MODELS SEE MORE THAN THEY SAY: THE ANSWER BOTTLENECK IN MULTIMODAL LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal language models can encode useful visual information and still answer incorrectly. We investigate this gap through the *Guided Answer Space (GAS)*, an explicit set of candidate answers specifying the alternatives the model must distinguish. By varying this space while keeping the image, question, and target answer fixed, and intervening on intermediate representations, we provide evidence for an *Answer Bottleneck* in determining correct answers from encoded visual evidence. Crucially, intermediate visual representations already provide substantial support for accurate open-ended answers: they help recover accuracy lost to image corruption and retain much of their contribution even when their states are held fixed in later layers. Across eleven models on TextVQA and GQA, two-choice selection depends on visual tokens through fewer layers than open-ended answering, with a median difference equivalent to 42% of model depth; expanding the GAS extends this dependence to later layers. Building on these findings, we develop a two-stage inference procedure that constructs a GAS from other models' answers and then selects the final answer from it, without ground-truth access during either stage. Using three of the models, this procedure corrects 23.7% of original open-ended errors and improves overall accuracy by 3.56 percentage points. Our findings show that answer requirements shape the decoder's reliance on visual evidence and that the GAS can improve accuracy through two-stage inference.

1 INTRODUCTION

Multimodal language models combine visual encoders with language models to perform a range of visual tasks through natural language, including scene understanding, text recognition, and document understanding (Li et al., 2025; Bai et al., 2025). Advances in visual encoding and multimodal training have improved their ability to extract information from images. However, accurate answering also requires the decoder to relate that information to the question and generate the correct response. Understanding how visual representations contribute to model predictions is therefore central to improving multimodal capabilities: it informs which components to improve during training, how to elicit accurate responses at inference time, and what visual capabilities benchmark scores actually measure (Liu et al., 2025b; Chen et al., 2024).

Yet recent studies reveal a gap between visual representations and final answers: readouts from a model's visual representations can identify the correct answer on perceptual tasks even when the model itself answers incorrectly (Fu et al., 2025; Liu et al., 2025a). This discrepancy motivates the **Answer Bottleneck** hypothesis: even when useful visual information has been encoded, the decoder may still fail to use it effectively to determine the correct answer. Figure 1 illustrates the role of answer requirements: for the same image and question, Qwen2.5-VL-7B produces an incorrect open-ended answer but selects the correct answer from supplied candidates. Improving multimodal answering therefore requires understanding not only what the model represents, but **how it uses those representations to determine an answer**. Reliably establishing the contribution of existing evidence is particularly challenging because later decoder layers can both refine visual representations and use them to determine a response (Neo et al., 2025; Yu & Lee, 2025).

To investigate this limitation, we introduce the **Guided Answer Space (GAS)**, an explicit set of candidate answers that specifies the alternatives the model must distinguish. By varying whether the GAS is supplied and how many candidates it contains, we change the answer requirements while keeping the image, question, and target answer fixed. Prior comparisons of multiple-choice

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

and open-ended evaluation show that answer requirements affect measured performance (Rosenthal et al., 2026; Chandak et al., 2025). However, these behavioral comparisons leave unresolved whether restricting the answer space also changes how visual representations contribute to answering. We therefore ask: *how does the GAS change the decoder's reliance on visual evidence, and can it help the decoder use that evidence more effectively to determine the correct answer?*

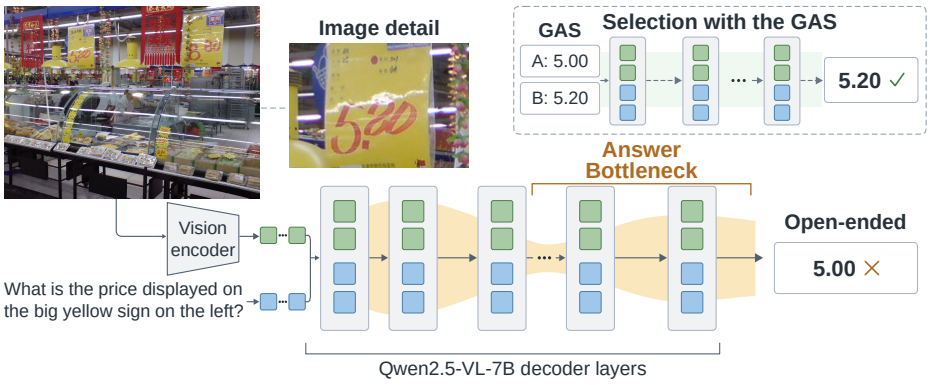

Figure 1: **The Answer Bottleneck.** The later decoder layers under the amber bracket illustrate difficulty in determining a correct answer from encoded visual evidence. On the same TextVQA image and question, Qwen2.5-VL-7B gives "5.00" in open-ended answering and selects "5.20" from the GAS containing its original answer and the reference answer (inset). Responses are recorded; the decoder is schematic.

We measure this reliance by zeroing visual token states from different decoder layers onward and evaluating answer accuracy. We find that two-choice selection depends on visual tokens through substantially fewer layers than open-ended answering, with a median difference equivalent to **42% of model depth** across eleven models on TextVQA and GQA. Moreover, expanding the GAS extends this dependence to later layers, showing that its size affects visual dependence even within selection. Using the original open-ended prompt, we then quantify the contribution of visual representations at these earlier depths. Crucially, transferring these states recovers much of the accuracy lost to image corruption, establishing their causal contribution to open-ended answering. Remarkably, later layers preserve much of the visual contribution to open-ended accuracy **even when the intermediate visual states are held fixed**. These results support the Answer Bottleneck hypothesis: accurate answering depends not only on the evidence encoded in visual representations, but also on how effectively the decoder uses that evidence.

Motivated by these findings, we develop a **two-stage inference** procedure with the GAS: first constructing a candidate set from other models' answers, then selecting an answer from that set using the image and question. Without ground-truth access during either stage, this procedure corrects 23.7% of original open-ended errors and improves overall accuracy by **3.56 percentage points (pp)**. Direct selection exceeds frequency voting by 2.56 pp on identical candidate pools, demonstrating the contribution of the selection stage under a fixed budget for candidate generation.

2 Background and Hypothesis

To investigate how the GAS affects the decoder's reliance on visual representations, we first describe how visual information enters multimodal autoregressive generation (Section 2.1), then define visual token dependence and its layer-wise measurement (Section 2.2). With this measurement established, we formulate the Answer Bottleneck hypothesis and two conjectures that guide the subsequent experiments (Section 2.3).

2.1 Multimodal Autoregressive Generation

The models in our study follow a widely used design that connects a vision encoder to a decoder-only language model (Liu et al., 2023; Li et al., 2025; Bai et al., 2025). Given an image x and a text prompt u containing the question and any supplied GAS, the vision encoder E_{vis} extracts image features, and a connector P maps them into the language model's embedding space:

$$Z_{\text{vis}}(x) = P(E_{\text{vis}}(x)) \,. \tag{1}$$

The sequence $Z_{\text{vis}}(x)$ forms the visual tokens in the input to the language model backbone. We refer to the hidden states at these positions as *visual token representations*. For a response $y = (y_1, \ldots, y_T)$, the model factorizes the conditional probability as

$$p_\theta(y \mid x, u) = \prod_{t=1}^{T} p_\theta(y_t \mid Z_{\text{vis}}(x), u, y_{<t}) \,, \tag{2}$$

where $y_{<t} = (y_1, \ldots, y_{t-1})$ denotes the previously generated tokens and θ denotes the model parameters. We index the L decoder layers by ℓ. Here, t indexes generation steps, whereas ℓ indexes layers. At each generation step, text tokens can attend to preceding visual tokens at every decoder layer, allowing the model to use visual information throughout response generation.

For a given image and question, we define a Guided Answer Space (GAS) as a finite set of candidate answers explicitly supplied in the prompt. In *selection*, the model chooses from this set and returns either the selected answer text or its option label; in *open-ended answering*, it generates an answer without supplied candidates. The candidates may be supplied by the benchmark, constructed for controlled experiments, or generated by models.

2.2 Measuring Visual Token Dependence

To compare the decoder's reliance on visual representations under different answer requirements, we measure how answer accuracy changes when visual token states are set to zero from different starting layers onward. We refer to this sensitivity as **visual token dependence**, or visual dependence for brevity. Let $\mathcal{V}$ denote the visual token positions and $h_k[\mathcal{V}]$ their hidden states at the input to decoder layer k, where $k \in \{0, \ldots, L-1\}$. Starting at layer ℓ, we zero these states at every remaining decoder layer:

$$h_k[\mathcal{V}] \leftarrow 0, \qquad k = \ell, \ldots, L-1. \tag{3}$$

Let A^{clean} denote the baseline accuracy and $A^{\text{remove}}(\ell)$ the accuracy when zeroing starts at layer ℓ. We summarize this dependence by the **visual token dependency depth**: the earliest tested layer from which visual token states can be zeroed throughout the remaining layers while keeping the accuracy loss within a specified tolerance:

$$\ell_{\text{dep}} = \min \left\{ \ell \in \mathcal{K} \;\middle|\; A^{\text{clean}} - A^{\text{remove}}(\ell) \le \varepsilon \right\}, \tag{4}$$

where $\mathcal{K}$ is the set of evaluated starting positions and ε is the allowed drop in accuracy. A larger ℓ_{dep} means that accurate answering depends on visual token representations at later layers. To compare models with different numbers of layers, we use the normalized depth $\bar{\ell}_{\text{dep}} = \ell_{\text{dep}}/L$. We write ℓ_{sel} and ℓ_{open} for selection and open-ended dependency depths, respectively. Appendix A details the measurement settings and sensitivity analyses.

2.3 The Answer Bottleneck Hypothesis

Because later decoder layers can both refine visual representations and use existing evidence to determine an answer, continued visual token dependence may reflect either process.

Answer Bottleneck Hypothesis. *A model's ability to answer accurately can be limited by how effectively it uses encoded visual evidence to determine the correct answer, even when intermediate representations already provide useful evidence. We call this limitation the Answer Bottleneck.*

Supplying the GAS reformulates answering as selection from a specified set of candidates, allowing us to vary the alternatives the model must distinguish. If fewer alternatives require less processing of visual evidence, a smaller GAS should allow visual token states to be zeroed from earlier layers onward with little loss in accuracy.

Conjecture 1 (Answer Space and Visual Dependence) *For the same image, question, and target answer, supplying a small GAS should reduce visual token dependency depth relative to open-ended answering; expanding the GAS should increase this depth.*

Conjecture 1 concerns how answer requirements affect visual dependence. Because the Answer Bottleneck concerns determining answers from already encoded evidence, evaluating this hypothesis also requires establishing whether earlier representations computed under the original open-ended prompt contribute to accurate answers.

Conjecture 2 (Early Visual Support for Open-Ended Answers) *Under the original open-ended prompt, intermediate visual representations should already provide substantial support for accurate answers at depths earlier than the open-ended dependency depth ℓ_{open}. We further expect these representations to preserve much of their contribution to answer accuracy even when their states are held fixed in later layers.*

We next examine how the GAS changes visual dependence (Conjecture 1; Section 3), then ask how much support for open-ended answering is already present in intermediate representations (Conjecture 2; Section 4). Building on these findings, we study candidate count and answer diversity as complementary aspects of the answer space and test whether model-generated candidates can improve answer accuracy (Section 5).

