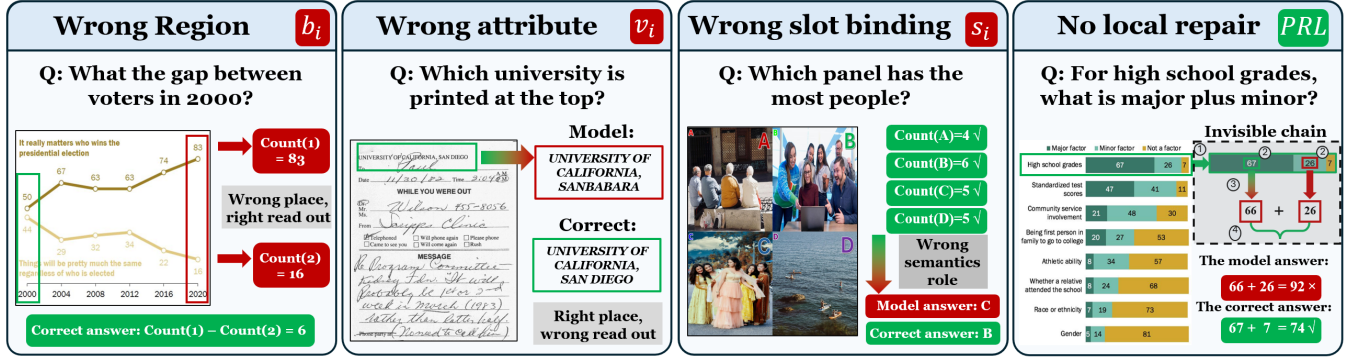


Figure 1: Four failure modes in current MLLMs (upper rows) and their anchor-level decomposition (lower rows). Columns 1–3 show structurally independent error types: wrong region (b_i), wrong attribute (v_i), wrong slot binding (s_i). Column 4 shows a consistent-but-wrong chain that lies outside the reach of local repair.

directly constrains this geometry by penalizing deviations from the reference grounding direction.

Annotating anchor trajectories by hand is not feasible at training scale, which motivates a self-training approach. The natural baseline is to filter the teacher’s candidate chains by answer correctness, as in STaR [23], but answer-correct chains can still contain misgrounded intermediate steps. *Confidence-Weighted Trajectory Bootstrapping* (CTB, §3.3) instead retains only chains where multiple independent samples from a frozen 72B teacher converge, filtering at the level of intermediate anchors rather than final answers, without any human annotation.

Even a well-trained model will occasionally produce an anchor chain whose intermediate evidence is inconsistent with its own provisional answer. The *Perceptual Restoring Loop* (PRL, §3.4) catches these cases at inference time, re-localizes only the conflicting anchors, and commits a corrected answer without any retraining.

Contributions.

- We show that over 99% of calibration inputs remain in a non-attenuating drift regime after format-only fine-tuning, establishing that output-level grounding compliance and representational grounding quality are structurally decoupled.
- We propose AVE, which makes every grounding decision in a reasoning chain independently checkable and localizable.
- We propose a stability loss that constrains hidden-state geometry at anchor positions and induces transversal contraction across all evaluated settings.
- We construct high-quality teacher trajectories without human annotation by filtering at the intermediate anchor level via self-consistency over a frozen 72B VLM.
- We propose the *Perceptual Restoring Loop* (PRL), which treats evidence-answer inconsistency as an inference-time-correctable failure mode dissociable from localization accuracy, and corrects it without retraining.

Results. PERCEPTION CHAINS achieves the highest average accuracy across *eight seen* benchmarks against six baselines. On DIOR-RSVG, grounding accuracy rises from 61.7% to 72.6% over the strongest spatial baseline. The anchor format alone does not generalize zero-shot; the stability loss does, and its gains carry to *seven held-out* benchmarks unseen during training.

2 Problem Formulation

Three failure modes and one boundary. Figure 1 gives concrete instances. The first three columns show that errors in region selection, attribute extraction, and slot assignment are structurally independent: each is a distinct failure at a different component of the grounding step, and each is opaque under text-only intermediate representations. The fourth column states a boundary that must be acknowledged: a chain whose every step is internally consistent can still produce a wrong answer, a case that no per-step check can detect. This boundary determines the scope of any inference-time correction mechanism.

Formalizing the structure. The three failure modes become formally statable once each reasoning step is required to name a spatial region before drawing a language conclusion.

Definition 2.1 (Visual Anchor Node). A visual anchor node is a triple $a_i = (b_i, v_i, s_i)$, where $b_i \in \mathcal{B}$ is a spatial region in I (concretely, an axis-aligned bounding box or a soft attention mask over the image tokens), $v_i \in \mathcal{V}$ is a structured tuple of visual attributes extracted from the sub-image delimited by b_i and $s_i \in \mathcal{S}$ is a reasoning slot in the template obtained by decomposing question q into K sub-questions, each assigned exactly one anchor node.

The sets $\mathcal{B}, \mathcal{V}, \mathcal{S}$ are thus not abstract: $\mathcal{B}$ is determined by the image resolution and localization mechanism; $\mathcal{V}$ is determined by the task-specific attribute vocabulary; $\mathcal{S}$ is determined by the question decomposition template. An *anchor chain* $C = (a_1, \dots, a_K)$ is an ordered sequence of anchor nodes produced before $\hat{y}$. Each node is independently checkable: one can verify whether b_i supports v_i and whether v_i is consistent with s_i , without re-running any other step. The three columns of Figure 1 correspond exactly to errors in b_i, v_i , and s_i respectively—each localizable, each correctable through its own entry point. Section 3 addresses both levels in sequence.

