

# See It, Say It, Sorted: An Iterative Training-Free Framework for Visually-Grounded Multimodal Reasoning in LVLMs

## Supplementary Material

### 1. Related Work

#### 1.1. Large Vision-Language Models

Projector-based pipelines such as BLIP-2 align frozen visual encoders with LLMs by mapping image features into the language space [9]. Subsequent instruction-tuned families such as LLaVA and LLaVA-OneVision strengthened multimodal following and efficiency while supporting higher input resolutions [7, 11]. Recent open-source series continue to scale capability and context handling across sizes, notably Qwen2.5-VL [1] and InternVL3 [32]. Language-only Chain-of-Thought often helps [21], but long textual prefixes can dominate attention during decoding, weakening token–evidence links and inducing hallucinations [10]. Modern evaluations make this gap explicit: TreeBench separates first-order perception from second-order reasoning to measure “thinking with images” [19], and RH-Bench quantifies the trade-off between longer reasoning and hallucination via RH-AUC [10]. Our method is designed for these backbones and addresses decoding-time drift without retraining.

#### 1.2. Mitigating hallucinations and thinking with images

Training-time grounding improves perception or alignment via larger corpora, refined instruction tuning, region-aware objectives, or RL, but ties improvements to data and architecture [9, 11]. Post-hoc detection flags ungrounded statements after generation [14] but leaves token choices unchanged. A complementary trend equips models with the ability to think with images by interleaving visual operations and intermediate representations at inference or by training policies that decide when and how to look.

**Training-free inference-time interventions.** Beyond RL/SFT pipelines, several training-free methods also improve visual grounding at test time. Woodpecker [26] performs inference-time hallucination correction by verifying and revising inconsistent visual claims. IBD [33] mitigates hallucinations by image-biased decoding that steers token selection toward visual evidence without retraining. ControlMLLM [23] provides training-free visual prompt learning to bias multimodal decoding via lightweight control signals, while ViperGPT [18] executes Python programs to support compositional visual reasoning. A recent survey [17] further systematizes this line of “thinking with images” methods and perspectives.

**Externalized visual thoughts.** Several works materialize intermediate visual reasoning steps instead of purely textual CoT. Visual sketchpad treats sketches (lines/boxes/marks) as a visual chain-of-thought that guides problem solving [4]. MVoT renders explicit visualization-of-thought images and aligns them with language via a token-discrepancy objective [8]. Visual Planning proposes a two-stage VPRL scheme for image-based planning without explicit tools [25]. Thinking with Generated Images couples step-wise subgoals with self-critique using generated pictures to steer complex reasoning [2]. GRIT trains grounded CoT with box-level evidence, improving traceable reasoning [3].

**Tool-/operation-augmented reasoning.** Another line uses tools or pixel operations during reasoning, often optimized with RL or preference learning. Pixel-Reasoner incentivizes zoom/selection while reasoning [15]; DeepEyes learns when to crop/zoom end-to-end [30]; Chain-of-Focus (CoF) adaptively searches and enlarges key regions [28]; OpenThinkIMG and VisTA train policies to select and compose visual tools [6, 16]. SAM-R1 leverages segmentation rewards to improve multimodal segmentation and reasoning [5]. VTool-R1 and PyVision expose programmable visual editing/plotting tools or dynamically generated Python tools for agentic multimodal reasoning [24, 29]. MMSearch-R1 learns web/image search in-the-wild [22]. Beyond static imagery, VITAL and VR-Thinker extend tool-augmented reasoning to long videos and video reward models [20, 27]. System-level releases such as OpenAI o3 natively integrate visual tool manipulation with broader tool use [13]. Recent frameworks also bridge high-level reasoning with fine-grained perception (VLM-FO1) [12] and explore RL policies that select from a palette of perception tools (ReVPT) [31]. Traceable Evidence Enhanced Visual Grounded Reasoning (TreeVGR) pairs a methodology with TreeBench for fine-grained evaluation [19].

**Position of our approach.** The above systems achieve impressive grounded reasoning but typically require curated traces, RL or SFT, and model-specific engineering; frequent cropping or tool calls can also raise inference cost and couple pixel operations with the whole sequence. We take an orthogonal, training-free route: keep the base VLM frozen, intervene only at ambiguous steps, and negotiate the next-token distribution with textual evidence produced on demand by a lightweight decider. This design aims to re-

duce hallucination propagation while remaining backbone-agnostic and scalable across model sizes.