3 The Guided Answer Space Shapes Visual Dependence

We first compare visual dependence across benchmarks with different answer requirements (Section 3.1). Because their images and questions also differ, we then test Conjecture 1 on the same examples with and without the GAS, isolating its effect for a fixed visual target (Section 3.2). Finally, we compare the effects of changing answer requirements within a task with differences between tasks (Section 3.3).

Experimental scope. Unless otherwise stated, experiments use eleven models from five families: LLaVA-OV-Qwen2 (0.5B, 7B), LLaVA-OV-1.5 (4B, 8B), Qwen2-VL (2B, 7B), Qwen2.5-VL (3B, 7B), and InternVL3 (1B, 8B, 14B). Model specifications and decoder depths are reported in Appendix A.1.

3.1 Visual Dependence across Answering Settings

We begin with eight benchmarks under their original answer requirements: POPE requires yes/no answers; BLINK-spatial and BLINK-depth use two-choice selection; MMBench and ScienceQA use multiple-choice selection; GQA requires short open-ended answers; and TextVQA and DocVQA have a broader range of possible answers. For each model and benchmark, we vary the layer from which visual token states are zeroed and record answer accuracy. Figure 2a shows the resulting curves for LLaVA-OV-1.5-8B; results for all eleven models appear in Appendix A.3.

For LLaVA-OV-1.5-8B, the normalized dependency depth $\bar{\ell}_{\text{dep}}$ (Section 2.2) ranges from 0.42 to 0.53 across the five benchmarks requiring yes/no or multiple-choice responses. By contrast, the open-ended depth $\bar{\ell}_{\text{open}}$ reaches 0.75 on GQA and 0.92 on both TextVQA and DocVQA. Across these benchmarks, broader answer spaces are associated with deeper visual dependence, a pattern that also appears across the eleven models (Appendix A.3).

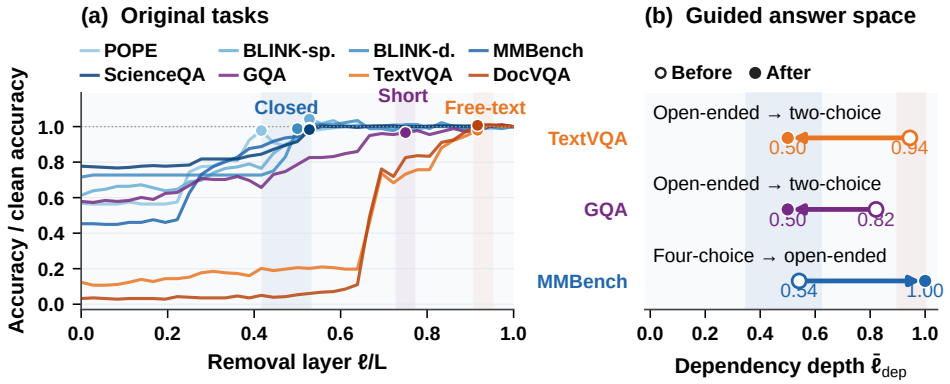

Figure 2: **The GAS changes visual dependency depth.** **(a)** Visual token removal on LLaVA-OV-1.5-8B; markers indicate ℓ_{dep}. **(b)** Median $\bar{\ell}_{\text{dep}}$ when adding candidates to TextVQA/GQA or removing them from MMBench. MMBench compares full-set selection with open-ended evaluation on initially correct examples.

3.2 Supplying the GAS for the Same Visual Target

To isolate the effect of the GAS from differences in images and questions across benchmarks, we compare open-ended answering and two-choice selection on the same 300 examples per task from

TextVQA and GQA. Specifically, we construct the GAS from the correct answer and one distractor while keeping the image, question, and target answer unchanged. The two candidates are presented as options A and B in randomized order, and the model is asked to return the option label. For MMBench, we make the reverse change by removing the original answer choices and asking the model to generate the same target answer directly.

Replacing open-ended answering with two-choice selection reduces dependency depth on both tasks (Figure 2b). Specifically, the median normalized depth decreases from 0.94 to 0.50 on TextVQA and from 0.82 to 0.50 on GQA. Importantly, the same ordering holds for all eleven models on both tasks. The median difference in normalized depth is 0.42, equivalent to approximately fifteen layers in a 36-layer model (Appendix Table 6). Conversely, removing the options from MMBench increases visual dependency depth. Across the eleven models, the median normalized depth rises from 0.54 for selection on the full set to 1.00 for open-ended answering on examples initially answered correctly in that setting. The same ordering also holds when three models are evaluated on identical subsets in both settings (Appendix Table 10).

To quantify the image contribution under two-choice selection, we replace the image with Gaussian noise while leaving the question and candidates unchanged. This reduces selection accuracy in all eleven models on both tasks, by 17.7 pp on average (Appendix Table 8). We further evaluate semantically harder AutoConverter distractors (Zhang et al., 2025b) on three models and observe the same ordering between selection and open-ended dependency depths ℓ_{sel} and ℓ_{open} (Appendix Figure 10). We also examine whether the depth difference observed with the GAS remains under the original open-ended prompt. To do so, we rank the correct answer and one distractor by likelihood without listing them in the prompt. Across three models on both tasks, ranking yields an earlier dependency depth than free generation, further supporting the link between answer requirements and visual token dependence (Appendix C.4).

These controlled comparisons show that supplying the GAS systematically **reduces visual token dependency depth** even when the image, question, and target answer remain unchanged (Conjecture 1).

3.3 Comparing GAS Effects with Task Differences

Having established that supplying the GAS changes visual dependence, we next assess the magnitude of this effect relative to task differences. TextVQA and GQA differ in both visual content and answer diversity. We therefore compare their absolute difference in normalized dependency depth $\bar{\ell}_{\text{dep}}$ under the original prompts and when both tasks use two-choice selection, alongside the changes caused by supplying the GAS within each task (Figure 3).

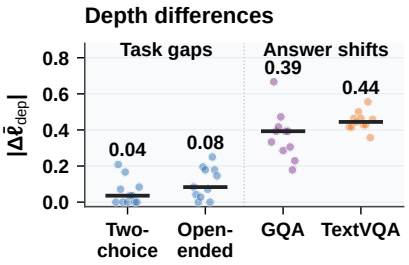

Figure 3: **Effects of supplying the GAS exceed the task difference.** Absolute differences in $\bar{\ell}_{\text{dep}}$ between TextVQA and GQA (left) and between open-ended answering and selection within each task (right). Dots: models; bars: medians.

The median difference between tasks is 0.08 under the original prompts and 0.04 under two-choice selection. By comparison, supplying the GAS within a task changes the normalized depth by a median of 0.44 on TextVQA and 0.39 on GQA (Appendix Table 7). These changes are **approximately five times** the task difference under the original prompts and **ten times** the difference under two-choice selection.

For TextVQA and GQA, the answer space is therefore a major factor in visual dependence: changing the answer requirement for a fixed visual target produces larger depth shifts than the difference between the two tasks.

4 Causal Contributions of Intermediate Visual Representations

The earlier selection depths ℓ_{sel} raise a central question: how much support for open-ended answering is already present at those depths? To test Conjecture 2, we use representations computed under the original open-ended prompt, first measuring their causal contribution (Section 4.1) and then determining how much of that contribution later layers retain when the visual states are held fixed (Section 4.2).

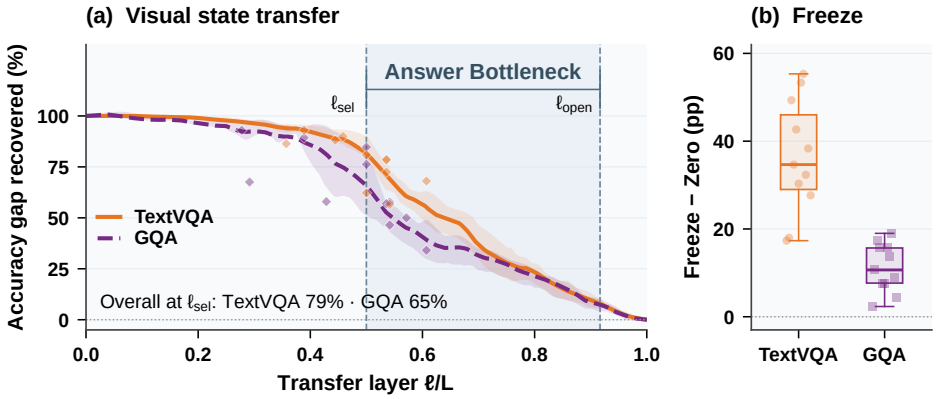

Figure 4: **Earlier visual representations support open-ended answers.** **(a)** Transfer recovery across layers: mean and interquartile range over eleven models per task, with diamonds at ℓ_{sel}. Shading spans the median $\bar{\ell}_{\text{sel}}$ to the median $\bar{\ell}_{\text{open}}$ across both tasks. **(b)** Freeze − Zero accuracy at ℓ_{sel}. Boxplots show medians and interquartile ranges, with min–max whiskers.

4.1 Causal Support for Open-Ended Answers

To establish the causal contribution of intermediate visual representations, we design a *Transfer* experiment that first corrupts the image input and then restores the visual states at a chosen decoder layer. With the original question and open-ended prompt held fixed, the resulting recovery in accuracy measures how much the restored representations contribute to answering.

Specifically, we compare three runs on the same examples. *Clean* uses the original image, whereas *Noise* replaces the image with Gaussian noise. *Transfer* follows the Noise run up to layer ℓ, replaces the visual token representations once with those from the corresponding Clean run, and then continues the remaining computation and autoregressive generation. We perform Transfer at the selection dependency depth ℓ_{sel} measured independently in Section 3.2.

Let A^{clean}, A^{noise}, and $A^{\text{transfer}}(\ell)$ denote the open-ended accuracies of these runs on the same examples. We report the fraction of the accuracy gap between Clean and Noise recovered by Transfer:

$$R(\ell) = \frac{A^{\text{transfer}}(\ell) - A^{\text{noise}}}{A^{\text{clean}} - A^{\text{noise}}}. \tag{5}$$

Already at the earlier selection depth ℓ_{sel}, Transfer improves open-ended accuracy over Noise for all eleven models on both tasks. Overall recovery is **79% for TextVQA and 65% for GQA**, computed as the summed Transfer gains divided by the summed Clean–Noise accuracy gaps across models (Appendix C.1). Crucially, representations obtained under the original open-ended prompt already provide substantial causal support for answering at the earlier selection depth ℓ_{sel}. The layer-wise scan further shows substantial recovery near ℓ_{sel}, followed by a pronounced decline toward the final decoder layers (Figure 4a).

A same-layer control on LLaVA-OV-1.5-8B with TextVQA further reveals a pronounced difference between visual and text transfer. Specifically, we repeat Transfer at the text token positions, replacing their states with the corresponding Clean states from the same layer. Visual transfer recovers 93.0% of the Clean–Noise accuracy gap, whereas text transfer recovers 0.5%, identifying the visual states as the main source of recovery in this comparison.

4.2 Continued Access to Intermediate Visual Representations

Building on the recovery results, we further quantify how much of this visual contribution later layers can retain when the intermediate states are held fixed. We therefore design a *Freeze* experiment that separates continued access to visual evidence from further updates to its representations: at every subsequent layer, we restore the same visual states from layer ℓ while allowing the text states to continue updating.

For comparison, *Zero* sets the visual states to zero from the same layer onward, using the operation in Section 2.2. Both conditions retain the original image and open-ended prompt, so their accuracy difference measures the support retained in the fixed representations. We express this gain relative to the Clean–Zero accuracy gap to quantify how much of the original visual contribution is preserved.