3 The PERCEPTION CHAINS Framework

Figure 2 illustrates the complete pipeline. *Anchored Visual Evidence* (AVE, §3.1) makes every grounding decision checkable at the output level; the *Stability Loss* (§3.2) constrains hidden-state geometry at anchor positions, enforcing transversal contraction that standard fine-tuning does not produce in our experiments;

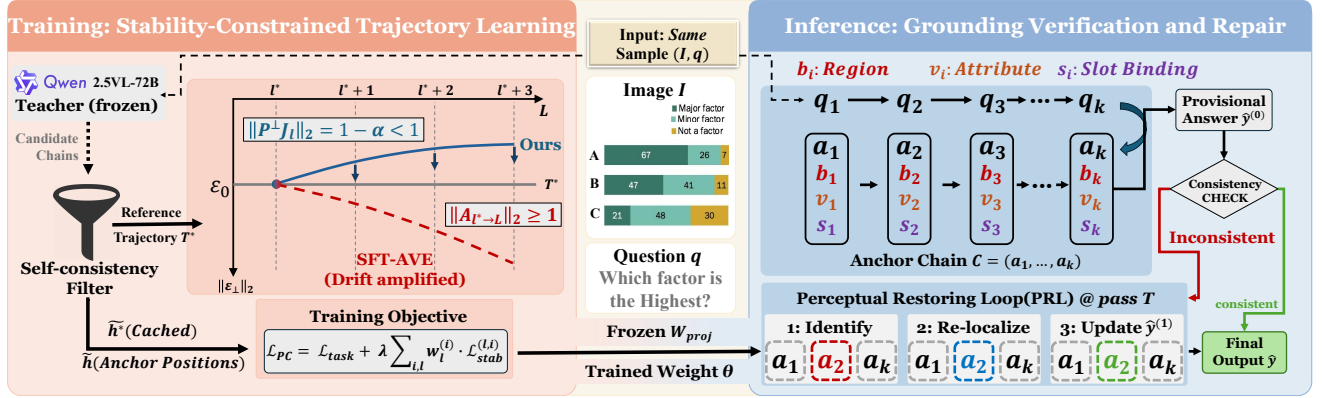


Figure 2: PERCEPTION CHAINS overview. Left (Training): A frozen 72B teacher’s self-consistent trajectories are cached offline. The stability loss drives $\|P_{\perp}^L J_L\|_2 < 1$ at each active anchor layer (blue); SFT-AVE leaves $\|A_{l^* \rightarrow L}\|_2 \geq 1$ throughout training (red, drift amplified). Right (Inference): AVE assembles anchor chain $C=(a_1, \dots, a_K)$; PRL flags inconsistent anchors I and re-localizes before committing the final answer.

Confidence-Weighted Trajectory Bootstrapping (CTB, §3.3) supplies high-quality teacher trajectories from a frozen 72B VLM without human annotation; and the *Perceptual Restoring Loop* (PRL, §3.4) corrects residual evidence-answer inconsistency at inference time without retraining. CTB caching is entirely offline; the stability loss is a training-time objective; PRL operates on the finished model with no access to teacher states.

3.1 Anchored Visual Evidence

Section 2 defines the anchor triple $a_i = (b_i, v_i, s_i)$ as a structural requirement; here we describe its implementation.

Question decomposition and slot template. A fixed instruction prefix prompts the model to decompose question q into K sub-questions $q_1, \dots, q_K$, each assigned one reasoning slot $s_i \in \mathcal{S}$. The slot schema $\mathcal{S}$ is task-specific (full schemas in supplementary, Table S2). The number of slots is fixed at $K=4$ (Table S3).

Region localisation (b_i). For sub-question q_i , the model produces a bounding box $b_i = [x_1, y_1, x_2, y_2]$ as four normalised coordinates in $[0, 1000]$, following the Qwen2.5-VL format [2]. The box token appears immediately after the slot label and *before* the attribute value token; this ordering ensures spatial commitment precedes any attribute extraction.

Attribute extraction (v_i). The model extracts a structured attribute tuple $v_i \in \mathcal{V}$ from within region b_i ; the schema is slot-specific and enforced by the instruction template via constrained decoding.

Chain assembly. The extracted v_i is declared to populate slot s_i , completing triple $a_i = (b_i, v_i, s_i)$. The chain $C = (a_1, \dots, a_K)$ is assembled in order; the final answer $\hat{y}$ is generated from C and the full image. Each node is independently checkable: one can verify whether b_i supports v_i and whether v_i is consistent with s_i without re-running any other step. This verifiability is a structural property of the output format; it holds regardless of whether the underlying representations are well-grounded. The gap that the format leaves open is addressed next.

3.2 Stability Loss

Even when the model produces a correctly structured C , the hidden states at anchor positions can drift away from the teacher’s

trajectory across layers. Standard SFT minimises $\mathcal{L}_{task}$ at the final answer token and does not explicitly constrain hidden-state geometry at intermediate anchor positions. We establish this decoupling empirically by measuring the cosine similarity between $\nabla_{\theta} \mathcal{L}_{task}$ and $\nabla_{\theta} \mathcal{L}_{stab}$ at anchor positions versus at the final answer token; the similarity at anchor positions is near zero and statistically non-significant, showing that this decoupling persists throughout the measured training trajectory (full gradient statistics in Section 4.4).

Transversal subspace and composite operator. The *grounding-relevant transversal subspace* at layer l , anchor i is the orthogonal complement of the teacher’s reasoning direction $u_l^{*(i)}$ (the unit vector along $\tilde{\Delta}_l^{*(i)} / \|\tilde{\Delta}_l^{*(i)}\|_2$, i.e. the normalised teacher residual update at anchor i , layer l): with projector $P_{\perp}^{(l,i)} = I - u_l^{*(i)} (u_l^{*(i)})^{\top}$. Concretely, $\tilde{\Delta}_l^{*(i)} = W_{proj} (h_{l+1}^{*(i)} - h_l^{*(i)})$ is the projected teacher residual between layers l and $l+1$, read directly from the cached trajectory T^* . Transversal deviations constitute pure grounding error; deviations along $u_l^{*(i)}$ are aligned with the teacher and are not penalised. The *composite transported transversal operator*

$$A_{l^* \rightarrow L} \triangleq P_{\perp}^{(L)} J_{L-1} P_{\perp}^{(L-1)} \dots J_{l^*} P_{\perp}^{(l^*)} \quad (1)$$

tracks how a transversal perturbation at the first active anchor layer l^* propagates through all subsequent layers. When $\|A_{l^* \rightarrow L}\|_2 \geq 1$, a direction exists in $\mathcal{T}^{(l^*, i)}$ that no subsequent layer attenuates, permanently displacing the student’s hidden state from the teacher trajectory (proof in supplementary). Section 4.4 reports that vanilla Qwen2.5-VL-7B and SFT-AVE both leave more than 99% of calibration inputs in the non-attenuating regime.

