SMALL DRAFTS, BIG VERDICT: INFORMATION-INTENSIVE VISUAL REASONING VIA SPECULATION

Anonymous authors

Paper under double-blind review

ABSTRACT

Large Vision-Language Models (VLMs) have achieved remarkable progress in multimodal understanding, yet they struggle with information-intensive images that densely interleave textual annotations with fine-grained graphical elements. The main challenges lie in precisely localizing critical cues in dense layouts and multi-hop reasoning to integrate dispersed evidence. We propose Speculative Verdict (SV), a training-free framework inspired by speculative decoding that combines multiple lightweight draft experts with a large verdict model. In the draft stage, small VLMs act as draft experts to generate reasoning paths that provide diverse localization candidates; in the verdict stage, a strong VLM synthesizes these paths to produce the final answer, minimizing computational cost while recovering correct answers. To further improve both efficiency and accuracy, SV introduces a consensus expert selection mechanism that forwards only high-agreement reasoning paths to the verdict. Empirically, SV achieves consistent gains on challenging information-intensive and high-resolution visual question answering benchmarks, including Infographic VQA, ChartMuseum, ChartQAPro, and HR-Bench 4K. By synthesizing correct insights from partially accurate reasoning paths, SV achieves both error correction and cost-efficiency compared to large proprietary models or training pipelines.

1 Introduction

Recent advances in large vision—language models (VLMs) have delivered impressive performance on tasks such as image captioning and general visual question answering (VQA) (Li et al., 2025; Fu et al., 2024). However, these models encounter challenges in information-intensive images that densely interleave diverse textual annotations (legends, labels, captions) with fine-grained graphical elements (charts, diagrams, plots) across multiple scales and formats (Su et al., 2025b). Addressing this task requires two interdependent capabilities (Figure 1; Ke et al., 2025): (i) comprehensive and precise localization, which involves not only pinpointing the exact positions of critical cues in densely populated layouts but also ensuring that all query-relevant regions are identified; (ii) multi-hop reasoning, which chains visual analysis—encompassing colors, shapes, and spatial relationships—with textual evidence, thereby integrating dispersed cues into a coherent and complete answer. As each reasoning step builds on the accuracy of the previous one, any intermediate error can propagate through the entire chain, making the overall process highly error-sensitive and difficult to correct retrospectively.

Existing work tackles information—intensive visual reasoning with search-based zoom-in pipelines that enlarge local regions for detailed reasoning. Specifically, learning-based methods train reinforcement learning policies to guide zoom operations iteratively (Zheng et al., 2025; Su et al., 2025a; Fan et al., 2025; Zhang et al., 2025b). Enhancing its performance would demand costly fine-grained supervision. Moreover, training-free methods perform cropping based on internal attention or confidence scores (Zhang et al., 2025a; Shen et al., 2024; Wang et al., 2025b). Yet in dense layouts, we find these signals correlate weakly with true relevance, misleading the model into visually similar but irrelevant areas. Consequently, these tool-driven designs fail to capture all evidence for multi-hop reasoning, leaving the core challenges of information-intensive visual reasoning unsolved.

To overcome these limitations, we propose Speculative Verdict (SV), a training-free framework inspired by speculative decoding that combines small draft visual experts with a large verdict

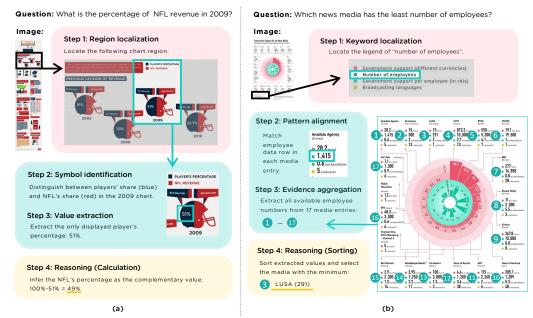


Figure 1: Examples of correct reasoning paths for information-intensive image VQA tasks. They illustrate distinct trajectories: (a) focuses on the localization of a specific chart, symbol identification, and complementary reasoning from a single percentage value; (b) focuses on keyword-based localization, evidence aggregation from multiple entries across the entire image, and cross-entity sorting to select the minimum.

model (Leviathan et al., 2023). The framework operates in two stages (Figure 2): (1) Draft stage: multiple lightweight VLMs serve as draft experts, each generating a reasoning path that offers diverse localization candidates; (2) Verdict stage: a large VLM acts as a strong verdict, which receives the reasoning paths as contextual evidence, distinguishes the correct information, and outputs the final answer tokens. SV directly tackles core challenges through complementary strengths: draft experts expand evidence coverage across scattered regions, while the verdict prevents error propagation by synthesizing these multiple perspectives. Importantly, unlike using a large proprietary model to reason over every image section, SV invokes the verdict only once to yield a concise final answer, thereby minimizing computational cost while effectively recovering correct answers. To further balance accuracy and efficiency, SV introduces a consensus expert selection mechanism in the draft stage, ensuring that only reasoning paths with strong agreement are forwarded to the verdict.

We evaluate Speculative Verdict on information-intensive VQA benchmarks, including InfographicVQA (Mathew et al., 2021), ChartMuseum (Tang et al., 2025), and ChartQAPro (Masry et al., 2025), which demand reasoning over dense textual and visual content. As a training-free framework, SV consistently outperforms strong open-source models, large proprietary models, and perception-focused search methods while remaining cost-efficient. In particular, SV yields average gains of 4% over small VLMs as draft experts and 10% over GPT-40 (Hurst et al., 2024) as verdict. Beyond overall gains, SV successfully corrects 47-53% of cases where majority voting or the verdict model alone fails, thereby reducing vulnerability to error propagation in information-intensive visual reasoning. Furthermore, SV surpasses all baselines on HR-Bench 4K (Wang et al., 2025a), a benchmark for high-resolution visual perception, underscoring its effectiveness in challenging multimodal reasoning scenarios.