Even with visual states held fixed at $\ell = \ell_{\text{sel}}$, Freeze achieves higher open-ended accuracy than Zero in all eleven models on each task, with median gains of 34.7 pp on TextVQA and 10.7 pp on GQA (Appendix C.2, Table 13). Figure 4b shows all model-task gains; the full layer scan appears in Appendix Figure 11.

Freeze recovers a median of **71% of the Clean–Zero accuracy gap** across the 22 model-task combinations. By showing that intermediate visual representations already contribute to open-ended answers and retain much of that contribution when held fixed, these experiments provide causal evidence for Conjecture 2. The interval $[\ell_{\text{sel}}, \ell_{\text{open}})$ places this early support in relation to the decoder's continued visual dependence.

5 The Guided Answer Space: Analysis and Two-Stage Inference

The preceding results show that answer requirements shape visual dependence and that intermediate representations already support open-ended answering. We now examine how the range of possible answers relates to this dependence through two complementary analyses: controlled changes in GAS size within a task (Section 5.1) and variation in answer diversity across benchmarks (Section 5.2). Building on this analysis, we develop two-stage inference with model-generated candidates to test whether the GAS can improve answer accuracy (Section 5.3).

5.1 Varying the Size of the GAS

Whereas Section 3.2 contrasts selection with open-ended answering, we now vary the size of the GAS while holding the output format fixed to isolate the effect of candidate count within selection. Starting with two choices, we increase the candidate set to $N = 4$ and 8 on the same TextVQA and GQA examples, keeping the image, question, and target answer fixed. We use 300 examples per task with LLaVA-OV-1.5-8B, Qwen2.5-VL-7B, and InternVL3-8B. To distinguish candidate count from the required response, we repeat each comparison with two output formats, one returning the option label and the other returning the answer text, holding the candidates and their order fixed across formats.

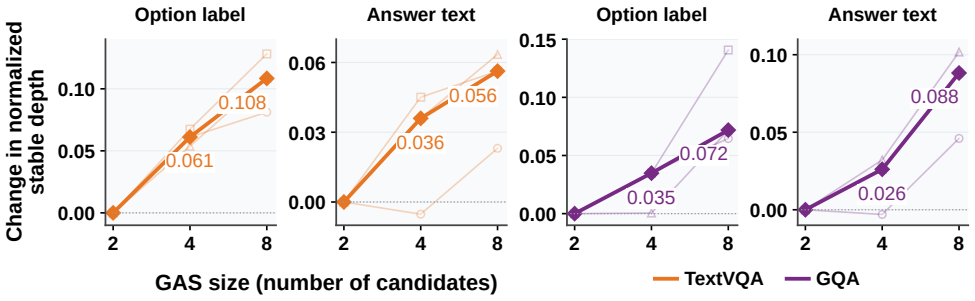

Figure 5: **Expanding the GAS extends visual dependence to later layers.** Mean within-example change in stable depth s_i from $N = 2$, normalized by L. Thin curves show models; thick curves and labels show their median.

Because dataset-level accuracy thresholds can obscure changes in individual examples, we complement the aggregate analysis with each example's *stable depth* s_i, defined as the earliest starting layer such that removal at that layer or any later evaluated layer yields a correct answer. For each change in candidate count, we evaluate the same examples for which this depth is defined under both conditions. Increasing the number of candidates from two to eight increases the mean depth in all twelve combinations of model, task, and output format. After normalization by model depth, the median of these twelve mean increases is 0.077 (Figure 5). Thus, **expanding the GAS increases mean stable depth under both output formats**, linking visual dependence to the number of answer alternatives the model must distinguish (Conjecture 1). Moreover, candidate count has a more consistent effect than switching between the two output formats in this experiment. Appendix D.1 details candidate construction, paired analyses, and controls for the output format.

5.2 Answer Diversity across Benchmarks

Section 5.1 changes the answer space by varying GAS size within a task; here we examine the corresponding variation in answer distributions across the original benchmarks. We quantify answer

diversity using entropy, $H = -\sum_z p(z) \log_2 p(z)$, where $p(z)$ is the probability of answer z under the benchmark-specific distribution specified in Appendix D.2.

Across eight task means computed from 81 model-task combinations, more diverse answer distributions are associated with greater normalized visual token dependency depths ($R^2 = 0.936$; Figure 6). The yes/no and multiple-choice benchmarks have lower answer entropy and shallower mean dependency depths, whereas GQA, TextVQA, and DocVQA have more diverse answers and greater mean dependency depths. Thus, the effect of GAS size within a task has a corresponding pattern across the original benchmarks: broader answer spaces are associated with visual dependence at later layers. Regression and robustness analyses appear in Appendix D.4.

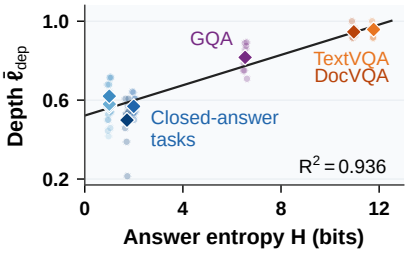

Figure 6: **Answer diversity and visual dependence.** Points show 81 model-task combinations; diamonds show eight task means; the line fits these means.

5.3 Two-Stage Inference with the GAS

The effect of the GAS raises a practical question: can guiding the model with candidate answers improve accuracy? To assess the potential of the GAS for correcting open-ended errors, we supply the correct answer as one of two candidates. Selection corrects 90.9% of original errors across six model-task combinations (Appendix E.1).

To realize this opportunity without access to ground-truth answers, we develop a two-stage inference procedure with the GAS. In **GAS construction**, each of the three models in Section 5.1 serves as the target, while the other two each generate four answers to the same image and question. We deduplicate these eight answers into a candidate set. In **answer selection**, when multiple candidates remain, we present them as labeled options with the image and question. We use direct selection, returning the answer whose label has the highest output logit. For comparison, we apply the same procedure to eight answers generated by the target itself. Each target is evaluated on 300 TextVQA and 300 GQA examples.

Crucially, with candidates from other models, overall accuracy increases from 65.3% to 68.9%, a net gain of **3.56 pp** (Figure 7a,b). On the original open-ended errors, these candidate pools contain a correct answer in 42.1% of cases, and the procedure corrects 23.7% of all original errors. By comparison, candidates from the target itself yield 24.2% coverage and 12.2% error recovery (Figure 7c).

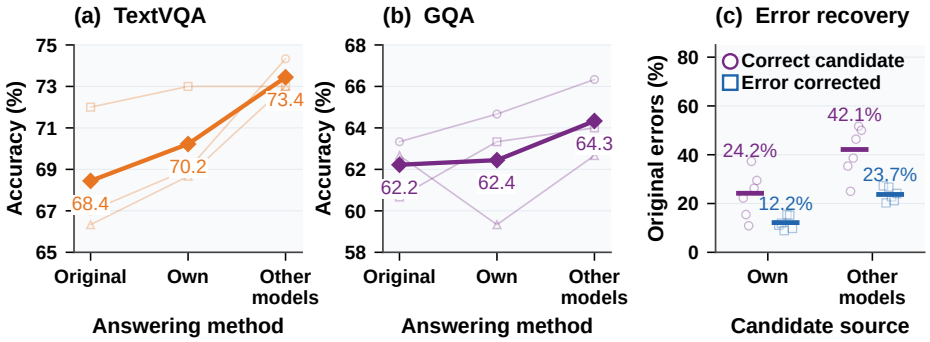

Figure 7: **Two-stage inference with the GAS improves answer accuracy. (a,b)** Accuracy with the original prompt or candidates from the target model (Own) or other models. Thin lines show individual models; thick lines show overall accuracy. **(c)** Correct-candidate coverage and error recovery among original errors, by model and task (points) and overall (bars).

To distinguish the contribution of selection from candidate generation, we compare four methods on identical candidate pools (Table 1). Frequency voting selects the most frequent answer, while candidate likelihood ranks candidates by mean per-token log-likelihood under the original open-ended prompt. Both answer text generation and direct selection receive the GAS: the former generates the answer text, whereas the latter scores option labels as described above. Importantly, direct selection achieves the highest accuracy, exceeding frequency voting by **2.56 pp**. Among the 827 predictions

requiring a choice between multiple distinct candidates, this advantage reaches 5.56 pp. The gains therefore depend on both which answers are supplied and how the target model selects among them. Appendix E.1 reports full results, uncertainty estimates, and ablations for candidate selection.

Table 1: **Direct selection achieves the highest accuracy on identical candidate pools.** All methods use identical candidate pools from other models and the same budget for candidate generation, with 1,800 predictions. Selection costs may differ. Original open-ended accuracy is 65.3%.

Method	Final accuracy (%)	Net gain (pp)
Frequency voting	66.3	+1.00
Candidate likelihood	65.0	−0.33
Answer text generation	67.7	+2.33
Direct selection	**68.9**	**+3.56**

6 Related Work

Visual representations and information flow. Layer-wise analyses trace visual information through multimodal models (Neo et al., 2025; Basu et al., 2024; Huang et al., 2025; Shi et al., 2025), while probes and internal readouts recover visual content from intermediate states (Jiang et al., 2025; Krojer et al., 2026; Liu et al., 2025a). Related studies examine discrepancies between encoded visual information and final answers (Fu et al., 2025; Nooralahzadeh et al., 2026). Yu & Lee (2025) further vary prompts and output formats to study visual grounding, reasoning, and answer decoding.

Efficient computation with visual tokens. Visual token pruning reduces inference cost by removing or compressing tokens (Chen et al., 2025; Lin et al., 2025; Tong et al., 2025; Wu et al., 2026; Zhang et al., 2025a), with behavior varying across tasks (Wen et al., 2025; Endo et al., 2025; Wang et al., 2026). ShortV (Yuan et al., 2025) instead freezes visual token updates to reduce computation. We use freezing to examine whether intermediate visual representations continue to support open-ended answering without further updates.

Answer spaces and answer selection. AutoConverter constructs challenging multiple-choice VQA questions from open-ended examples (Zhang et al., 2025b), while evaluation studies examine sensitivity to option labels and prompt formats (Rosenthal et al., 2026) and alternatives to multiple-choice scoring (Chandak et al., 2025). For inference, Fu et al. (2023) combine answer generation and selection in VQA. We examine how the GAS affects visual dependence and evaluate answer selection on identical candidate pools to separate its contribution from candidate generation.

7 Discussion

Our findings support the Answer Bottleneck hypothesis: answer accuracy depends not only on the information encoded in visual representations, but also on how effectively the decoder uses it to determine an answer. Crucially, the GAS changes visual dependence while the image, question, and target answer remain fixed. Moreover, **intermediate visual representations preserve substantial support for open-ended answering even when their states are held fixed** in later layers. Their continued contribution therefore motivates improving how the decoder uses existing evidence, alongside improving the representations themselves.

Building on this insight, two-stage inference with the GAS improves accuracy through candidate construction and answer selection. The higher coverage of candidates from other models creates more opportunities to correct open-ended errors, while direct selection outperforms frequency voting on identical candidate pools, establishing the contribution of the selection stage. Consequently, improving candidate coverage and the model's ability to distinguish the available alternatives are complementary directions for more accurate multimodal answering.

Limitations. Our study focuses on decoder-only multimodal models and primarily on short answers to single-image questions. Extending the analysis to longer responses, video, and multi-turn interaction would help determine how the effect of the GAS on visual dependence generalizes to more complex multimodal tasks.

AI USE STATEMENT

We used generative AI tools to assist with experimental code implementation, synthetic data generation, language polishing, figure editing, and literature search. The authors reviewed all AI-assisted work and take responsibility for the final paper.