Loss definition. Let $\tilde{h}_l^{(i)} \in \mathbb{R}^{d_s}$ denote the student’s post-LayerNorm hidden state at the first region-coordinate token of anchor a_i^* , at layer l ($d_s=3584$ for Qwen2.5-VL-7B). Let $\tilde{h}_l^{*(i)} = W_{proj} h_l^{*(i)} \in \mathbb{R}^{d_s}$ be the teacher’s projected state, where $W_{proj} \in \mathbb{R}^{d_s \times d_t}$ ($d_t=7168$, Qwen2.5-VL-72B) is fitted by PCA on 1,000 samples drawn from the training distribution and set aside before training begins (excluded from all gradient updates); W_{proj} is frozen throughout. Specifically, W_{proj} retains the top- d_s principal components of the teacher’s

anchor-position hidden states $\{h_l^{*(i)}\}$ collected on the held-out calibration set. For perturbation $\delta \sim \mathcal{N}(0, \sigma^2 I)$, $\sigma=0.01$:

$$\mathcal{L}_{\text{stab}}^{(l,i)} = \mathbb{E}_{\delta} \left\| P_{\perp}^{(l,i)} (\tilde{\Delta}_l^{(i),\delta} - \tilde{\Delta}_l^{*(i)}) + \alpha P_{\perp}^{(l,i)} \delta \right\|^2 \quad (2)$$

where $\tilde{\Delta}_l^{(i),\delta} = f_{\theta}^{(l)}(\hat{h}_l^{(i)} + \delta)$ is the student's perturbed residual update and $\tilde{\Delta}_l^{*(i)} = W_{\text{proj}} f_{\theta^*}^{(l)}(h_l^{*(i)})$ is the teacher's projected residual update. Under first-order Taylor expansion ($\sigma \|J_f\|_2 \ll 1$, verified by $\sigma=0.01$), Eq. (2) decomposes into a *drift term* $\|P_{\perp}^{(l,i)} r_l^{(i)}\|^2$ and a *spectral term* $\sigma^2 \|P_{\perp}^{(l,i)} (J_f^{(l,i)} + \alpha I)\|_F^2$, penalising the transversal Jacobian norm toward $1-\alpha$.

THEOREM 3.1 (STRICT TRANSVERSAL CONTRACTION). *Under the single-token independence assumption ($\beta \triangleq \|J_{\text{off}}\|_F / \|J_{\text{diag}}\|_F \ll 1$; measured 0.083 ± 0.029), if $\mathcal{L}_{\text{stab}}^{(l,i)} = 0$ with $\sigma > 0$ and $\alpha \in (0, 1)$, then $\|P_{\perp}^{(l,i)} J_f^{(l,i)}\|_2 = 1 - \alpha < 1$.*

Proof and the finite-loss extension are in the supplementary; $\alpha=0.10$ is selected by the α -sweep (supplementary Table S1).

Overall training objective.

$$\mathcal{L}_{\text{PC}} = \mathcal{L}_{\text{task}} + \lambda \sum_{i=1}^K \sum_{l \in \mathcal{L}_{\text{active}}^{(i)}} w_l^{(i)} \cdot \mathcal{L}_{\text{stab}}^{(l,i)} \quad (3)$$

where $\lambda=1.0$. Layer l is *active* for anchor i if the teacher update magnitude $\|\tilde{\Delta}_l^{*(i)}\|_2$ exceeds the 5th percentile of $\{\| \cdot \|_2\}_{l=1, \dots, L}$ for that anchor on the calibration set ($w_l^{(i)}=1$; else 0). Section 4.4 reports the operator norm distributions before and after training.

3.3 Confidence-Weighted Trajectory Bootstrapping

The quality ceiling of the stability loss is set by the quality of T^* : the same objective applied to a misgrounded trajectory produces a well-constrained but incorrectly oriented student. CTB constructs T^* from a frozen 72B VLM without human annotation. All CTB computation is offline: teacher inference and hidden-state caching complete before student training begins.

Phase 1: Cold start. A frozen Qwen2.5-VL-72B teacher generates $k=8$ anchor chain candidates per sample for seed set $\mathcal{D}_{\text{cold}}$ ($|\mathcal{D}_{\text{cold}}|=5 \times 10^3$, $\tau_{\text{samp}}=0.7$, top- $p=0.9$). A chain is retained only if all k samples agree on the final answer via exact match. Hidden states of retained chains are cached (≈ 110 GB, float16) before training begins.

Phase 2: Self-bootstrap. Starting from the Phase 1 checkpoint, the 7B student generates $k=8$ chains per sample ($\tau_{\text{samp}}=0.7$) and scores each by:

$$\text{conf}(C_j) = \bar{a}(C_j) \cdot \text{crr}(C_j) \cdot \text{tsc}(C_j) \quad (4)$$

where $\text{crr} \equiv \text{correct_rate}$ and $\text{tsc} \equiv \text{teacher_score}$, defined as:

- $\bar{a}(C_j) = \frac{1}{K} \sum_{i=1}^K (1 - H_i / \log |V_i|)$: normalised slot-level anchor agreement (H_i = empirical entropy of k sampled values at slot i).
- $\text{crr}(C_j) = 1[\hat{y}(C_j)=y^*]$: ground-truth match; 1.0 for unlabelled samples.
- $\text{tsc}(C_j) \in \{0, 1\}$: 72B verifier's binary judgment on anchor-region correctness (prompt template in supplementary).

Chains with $\text{conf}(C_j) < \tau_{\text{conf}}=0.5$ are excluded from the stability loss. Phase 2 runs for two rounds. In round 1, the student trains on a merged pool of Phase 1 seed trajectories and its own filter-passing chains from the current round. Round 2 re-generates chains from the updated checkpoint, applies the same filter, and incorporates $\approx 3 \times 10^4$ additional trajectories. The three factors in Eq. (4) are ablated in Section 4.5 (A4c, A4b, A4).

3.4 Perceptual Restoring Loop

Even after training, individual inference passes may produce anchor chains inconsistent with the provisional answer. PRL detects and corrects this residual in-pass without access to teacher hidden states and without retraining.

Algorithm.