2 Related Work

Vision-Language Model Reasoning with Tools. Recent research has explored enhancing VLM perception by manipulating input images with zooming operations to locate relevant regions (Hu et al., 2024). (1) Prompting-based methods exploit internal signals of VLMs to decide where to zoom. ViCrop (Zhang et al., 2025a) leverages models' attention maps to highlight query-related

regions, thereby generating automatic visual crops. Other works perform tree-based search, where models evaluate candidate sub-images with confidence scores to iteratively narrow down to relevant regions (Shen et al., 2024; Wang et al., 2025b). However, such signals align poorly with the required evidence in information-intensive images, since queries often require reasoning across multiple dispersed regions. (2) Reinforcement learning approaches instead optimize policies that interleave visual zooming with textual reasoning (Zheng et al., 2025; Su et al., 2025a; Fan et al., 2025; Zhang et al., 2025b). By calling zooming tools within the agentic framework, these methods adaptively crop regions and concatenate them into the reasoning trajectory, enabling more active evidence gathering. Yet these methods still fall short on information-intensive images, requiring costly task-specific training to scale.

Speculative Decoding. Speculative decoding is a draft-then-verify decoding paradigm to accelerate LLM inference (Xia et al., 2024). Specifically, it utilizes a draft model to generate future tokens, and a larger target model verifies them via parallel rejection sampling. Beyond the vanilla setting, recent work extends acceptance from token-level equivalence to step-level semantic similarity to speed up reasoning (Yang et al., 2025; Pan et al., 2025; Fu et al., 2025b; Liao et al., 2025). Collaborative decoding via Speculation (Fu et al., 2025a) further applies speculative decoding with multiple draft LLMs by verifying proposals against a combined distribution of the drafts and the target, yielding greater speedups than standard ensembling. However, these adaptations primarily target speed in LLM inference and also do not address the challenges of vision–language reasoning.

Large Language Model Ensemble. Majority voting aggregates answers by frequency, but fails when the correct solution is produced by a minority. Universal Self-Consistency (Chen et al., 2023) mitigates this failure mode by prompting the LLM to select the most consistent candidate across samples. Further, learned aggregators read multiple rationales and synthesize them to recover minority-correct information (Qi et al., 2025; Zhao et al., 2025). However, these approaches focus on text-only ensembling. In visual-language reasoning, supervision of ensembling is not cost-effective since multimodal complexity requires costly, fine-grained annotations.

3 SPECULATIVE VERDICT

Speculative decoding is an inference-time optimization originally developed to mitigate the latency of autoregressive generation (Leviathan et al., 2023). The approach employs a draft-then-verify paradigm: (i) a small, fast draft model proposes one or more future tokens speculatively, and (ii) a large, accurate base model verifies these proposals in parallel, accepts or revises the proposals, and generates output that is consistent with the base model's distribution (Xia et al., 2024; Zhang et al., 2024). This token-level process speeds up inference by committing several tokens at once, while maintaining quality by discarding continuations that diverge from the base model's distribution.

The key insight is that draft models expand coverage quickly, while the verifier ensures correctness. Although this idea has been mainly applied to accelerate text generation, its high-level principle is also well-suited for information-intensive multimodal reasoning.

3.1 METHOD OVERVIEW

Information-intensive visual question answering (VQA) requires models to localize query-relevant regions, perceive diverse fine-grained textual and visual details, and integrate dispersed evidence into a single correct answer. These tasks are highly error-sensitive as elaborated in Section 1: a single misread or mislocalized element often leads to a completely wrong prediction.

To address this challenge, we adapt the draft—then—verify paradigm of speculative decoding to multimodal reasoning. Unlike its original use for inference acceleration, we repurpose the paradigm to improve robustness and error correction in information-intensive visual reasoning. On a high level, our Speculative Verdict (SV) framework operates in two stages (Figure 2):

- (i) **Draft stage**, where multiple lightweight VLMs serve as experts to provide diverse reasoning paths (Section 3.2);
- (ii) **Verdict stage**, where a large VLM acts as verdict to verify, refine, and synthesize these reasoning paths into the final prediction (Section 3.3).

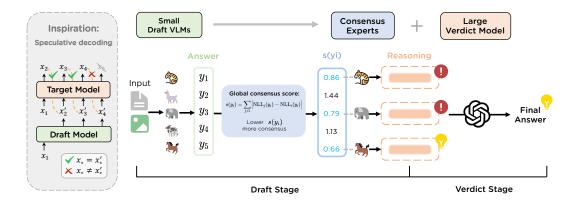


Figure 2: Overview of Speculative Verdict (SV). Inspired by speculative decoding, SV operates in two stages. In the draft stage, given an input question-image pair, multiple small draft VLMs first generate candidate answers. A global consensus score $s(y_i)$, computed based on pairwise NLL differences, is used to select consensus experts for reasoning path generation. In the verdict stage, these reasoning paths are verified and corrected by a large verdict model to produce the final answer.

3.2 Draft Stage

Chain-of-Thought (CoT) prompting exposes models' intermediate reasoning steps in an explicit, stepwise form (Wei et al., 2022). This is critical for information-intensive VQA, where solving a question requires a sequence of localization, evidence extraction, and analytic operations (Figure 1). However, current VLM often lacks fine-grained perception and localization on densely annotated images, and existing tool-driven zoom-in methods are ineffective as elaborated in Section 2. We therefore utilize multiple VLMs to produce reasoning paths rather than a single direct answer, so that the subsequent verdict can verify and synthesize structured evidence. Concretely, given an image–question pair (x,q), several lightweight VLMs $\{M_1,\ldots,M_m\}$ are selected as draft experts and are prompted with a CoT prompt (details in Appendix G). Each expert M_i outputs a candidate answer $y_i = M_i(x,q)$ and a reasoning path r_i .

We observe that each reasoning path r_i provided by draft experts typically includes: (i) global scan and localization proposals that identify query-related regions, sections, or subplots, often referencing axes, titles, or captions; (ii) evidence extraction, which transforms visual or textual elements into structured cues, including reading legends, mapping colors to series, parsing axis labels, or assembling lists of values or tokens for subsequent operations; (iii) analytic and reasoning operations, which operate over the extracted cues to derive higher-level conclusions, such as filtering or selecting relevant entities, computing differences, sorting across panels, and cross-referencing dispersed cues. As shown in the running case (Figure 3), different experts may match legends to charts differently; some correctly gather the required cues while others misread adjacent values. This diversity yields a complementary but potentially noisy pool of reasoning signals.

To reduce noise and cost of the pipeline, we additionally apply a consensus-based selection mechanism that forwards only the k most reliable draft experts, detailed in Section 3.4.