ETHICS STATEMENT

Our work uses publicly available benchmarks and open model checkpoints and involves no human subjects or private data. We do not foresee direct negative societal impacts beyond the general risks of improving multimodal language models.

REPRODUCIBILITY STATEMENT

All interventions, thresholds, and evaluation settings are specified in Appendices A–F, with complete per-model-task results tabulated there. Anonymized code is provided in the supplementary material.

REFERENCES

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. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Samyadeep Basu, Martin Grayson, Cecily Morrison, Besmira Nushi, Soheil Feizi, and Daniela Massiceti. Understanding information storage and transfer in multi-modal large language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 7400–7426. Curran Associates, Inc., 2024. doi: 10.52202/079017-0237. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/0dfe31d6e703e138d46a7d2fced38b7c-Paper-Conference.pdf.

Nikhil Chandak, Shashwat Goel, Ameya Prabhu, Moritz Hardt, and Jonas Geiping. Answer matching outperforms multiple choice for language model evaluation, 2025. URL https://arxiv.org/abs/2507.02856.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 19–35, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73004-7.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 27056–27087. Curran Associates, Inc., 2024. doi: 10.52202/079017-0850. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/2f8ee6a3d766b426d2618e555b5aeb39-Paper-Conference.pdf.

Mark Endo, Xiaohan Wang, and Serena Yeung-Levy. Feather the throttle: Revisiting visual token pruning for vision-language model acceleration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22826–22835, October 2025.

Stephanie Fu, tyler bonnen, Devin Guillory, and Trevor Darrell. Hidden in plain sight: VLMs overlook their visual representations. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=qQb1JLrwol.

Xingyu Fu, Sheng Zhang, Gukyeong Kwon, Pramuditha Perera, Henghui Zhu, Yuhao Zhang, Alexander Hanbo Li, William Yang Wang, Zhiguo Wang, Vittorio Castelli, Patrick Ng, Dan Roth, and Bing Xiang. Generate then select: Open-ended visual question answering guided by world knowledge. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 2333–2346. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.147. URL https://aclanthology.org/2023.findings-acl.147/.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Jiaqi Wang, Weiming Zhang, and Nenghai Yu. Deciphering cross-modal alignment in large vision-language models via modality integration rate. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 218–227, October 2025.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25004–25014, June 2025.

Benno Krojer, Shravan Nayak, Oscar Mañas, Vaibhav Adlakha, Desmond Elliott, Siva Reddy, and Marius Mosbach. Latentlens: Revealing highly interpretable visual tokens in LLMs. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=qJFatzEGQc.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-onevision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=zKv8qULV6n.

Zhihang Lin, Mingbao Lin, Luxi Lin, and Rongrong Ji. Boosting multimodal large language models with visual tokens withdrawal for rapid inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5334–5342, 2025.

Benlin Liu, Amita Kamath, Madeleine Grunde-McLaughlin, Winson Han, and Ranjay Krishna. Visual representations inside the language model. In *Second Conference on Language Modeling*, 2025a. URL https://openreview.net/forum?id=99e72TkWTi.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916. Curran Associates, Inc., 2023. doi: 10.52202/075280-1516. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 216–233, Cham, 2025b. Springer Nature Switzerland. ISBN 978-3-031-72658-3.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 57172–57189, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/900fb3439e4968df79a6f2bfedec49cd-Paper-Conference.pdf.

Farhad Nooralahzadeh, Omid Rohanian, Yi Zhang, Jonathan Fürst, and Kurt Stockinger. Arbitration failure, not perceptual blindness: How vision-language models resolve visual-linguistic conflicts, 2026. URL https://arxiv.org/abs/2604.09364.

Fabio Rosenthal, Sebastian Schmidt, Thorsten Graf, Thorsten Bagdonat, Stephan Günnemann, and Leo Schwinn. Unexplored flaws in multiple-choice vqa make benchmarking unreliable, 2026. URL https://arxiv.org/abs/2511.22341.

Cheng Shi, Yizhou Yu, and Sibei Yang. Vision function layer in multimodal llms. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 27510–27537. Curran Associates, Inc., 2025. doi: 10.52202/085713-0926. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/27c4e15d9af120d7fef04432c7db577f-Paper-Conference.pdf.

Jintao Tong, Wenwei Jin, Pengda Qin, Anqi Li, Yixiong Zou, Yuhong Li, Yuhua Li, and Ruixuan Li. Flowcut: Rethinking redundancy via information flow for efficient vision-language models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 94946–94973. Curran Associates, Inc., 2025. doi: 10.52202/085713-3174. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/88441fee0c0dacceb95e4ad77ece5d0a-Paper-Conference.pdf.

Yahong Wang, Juncheng Wu, Zhangkai Ni, Longzhen Yang, Yihang Liu, Chengmei Yang, Ying Wen, Lianghua He, Xianfeng Tang, Hui Liu, and Yuyin Zhou. When token pruning is worse than random: Understanding visual token information in vllms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 31910–31919, June 2026.

Zichen Wen, Yifeng Gao, Weijia Li, Conghui He, and Linfeng Zhang. Token pruning in multimodal large language models: Are we solving the right problem? In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 15537–15549, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.802. URL https://aclanthology.org/2025.findings-acl.802/.

Hao Wu, Yingqi Fan, Dai Jinyang, Junlong Tong, Yunpu Ma, and Xiaoyu Shen. Hidrop: Hierarchical vision token reduction in mllms via late injection, concave pyramid pruning, and early exit. In *International Conference on Learning Representations*, volume 2026, pp. 97296–97316, 2026.

Zhuoran Yu and Yong Jae Lee. How multimodal LLMs solve image tasks: A lens on visual grounding, task reasoning, and answer decoding. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=kjNJYWvfPA.

Qianhao Yuan, Qingyu Zhang, Yanjiang Liu, Jiawei Chen, Yaojie Lu, Hongyu Lin, Jia Zheng, Xianpei Han, and Le Sun. Shortv: Efficient multimodal large language models by freezing visual tokens in ineffective layers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 329–339, October 2025.

Qizhe Zhang, Aosong Cheng, Ming Lu, Renrui Zhang, Zhiyong Zhuo, Jiajun Cao, Shaobo Guo, Qi She, and Shanghang Zhang. Beyond text-visual attention: Exploiting visual cues for effective token pruning in vlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20857–20867, October 2025a.

Yuhui Zhang, Yuchang Su, Yiming Liu, Xiaohan Wang, James Burgess, Elaine Sui, Chenyu Wang, Josiah Aklilu, Alejandro Lozano, Anjiang Wei, Ludwig Schmidt, and Serena Yeung-Levy. Automated generation of challenging multiple-choice questions for vision language model evaluation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29580–29590, 2025b. doi: 10.1109/CVPR52734.2025.02754.

APPENDIX OVERVIEW

The appendices follow the main argument. First, Appendix A establishes the measurement protocol and its sensitivity; Appendix B details the controlled comparisons with and without the GAS. Appendix C then examines the contribution of intermediate representations through Transfer, Freeze, and complementary readouts. Appendix D extends the analysis to candidate count and answer diversity, while Appendix E evaluates error recovery and a pruning diagnostic. Finally, Appendix F summarizes reproducibility and interpretation.

A EXPERIMENTAL SETTINGS AND MEASUREMENT

A.1 MODELS, EXAMPLES, AND NOTATION

We evaluate eleven decoder-only multimodal models from five families: LLaVA-OV-Qwen2, LLaVA-OV-1.5, Qwen2-VL, Qwen2.5-VL, and InternVL3. Table 2 lists the model names and decoder depths used throughout the paper. A model-task pair denotes one model evaluated on one task. Within a comparison of answer requirements, the same examples are used across answering settings; table captions specify any differences in evaluated subsets. Across experiments, however, samples from the same benchmark can differ.

Table 2: **Models and decoder depths.** L is the number of decoder layers.

Model	Family	L
LLaVA-OV-Qwen2-0.5B	LLaVA-OV-Qwen2	24
LLaVA-OV-Qwen2-7B	LLaVA-OV-Qwen2	28
LLaVA-OV-1.5-4B	LLaVA-OV-1.5	36
LLaVA-OV-1.5-8B	LLaVA-OV-1.5	36
Qwen2-VL-2B	Qwen2-VL	28
Qwen2-VL-7B	Qwen2-VL	28
Qwen2.5-VL-3B	Qwen2.5-VL	36
Qwen2.5-VL-7B	Qwen2.5-VL	28
InternVL3-1B	InternVL3	24
InternVL3-8B	InternVL3	28
InternVL3-14B	InternVL3	48

The original benchmark evaluation contains 88 model-task pairs across eight tasks. Sample sizes vary from 124 to 300: 52 pairs use 150 examples, six use 200, and eight use 300. The eleven BLINK-spatial pairs each use 143 examples, and the eleven BLINK-depth pairs each use 124 examples. By comparison, the TextVQA and GQA comparisons of answer requirements use 300 examples per condition, while the larger error recovery evaluation uses 1,000 per pair. Each table identifies any further restriction, such as selecting examples answered correctly before intervention.

In open-ended answering, the model generates an answer without supplied candidates. By contrast, selection provides candidate answers and asks for either an option label or the answer text. Benchmarks whose original protocol supplies a fixed candidate set (POPE, BLINK-spatial, BLINK-depth, MMBench, ScienceQA) are called *fixed-choice* in the remainder of this appendix. Both settings nevertheless use the same autoregressive model. Accuracies and recovery rates are proportions unless a percentage sign is shown; accuracy differences are reported in percentage points (pp) where indicated.

Answer generation and scoring. Option-label selection chooses the option label with the highest logit among the available options. For comparisons with a supplied candidate set, correctness is determined by the target candidate's position. For open-ended answers, we instead use greedy decoding, with a limit of 16 new tokens in the benchmark evaluations and comparisons of answer requirements. We report binary answer accuracy: after lowercasing and removing surrounding whitespace, a response is correct if it contains a reference answer as a substring. The original TextVQA and DocVQA evaluations accept any provided reference answer; GQA and the controlled

GAS experiments use the target answer for that example. The same rule is applied before and after intervention.

POPE responses are scored by their initial yes/no answer. Similarly, verification extracts the first yes/no response and compares it with the proposal's correctness. The factorial experiment additionally requires answer text to match a complete candidate; its scoring sensitivity is described in Appendix D.1.

A.2 Dependency depth and related measurements

Decoder-layer inputs are indexed from 0 to $L-1$, where L is the number of decoder layers. Removal sets the hidden states at visual token positions to zero at the input of every layer from ℓ onward (Eq. (3)). The visual positions remain in the sequence. Importantly, information already transferred to text positions can remain available after removal. The intervention therefore measures dependence on the visual token states at the affected layers.

The visual token dependency depth ℓ_{dep} is the first evaluated layer whose accuracy drop is within the tolerance in Eq. (4). We use $\varepsilon = \max(0.02, 1.5/n)$, with n equal to the number of examples included in that measurement. The second term allows for discrete accuracy changes on small subsets. A bar denotes division by L; ℓ_{sel} and ℓ_{open} specify the answering setting. Comparisons of normalized depths use the depth of the corresponding model.

The evaluated positions $\mathcal{K}$ include the boundary L. At $\ell = L$, the set of layers to zero is empty, so $A^{\text{remove}}(L) = A^{\text{clean}}$. Thus, a complete scan has depth L when no earlier position meets the tolerance. The same boundary convention applies to accuracy-retention thresholds.