- (1) **Vote.** Sample $N_{\text{vote}}=4$ chains at $\tau_{\text{samp}}=0.7$; take the majority-vote answer as $\hat{y}^{(0)}$.
- (2) **Select.** Choose the chain with the highest $\bar{a}(C)$ consistent with $\hat{y}^{(0)}$.
- (3) **Identify.** Check chain consistency with $\hat{y}^{(0)}$ (criterion below). If inconsistent, identify $\mathcal{I} \subseteq \{1, \dots, K\}$ —slots where v_i contradicts $\hat{y}^{(0)}$. If consistent, output $\hat{y}^{(0)}$ directly.
- (4) **Repair.** Re-localise only anchors in $\mathcal{I}$, using $\hat{y}^{(0)}$ as prior, to produce a corrected chain and $\hat{y}^{(1)}$.

Steps (3)–(4) repeat for at most $T_{\text{max}}=2$ iterations.

Consistency criterion. The criterion is matched to the answer space of each task. For closed-ended tasks (ChartQA, POPE, RSVQA), where the answer space is discrete and fixed, consistency is exact string match between the derived anchor value and the claimed answer. For open-ended tasks (DocVQA), where correct answers vary in surface form, consistency is NLI entailment scored by DeBERTa-v3-large [5]; inconsistency is declared when the contradiction label probability exceeds 0.5 (checkpoint details in supplementary).

Scope. The check is sufficient, not necessary: coincidentally consistent wrong chains are not detected. Section 4.5 reports trigger rate, correction rate, and comparisons against exhaustive re-localization (A2c) and random selection (A2b). Removing PRL (A2) raises hallucination rate while leaving grounding accuracy nearly unchanged, confirming that PRL addresses output-level inconsistency rather than the localization precision that training determines.

4 Experiment

The experiments address four questions: whether explicit grounding commitments are *necessary* (Stage 1); whether anchor-format supervision alone is *sufficient* to suppress hidden-state drift (Stage 2); whether CTB and PRL contribute *independently* (Stage 3); and whether gains *generalize* across architectures and unseen task families (Stage 4). Each subsection states the prediction implied by an affirmative answer, then presents the evidence.

4.1 Setup

Table 1 summarizes the evaluation benchmarks and comparison methods. Evaluation is separated into a *core* suite of eight seen benchmarks used for both training and in-distribution testing, and an *extended* suite of seven unseen benchmarks held out entirely and evaluated zero-shot. **Backbone and training data.** All methods

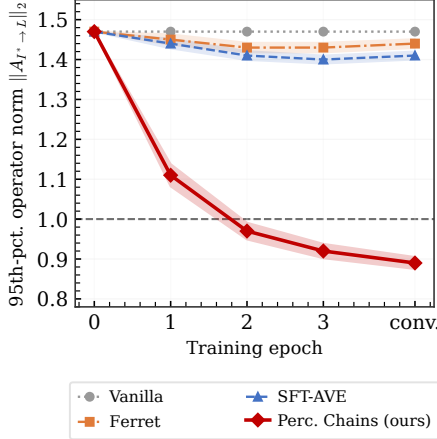


Figure 3: Training dynamics of the 95th-percentile composite transversal operator norm $\|A_{I^* \rightarrow L}\|_2$.

Table 1: Experimental setup.

(a) Benchmarks

Regime	Capability	Benchmark	Metric
Seen	Chart understanding	ChartQA [10]	Acc. $\uparrow$
	Document QA	DocVQA [11]	ANLS $\uparrow$
	Remote-sensing QA	RSVQA [9]	Acc. $\uparrow$
	Remote-sensing grounding	DIOR-RSVG [24]	GA@IoU $\geq 0.5 \uparrow$
	Multi-image QA	MuMuQA [14]	Acc. $\uparrow$
	Multi-image reasoning	MMIU [12]	Acc. $\uparrow$
	Object hallucination	POPE [7]	HR $\downarrow$
	Visual hallucination	HallusionBench [4]	HR $\downarrow$
Unseen	General VQA	GQA [1]	Acc. $\uparrow$
	Scene-text VQA	TextVQA [18]	Acc. $\uparrow$
	Diagram reasoning	AI2D [6]	Acc. $\uparrow$
	Object hallucination	Object HalBench [15]	F1 $\uparrow$
	Referring grounding	RefCOCO [22]	Acc@IoU $\geq 0.5 \uparrow$
	Phrase grounding	Flickr30k [13]	Recall@1 $\uparrow$
	Hallucination quality	Hallucinogen [16]	Avg. $\uparrow$

(b) Baselines

Category	Method	Role
Base VLM	Vanilla Qwen2.5-VL-7B	Base reference
CoT reasoning	Visual CoT [17]	Reasoning comparator
Spatial grounding	Ferret [21]	Grounding comparator
Hallucination	TPO [8]	Hallucination comparator
Post-hoc	Woodpecker [20]	Post-hoc verifier
Ablation control	SFT-AVE	Same format, no stab. loss

are fine-tuned from the same **Qwen2.5-VL-7B** [2] checkpoint using the official training splits of the eight core benchmarks. To avoid over-representing any single task, we construct a balanced mixture of 1.25×10^4 instances per benchmark, yielding 10^5 training examples in total. The seven extended benchmarks are never seen during training and are used solely for zero-shot evaluation.

Baseline re-implementation. All training-time baselines are ported to Qwen2.5-VL-7B under the same data mixture and schedule; observed differences therefore reflect the training objective rather than backbone capacity or data volume. Woodpecker [20] requires no re-training and is applied post-hoc. Per-method details are in the supplementary.

CTB trajectory source. For the full method, anchor-chain candidates are generated offline by a frozen **Qwen2.5-VL-72B** teacher. The teacher is used only to construct the training cache and is absent during student inference and PRL. We additionally evaluate a self-distilled variant in which the 7B model generates its own reference trajectories and no 72B model is used at any stage.

Implementation details. All baselines use AdamW (lr 2×10^{-5} , cosine decay, effective batch 256, $8 \times A100$) for three epochs. For PERCEPTION CHAINS: $\lambda=1.0$, $\sigma=0.01$, $\alpha=0.10$; W_{proj} frozen after PCA on 1,000 held-out calibration samples. All evaluations use greedy decoding; sensitivity analyses for λ , σ , α , τ_{conf} are in the supplementary. Results are means over seeds $\{42, 0, 2\}$ with a shared teacher cache; the primary paired GA comparison is significant at $p < 0.01$.

4.2 Metrics

Three metrics appear across ablation comparisons. **Acc**: macro-average over the eight seen benchmarks (relaxed $\pm 5\%$ for ChartQA, ANLS for DocVQA, standard accuracy otherwise). **GA**: grounding accuracy on DIOR-RSVG, $\text{Acc@IoU} \geq 0.5$, reported only for methods producing bounding-box outputs. **HR**: hallucination rate $1-s$ on POPE ($s = F1$). Extended benchmarks use metrics as in Table 1.