## 2. Extended Analysis

We first quantify the **effect of the visual decider** via outcome flips, and then analyze **failure cases** to distinguish its side effects ( $T \rightarrow F$ ) from intrinsic limitations ( $F \rightarrow F$ ). On TreeBench with Qwen2.5-VL-7B (405 questions), the decider is triggered in 64.9% of cases (263/405). On these triggered samples, the outcome changes are:  $F \rightarrow T$  11.8% (31/263),  $T \rightarrow F$  0.8% (2/263),  $T \rightarrow T$  14.8% (39/263), and  $F \rightarrow F$  72.6% (191/263).

*Why does  $T \rightarrow F$  happen?* All  $T \rightarrow F$  cases are “accidental-correct” outcomes of the base model: early visual grounding is incorrect, but the final option is guessed correctly. When the decider fixes this early visual fact, downstream textual reasoning errors remain, causing the flip.

*Why does this not cause degradation in general?* Such cases are rare because the visual decider is limited to single-step grounding. When the base model is already correct, it rarely injects misleading evidence that overturns the final answer. Thus, the decider introduces little negative effect in aggregate, while its benefit is bounded by the base model’s reasoning capacity.

*Why are many cases  $F \rightarrow F$ ?* Most  $F \rightarrow F$  cases fall into three categories: (i) **59** questions with purely textual reasoning errors, where visual evidence cannot help; (ii) **11** questions where the image is difficult to recognize; and (iii) **121** questions where the base model is far from the correct solution, e.g., assigning high confidence to an incorrect token, leaving little room for the decider to intervene. These results indicate that the decider mainly refines the reasoning chain rather than fundamentally upgrading the base model’s reasoning capability.

## 3. Limitation

Our method is incompatible with tool-calling traces that replace natural language such as TreeVGR [19]. Our decoding assumes that both the evidence and the next token live in the same lexical space and carry comparable semantics. Concretely, the supervisor computes an evidence score by averaging  $p(w \mid e_{<j})$  over textual prefixes and then forms an evidence-induced distribution that is negotiate with the base model. In TreeVGR-style pipelines, however, the chain is interleaved with non-linguistic tool tokens such as `<box>(x1,y1,x2,y2)</box>`. These coordinates (often split into digits and punctuation by the tokenizer) encode grounding outside natural language. As a result, there is no meaningful distributional overlap between such tags and ordinary answer tokens. The KL-based supervision becomes ill-posed: when this misaligned signal is mixed back into the base distribution, the supervisor cannot provide a

reliable reweighting, and performance can degrade below the frozen base model. Put differently, TreeVGR’s traceable evidence improves grounding by replacing parts of the CoT with tool tags, whereas our method improves grounding by negotiating with textual evidence; the former breaks the assumption our supervisor relies on. Therefore, we do not apply ECRD to TreeVGR-style traces in this paper.