To distinguish a first crossing from stable performance at later layers, we use two additional measurements. A *sustained-crossing depth* requires the accuracy drop to remain within tolerance at every later evaluated layer. A *per-example stable depth* requires an individual answer to remain correct at all later evaluated layers; its definition and paired analysis are given in Appendix D.1. The main benchmark comparison uses the *first-crossing depth*, defined as the earliest evaluated layer whose accuracy drop is within the tolerance.

To assess the preservation of correct answers separately, some controls track retention on initially correct examples. If $\mathcal{C}$ is the set of examples answered correctly without intervention, their retained accuracy at layer k is

$$A_{\text{ret}}(k) = \frac{1}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} \mathbf{1}[\hat{y}_i(k) \text{ is correct}]. \tag{6}$$

Its threshold is $1 - \max(0.02, 1.5/|\mathcal{C}|)$. We identify this subset explicitly in the verification and Freeze layer scan results. In contrast, Transfer and Freeze gains at ℓ_{sel} use all 300 examples and their own experiment's clean baseline.

A.3 Dependency depths across the original benchmarks

To assess the breadth of the benchmark pattern, Figure 8 presents all 88 model-task pairs, including the two excluded for clean accuracy below 0.20. The matrix shows how the separation between fixed options and open-ended answering varies across models. The entropy analysis uses an additional accuracy filter specified in Appendix D.2.

A.4 Text token removal

Removing text token states provides a reference for the visual token intervention. On LLaVA-OV-1.5-8B, we zero the prompt token states after the image at every layer from the selected starting layer onward. The span starts immediately after the visual tokens and ends at the last prompt token. System and chat header tokens before the image are excluded. Each of the four settings uses 300 examples.

Starting at layer 0 sharply reduces accuracy in all four settings (Table 3). MMBench accuracy is close to the uniform four-choice baseline, while POPE accuracy falls below the uniform yes/no baseline. Open-ended GQA and TextVQA have no fixed uniform chance accuracy; their remaining accuracies are low. The text token dependency depth is L in all four settings.

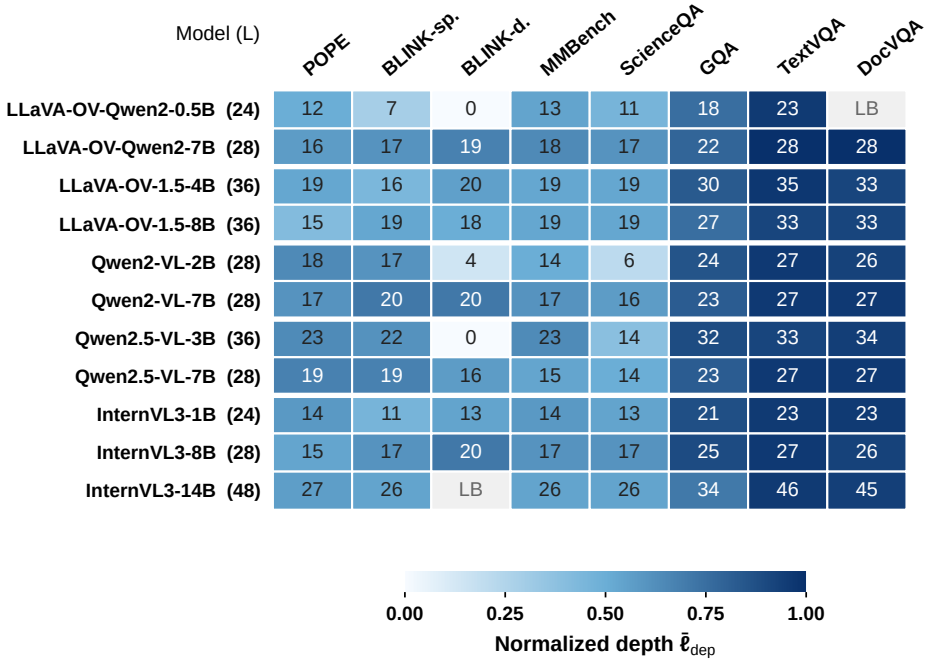

Figure 8: **Visual dependency depth across models and tasks.** Cells give ℓ_{dep}; colors show $\bar{\ell}_{\text{dep}}$. "LB" marks clean accuracy below 0.20. The nominal tolerance is 0.02.

Table 3: **Text token removal on LLaVA-OV-1.5-8B.** Each task uses 300 examples. The uniform reference applies to fixed-choice tasks only; — indicates that no fixed uniform chance level is defined. No intervention at L reproduces the clean baseline.

Task	clean	zero from 0	uniform chance	no intervention (L)
MMBench	0.890	0.243	0.25	0.890
POPE	0.887	0.000	0.50	0.887
GQA	0.593	0.003	—	0.593
TextVQA	0.810	0.027	—	0.810

A.5 SENSITIVITY TO THE ACCURACY THRESHOLD

Because dependency depth is determined by an accuracy threshold, we check how it changes under nearby tolerances. The nominal values range from 0.005 to 0.03, with the same finite-sample adjustment used above. Table 4 reports changes in depth relative to the 0.02 setting and the number of depths equal to L.

Table 4: **Sensitivity to the nominal accuracy tolerance.** Depth changes are relative to $\varepsilon_0 = 0.02$ on 86 model-task pairs. The nominal tolerance is adjusted to $\max(\varepsilon_0, 1.5/n)$ for each sample size n.

ε_0	Pairs	$\ell_{\text{dep}} = L$	Median $\lvert\Delta\ell_{\text{dep}}\rvert$
0.005	86	7	1 layer
0.01	86	7	1 layer
0.015	86	2	0 layers
0.02	86	2	n/a
0.025	86	2	0 layers
0.03	86	2	0 layers

The median absolute change from the 0.02 setting is at most one layer. However, stricter tolerances increase the number of depths equal to L from two to seven. Thus, the threshold affects individual depth estimates, while the overall ordering of answering settings remains similar.

A.6 SENSITIVITY TO THE REPLACEMENT METHOD

Zeroing can change the distribution of hidden states. To assess whether the benchmark pattern depends on this replacement, we also replace visual states with a mean state from clean examples or with a mean state from a different image at the same layer. All methods intervene at the same visual positions and layers.

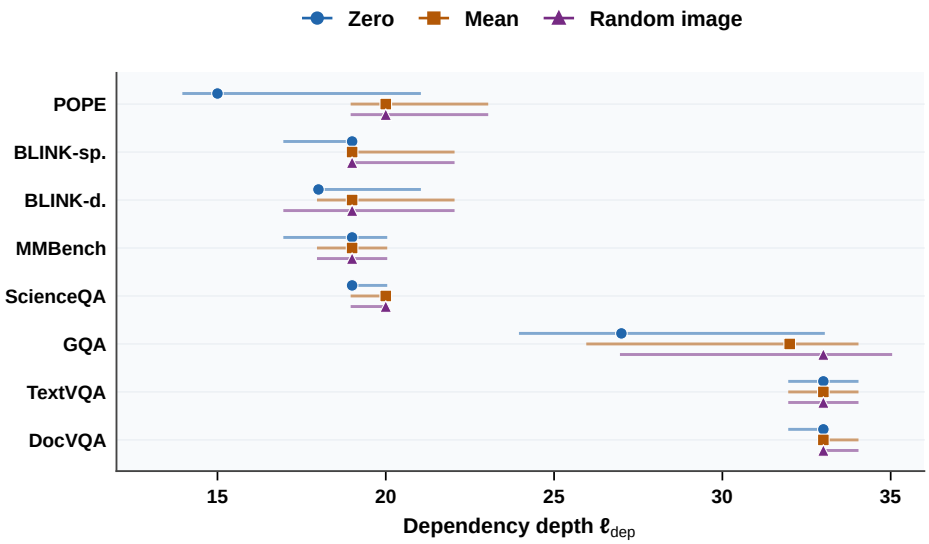

Figure 9: **Dependency depth under different replacement methods.** Estimates of ℓ_{dep} for LLaVA-OV-1.5-8B with paired 95% bootstrap intervals (1,000 resamples). Sample sizes: 143 for BLINK-spatial, 124 for BLINK-depth, and 300 for the other tasks.

The five tasks with fixed answer options have depths between layers 15 and 20 across these methods. By contrast, GQA, TextVQA, and DocVQA have later depths (Figure 9; Table 5). GQA is the most

Table 5: **Replacement method results on LLaVA-OV-1.5-8B.** Entries give layer input depths and paired 95% bootstrap intervals ($B = 1{,}000$). Sample sizes are 143 for BLINK-spatial, 124 for BLINK-depth, and 300 for the other tasks. Accuracies are proportions.

Task	baseline acc.	zero ℓ_{dep}	mean ℓ_{dep}	random-image ℓ_{dep}
POPE	0.887	L15 $[14, 21]$	L20 $[19, 23]$	L20 $[19, 23]$
BLINK-spatial	0.832	L19 $[17, 19]$	L19 $[19, 22]$	L19 $[19, 22]$
BLINK-depth	0.710	L18 $[18, 21]$	L19 $[18, 22]$	L19 $[17, 22]$
MMBench	0.890	L19 $[17, 20]$	L19 $[18, 20]$	L19 $[18, 20]$
ScienceQA	0.987	L19 $[19, 20]$	L20 $[19, 20]$	L20 $[19, 20]$
GQA	0.593	L27 $[24, 33]$	L32 $[26, 34]$	L33 $[27, 35]$
TextVQA	0.810	L33 $[32, 34]$	L33 $[32, 34]$	L33 $[32, 34]$
DocVQA	0.937	L33 $[32, 33]$	L33 $[33, 34]$	L33 $[33, 34]$

sensitive to the replacement: its depth is 27 under Zero, 32 under Mean, and 33 under Random image. The exact depth therefore depends on the intervention, but the separation between the two groups persists.

B GUIDED ANSWER SPACE ON THE SAME EXAMPLES

B.1 CANDIDATE CONSTRUCTION AND PAIRED COMPARISONS

The controlled TextVQA and GQA comparisons keep the image, question, and target answer fixed. Each condition uses the same 300 examples per task, sampled with seed 20260525 and paired by source ID. Two-choice selection presents the correct answer and one distractor. For TextVQA, the distractor is sampled from normalized answers to other images; for GQA, it is sampled from answers to other questions. The correct candidate is assigned to A or B at random, with equal probability. The model returns the option label, which is scored against that assignment.

The two-choice prompt is:

```
Question: {question}
A) {answer A}
B) {answer B}
Answer with a single letter from A to B.
```

Table 6: **The GAS and normalized dependency depth.** TextVQA and GQA use the same 300 examples in both conditions. MMBench compares full set selection with open-ended answering on initially correct examples; Table 10 holds that subset fixed. Δ is after minus before.