4.3 Stage 1

Prediction. Removing anchor commitments (A3) should collapse spatial accuracy; text-only methods should show near-zero spatial gains; gains should concentrate where region-attribute binding is most consequential.

Evidence. A3 (text CoT only) eliminates grounding accuracy and produces the highest hallucination rate among all trained variants, worse than the vanilla baseline on spatial tasks. Visual CoT [17], despite carefully structured reasoning steps, shows negligible spatial gains on both seen and unseen benchmarks: text structure cannot substitute for explicit region binding. The spatial-demand gradient in Tables 2 and 3 provides the positive evidence: gains are largest on remote-sensing grounding and multi-image reasoning, and smallest where text statistics dominate—the expected signature of a region-attribute binding mechanism.

4.4 Stage 2

Stage 2 asks whether anchor-format supervision alone is sufficient to suppress hidden-state drift.

Prediction. A model with the identical anchor format but without the stability loss should show persistent non-attenuating drift and substantially lower grounding accuracy despite producing identically structured outputs.

Evidence. Table 4, Table 5, and Figure 3 provide three levels of evidence.

At the output level, removing the stability loss while keeping the identical anchor format (A1, SFT-AVE) drops grounding accuracy by more than nine points on DIOR-RSVG and raises hallucination rate substantially. At the representation level, Table 5 shows format training alone barely moves the operator norm: over 99% of inputs still exhibit non-attenuating drift after SFT-AVE, nearly identical to the pre-trained model. Figure 3 confirms this holds throughout training; the stability loss collapses the fraction above 1.0 to under 1%.

Table 2: In-distribution results on the eight seen benchmarks. AVE: anchor-visual evidence triple (named region, extracted attribute, answer contribution). A-IDs cross-reference the ablation tables.

ID	Category	Method	ChartQA $\uparrow$	DocVQA $\uparrow$	RSVQA $\uparrow$	DIOR-RSVG $\uparrow$	MuMuQA $\uparrow$	MMIU $\uparrow$	POPE HR $\downarrow$	HalBench HR $\downarrow$
	Base	Vanilla Qwen2.5-VL-7B	88.6	95.9	87.4	N/A	66.8	57.9	12.8	47.1
	CoT	Visual CoT [17]	88.8	95.6	87.6	N/A	67.2	58.1	12.2	45.8
	Grounding	Ferret [21]	88.7	95.5	88.7	61.7	68.8	59.6	11.2	44.8
	Hallucination	TPO [8]	88.4	95.7	87.3	N/A	66.7	57.8	9.2	33.2
	Post-hoc	Woodpecker [20]	88.6	95.9	87.4	N/A	66.8	57.9	9.8	45.8
A1	Ablation	SFT-AVE	89.8 $\pm$ 0.4	96.2 $\pm$ 0.3	89.5 $\pm$ 0.6	63.4 $\pm$ 0.9	71.2 $\pm$ 0.6	61.5 $\pm$ 0.7	10.2 $\pm$ 0.3	43.1 $\pm$ 1.4
A3	Ablation	w/o AVE (text CoT only)	88.1	95.4	87.2	N/A	66.3	57.6	11.6	46.2
	Ours	PERCEPTION CHAINS	91.4$\pm$0.3	96.8$\pm$0.2	92.6$\pm$0.5	72.6$\pm$0.7	74.3$\pm$0.5	64.8$\pm$0.6	7.3$\pm$0.2	31.8$\pm$1.2

Table 3: Zero-shot results on the seven unseen benchmarks.

Category	Method	GQA $\uparrow$	TextVQA $\uparrow$	AI2D $\uparrow$	ObjHal F1 $\uparrow$	Hallucinogen $\uparrow$	RefCOCO $\uparrow$	Flickr30k $\uparrow$
Base	Vanilla Qwen2.5-VL-7B	63.4	73.8	79.3	85.7	35	N/A	N/A
CoT	Visual CoT [17]	63.8	74.1	79.6	87.1	34	N/A	N/A
Grounding	Ferret [21]	63.1	73.5	79.1	87.5	36	63.0	67.0
Halluc.	TPO [8]	64.0	74.1	79.6	91.2	43	N/A	N/A
Post-hoc	Woodpecker [20]	63.5	73.9	79.3	89.2	36	N/A	N/A
Ablation	SFT-AVE	64.9	75.0	80.6	90.1	44	58.0	63.0
Ours	PERCEPTION CHAINS	66.1	76.9	82.1	92.3	51	66.0	70.0

Table 4: Stage 2 ablation: effect of the stability loss and distillation geometry.

ID	Variant	ChartQA $\uparrow$	DocVQA $\uparrow$	RSVQA $\uparrow$	DIOR-RSVG $\uparrow$	MuMuQA $\uparrow$	MMIU $\uparrow$	POPE HR $\downarrow$	HalBench HR $\downarrow$
	PERCEPTION CHAINS	91.4	96.8	92.6	72.6 $\pm$ 0.7	74.3	64.8	7.3 $\pm$ 0.2	31.8
A1	w/o stability loss (SFT-AVE)	89.8	96.2	89.5	63.4 $\pm$ 0.9	71.2	61.5	10.2 $\pm$ 0.3	43.1
A0	Full-HS distill (no $P_{\perp}$)	90.4	96.5	90.8	67.8	72.4	62.8	9.1	38.2

Table 5: Composite transversal operator norm $\|A_{l^* \rightarrow L}\|_2$ before and after training, where l^* is the first anchor layer and L the final layer (1,000 calibration inputs, 200 per task family). Values above 1 indicate non-attenuating drift.

Checkpoint	Mean	Median	95th pct	Frac. > 1
Vanilla (pre-trained)	1.32	1.31	1.47	>99%
SFT-AVE (format, no stab. loss)	1.26	1.25	1.41	>99%
PERCEPTION CHAINS	0.79	0.78	0.89	<1%

The gradient-level picture closes the argument. The cosine similarity between $\nabla_{\theta} \mathcal{L}_{\text{task}}$ and $\nabla_{\theta} \mathcal{L}_{\text{stab}}$ at anchor positions is 0.031 ± 0.019 ($p=0.14$, non-significant) vs. 0.183 ± 0.041 ($p<0.01$) at the final answer token, localizing the decoupling specifically to anchor positions.