3.3 VERDICT STAGE

The set $\{r_i\}$ captures diverse cues, offering richer evidence but also introducing contradictions, which motivates the need for a verdict stage to verify and integrate them. Answer-level ensembling, such as majority voting, often fails in minority-correct scenarios where many experts converge on the same incorrect decision, such as mislocalizing the query-related region or misreading fine-grained textual details, even after correct localization. This failure mode is frequently observed in information-intensive reasoning (as illustrated in Figure 3). Rather than discarding minority-correct signals, we leverage a stronger model as a verdict to validate grounding, resolve conflicts, and synthesize coherent reasoning from the draft paths.

Specifically, given the image–question pair (x,q) and the drafts' reasoning paths $\{r_i\}_{i=1}^k$, we concatenate these paths with (x,q) and present them to a large verdict model J. The original image

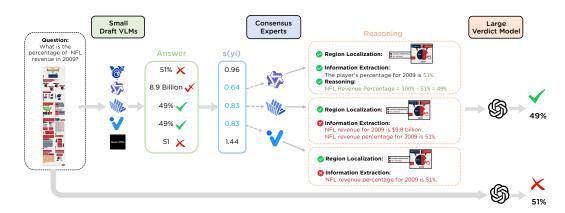


Figure 3: An illustration of Speculative Verdict on Infographic VQA. Five draft VLMs first produce candidate answers, with only two providing the correct result. Consensus scoring ranks answers by agreement, and the three with the lowest scores are selected as consensus experts. Although some experts commit extraction errors (confusing player's share with NFL revenue), the verdict synthesizes their reasoning paths and successfully recovers the correct answer (49%). This illustrates SV's ability to identify reliable experts and achieve error correction.

x is retained as input, as the verdict is expected to perform analysis rather than merely text-based ensembling. The verdict processes all evidence in a single forward pass and outputs the final answer:

$$y = J(x, q, \{r_i\}_{i=1}^k).$$

In this design, the verdict acts not as a voter but as a synthesizer. It evaluates grounding consistency, identifies contradictions across reasoning paths, and integrates consistent cues into a coherent prediction. The case in Figure 3 illustrates this intended role: when only one draft extracts the correct evidence, the verdict is designed to recover it by contrasting against competing but inconsistent paths.

This setup enables us to leverage the reasoning capabilities of large models while keeping the inference cost manageable. Our verdict stage maximizes prefill utilization: the verdict processes multiple draft reasoning paths as prefill input while generating only short answer tokens through minimal autoregressive decode. This design also avoids relying on large closed-source models for iterative per-section analysis and long rationale generation, both of which are token-expensive.

3.4 Consensus Expert Selection

To keep the verdict input both efficient and accurate, we introduce a training-free expert selection mechanism at the beginning of the draft stage (Section 3.2). Since each question in information-intensive VQA admits a unique correct answer, consensus among model answers provides a principled criterion for identifying reasoning paths with higher reliability. Therefore, the key idea here is to measure agreement among draft experts' answers and retain only those with stronger peer consensus. This mechanism is computed efficiently by prefilling the question and answer tokens, with each draft decoded only once, making it plug-and-play with minimal overhead.

Consensus Score. We define a consensus score that measures how strongly a draft expert's answer is agreed by its peers. Formally, let x be the input image and $q=(q_1,\ldots,q_n)$ the question tokens. From the pool of candidate draft experts $\{M_i\}_{i=1}^m$, each model produces a candidate answer $y_i=(y_{i,1},\ldots,y_{i,T})$. For a peer model M_j $(j\neq i)$ in the pool, we measure how plausible it finds y_i by computing the negative log-likelihood (NLL) of the concatenated input (x,q,y_i) , i.e., the original image together with the question tokens followed by the candidate answer tokens:

$$NLL_j(y_i) = -\frac{1}{T} \sum_{t=1}^{T} \log p_{M_j}(y_{i,t} \mid x, q_{\leq n}, y_{i, < t}).$$

To account for calibration differences, we normalize against M_j 's own answer y_j , thus relative consensus score from M_j 's perspective is:

$$s_j(y_i) = |NLL_j(y_i) - NLL_j(y_j)|, \quad j \neq i,$$

where a smaller $s_j(y_i)$ indicates stronger agreement, as M_j finds y_i nearly as plausible as its own answer y_j .

To capture overall agreement rather than pairwise consistency, we define the *global consensus score* of candidate y_i by summing across all peers:

$$s(y_i) = \sum_{j \neq i} s_j(y_i),$$

which quantifies the overall level of peer consensus for M_i 's answer, and a lower $s(y_i)$ indicates stronger agreement and thus higher reliability.

Consensus Expert Selection Strategy. We adopt a cross-all strategy that selects the k draft experts with the strongest consensus, measured by the lowest consensus scores, from the pool of m VLMs. By aggregating agreement across all peers, this strategy provides a holistic measure of reliability. It thus yields a subset of reasoning paths that are well-grounded and compact in size, balancing informativeness and efficiency. An illustrative example of this selection process is shown in Figure 3.

4 EXPERIMENTS

4.1 SETUPS

Configuration Details. We set the draft pool size to k=5 considering efficiency and select m=3 consensus experts in our main experiments. Ablation studies over different m values are reported in Section 4.4. The draft pool consists of the following VLMs for expert selection: Qwen2.5-VL-7B-Instruct (Bai et al., 2025), MiMo-VL-7B-RL (Team, 2025b), InternVL3-8B (Zhu et al., 2025), GLM-4.1V-9B-Thinking (Team, 2025a), Ovis2.5-9B (Lu et al., 2025). These models are chosen as draft visual experts for their strong multimodal perception and reasoning on information-intensive and high-resolution images. For the verdict models, we employ GPT-40 (Hurst et al., 2024) and Qwen2.5-VL-72B-Instruct respectively, given their superior ability in visual reasoning. In particular, for information-intensive image benchmarks, we preprocess images with PP-StructureV3 (Cui et al., 2025) to produce a layout-preserving structured format, provided together with the original image as auxiliary input to the verdict model.

Baselines. We compare SV with proprietary models GPT-40 and GPT-40-mini, and the large open-source model Qwen2.5-VL-72B-Instruct as it is one of our verdicts. We also evaluate SV against draft experts mentioned above. These baselines are evaluated under the same chain-of-thought prompting template in Appendix G. Additionally, we include DeepEyes (Zheng et al., 2025) as a representative tool-driven baseline with zoom-in operations.