	TextVQA open-ended → two-choice			GQA open-ended → two-choice			MMBench MC → open-ended		
Model	before	after	Δ	before	after	Δ	before	after	Δ
LLaVA-OV-Qwen2-0.5B	0.96	0.54	-0.42	0.88	0.54	-0.33	0.54	0.96	0.42
LLaVA-OV-Qwen2-7B	1.00	0.54	-0.46	0.82	0.54	-0.29	0.57	1.00	0.43
LLaVA-OV-1.5-4B	0.94	0.44	-0.50	0.75	0.28	-0.47	0.53	0.86	0.33
LLaVA-OV-1.5-8B	0.94	0.39	-0.56	0.69	0.39	-0.31	0.53	0.75	0.22
Qwen2-VL-2B	0.82	0.36	-0.46	0.82	0.43	-0.39	0.50	1.00	0.50
Qwen2-VL-7B	0.96	0.54	-0.43	0.89	0.50	-0.39	0.57	1.00	0.43
Qwen2.5-VL-3B	0.94	0.50	-0.44	0.92	0.50	-0.42	0.61	0.94	0.33
Qwen2.5-VL-7B	0.96	0.54	-0.43	0.96	0.57	-0.39	0.54	1.00	0.46
InternVL3-1B	0.92	0.50	-0.42	0.96	0.29	-0.67	0.54	1.00	0.46
InternVL3-8B	0.96	0.61	-0.36	0.79	0.61	-0.18	0.61	0.96	0.36
InternVL3-14B	0.92	0.46	-0.46	0.77	0.54	-0.23	0.54	1.00	0.46
Median	**0.94**	**0.50**	**-0.44**	**0.82**	**0.50**	**-0.39**	**0.54**	**1.00**	**0.43**

On the same TextVQA and GQA examples in both answering settings, open-ended answering has the later depth in all 22 model-task combinations (Table 6). The median normalized difference is 0.42, supporting Conjecture 1 in the comparisons of Section 3.2. We compute each model's difference before taking the median; this summary can therefore differ from the difference between the two condition medians.

Table 7: **GAS effects and task differences.** Values are absolute differences in normalized depth. GAS effects compare open-ended answering and two-choice selection within a task; task effects compare TextVQA and GQA within a setting. Each median uses all eleven models.

	GAS effect (same task)		Task effect (same setting)	
Model	TextVQA	GQA	under two-choice selection	under original
LLaVA-OV-Qwen2-0.5B	0.42	0.33	0.00	0.08
LLaVA-OV-Qwen2-7B	0.46	0.29	0.00	0.18
LLaVA-OV-1.5-4B	0.50	0.47	0.17	0.19
LLaVA-OV-1.5-8B	0.56	0.31	0.00	0.25
Qwen2-VL-2B	0.46	0.39	0.07	0.00
Qwen2-VL-7B	0.43	0.39	0.04	0.07
Qwen2.5-VL-3B	0.44	0.42	0.00	0.03
Qwen2.5-VL-7B	0.43	0.39	0.04	0.00
InternVL3-1B	0.42	0.67	0.21	0.04
InternVL3-8B	0.36	0.18	0.00	0.18
InternVL3-14B	0.46	0.23	0.08	0.15
Median	**0.44**	**0.39**	**0.04**	**0.08**

The magnitude of this effect is also informative. In Table 7, the median TextVQA–GQA difference is 0.08 under original prompts and 0.04 under two-choice selection. By comparison, supplying the GAS within a task produces larger median shifts: 0.44 for TextVQA and 0.39 for GQA.

The exact task ordering, however, varies with the evaluation sample. In the comparison with and without the GAS, TextVQA has the later open-ended depth ℓ_{open} in eight of the eleven task comparisons, an equal depth in two, and an earlier depth in one. The original benchmark samples in Appendix A.3 give a later TextVQA depth in all eleven comparisons.

B.2 Image dependence and harder distractors

To quantify the image's contribution to two-choice selection, we replace each image with Gaussian noise while keeping the question, candidates, option order, and scoring unchanged. The noise image has the same dimensions as the original. Pixel values are drawn from a Gaussian distribution with mean 127.5 and standard deviation 52.0, clipped to $[0, 255]$, and converted to unsigned 8-bit RGB values. A fixed seed is used for each example. Input and feature norms are not rescaled. The same noise construction is used for Transfer.

Importantly, original-image accuracy exceeds noise-image accuracy in all 22 pairs (Table 8). The mean difference is 17.7 pp, with a 95% bootstrap interval of $[14.2, 21.2]$ obtained by resampling model-task pairs. The task means are 24.8 points for TextVQA ($[21.6, 28.0]$) and 10.5 for GQA ($[9.0, 12.2]$). These accuracy differences establish that visual input contributes substantially to two-choice selection on both tasks.

We further test whether the depth ordering persists with harder alternatives. The AutoConverter condition uses a distractor generated by Qwen3.6-27B, conditioned on the image content, question, and correct answer. Generation uses vLLM with a maximum of 1,024 tokens, temperature 0.7, top-p 0.8, and top-k 20. The examples, two-choice prompt, and option order randomization match the in-dataset distractor condition. On the three models in Figure 10, these distractors reduce clean selection accuracy; nevertheless, selection still has an earlier dependency depth than open-ended answering.

Table 8: **Image contribution to two-choice selection.** Each pair uses the same 300 examples with the original image and Gaussian noise. Δ is clean minus noise accuracy. n_{img} counts answers correct with the original image and incorrect with noise. Accuracies and differences are proportions.

TextVQA				
Model	clean	noise	Δ	n_{img}
LLaVA-OV-Qwen2-0.5B	0.820	0.523	+0.297	100
LLaVA-OV-Qwen2-7B	0.973	0.687	+0.287	90
LLaVA-OV-1.5-4B	0.983	0.763	+0.220	68
LLaVA-OV-1.5-8B	0.977	0.720	+0.257	80
Qwen2-VL-2B	0.973	0.633	+0.340	106
Qwen2-VL-7B	0.983	0.723	+0.260	81
Qwen2.5-VL-3B	0.967	0.797	+0.170	53
Qwen2.5-VL-7B	0.967	0.810	+0.157	50
InternVL3-1B	0.950	0.673	+0.277	88
InternVL3-8B	0.987	0.710	+0.277	84
InternVL3-14B	0.967	0.777	+0.190	62
GQA				
Model	clean	noise	Δ	n_{img}
LLaVA-OV-Qwen2-0.5B	0.923	0.777	+0.147	51
LLaVA-OV-Qwen2-7B	0.960	0.863	+0.097	34
LLaVA-OV-1.5-4B	0.970	0.887	+0.083	29
LLaVA-OV-1.5-8B	0.960	0.853	+0.107	36
Qwen2-VL-2B	0.957	0.797	+0.160	52
Qwen2-VL-7B	0.980	0.873	+0.107	35
Qwen2.5-VL-3B	0.970	0.890	+0.080	30
Qwen2.5-VL-7B	0.980	0.923	+0.057	18
InternVL3-1B	0.957	0.857	+0.100	36
InternVL3-8B	0.973	0.867	+0.107	38
InternVL3-14B	0.977	0.860	+0.117	37

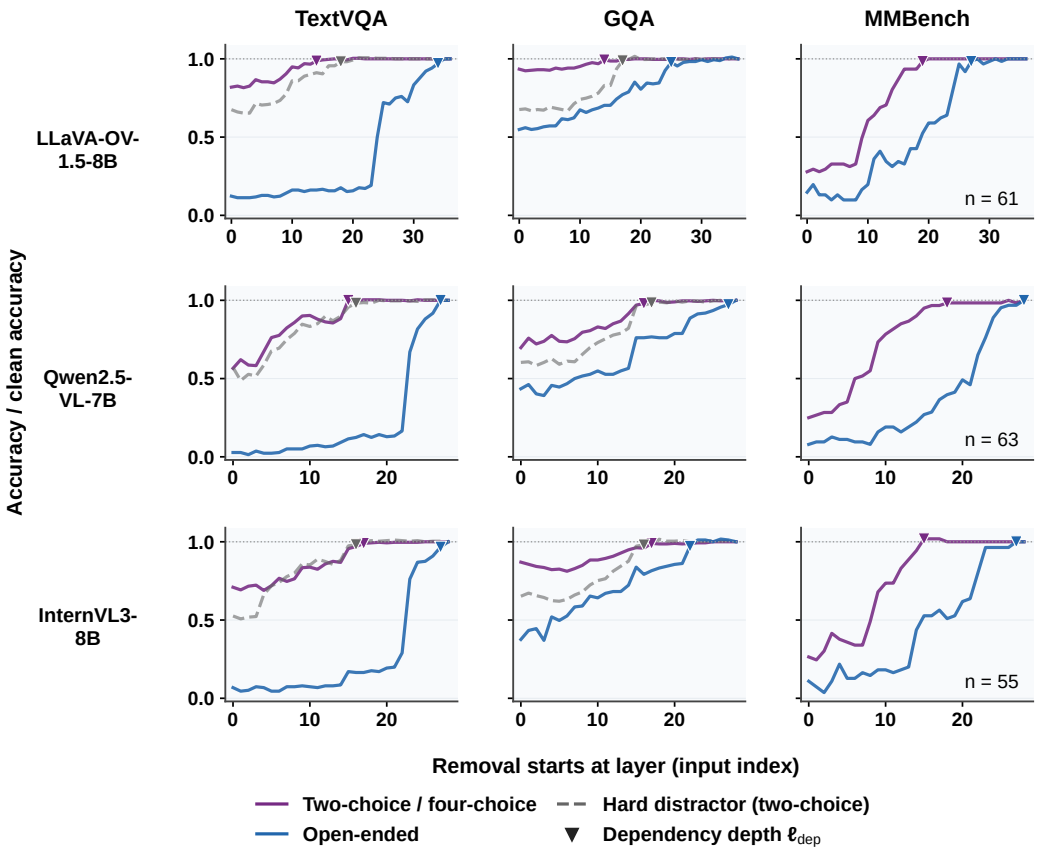

Figure 10: **Visual token removal under different answer requirements.** Accuracy is relative to each condition's clean baseline; markers indicate ℓ_{dep}. TextVQA/GQA use the same 300 examples across conditions. MMBench uses the same initially correct open-ended examples.

B.3 SELECTION WITH ANSWER-TEXT OUTPUT

Providing candidates and requiring an option label are distinct changes to the answering task. To examine selection with a textual response, the answer-text condition provides the correct answer and a distractor but asks for the selected answer in words:

```
Question: {question}. The correct answer is one of:
{correct answer}, {distractor}. Reply with the answer
in full words.
```

This experiment uses 300 examples per task and model, with the same substring matching rule as its open-ended evaluation. Its answer-text and label comparisons are reported in Appendix E.3. These samples are evaluated separately from the paired two-choice experiment above, so depths from the two experiments are not treated as paired effects of output format. To isolate that effect, the factorial experiment in Appendix D.1 provides a controlled comparison.

Table 9: **Answer-text selection with candidates in the prompt.** Accuracies are proportions. A^B is the matching no-image accuracy; n_{img} counts examples correct with the image and incorrect without it. Depths are layer indices on all 300 examples or this subset. † marks a clean/no-image gap below 0.05.

TextVQA Model	clean	A^B	n_{img}	ℓ_{dep}(full)	ℓ_{dep}(image-dep)
LLaVA-OV-Qwen2-0.5B	0.40	0.23	76	22	24
LLaVA-OV-Qwen2-7B	0.82	0.45	129	19	27
LLaVA-OV-1.5-4B	0.84	0.49	120	13	21
LLaVA-OV-1.5-8B	0.90	0.73	68	10	21
Qwen2-VL-2B	0.87	0.54	114	18	26
Qwen2-VL-7B	0.84	0.59	99	15	27
Qwen2.5-VL-3B	0.83	0.72	56	33	33
Qwen2.5-VL-7B	0.72	0.55	61	15	28
InternVL3-1B	0.84	0.65	74	4	15
InternVL3-8B	0.82	0.58	88	15	27
InternVL3-14B	0.95	0.81	51	21	25
GQA Model	clean	A^B	n_{img}	ℓ_{dep}(full)	ℓ_{dep}(image-dep)
LLaVA-OV-Qwen2-0.5B†	0.57	0.53	48	5	24
LLaVA-OV-Qwen2-7B	0.91	0.77	58	10	26
LLaVA-OV-1.5-4B	0.93	0.78	51	9	22
LLaVA-OV-1.5-8B	0.94	0.86	33	0	17
Qwen2-VL-2B	0.90	0.72	61	15	16
Qwen2-VL-7B	0.90	0.80	47	15	27
Qwen2.5-VL-3B†	0.92	0.90	17	22	22
Qwen2.5-VL-7B†	0.88	0.87	28	15	21
InternVL3-1B	0.88	0.79	38	14	14
InternVL3-8B	0.95	0.78	55	2	17
InternVL3-14B	0.96	0.89	29	19	25

Table 9 also restricts evaluation to examples answered correctly with the image but incorrectly in the corresponding no-image condition. Depth can increase on this subset because answers that can be produced without the image no longer contribute to the measurement. Because subset sizes vary from 17 to 129 examples, the finite-sample threshold adjustment is relevant. The no-image accuracies in this table belong to this answer-text experiment and are separate from the benchmark controls in Appendix D.2.