A0 tests geometric specificity: replacing $P_{\perp}$ with full hidden-state L2 distillation recovers 4.4 of the 9.2-point gap, but the remaining 4.8 points show that constraining irrelevant directions imposes a cost that geometric specificity avoids.

4.5 Stage 3

Stage 3 tests whether CTB and PRL address genuinely distinct failure modes and are each independently necessary.

Prediction. Replacing CTB with random trajectories should sharply degrade grounding accuracy while leaving format and loss unchanged. Removing PRL should raise hallucination rate while leaving grounding accuracy nearly unchanged.

Evidence. Table 6 bears out both predictions.

CTB trajectory quality (upper group). The four variants form a monotone quality ladder. A4c removes the slot-level anchor-agreement filter while retaining the 72B teacher: the 71.1 vs. 72.6 GA gap shows the filter adds value beyond teacher capacity alone. A4b (STaR-style, answer-filtered only) recovers fewer errors (68.7 GA) because answer-correct chains can still contain misgrounded intermediate anchors. A4 (random trajectories) collapses accuracy by more than thirteen points while format and loss remain unchanged, confirming the constraint quality is bounded by trajectory quality. *Dependence on the 72B teacher.* When the 72B teacher is removed entirely, the 7B self-distilled variant reaches 70.2 GA, 7.9 POPE HR, and 37.5 HallusionBench HR. It therefore retains 6.8 of the full method’s 9.2-point GA improvement over SFT-AVE, or approximately 74%. The stability objective thus remains effective without a larger teacher, while the remaining 2.4-point difference is consistent with the higher quality of the 72B reference trajectories. *PRL inference-time mechanism (lower group).* A2c (re-localize all K anchors) achieves marginally lower HR (7.0 vs. 7.3) at roughly twice the overhead, confirming the correction is valid in principle. A2b

Table 6: Stage 3 ablation: CTB trajectory quality (upper group) and PRL inference-time correction (lower group).

ID	Variant	ChartQA↑	DocVQA↑	RSVQA↑	DIOR-RSVG↑	MuMuQA↑	MMIU↑	POPE HR↓	HalBench HR↓
	PERCEPTION CHAINS (full)	91.4	96.8	92.6	72.6$\pm$0.7	74.3	64.8	7.3$\pm$0.2	31.8
<i>CTB trajectory source — stability loss and format unchanged</i>									
A4c	72B unfiltered (no quality filter)	91.0	96.6	91.8	71.1	73.8	64.3	7.9	33.5
A4b	STaR (answer-filtered; no anchor agreement)	90.7	96.4	91.1	68.7	73.4	64.0	8.5	36.2
A4	Random trajectories	89.1	95.8	88.6	59.1	68.9	59.4	10.8	44.6
<i>PRL inference-time mechanism — training unchanged</i>									
A2c	PRL-all (re-localize all K anchors)	91.4	96.8	92.6	72.7	74.3	64.8	7.0	30.8
A2b	PRL-random (same $ I $, random selection)	91.3	96.7	92.4	72.2	74.1	64.6	8.4	35.4
A2	w/o PRL	91.3	96.7	92.4	71.8	74.1	64.6	9.4	38.4

Table 7: Cross-architecture results.

Architecture	Pre-norm [†]	β	Δ CQA↑	Δ HR↓	Δ GA↑	CF
Qwen2.5-VL-7B	1.47	0.083	+2.8	−5.5	+9.2	0.903 $\pm$ 0.041
InternVL2-8B	1.44	0.079	+2.6	−4.8	+8.4	0.908 $\pm$ 0.038
LLaVA-NeXT-8B	1.52	0.091	+3.1	−5.9	+8.6	0.897 $\pm$ 0.044
Idefics3-8B	1.58	0.097	+3.4	−6.2	+9.4	0.891 $\pm$ 0.048

[†] 95th-percentile $\|A_{I^* \rightarrow L}\|_2$ on 1,000 calibration inputs.

(same $|I|$ anchors, random selection) recovers far fewer cases (8.4 HR), showing the value of PRL lies in identifying *which* anchors are inconsistent, not in re-localizing per se. A2 (no PRL) raises hallucination rate while leaving grounding accuracy nearly unchanged, confirming that PRL corrects output-level inconsistency rather than localization quality.

4.6 Stage 4

Prediction. If drift is a general transformer property, pre-training operator norm should exceed 1.0 across architectures, gains should be monotone in pre-norm, and the contraction factor should converge to $1 - \alpha = 0.90$ everywhere. If gains reflect representation improvement, they should transfer with the stability loss as the transferable component.

Architecture generalization (Table 7). Pre-training operator norm exceeds 1.0 in every evaluated architecture, showing that the observed drift is not specific to Qwen2.5-VL. Grounding gains are monotone in pre-norm, and the post-training contraction factor converges to target across all four designs. The β column ($\beta = \|J_{\text{off}}\|_F / \|J_{\text{diag}}\|_F$) remains below 0.10 everywhere, supporting the diagonal-block approximation across the evaluated architectures.

Task-distribution transfer (Table 8). All Δ Acc values use same-setting vanilla baselines. Set A contains partially overlapping task families, Set B held-out distributions of seen task types, and Set C unseen task types. SFT-AVE+PRL isolates PRL’s contribution from that of the stability loss.

SFT-AVE and SFT-AVE+PRL transfer nearly identically, whereas PERCEPTION CHAINS retains positive gains through Set C. This isolates the stability loss, rather than PRL or the anchor format, as the transferable component. Without referring- or phrase-grounding training, it also exceeds Ferret on RefCOCO (66.0 vs. 63.0) and Flickr30k (70.0 vs. 67.0).

Table 8: Out-of-distribution transfer.

Method	In-dist	Set. A	Set. B	Set. C
Visual CoT [17]	+0.1	−0.1	−0.2	+0.1
Woodpecker [20]	−0.2	−0.3	−0.4	−0.1
Ferret [21]	+0.8	+0.6	+0.4	+0.3
SFT-AVE	+1.5	+0.9	+0.7	+0.6
SFT-AVE + PRL	+1.7	+1.0	+0.8	+0.7
PERCEPTION CHAINS	+3.4	+2.2	+1.7	+1.4

5 Analysis

The four stages of Section 4 showed that the mechanism works. What they do not explain is why the gains concentrate where they do, what they cost in practice, and what the method cannot fix even when everything else is working correctly. These questions are taken up in turn.