Benchmarks. We evaluate SV on three information-intensive benchmarks and extend the evaluation to a representative high-resolution benchmark, providing a comprehensive assessment of fine-grained visual reasoning: InfographicVQA (Mathew et al., 2021), ChartMuseum (Tang et al., 2025), ChartQAPro (Masry et al., 2025) and HR-Bench 4K (Wang et al., 2025a). InfographicVQA collects infographics with an average high resolution over 2k, designed to test reasoning over layout, graphical and textual content, including operations such as counting, sorting, and basic arithmetic. Chart-Museum and ChartQAPro introduce substantially greater visual reasoning complexity by covering a broad spectrum of real-world chart types and question formats, revealing a large performance gap between current Large VLMs and humans. These benchmarks require models to visually ground relevant regions, extract information, and conduct reasoning to answer queries.

We further assess generalization to high-resolution images on HR-Bench 4K. It comprises two subtasks: FSP (Fine-grained Single-instance Perception) and FCP (Fine-grained Cross-instance Perception), stressing small-object perception and cross-instance reasoning under high-resolution inputs.

4.2 RESULTS ON INFORMATION-INTENSIVE BENCHMARKS

As shown in Table 1, SV demonstrates superior performance across all benchmarks, outperforming a wide range of baselines. Based on the results, we have the following key observations:

Table 1: Results on test sets of four benchmarks. The three benchmarks on the left are our main information-intensive VQA tasks, while the benchmark on the right evaluates high-resolution perception. We compare SV against closed-source VLMs, open-source VLMs, and the tool-driven method, with all results reproduced by ourselves. The best results for each benchmark are high-lighted in **bold** and the second-best results are <u>underlined</u>.

Model	Param Size	InfographicVQA ANLS	ChartMuseum Acc	ChartQAPro Acc	HR-Bench 4K Acc		
Closed-source VLMs							
GPT-4o	=	76.5	42.7	52.6	67.4		
GPT-4o-mini	_	67.2	31.5	44.1	53.8		
		Open-source VLMs					
Qwen2.5-VL-Instruct	7B	79.8	29.5	51.0	73.0		
MiMO-VL-RL (think)	7B	83.5	29.0	57.3	72.3		
InternVL3	8B	72.3	25.9	45.1	68.0		
GLM-4.1V-Thinking	9B	84.8	48.0	56.2	72.3		
Ovis2.5	9B	81.7	34.0	55.9	69.5		
Qwen2.5-VL-Instruct	72B	84.2	40.7	60.7	<u>73.1</u>		
Tool-driven method							
DeepEyes	7B	75.5	28.0	48.7	73.0		
SV w/ GPT-4o Verdict	-	88.4	49.3	64.0	71.4		
SV w/ Qwen2.5-VL-72B-Instruct Verdict	-	<u>86.7</u>	<u>48.2</u>	<u>63.0</u>	75.6		

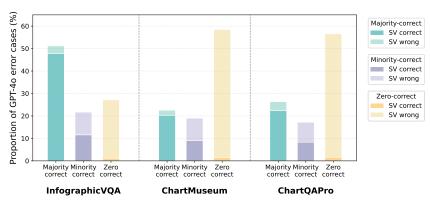


Figure 4: SV's correction ability on verdict's error cases across information-intensive benchmarks (GPT-40 as verdict). We consider only cases where the verdict itself fails, to isolate SV's independent correction capacity. For each benchmark, three bars denote expert correctness categories (majority-correct, minority-correct, and zero-correct), defined by how many selected experts provide the correct answer. Within each category, the bars are split into the proportion corrected by SV (dark) versus not corrected (light). More details can be found in Appendix C.

- (i) **SV** shows consistent gains over all strong draft experts' baselines, with improvements of 3.6% on Infographic VQA, 1.3% on ChartMuseum, and 6.6% on ChartQAPro with GPT-40 as verdict. SV also achieves comparable gains with Qwen-2.5-VL-72B-Instruct as a verdict.
- (ii) Importantly, SV enables strong error correction beyond simple answer aggregation. Across benchmarks, SV recovers 47–53% of minority-correct cases, where few draft experts are correct and the verdict alone fails (see Figure 4 and case in Figure 3). Moreover, SV even recovers 2.5-4.5% of zero-correct cases, where neither the drafts nor the verdict answers correctly (see the case in Appendix F). In these settings, SV succeeds by extracting partially correct pieces from different draft reasoning paths while rejecting misleading cues, thus reconstructing them into a correct answer. These findings indicate that errors in information-intensive visual reasoning are often structured: even incorrect paths contain informative subparts. In this way, SV exploits this structure by synthesizing multi-perspective reasoning paths, enabling effective correction where traditional ensemble methods fail.
- (iii) **SV strengths large verdict models significantly**, and using GPT-40 as verdict delivers stronger results due to its reasoning advantage on information-intensive benchmarks. Specifically, when GPT-40 is used as verdict, SV surpasses the GPT-40 baseline by 11.9% on InfographicVQA, 6.6%

on ChartMuseum, and 11.4% on ChartQAPro. These improvements come with reduced inference cost for the large verdict model, demonstrating that SV can outperform much larger or proprietary LVLMs in a cost-efficient manner.

(iv) SV substantially outperforms representative tool-driven pipeline DeepEyes, with gains of +12.9% on InfographicVQA, +21.3% on ChartMuseum, and +11.3% on ChartQAPro. This gap arises because DeepEyes is strong in local grounding but weak in reasoning over dense textual and visual content. For example, it often locates on text spans or legends rather than full regions need for analytical operations, and its zoom-in calls are sometimes redundant or misdirected (see Appendix D for error analysis). As a result, it struggles with global comparison and dispersed evidence synthesis. In contrast, SV's reasoning-path synthesis enables it to integrate evidence across regions reliably without relying on predefined tool-based visual search.

4.3 RESULTS ON HIGH-RESOLUTION BENCHMARK

We further assess generalization to high-resolution images using HR-Bench 4K to evaluate whether SV can enhance fine-grained visual perception. The key observations are as follows (Table 1):

- (i) With Qwen2.5-VL-72B-Instruct as verdict, SV achieves its largest margin, surpassing the best-performing draft expert by 2.6% and even outperforming the verdict itself by 2.5%. The superior performance of Qwen2.5-VL-72B as verdict on this task correlates with its stronger visual localization capabilities, indicating verdict selection should align with task-specific requirements.
- (ii) SV also exceeds DeepEyes, which is explicitly trained with zoom-in tools for iterative visual search on high-resolution perception. This highlights SV's ability to generalize to high-resolution tasks, where accurate recognition of small objects is critical. Aligning perceptually strong draft experts with a verdict thus provides a simpler yet effective solution for high-resolution reasoning.