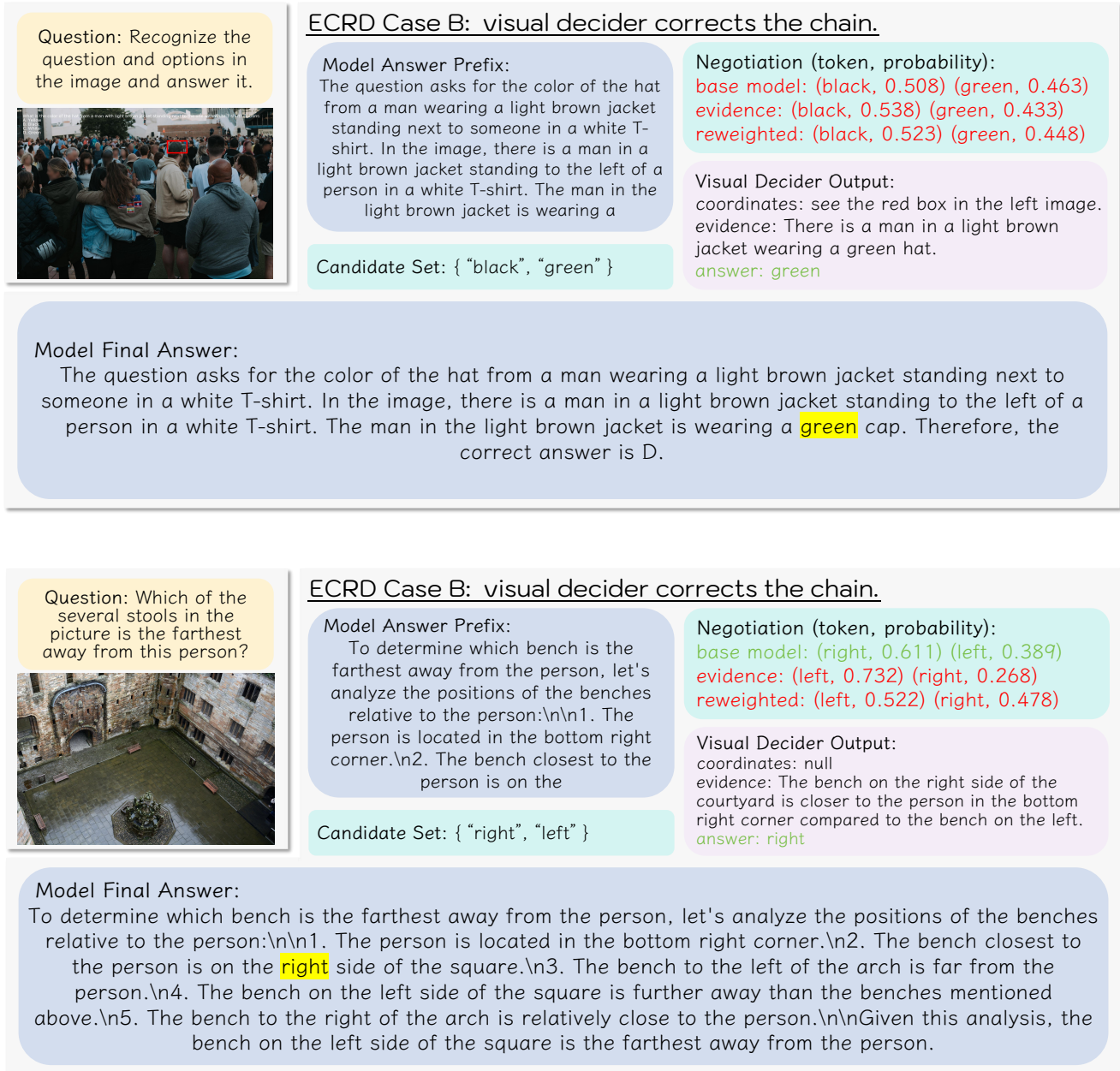
## 4. More execution examples of ECRD Case B

As shown in Figs. 1 and 2, we provide additional examples of ECRD at work in Case B, further demonstrating the effectiveness of our method.

In particular, because the spatial relations in the figure are relatively complex, we provide an interpretation of the second example in Fig. 2. At the key step, the candidate set is {“blue”, “white”}. The base slightly prefers “blue”, while the evidence-induced distribution prefers “white”, and the reweighted gap remains small, so the trigger fires. This choice is crucial: if the model selects “blue”, it will treat the second ship as the blue boat in front of the pier, because that blue boat has no cockpit, the chain will wrongly conclude that the first ship without a cockpit is the second ship. The visual decider reads the image with the prefix tail and returns the evidence sentence: “The second ship from the left is white and located behind the pier.” ECRD forces the token “white” for the current step and appends this sentence to the pool. Subsequent steps then follow this micro-observation, enabling the model to locate the ships correctly and derive that the first ship from left to right without a cockpit is the third one.

## References

- [1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1
- [2] Ethan Chern, Zhulin Hu, Steffi Chern, Siqi Kou, Jiadi Su, Yan Ma, Zhijie Deng, and Pengfei Liu. Thinking with generated images. *arXiv preprint arXiv:2505.22525*, 2025. 1
- [3] Yue Fan, Xuehai He, Diji Yang, Kaizhi Zheng, Ching-Chen Kuo, Yuting Zheng, Sravana Jyothi Narayanaraju, Xinze Guan, and Xin Eric Wang. Grit: Teaching mllms to think with images. *arXiv preprint arXiv:2505.15879*, 2025. 1
- [4] Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *Advances in Neural Information Processing Systems*, 37:139348–139379, 2024. 1
- [5] Jiaqi Huang, Zunnan Xu, Jun Zhou, Ting Liu, Yicheng Xiao, Mingwen Ou, Bowen Ji, Xiu Li, and Kehong Yuan. Sam-r1: Leveraging sam for reward feedback in multimodal segmentation via reinforcement learning. *arXiv preprint arXiv:2505.22596*, 2025. 1



Question: Which of the several stools in the picture is the farthest away from this person?