B.4 REMOVING THE OPTIONS FROM MMBENCH

MMBench tests the reverse change in the answer space: we remove the four options and ask for the same target answer in words. To compare dependence on visual states for questions the model can answer in this setting, we identify the examples answered correctly under the clean open-ended prompt. For LLaVA-OV-1.5-8B, this subset contains 61 of 300 examples (accuracy 0.203) under the substring matching rule.

The MMBench columns from the eleven models in Table 6 compare selection on the full set with open-ended answering restricted to initially correct answers. To isolate the comparison with and without the GAS from this difference in evaluation subsets, Table 10 evaluates both settings on the same examples answered correctly under the clean open-ended prompt for each of three models.

Table 10: **MMBench comparison on identical subsets.** Both conditions use the examples answered correctly under the clean open-ended prompt.

Model	n_{correct}	L	original MC	open-ended	shift	shift$/L$
LLaVA-OV-1.5-8B	61	36	L19	L27	+8	0.22
Qwen2.5-VL-7B	63	28	L18	L28	+10	0.36
InternVL3-8B	55	28	L15	L27	+12	0.43

Importantly, the later open-ended dependence persists on the shared subsets. Selection depths are 19, 18, and 15, compared with 19, 15, and 17 on the full selection sets. Open-ended depths are 27 for LLaVA-OV-1.5-8B and InternVL3-8B, and 28 for Qwen2.5-VL-7B.

B.5 EXTENSION TO DOCVQA

We repeat the open-ended versus two-choice comparison on DocVQA to examine whether the pattern extends to document questions. Each condition uses the same 300 examples, with a distractor sampled from other DocVQA answers. The image, question, and target answer remain unchanged.

Table 11: **The GAS on DocVQA.** Depths are divided by L. Each DocVQA comparison uses the same 300 examples in both conditions. TextVQA selection depths are shown as a reference from the separate comparison with and without the GAS.

Model	DocVQA two-choice	DocVQA open-ended	TextVQA two-choice
LLaVA-OV-Qwen2-0.5B	1.00	1.00	0.54
LLaVA-OV-Qwen2-7B	0.61	1.00	0.54
LLaVA-OV-1.5-4B	0.44	0.94	0.44
LLaVA-OV-1.5-8B	0.44	0.92	0.39
Qwen2-VL-2B	0.57	0.93	0.36
Qwen2-VL-7B	0.54	1.00	0.54
Qwen2.5-VL-3B	0.58	0.94	0.50
Qwen2.5-VL-7B	0.61	1.00	0.54
InternVL3-1B	0.54	1.00	0.50
InternVL3-8B	0.61	1.00	0.61
InternVL3-14B	0.65	1.00	0.46

Selection has an earlier dependency depth in ten of the eleven models; LLaVA-OV-Qwen2-0.5B has depth L in both settings. In the other ten models, selection depths range from $0.44L$ to $0.65L$, compared with open-ended depths between $0.92L$ and L. These results therefore extend the effect of the GAS to document questions.

C EARLIER VISUAL REPRESENTATIONS SUPPORT LATER ANSWERS

Conjecture 2 concerns how much visual support for open-ended answering is already available in intermediate representations. We test this contribution at the selection dependency depth ℓ_{sel} using representations from the original open-ended prompt. Transfer measures their contribution, and Freeze tests the benefit of retaining them without further updates. Verification, answer likelihoods, and linear probes provide complementary readouts of information relevant to answering.

C.1 TRANSFER UNDER THE ORIGINAL OPEN-ENDED PROMPT

To measure the contribution of earlier visual states to open-ended answering, we compare Clean, Noise, and Transfer. For each example, Clean uses the original image, whereas Noise uses a Gaussian noise image of the same dimensions. Both use the original open-ended prompt. At the input to ℓ_{sel}, Transfer replaces the Noise run's visual states once with those recorded from the corresponding Clean run. The remaining computation and answer generation then continue normally. Clean and Noise visual spans must have matching positions and shapes. As implementation checks, transfers from Clean to Clean and from Noise to Noise reproduce the original outputs.

Table 12: **Transfer improves open-ended answers with noisy images.** Clean visual states are transferred once at ℓ_{sel}. All conditions use the same 300 examples per pair. $\Delta = A^{\text{transfer}} - A^{\text{noise}}$; recovery is $\Delta/(A^{\text{clean}} - A^{\text{noise}})$. Values are proportions.

TextVQA Model	ℓ_{sel}	clean	noise	transfer	Δ	recovery
LLaVA-OV-Qwen2-0.5B	13	0.373	0.020	0.220	+0.200	0.57
LLaVA-OV-Qwen2-7B	15	0.560	0.023	0.443	+0.420	0.78
LLaVA-OV-1.5-4B	16	0.660	0.043	0.587	+0.543	0.88
LLaVA-OV-1.5-8B	14	0.680	0.060	0.637	+0.577	0.93
Qwen2-VL-2B	10	0.660	0.023	0.573	+0.550	0.86
Qwen2-VL-7B	15	0.700	0.047	0.520	+0.473	0.72
Qwen2.5-VL-3B	18	0.683	0.057	0.563	+0.507	0.81
Qwen2.5-VL-7B	15	0.727	0.070	0.587	+0.517	0.79
InternVL3-1B	12	0.527	0.050	0.347	+0.297	0.62
InternVL3-8B	17	0.587	0.033	0.410	+0.377	0.68
InternVL3-14B	22	0.613	0.050	0.557	+0.507	0.90
GQA Model	ℓ_{sel}	clean	noise	transfer	Δ	recovery
LLaVA-OV-Qwen2-0.5B	13	0.490	0.293	0.407	+0.113	0.58
LLaVA-OV-Qwen2-7B	15	0.590	0.303	0.467	+0.163	0.57
LLaVA-OV-1.5-4B	10	0.587	0.303	0.567	+0.263	0.93
LLaVA-OV-1.5-8B	14	0.583	0.303	0.553	+0.250	0.89
Qwen2-VL-2B	12	0.580	0.287	0.457	+0.170	0.58
Qwen2-VL-7B	14	0.600	0.333	0.537	+0.203	0.76
Qwen2.5-VL-3B	18	0.600	0.360	0.563	+0.203	0.85
Qwen2.5-VL-7B	16	0.613	0.387	0.500	+0.113	0.50
InternVL3-1B	7	0.567	0.320	0.487	+0.167	0.68
InternVL3-8B	17	0.577	0.283	0.383	+0.100	0.34
InternVL3-14B	26	0.640	0.310	0.463	+0.153	0.46

Crucially, Transfer improves open-ended accuracy in all 22 pairs (Table 12). The mean gain is 45.2 pp for TextVQA and 17.3 for GQA. Their 95% bootstrap intervals, resampling model-task pairs, are $[38.0, 51.2]$ and $[14.3, 20.3]$. Recovery divides each gain by that pair's Clean–Noise accuracy gap, as defined in Eq. (5); its median across pairs is 0.74.

Pooling and layer-wise results. For each task, pooled recovery divides the sum of Transfer gains by the sum of Clean–Noise gaps across models:

$$R_{\text{pool}} = \frac{\sum_m \left[A_m^{\text{transfer}}(\ell_{\text{sel},m}) - A_m^{\text{noise}}\right]}{\sum_m \left[A_m^{\text{clean}} - A_m^{\text{noise}}\right]}. \tag{7}$$

With 300 examples in every pair, pooled recovery is 79% for TextVQA and 65% for GQA. These estimates differ from the mean of individual recovery ratios because the Clean–Noise gaps have different sizes.

To locate the recovery relative to ℓ_{sel}, we also transfer at every decoder-layer input. Transfer at layer 0 reproduces Clean; the endpoint L applies no transfer and reproduces Noise. Values at ℓ_{sel} match Table 12. At that depth, recovery is at least 50% for all eleven TextVQA pairs and nine GQA pairs. The layer scan 50% recovery boundary lies after ℓ_{sel} in 20 of 22 pairs and before ℓ_{open} in all 22 pairs. The full curves appear in Figure 4a. Recovery generally decreases at later transfer layers.

Text token transfer. To compare visual and text contributions at the same layer, we transfer the Clean states of all prompt tokens after the image at layer 14 on LLaVA-OV-1.5-8B with TextVQA. Across 300 examples, Clean, Noise, text token transfer, and visual token transfer have accuracies of 0.680, 0.060, 0.063, and 0.637. The text transfer gain is 0.003, with a paired example bootstrap interval of $[-0.020, 0.027]$. It recovers 0.5% of the Clean–Noise gap, compared with 93.0% for visual transfer. Thus, in this setting, transferring visual states accounts for the large recovery, while transferring text states at the same layer has little effect.

C.2 Freeze and continued access to visual representations

Having established the contribution of earlier visual states through Transfer, we use Freeze to test the benefit of retaining them without further updates. Starting at layer k, we record its input visual states and supply these same values at each subsequent layer input. Meanwhile, the text states continue to update. For comparison, Zero replaces the visual states with zeros at the same layers. Both conditions use the original image and open-ended prompt.

At the independently measured ℓ_{sel}, Freeze outperforms Zero in all 22 pairs (Table 13). The median gain across pairs is 17.7 pp, and the median recovered fraction of the Clean–Zero gap is 0.71. Task-specific median recovery is 66.4% for TextVQA and 74.6% for GQA. Absolute gains are larger on TextVQA because Zero causes a larger initial loss. Resampling pairs gives a 95% interval of $[0.173, 0.305]$ for the overall mean gain; task intervals are $[0.290, 0.437]$ and $[0.081, 0.142]$. These comparisons use each experiment's own clean baseline and all 300 examples. Thus, continued access to fixed intermediate states supports open-ended answering across the evaluated models and tasks.

To complement the full set gains at ℓ_{sel}, the layer scan control tracks retention on initially correct answers. Freeze reaches this retention threshold before L in all 22 pairs, compared with 12 for Zero. Before the final decoder-layer input $L-1$, Freeze reaches the threshold in 20 pairs, compared with five for Zero. Figure 11 shows the full depth distributions.

C.3 Verification of proposed answers

Verification extends the analysis to judging a proposed answer. The prompt asks whether the proposed answer is correct and requests Yes or No. For the balanced verification experiment, the proposal is randomly chosen to be the correct answer or an in-dataset distractor with equal probability. The prompt is:

```
Question: {question}
Is the answer "{proposed answer}"? Reply Yes or No.
```

The actual counts can differ slightly from an exact 50/50 split. We evaluate 300 examples per pair and measure retention on the subset answered correctly without intervention. Table 14 reports its depth, $\ell_{\text{ver}}^{\text{ret}}$, and places the selection and open-ended depths, ℓ_{sel} and ℓ_{open}, beside it as references. These reference depths use the full set measurement from Appendix B.