4.4 ABLATION STUDY

To better understand the effectiveness of SV, we conduct ablation studies on information-intensive benchmarks to analyze the impact of individual components. In these experiments, the reasoning baseline refers to the best-performing draft VLM in our pool for each benchmark (Table 1).

Number of Consensus Experts. Our setting with m=3 consensus experts yields a favorable trade-off between accuracy and efficiency, as it determines the number of reasoning paths forwarded to the verdict. As shown in Figure 5, we observe that the performance improves nearly linearly up to three draft models and then saturates, while inference cost grows roughly linearly with size.

Selection Strategies of Consensus Experts. We confirm the effectiveness of our cross-all selection strategy by comparing it with a best-reference strategy. In the best-reference variant, the top-performing draft VLM serves as reference and the two most consistent experts are selected with it. While best-reference is expected to be the strongest criterion, cross-all achieves comparable gains while remaining reference-free (Figure 6).

Selection Criteria. Selecting consensus-based experts consistently improves performance, while divergent selection can even fall below the single-draft reasoning baseline (Table 7). These results support that, for information-intensive tasks, consensus-based selection more reliably identifies the correct reasoning trajectory than enforced diversity.

Impact of Verdict Stage. The verdict stage yields higher performance than majority voting across information-intensive benchmarks (Figure 8). Notably, majority voting with all five draft experts performs comparably as majority voting with three consensus experts, consistent with our finding that consensus selection can match the performance of all drafts at a lower cost (Table 5). SV further surpasses both by leveraging the verdict's error correction ability, successfully capturing minority-correct cases that majority voting discards (Figure 4 and Figure 3).

Choice of Verdict Input. Providing full reasoning paths to the verdict yields substantially better performance than passing only final answers (Table 2), with improvements of 15% on InfographicVQA, and 4.8% on ChartQAPro. These results highlight that rich contextual evidence is essential for the verdict to recover correct reasoning, whereas final predictions alone are insufficient.

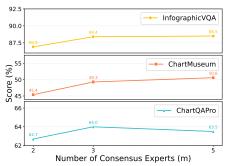


Figure 5: Ablations on the number of consensus experts m.

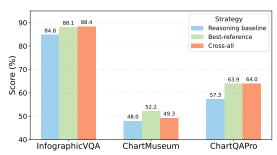


Figure 6: Ablations on different consensus expert selection strategies.

Choice of Verdict Scale. Using a large verdict model yields stronger gains than a small verdict model. For ablations, we select GLM-4.1V-9B-Thinking as the small verdict because it is the strongest reasoning model among the baselines. However, results in Table 3 show that it brings only modest improvements, while GPT-40 delivers additional gains of 3.4% on InfographicVQA and 1.3% on ChartMuseum compared to this small verdict. These results indicate that even reasoning-strong small verdicts offer limited benefit in synthesizing correct answers, validating SV's design principle of invoking a strong verdict only once to achieve robust and efficient error correction.

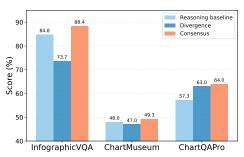


Figure 7: Ablations on selection criteria.

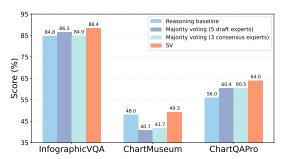


Figure 8: Performance comparison on SV and majority voting with different model sets.

Table 2: Ablations on verdict input.

Textual input	InfographicVQA ANLS	ChartQAPro Acc
Reasoning baseline	84.8	57.3
Answers only Reasoning paths	73.4 88.4	59.2 64.0

Table 3: Ablations on verdict scale. A subset of 1000 samples is tested on Infographic VQA.

Verdict Choice	InfographicVQA	ChartMuseum	ChartQAPro	
	ANLS	Acc	Acc	
Reasoning baseline	84.5	<u>48.0</u>	57.3	
GLM-4.1V-9B-Thinking Verdict	86.0	48.0	59.4	
GPT-40 Verdict	89.4	49.3	64.0	

5 CONCLUSION

This paper introduces Speculative Verdict (SV), a training-free framework to address challenges of information-intensive visual reasoning. Inspired by speculative decoding, SV repositions large models as efficient synthesizers rather than computationally expensive step-by-step reasoners. By integrating diverse reasoning paths from lightweight experts, the verdict can distinguish informative cues and recover correctness from structured errors. Experiments show that SV consistently outperforms strong proprietary, open-source, and tool-driven methods, establishing a cost-efficient paradigm for reasoning on information-intensive images. A limitation of our study is that the evaluation is restricted to a fixed draft model pool. Future work could examine more diverse pools to provide further insights into its generalizability.

ETHICS STATEMENT

All authors have read and commit to adhering to the ICLR Code of Ethics. This work does not involve human subjects, sensitive personal data, biometrics, or medical information. All datasets used are publicly available under permissible licenses and are not privacy-sensitive. We recognize that any automated reasoning system may produce incorrect or misleading outputs. To ensure responsible use, we emphasize that our method is intended for research and analysis rather than deployment in high-stakes settings. Users are encouraged to verify model outputs and apply human oversight when necessary. We take full responsibility for all reported results, analyses, and claims, and we welcome community scrutiny and feedback.

7 REPRODUCIBILITY STATEMENT

To support reproducibility, we provide comprehensive implementation details throughout our paper. Key experimental configurations, such as draft expert selection, consensus scoring computation, and verdict model specifications, are documented in Section 3.4 and Section 4.1. Detailed prompt templates are presented in Appendix G. The supplementary material includes anonymized source code to further clarify the implementation steps and enable faithful reproduction of our results.