5.1 Why Spatial Tasks Benefit Most

Tables 2 and 3 show that gains on Remote Sensing (+5.2 on RSVQA) are more than five times those on Document Understanding (+0.9 on DocVQA), with Multi-Image Reasoning and Chart QA falling between them. Section 4.3 attributed this disparity to the role of explicit region-attribute binding; Section 4.6 showed it persists under out-of-distribution transfer. Neither result explains why the ordering follows this particular sequence or why the ratio is this large.

Theorem 3.1 offers a prediction: the stability loss yields larger returns on inputs where non-attenuating drift is already more severe before training, because those are the inputs where the correction is most needed. Table 9 tests this prediction directly by grouping the operator norm measurements (same 1,000-sample calibration set as Table 5, 200 inputs per family) by task family alongside the observed accuracy gain.

The pre-norm ordering matches the accuracy-gain ordering exactly: Remote Sensing highest (1.41 $\pm$ 0.09), Document lowest (1.24 $\pm$ 0.06). Tasks where text-based pre-training provides little signal accumulate more drift, leaving more for the stability loss to correct; the disparity in the result tables reflects this pre-existing condition, not a side effect of the method. Pre-norm is therefore measurable before fine-tuning begins and serves as a predictor of per-family gain.

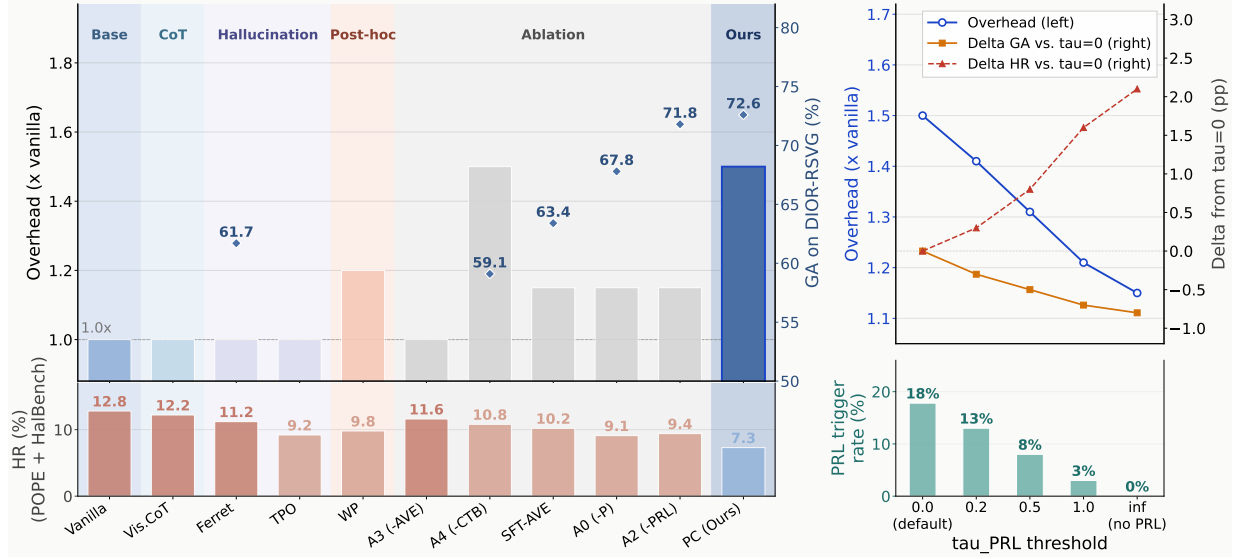
Figure 4: Inference overhead and accuracy across all methods and ablations (left), and the accuracy–overhead trade-off as τ_{PRL} varies (right).

Table 9: Pre-training operator norm by task family.

Task family	Pre-norm (mean $\pm$ std)	95th pct
Remote Sensing	1.41 $\pm$ 0.09	1.56
Multi-Image	1.34 $\pm$ 0.08	1.47
Chart QA	1.31 $\pm$ 0.07	1.43
Hallucination	1.28 $\pm$ 0.07	1.40
Document	1.24 $\pm$ 0.06	1.34

Table 11: PRL trigger rate and total overhead by task family.

Task family	PRL rate	AVE only	Total
Remote Sensing	24%	1.15 $\times$	1.62 $\times$
Multi-Image	20%	1.15 $\times$	1.55 $\times$
Hallucination	16%	1.15 $\times$	1.49 $\times$
Chart QA	15%	1.15 $\times$	1.48 $\times$
Document	13%	1.15 $\times$	1.44 $\times$
Average (all)	17.8%	1.15 $\times$	1.50 $\times$

Table 10: Accuracy–overhead trade-off as τ_{PRL} is varied.

τ_{PRL}	Trigger rate	Overhead	GA $\uparrow$	HR $\downarrow$
0.0 (default)	17.8%	1.50 $\times$	72.6	7.3
0.2	13%	1.41 $\times$	72.3	7.6
0.5	8%	1.31 $\times$	72.1	8.1
1.0	3%	1.21 $\times$	71.9	8.9
∞ (no PRL)	0%	1.15 $\times$	71.8	9.4

5.2 Inference Cost and Deployment Flexibility

Training adds 1.44 $\times$ the compute of vanilla fine-tuning (69 vs. 48 GPU-hours). Figure 4 breaks down the 1.50 $\times$ average inference overhead. A4, SFT-AVE, and PERCEPTION CHAINS share the same data and anchor format; their GA values are 59.1, 63.4, and 72.6 at overheads of 1.15 $\times$, 1.15 $\times$, and 1.50 $\times$ —the 9.2-point GA gap costs 0.35 $\times$, of which the main return is a 2.1-point HR reduction (PRL corrects evidence-answer inconsistency, not localization precision). Raising τ_{PRL} to 0.5 reduces overhead to 1.31 $\times$ with only 0.5 pp GA loss (Table 10), a flexibility no fixed-overhead baseline can match.

The AVE-only cost of 1.15 $\times$ is constant across all families; the remainder reflects the PRL trigger rate (Table 11), which follows the same ordering as pre-training operator norm—tasks with more drift also produce more evidence-answer inconsistency at inference. 82.2% of samples incur zero PRL overhead at the default threshold.