### ECRD Case B: visual decider corrects the chain.

**Model Answer Prefix:**  
To determine which bench is the farthest away from the person, let's analyze the positions of the benches relative to the person:\n\n1. The person is located in the bottom right corner.\n2. The bench closest to the person is on the

**Negotiation (token, probability):**  
base model: (right, 0.611) (left, 0.389)  
evidence: (left, 0.732) (right, 0.268)  
reweighted: (left, 0.522) (right, 0.478)

**Candidate Set:** { "right", "left" }

**Visual Decider Output:**  
coordinates: null  
evidence: The bench on the right side of the courtyard is closer to the person in the bottom right corner compared to the bench on the left.  
answer: right

**Model Final Answer:**  
To determine which bench is the farthest away from the person, let's analyze the positions of the benches relative to the person:\n\n1. The person is located in the bottom right corner.\n2. The bench closest to the person is on the right side of the square.\n3. The bench to the left of the arch is far from the person.\n4. The bench on the left side of the square is further away than the benches mentioned above.\n5. The bench to the right of the arch is relatively close to the person.\n\nGiven this analysis, the bench on the left side of the square is the farthest away from the person.

Figure 1. Execution examples of ECRD Case B.

- [6] Zeyi Huang, Yuyang Ji, Anirudh Sundara Rajan, Zefan Cai, Wen Xiao, Haohan Wang, Junjie Hu, and Yong Jae Lee. Visualtoolagent (vista): A reinforcement learning framework for visual tool selection. *arXiv preprint arXiv:2505.20289*, 2025. 1
- [7] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 1
- [8] Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. Imagine while reasoning in space: Multimodal visualization-of-thought. *arXiv preprint arXiv:2501.07542*, 2025. 1
- [9] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1
- [10] Chengzhi Liu, Zhongxing Xu, Qingyue Wei, Juncheng Wu, James Zou, Xin Eric Wang, Yuyin Zhou, and Sheng Liu. More thinking, less seeing? assessing amplified hallucination in multimodal reasoning models, 2025. 1
- [11] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information*



**Question:** From left to right, which is the first ship without a cockpit?



**ECRD Case B: visual decider corrects the chain.**

**Model Answer Prefix:**  
The question asks to find the first ship without a cockpit from left to right. First, identify the ships in the image. The ships are docked along the canal, and their positions are as follows: the first ship on the left is white and has a motor at the back, the second is a

**Negotiation (token, probability):**  
base model: (blue, 0.470) (white, 0.313)  
evidence: (white, 0.457) (blue, 0.326)  
reweighted: (blue, 0.394) (white, 0.389)

**Visual Decider Output:**  
coordinates: see the red box in the left image.  
evidence: The second ship from the left is white and located behind the pier.  
answer: white

**Candidate Set:** { "blue", "white" }

**Model Final Answer:**  
The question asks to find the first ship without a cockpit from left to right. First, identify the ships in the image. The ships are docked along the canal, and their positions are as follows: the first ship on the left is white and has a motor at the back, the second is a white boat, the third is a white boat, the fourth is a white boat, the fifth is a blue boat, and the sixth is a blue boat. The first and second ships are both white motorboats, and each of them has a cockpit. The third ship is the white boat in the middle. It has a motor at the back, but it does not have a cockpit. So, the first ship without a cockpit is the third one.

Figure 2. Execution examples of ECRD Case B.