REFERENCES

- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- Xinyun Chen, Renat Aksitov, Uri Alon, Jie Ren, Kefan Xiao, Pengcheng Yin, Sushant Prakash, Charles Sutton, Xuezhi Wang, and Denny Zhou. Universal self-consistency for large language model generation. *arXiv preprint arXiv:2311.17311*, 2023.
- Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, et al. Paddleocr 3.0 technical report. *arXiv preprint arXiv:2507.05595*, 2025.
- 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.
- Chaoyou Fu, Yi-Fan Zhang, Shukang Yin, Bo Li, Xinyu Fang, Sirui Zhao, Haodong Duan, Xing Sun, Ziwei Liu, Liang Wang, et al. Mme-survey: A comprehensive survey on evaluation of multimodal llms. *arXiv preprint arXiv:2411.15296*, 2024.
- Jiale Fu, Yuchu Jiang, Junkai Chen, Jiaming Fan, Xin Geng, and Xu Yang. Fast large language model collaborative decoding via speculation. *arXiv preprint arXiv:2502.01662*, 2025a.
- Yichao Fu, Rui Ge, Zelei Shao, Zhijie Deng, and Hao Zhang. Scaling speculative decoding with lookahead reasoning. *arXiv preprint arXiv:2506.19830*, 2025b.
- 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.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- Fucai Ke, Joy Hsu, Zhixi Cai, Zixian Ma, Xin Zheng, Xindi Wu, Sukai Huang, Weiqing Wang, Pari Delir Haghighi, Gholamreza Haffari, Ranjay Krishna, Jiajun Wu, and Hamid Rezatofighi. Explain before you answer: A survey on compositional visual reasoning, 2025.
- Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding. In *International Conference on Machine Learning*, pp. 19274–19286. PMLR, 2023.

- Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges. *arXiv preprint arXiv:2501.02189*, 2025.
 - Baohao Liao, Yuhui Xu, Hanze Dong, Junnan Li, Christof Monz, Silvio Savarese, Doyen Sahoo, and Caiming Xiong. Reward-guided speculative decoding for efficient llm reasoning. *arXiv preprint arXiv:2501.19324*, 2025.
 - Shiyin Lu, Yang Li, Yu Xia, Yuwei Hu, Shanshan Zhao, Yanqing Ma, Zhichao Wei, Yinglun Li, Lunhao Duan, Jianshan Zhao, et al. Ovis2. 5 technical report. *arXiv preprint arXiv:2508.11737*, 2025.
 - Ahmed Masry, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, Firoz Kabir, Aaryaman Kartha, Md Tahmid Rahman Laskar, Mizanur Rahman, Shadikur Rahman, Mehrad Shahmohammadi, et al. Chartqapro: A more diverse and challenging benchmark for chart question answering. arXiv preprint arXiv:2504.05506, 2025.
 - Minesh Mathew, Viraj Bagal, Rubèn Pérez Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V Jawahar. Infographicvqa. *arXiv preprint arXiv:2104.12756*, 2021.
 - Rui Pan, Yinwei Dai, Zhihao Zhang, Gabriele Oliaro, Zhihao Jia, and Ravi Netravali. Specreason: Fast and accurate inference-time compute via speculative reasoning. *arXiv* preprint arXiv:2504.07891, 2025.
 - Jianing Qi, Xi Ye, Hao Tang, Zhigang Zhu, and Eunsol Choi. Learning to reason across parallel samples for llm reasoning. *arXiv preprint arXiv:2506.09014*, 2025.
 - Haozhan Shen, Kangjia Zhao, Tiancheng Zhao, Ruochen Xu, Zilun Zhang, Mingwei Zhu, and Jianwei Yin. Zoomeye: Enhancing multimodal llms with human-like zooming capabilities through tree-based image exploration. *arXiv preprint arXiv:2411.16044*, 2024.
 - Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. *arXiv preprint arXiv:2505.15966*, 2025a.
 - 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*, 2025b.
 - Liyan Tang, Grace Kim, Xinyu Zhao, Thom Lake, Wenxuan Ding, Fangcong Yin, Prasann Singhal, Manya Wadhwa, Zeyu Leo Liu, Zayne Sprague, et al. Chartmuseum: Testing visual reasoning capabilities of large vision-language models. *arXiv preprint arXiv:2505.13444*, 2025.
 - GLM-V Team. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025a.
 - Xiaomi LLM-Core Team. Mimo-vl technical report. arXiv preprint arXiv:2506.03569, 2025b.
 - Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7907–7915, 2025a.
 - Wenbin Wang, Yongcheng Jing, Liang Ding, Yingjie Wang, Li Shen, Yong Luo, Bo Du, and Dacheng Tao. Retrieval-augmented perception: High-resolution image perception meets visual rag. *arXiv* preprint arXiv:2503.01222, 2025b.
 - 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.
 - Heming Xia, Zhe Yang, Qingxiu Dong, Peiyi Wang, Yongqi Li, Tao Ge, Tianyu Liu, Wenjie Li, and Zhifang Sui. Unlocking efficiency in large language model inference: A comprehensive survey of speculative decoding. *arXiv* preprint arXiv:2401.07851, 2024.

- Wang Yang, Xiang Yue, Vipin Chaudhary, and Xiaotian Han. Speculative thinking: Enhancing small-model reasoning with large model guidance at inference time. *arXiv* preprint *arXiv*:2504.12329, 2025.
- Chen Zhang, Zhuorui Liu, and Dawei Song. Beyond the speculative game: A survey of speculative execution in large language models. *arXiv preprint arXiv:2404.14897*, 2024.
- Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. Mllms know where to look: Training-free perception of small visual details with multimodal llms. *arXiv preprint arXiv:2502.17422*, 2025a.
- 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*, 2025b.
- Wenting Zhao, Pranjal Aggarwal, Swarnadeep Saha, Asli Celikyilmaz, Jason Weston, and Ilia Kulikov. The majority is not always right: Rl training for solution aggregation. *arXiv* preprint arXiv:2509.06870, 2025.
- Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv* preprint arXiv:2505.14362, 2025.
- 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.

A DATASET STATISTICS

Table 4 reports the statistics of the four evaluation benchmarks. All benchmarks are based on **real-world images** rather than synthetic renderings, ensuring the authenticity and diversity of the evaluation setting. In particular, InfographicVQA, ChartMuseum, and ChartQAPro are information-intensive benchmarks: they contain thousands of images and questions with dense textual and numerical content, collected from diverse sources spanning 2594, 157, and 184 distinct web domains respectively (Mathew et al., 2021; Tang et al., 2025; Masry et al., 2025). This diversity reduces source bias and reflects practical challenges in multimodal reasoning.