At $\tau_{PRL} = \infty$, the model retains 71.8% GA without PRL, confirming that the stability loss drives the representational gain. PRL cannot recover errors that produce no inconsistency signal.

6 Limitations and Future Work

Three classes of errors in the framework cannot be reached, each for a distinct structural reason. For Remote Sensing, PRL triggers at the highest rate (24%) but a portion of re-localizations still fail to find the correct region after triggering, exposing a ceiling imposed by the region-proposal mechanism rather than the consistency check itself. For Chart QA, the model localizes the correct region but extracts the wrong numerical value; because the extracted attribute is internally plausible, no evidence-answer inconsistency arises and PRL has no signal to act on for this error class despite a 15% overall trigger rate. For Document understanding, slot-assignment errors are internally consistent by construction—region correct, attribute correct, but bound to the wrong slot—which directly accounts for the lowest trigger rate (13%) and smallest gain (+0.9) of any task family.

References

- [1] Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. 2023. GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints. arXiv:2305.13245 [cs.CL] <https://arxiv.org/abs/2305.13245>
- [2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-VL Technical Report. arXiv:2502.13923 [cs.CV] <https://arxiv.org/abs/2502.13923>
- [3] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023. Shikra: Unleashing Multimodal LLM’s Referential Dialogue Magic. arXiv:2306.15195 [cs.CV] <https://arxiv.org/abs/2306.15195>
- [4] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. HallusionBench: An Advanced Diagnostic Suite for Entangled Language Hallucination and Visual Illusion in Large Vision-Language Models. arXiv:2310.14566 [cs.CV] <https://arxiv.org/abs/2310.14566>
- [5] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. arXiv:2006.03654 [cs.CL] <https://arxiv.org/abs/2006.03654>
- [6] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A Diagram Is Worth A Dozen Images. arXiv:1603.07396 [cs.CV] <https://arxiv.org/abs/1603.07396>
- [7] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating Object Hallucination in Large Vision-Language Models. arXiv:2305.10355 [cs.CV] <https://arxiv.org/abs/2305.10355>
- [8] Weinbin Liao, Xu Chu, and Yasha Wang. 2025. TPO: Aligning Large Language Models with Multi-branch Multi-step Preference Trees. arXiv:2410.12854 [cs.CL] <https://arxiv.org/abs/2410.12854>
- [9] Sylvain Lobry, Diego Marcos, Jesse Murray, and Devis Tuia. 2020. RSVQA: Visual Question Answering for Remote Sensing Data. *IEEE Transactions on Geoscience and Remote Sensing* 58, 12 (Dec. 2020), 8555–8566. doi:10.1109/tgrs.2020.2988782
- [10] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. arXiv:2203.10244 [cs.CL] <https://arxiv.org/abs/2203.10244>
- [11] Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021. DocVQA: A Dataset for VQA on Document Images. arXiv:2007.00398 [cs.CV] <https://arxiv.org/abs/2007.00398>
- [12] Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. 2024. MMIU: Multimodal Multi-image Understanding for Evaluating Large Vision-Language Models. arXiv:2408.02718 [cs.CV] <https://arxiv.org/abs/2408.02718>
- [13] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2016. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. arXiv:1505.04870 [cs.CV] <https://arxiv.org/abs/1505.04870>
- [14] Revanth Gangi Reddy, Xilin Rui, Manling Li, Xudong Lin, Haoyang Wen, Jaemin Cho, Lifu Huang, Mohit Bansal, Avirup Sil, Shih-Fu Chang, Alexander Schwing, and Heng Ji. 2022. MuMuQA: Multimedia Multi-Hop News Question Answering via Cross-Media Knowledge Extraction and Grounding. arXiv:2112.10728 [cs.CL] <https://arxiv.org/abs/2112.10728>
- [15] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object Hallucination in Image Captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun’ichi Tsuiji (Eds.). Association for Computational Linguistics, Brussels, Belgium, 4035–4045. doi:10.18653/v1/D18-1437
- [16] Ashish Seth, Dinesh Manocha, and Chirag Agarwal. 2025. HALLUCINOGEN: Benchmarking Hallucination in Implicit Reasoning within Large Vision Language Models. In *Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertainNLP 2025)*, Bryan Eikema, Raúl Vázquez, Jonathan Berant, Marie-Catherine de Marneffe, Barbara Plank, Artem Shelmanov, Swabha Swayamdipta, Jörg Tiedemann, Chrysoula Zerva, and Wilker Aziz (Eds.). Association for Computational Linguistics, Suzhou, China, 89–102. doi:10.18653/v1/2025.uncertainlp-main.10
- [17] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024. Visual CoT: Advancing Multi-Modal Language Models with a Comprehensive Dataset and Benchmark for Chain-of-Thought Reasoning. arXiv:2403.16999 [cs.CV] <https://arxiv.org/abs/2403.16999>
- [18] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards VQA Models That Can Read. arXiv:1904.08920 [cs.CL] <https://arxiv.org/abs/1904.08920>
- [19] Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. 2025. LLaVA-CoT: Let Vision Language Models Reason Step-by-Step. arXiv:2411.10440 [cs.CV] <https://arxiv.org/abs/2411.10440>
- [20] Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2024. Woodpecker: hallucination correction for multimodal large language models. *Science China Information Sciences* 67, 12 (Dec. 2024). doi:10.1007/s11432-024-4251-x
- [21] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. 2023. Ferret: Refer and Ground Anything Anywhere at Any Granularity. arXiv:2310.07704 [cs.CV] <https://arxiv.org/abs/2310.07704>
- [22] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. 2016. Modeling Context in Referring Expressions. In *Computer Vision – ECCV 2016*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer International Publishing, Cham, 69–85.
- [23] Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. 2022. STaR: Bootstrapping Reasoning With Reasoning. arXiv:2203.14465 [cs.LG] <https://arxiv.org/abs/2203.14465>
- [24] Yang Zhan, Zhitong Xiong, and Yuan Yuan. 2023. RSVG: Exploring Data and Models for Visual Grounding on Remote Sensing Data. *IEEE Transactions on Geoscience and Remote Sensing* 61 (2023), 1–13. doi:10.1109/TGRS.2023.3250471
- [25] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2024. Multimodal Chain-of-Thought Reasoning in Language Models. arXiv:2302.00923 [cs.CL] <https://arxiv.org/abs/2302.00923>