- processing systems*, 36:34892–34916, 2023. 1
- [12] Peng Liu, Haozhan Shen, Chunxin Fang, Zhicheng Sun, Jijia Liao, and Tiancheng Zhao. Vlm-fo1: Bridging the gap between high-level reasoning and fine-grained perception in vlms. *arXiv preprint arXiv:2509.25916*, 2025. 1
- [13] OpenAI. Openai-o3, 2025. 1
- [14] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018. 1
- [15] Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhui Chen. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. *arXiv preprint arXiv:2505.15966*, 2025. 1
- [16] Zhaochen Su, Linjie Li, Mingyang Song, Yunzhuo Hao, Zhengyuan Yang, Jun Zhang, Guanjie Chen, Jiawei Gu, Juntao Li, Xiaoye Qu, et al. Openthinking: Learning to think with images via visual tool reinforcement learning. *arXiv preprint arXiv:2505.08617*, 2025. 1
- [17] Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, et al. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. *arXiv preprint arXiv:2506.23918*, 2025. 1
- [18] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11888–11898, 2023. 1
- [19] Haochen Wang, Xiangtai Li, Zilong Huang, Anran Wang, Jiacong Wang, Tao Zhang, Jiani Zheng, Sule Bai, Zijian Kang, Jiashi Feng, et al. Traceable evidence enhanced visual

- grounded reasoning: Evaluation and methodology. *arXiv preprint arXiv:2507.07999*, 2025. 1, 2
- [20] Qunzhong Wang, Jie Liu, Jiajun Liang, Yilei Jiang, Yuanxing Zhang, Jinyuan Chen, Yaozhi Zheng, Xintao Wang, Pengfei Wan, Xiangyu Yue, et al. Vr-thinker: Boosting video reward models through thinking-with-image reasoning. *arXiv preprint arXiv:2510.10518*, 2025. 1
- [21] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 1
- [22] Jiming Wu, Zihao Deng, Wei Li, Yiding Liu, Bo You, Bo Li, Zejun Ma, and Ziwei Liu. Mmsearch-r1: Incentivizing llms to search. *arXiv preprint arXiv:2506.20670*, 2025. 1
- [23] Mingrui Wu, Xinyue Cai, Jiayi Ji, Jiale Li, Oucheng Huang, Gen Luo, Hao Fei, Guannan Jiang, Xiaoshuai Sun, and Rongrong Ji. Controlmllm: Training-free visual prompt learning for multimodal large language models. *Advances in Neural Information Processing Systems*, 37:45206–45234, 2024. 1
- [24] Mingyuan Wu, Jingcheng Yang, Jize Jiang, Meitang Li, Kaizhuo Yan, Hanchao Yu, Minjia Zhang, Chengxiang Zhai, and Klara Nahrstedt. Vtool-r1: Vllms learn to think with images via reinforcement learning on multimodal tool use. *arXiv preprint arXiv:2505.19255*, 2025. 1
- [25] Yi Xu, Chengzu Li, Han Zhou, Xingchen Wan, Caiqi Zhang, Anna Korhonen, and Ivan Vulić. Visual planning: Let’s think only with images. *arXiv preprint arXiv:2505.11409*, 2025. 1
- [26] Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12):220105, 2024. 1
- [27] Haoji Zhang, Xin Gu, Jiawen Li, Chixiang Ma, Sule Bai, Chubin Zhang, Bowen Zhang, Zhichao Zhou, Dongliang He, and Yansong Tang. Thinking with videos: Multimodal tool-augmented reinforcement learning for long video reasoning. *arXiv preprint arXiv:2508.04416*, 2025. 1
- [28] Xintong Zhang, Zhi Gao, Bofei Zhang, Pengxiang Li, Xiaowen Zhang, Yang Liu, Tao Yuan, Yuwei Wu, Yunde Jia, Song-Chun Zhu, et al. Chain-of-focus: Adaptive visual search and zooming for multimodal reasoning via rl. *arXiv preprint arXiv:2505.15436*, 2025. 1
- [29] Shitian Zhao, Haoquan Zhang, Shaoheng Lin, Ming Li, Qilong Wu, Kaipeng Zhang, and Chen Wei. Pyvision: Agentic vision with dynamic tooling. *arXiv preprint arXiv:2507.07998*, 2025. 1
- [30] Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deep-eyes: Incentivizing” thinking with images” via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025. 1
- [31] Zetong Zhou, Dongping Chen, Zixian Ma, Zhihan Hu, Mingyang Fu, Sinan Wang, Yao Wan, Zhou Zhao, and Ranjay Krishna. Reinforced visual perception with tools. *arXiv preprint arXiv:2509.01656*, 2025. 1
- [32] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 1
- [33] Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. Ibd: Alleviating hallucinations in large vision-language models via image-biased decoding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1624–1633, 2025. 1