HR-Bench 4K is used primarily to evaluate the generalization of our method, serving as a high-resolution benchmark with average sizes exceeding 4000×3500 pixels (Wang et al., 2025a). At the same time, one of our main benchmarks, InfographicVQA, also exhibits high-resolution characteristics. In particular, it frequently contains long-format images where diagrams span large vertical layouts (see the case in Figure 3), which further compounds the difficulty of grounding and multihop reasoning across dispersed regions.

Table 4: Statistics of the evaluation benchmarks. We report the number of images and questions, as well as the average image resolution (width \bar{W} and height \bar{H}).

Dataset	Real vs. Synthetic	#Images	#Questions	$ar{W}$	$ar{H}$
InfographicVQA (test)	Real	3288	579	1092	2771
ChartMuseum (test)	Real	1000	818	1551	1213
ChartQAPro	Real	1948	1341	1194	986
HR-Bench 4K	Real	800	200	4024	3503

B Costs

Table 5 reports the average inference cost of invoking GPT-40 as the verdict model per sample across benchmarks. Costs are estimated using the official GPT-40 pricing (version gpt-40-2024-08-06) as of September 2025. The small variation across benchmarks is mainly attributed to differences in reasoning path length, as more challenging tasks typically induce more complex reasoning. Overall, the inference cost of using GPT-40 as the verdict is under \$0.011 per sample across all benchmarks.

Table 5: Average inference cost of GPT-40 as verdict per sample across benchmarks. Costs are computed using GPT-40 (gpt-40-2024-08-06) pricing by September 2025.

Dataset	GPT-40 cost per sample
InfographicVQA	\$0.0068
ChartMuseum	\$0.0109
ChartQAPro	\$0.0071
HR-Bench 4K	\$0.0044

C SUPPLEMENTARY RECOVERY ANALYSIS ON INFORMATION-INTENSIVE BENCHMARKS

Table 6 and Figure 9 show the detailed recovery statistics across information-intensive benchmarks with GPT-40 as verdict. We break down SV performance under three scenarios: (i) cases where the majority of draft experts are correct (majority-correct), (ii) cases where only a minority are correct (minority-correct), and (iii) cases where none are correct (zero-correct). While the main paper focuses on the GPT-40's error cases to isolate SV's effectiveness, here we provide the full results for completeness.

Notably, in the zero-correct setting, recovery occurs rarely (2.6–24%), but it demonstrates verdict's surprising ability to infer the correct answer by synthesizing signal from entirely noisy reasoning.

Table 6: Recovery accuracy (%) with GPT-40 as verdict. Results are conditioned on whether GPT-40 itself can produce the correct answer.

	GPT-4o Correct			GF	T-40 Wrong	3
Dataset	Majority-	Minority-	Zero-	Majority-	Minority-	Zero-
Dataset	correct	correct	correct	correct	correct	correct
InfographicVQA	96.81	64.13	20.54	93.30	53.42	4.44
ChartMuseum	98.46	69.84	15.38	89.92	47.71	2.69
ChartQAPro	94.59	68.18	24.00	85.25	48.43	2.86

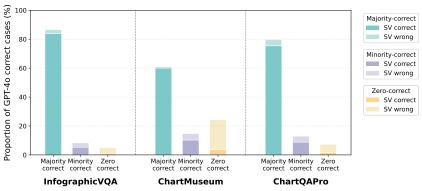


Figure 9: SV's correction ability on verdict's correct cases (GPT-40 as verdict), complementary to its error cases in the Figure 4.

ERROR ANALYSIS OF TOOL-DRIVEN PIPELINE

As mentioned in Section 2, tool-driven methods represent a line of work that augments visionlanguage reasoning with explicit zoom-in operations. The representative pipeline DeepEyes is designed to iteratively ground into image regions, and integrate them into the ongoing reasoning trajectory under an RL framework. This mechanism has proven effective on high-resolution benchmarks, where localized inspection of fine details is crucial.

However, DeepEyes is not specifically trained on our benchmarks, which require reasoning over information-intensive images with densely interleaved textual and visual elements. Its performance on Infographic VQA reveals the current limitations of such tool-based pipelines in this domain. We categorize the observed deficiencies into three core challenges:

- (i) Tendency toward literal grounding. DeepEyes is proficient at small-scale grounding but often focuses on literal text spans or legends rather than reasoning-critical regions. For example, when a question requires aligning numerical values with a chart axis, the model frequently grounds directly onto the answer text or nearby labels instead of the relevant data regions. This shortcut strategy works for simple queries but fails on complex reasoning on information-intensive images that require global comparison.
- (ii) Inefficient tool usage. Although DeepEyes is trained to iteratively apply zoom-in tools, we observe that it invokes only one zoom step in more than half of the test cases. Among the doublezoom cases, 92.8% duplicate the same bounding box, which serves only for verification rather than exploration. In some instances, the model zooms into empty areas or irrelevant regions.
- (iii) Lack of robustness on long and dense images. Information-intensive images often contain multipanel figures and dense annotations. DeepEyes cannot maintain a trajectory across multiple zoom steps, making it difficult to integrate dispersed evidence. As a result, tasks requiring cross-region synthesis, such as counting, sorting, or comparing across multiple subplots, remain challenging for it.

Overall, this analysis indicates that while tool-driven pipelines are promising for high-resolution inspection tasks, they face notable difficulties applying to information-intensive images without domain-specific supervision. In contrast, SV achieves strong performance without additional training, offering a simple and effective alternative for reasoning over complex multimodal inputs.

E ABLATION STUDY ON ADDITIONAL STRUCTURED IMAGE INPUT TO VERDICT

In our experimental setup in Section 4.1, we preprocess each image via PP-StructureV3, a document parsing model that generates Markdown representations capturing layout, textual blocks, and visual metadata (Cui et al., 2025). This structured representation is then converted into an additional image input for the verdict. To verify whether this input is critical or merely auxiliary, we conduct an ablation study (Table 7).

The results show that SV achieves substantial gains over the reasoning baseline even without structured input. With the structured input, performance is generally slightly improved, though the gain is negligible or even marginally lower in some cases. This pattern suggests that structured OCR-derived signals are not essential for SV's core performance, but may assist the verdict to distinguish among competing reasoning paths.

Table 7: Ablations on additional structured image input to the verdict.

InfographicVQA ChartMuseum ChartQAPro

Input	InfographicVQA	ChartMuseum	ChartQAPro
	ANLS	Acc	Acc
Reasoning baseline	84.8	48.0	57.3
w/o input	88.3	49.5 49.3	59.4
w input	88.4		64.0

F QUALITATIVE EXAMPLE

Figure 10 illustrates a case where all three draft experts produced incorrect reasoning paths, yet the verdict successfully corrected the answer. Specifically, the draft experts faced different types of failures: some mis-extracted information from the image, others extracted the key information correctly but failed to sort the values properly, and thus all generated wrong answers. Interestingly, the verdict itself, when asked directly, also tends to answer "Australia" incorrectly. However, when analyzing the noisy and conflicting reasoning paths together, the verdict was able to recover the correct answer (Portugal).

This example complements the main results section: while Figure 3 illustrates recovery from minority-correct experts, here we present a zero-correct case to show that SV can still synthesize the correct solution even when all drafts and the verdict individually fail.

G PROMPT TEMPLATES

G.1 CHAIN-OF-THOUGHT PROMPTS

As described in Section 4.1, we employ a Chain-of-Thought prompt for each consensus expert to generate reasoning paths and apply it identically when evaluating baselines. For InfographicVQA and HR-Bench 4K, we use the same CoT prompt. For ChartMuseum (Tang et al., 2025), we adopt its official reasoning prompt, and adapt that prompt strategy to ChartQAPro, given their similarity in task complexity. Since ChartQAPro requires different prompt templates tailored to question types (Masry et al., 2025), we first follow its official template per question type, then concatenate it with our reasoning prompt.

The reasoning prompts for these datasets are shown in Figure 11.

G.2 PROMPTS FOR VERDICT

The user prompts used in the verdict stage are identical across datasets except for the final instruction sentence, which is customized (see Figure 13). For GPT-40 as verdict, the system prompt is shown in Figure 12. For Qwen-2.5-VL-72B-Instruct as verdict, we prepend its system prompt at the beginning of the user prompt.

H THE USE OF LARGE LANGUAGE MODELS (LLMS)

In this work, we used LLMs solely for auxiliary tasks such as language polishing, prompt refining, and proofreading. Importantly, these interventions did not contribute any main scientific insight, experimental design, or methodological advance. All core ideas, theoretical developments, experiments, analyses, and claims in this paper are the work of the authors. Accordingly, authors assume full responsibility for all content, consistent with ICLR policy that LLMs are not eligible for authorship.

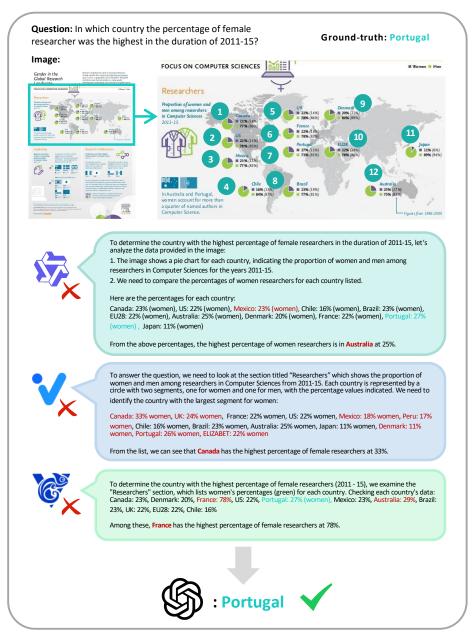


Figure 10: A qualitative zero-correct case corrected by verdict. All three draft experts fail due to errors in extracting or sorting visual information, yet the verdict synthesizes their noisy reasoning paths to recover the correct answer (i.e., Portugal).

InfographicVQA / HR-Bench 4K

Question: {QUESTION} Please think step-by-step about the image to answer the question using a single word or phrase enclosed within $\begin{tabular}{ll} \begin{tabular}{ll} \begin{tabul$

ChartMuseum

```
Please answer the question using the chart image.

Question: {QUESTION}

Please first generate your reasoning process and then provide the user with the answer. Use the following format:

<think>
... your thinking process here ...
</think>
<answer>
... your final answer (entity(s) or number) ...
</answer>
```

ChartQAPro

```
{PROMPT for a specific question type}

Please first generate your reasoning process and then provide the user with the answer. Use the following format:

<think>
... your thinking process here ...
</think>
<answer>
... your final answer (entity(s) or number) ...
</answer>
```

Figure 11: Prompt templates for reasoning.

All benchmarks

You are a vision-and-language judge. Follow the instructions strictly $\boldsymbol{\cdot}$

Figure 12: System prompt template for verdict.

919 920

921

922

923

924

925

926

927

928

929

930

931

932

933

934 935 936

937 938

939

940

941

942

943

944

945

946

947

948

949

951

952953954

955 956

957

958

959

960

961

962

963

964

965

966

967

968

969

```
InfographicVQA / ChartMuseum
Ouestion:
{OUESTION}
 -- Model 1 ---
Reasoning:
{Reasoning path 1}
Proposed Answer: {Answer 1}
--- Model 2 ---
Reasoning:
{Reasoning path 2}
Proposed Answer: {Answer 2}
 -- Model 3 --
Reasoning:
{Reasoning path 3}
Proposed Answer: {Answer 3}
Given the raw image, the layout-annotated image, the question, and
the reasoning from three models, please give the final answer using a
 single word or phrase enclosed within \\boxed{{}}.
```

```
ChartQAPro
Question:
{QUESTION}
 -- Model 1 ---
Reasoning:
{Reasoning path 1}
Proposed Answer: {Answer 1}
 -- Model 2 ---
Reasoning:
{Reasoning path 2}
Proposed Answer: {Answer 2}
--- Model 3 ---
Reasoning:
{Reasoning path 3}
Proposed Answer: {Answer 3}
Given the raw image, the layout-annotated image, the question, and
the reasoning from three models, please directly give the final
answer enclosed within \\boxed{{}}.
```

```
HR-Bench 4K
Question:
{QUESTION}
  - Model 1 ---
Reasoning:
{Reasoning path 1}
Proposed Answer: {Answer 1}
--- Model 2 ---
Reasoning:
{Reasoning path 2}
Proposed Answer: {Answer 2}
--- Model 3 ---
Reasoning:
{Reasoning path 3}
Proposed Answer: {Answer 3}
Given the image, the question, and the reasoning from three models,
please directly give the final answer with the option's letter
enclosed within \\boxed{{}}.
```

Figure 13: User prompt templates for verdict.