Table 13: **Freeze supports open-ended answering at the selection depth.** Freeze and Zero start at ℓ_{sel} and use the same 300 examples per pair. $\Delta = A^{\text{freeze}} - A^{\text{zero}}$; recovery is $\Delta/(A^{\text{clean}} - A^{\text{zero}})$. Values are proportions.

TextVQA Model	ℓ_{sel}	clean	zero	freeze	Δ	recovery
LLaVA-OV-Qwen2-0.5B	13	0.323	0.053	0.227	+0.173	0.64
LLaVA-OV-Qwen2-7B	15	0.520	0.063	0.367	+0.303	0.66
LLaVA-OV-1.5-4B	16	0.607	0.063	0.490	+0.427	0.79
LLaVA-OV-1.5-8B	14	0.610	0.087	0.410	+0.323	0.62
Qwen2-VL-2B	10	0.617	0.030	0.523	+0.493	0.84
Qwen2-VL-7B	15	0.660	0.057	0.590	+0.533	0.88
Qwen2.5-VL-3B	18	0.640	0.047	0.393	+0.347	0.58
Qwen2.5-VL-7B	15	0.660	0.063	0.617	+0.553	0.93
InternVL3-1B	12	0.543	0.060	0.240	+0.180	0.37
InternVL3-8B	17	0.547	0.083	0.467	+0.383	0.83
InternVL3-14B	22	0.560	0.073	0.350	+0.277	0.57
GQA Model	ℓ_{sel}	clean	zero	freeze	Δ	recovery
LLaVA-OV-Qwen2-0.5B	13	0.523	0.390	0.413	+0.023	0.17
LLaVA-OV-Qwen2-7B	15	0.583	0.460	0.537	+0.077	0.62
LLaVA-OV-1.5-4B	10	0.610	0.413	0.490	+0.077	0.39
LLaVA-OV-1.5-8B	14	0.640	0.410	0.600	+0.190	0.83
Qwen2-VL-2B	12	0.580	0.373	0.547	+0.173	0.84
Qwen2-VL-7B	14	0.627	0.420	0.577	+0.157	0.76
Qwen2.5-VL-3B	18	0.613	0.403	0.560	+0.157	0.75
Qwen2.5-VL-7B	16	0.593	0.447	0.583	+0.137	0.93
InternVL3-1B	7	0.570	0.317	0.423	+0.107	0.42
InternVL3-8B	17	0.607	0.493	0.583	+0.090	0.79
InternVL3-14B	26	0.600	0.527	0.570	+0.043	0.59

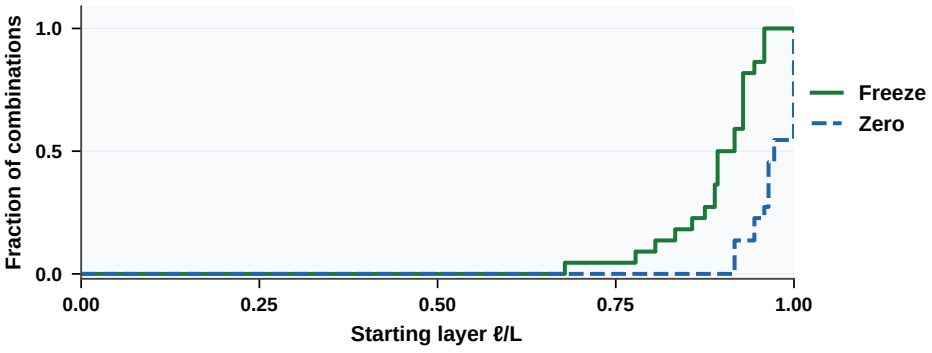

Figure 11: **Freeze retains initially correct answers from earlier layers.** Fraction of the 22 model-task combinations reaching the retention threshold (Eq. 6) by each starting depth ℓ/L. Each condition uses its own initially correct subset.

Table 14: **Balanced yes/no verification.** Each pair uses 300 examples. Verification depth uses retention on its n_{correct} initially correct examples. Selection and open-ended depths are full set references from the separate comparison with and without the GAS. All depths are layer indices; the different definitions preclude interpreting their difference as a paired format effect.

TextVQA Model	Clean	n_{correct}	$\ell_{\text{ver}}^{\text{ret}}$	ℓ_{sel}	ℓ_{open}
LLaVA-OV-Qwen2-0.5B	0.543	163	23	13	23
LLaVA-OV-Qwen2-7B	0.923	277	17	15	28
LLaVA-OV-1.5-4B	0.950	285	18	16	34
LLaVA-OV-1.5-8B	0.943	283	19	14	34
Qwen2-VL-2B	0.923	277	17	10	23
Qwen2-VL-7B	0.930	279	20	15	27
Qwen2.5-VL-3B	0.943	283	22	18	34
Qwen2.5-VL-7B	0.900	270	20	15	27
InternVL3-1B	0.893	268	14	12	22
InternVL3-8B	0.887	266	17	17	27
InternVL3-14B	0.913	274	26	22	44
GQA Model	Clean	n_{correct}	$\ell_{\text{ver}}^{\text{ret}}$	ℓ_{sel}	ℓ_{open}
LLaVA-OV-Qwen2-0.5B	0.610	183	23	13	21
LLaVA-OV-Qwen2-7B	0.823	247	18	15	23
LLaVA-OV-1.5-4B	0.867	260	19	10	27
LLaVA-OV-1.5-8B	0.877	263	19	14	25
Qwen2-VL-2B	0.680	204	21	12	23
Qwen2-VL-7B	0.730	219	19	14	25
Qwen2.5-VL-3B	0.767	230	32	18	33
Qwen2.5-VL-7B	0.800	240	19	16	27
InternVL3-1B	0.753	226	14	7	23
InternVL3-8B	0.803	241	17	17	22
InternVL3-14B	0.877	263	27	26	37

Correct answer and wrong answer controls. To distinguish confirmation from rejection, we separately evaluate prompts containing only correct proposals or only incorrect proposals. Incorrect proposals use the harder AutoConverter distractors. The *correct-rejection rate* is the accuracy on the wrong answer condition; the correct answer condition measures successful confirmation. Table 15 keeps these conditions separate from balanced verification.

Table 15: **Verification with incorrect or correct proposals.** Wrong answer accuracy is the correct-rejection rate; correct answer accuracy is the confirmation rate. These one-class controls are separate from balanced verification. Depths are divided by L. — denotes an unreported condition.

TextVQA	Wrong proposal		Correct proposal	
Model	Accuracy	ℓ_{dep}/L	Accuracy	ℓ_{dep}/L
LLaVA-OV-Qwen2-0.5B	0.878	0.96	0.214	0.54
LLaVA-OV-Qwen2-7B	0.701	1.00	0.915	0.64
LLaVA-OV-1.5-4B	0.827	0.58	0.915	0.50
LLaVA-OV-1.5-8B	0.823	0.64	0.929	0.25
Qwen2-VL-2B	0.779	0.04	0.884	0.61
Qwen2-VL-7B	0.759	0.14	0.932	0.61
Qwen2.5-VL-3B	0.789	0.14	0.935	0.61
Qwen2.5-VL-7B	0.905	0.07	—	—
InternVL3-1B	0.677	0.00	0.813	0.58
InternVL3-8B	0.850	0.00	0.844	0.61
InternVL3-14B	0.847	0.00	0.857	0.58
GQA	Wrong proposal		Correct proposal	
Model	Accuracy	ℓ_{dep}/L	Accuracy	ℓ_{dep}/L
LLaVA-OV-Qwen2-0.5B	0.819	0.29	0.308	0.54
LLaVA-OV-Qwen2-7B	0.689	0.11	0.722	0.71
LLaVA-OV-1.5-4B	0.672	0.61	0.739	0.56
LLaVA-OV-1.5-8B	0.652	0.53	—	—
Qwen2-VL-2B	0.595	0.00	0.582	0.93
Qwen2-VL-7B	0.662	0.00	0.649	0.75
Qwen2.5-VL-3B	0.786	0.08	0.562	0.61
Qwen2.5-VL-7B	—	—	0.589	0.68
InternVL3-1B	0.625	0.00	0.595	0.71
InternVL3-14B	0.649	0.15	0.739	0.54

With hard distractors, correct-rejection accuracy ranges from 0.60 to 0.91. LLaVA-OV-1.5-8B has normalized wrong answer depths of 0.53–0.64 on the two tasks. In contrast, Qwen2.5-VL-7B and InternVL3-8B have near-zero depths on the reported TextVQA controls: their accuracy changes by at most 0.002 under removal. However, a one-class control cannot by itself separate answer verification from a preference for Yes or No. We therefore interpret these depths together with balanced verification and the image interventions.

C.4 Answer likelihood under the original prompt

To complement the experiments with candidates included in the prompt, we test whether candidate likelihoods reveal answer information under the unchanged open-ended prompt: "Answer the question using a single word or phrase." At each removal layer, we record greedy answers without candidates in the prompt and separately score nested candidate sets of sizes $N \in \{2, 4, 8\}$, each containing the correct answer and $N - 1$ distractors. We score candidates as continuations of the same image and question prompt. Mean per-token log-likelihood defines the primary ranking, while summed log-likelihood checks sensitivity to answer length.

At $N = 2$, ranking has an earlier first-crossing depth than open-ended generation in all six model-task pairs for both scoring rules. Moreover, the ordering $\ell_2 \leq \ell_4 \leq \ell_8 \leq \ell_{\text{open}}$ holds in five

Table 16: **Candidate likelihoods under the unchanged open-ended prompt.** Top: first-crossing layer indices for mean per-token and summed log-likelihood. Bottom: ranking accuracy on original open-ended errors using the mean score. Each pair uses 300 examples; candidate sets contain the correct answer.

				Mean log-likelihood			Summed log-likelihood		
Model	Task	L	ℓ_{open}	N=2	4	8	N=2	4	8
LLaVA-OV-1.5-8B	TextVQA	36	34	30	30	32	30	30	32
Qwen2.5-VL-7B	TextVQA	28	27	24	26	26	26	26	27
InternVL3-8B	TextVQA	28	27	24	24	27	24	24	24
LLaVA-OV-1.5-8B	GQA	36	26	2	10	0	10	24	28
Qwen2.5-VL-7B	GQA	28	26	20	24	26	14	24	24
InternVL3-8B	GQA	28	24	12	22	22	16	20	24

Model	Task	Original errors	N=2	N=4	N=8
LLaVA-OV-1.5-8B	TextVQA	99	0.889	0.808	0.758
Qwen2.5-VL-7B	TextVQA	84	0.917	0.857	0.798
InternVL3-8B	TextVQA	101	0.901	0.812	0.762
LLaVA-OV-1.5-8B	GQA	110	0.873	0.755	0.555
Qwen2.5-VL-7B	GQA	118	0.932	0.822	0.720
InternVL3-8B	GQA	112	0.938	0.821	0.679

of six pairs for each rule. Importantly, on the original open-ended errors, two-candidate ranking accuracy is 0.873–0.938 under the primary score. Ranking remains above uniform chance at $N = 4$ and $N = 8$ as well. Because the correct answer is supplied to the ranking procedure, this analysis tests whether the model assigns useful relative likelihoods; the inference procedure without supplied correct answers is evaluated separately in Appendix E.1.

C.5 Linear probes of text representations

The preceding analyses assess the model's own responses and likelihoods. We additionally test whether proposal correctness is linearly accessible in text representations. The inputs are balanced yes/no verification prompts on 300 examples per pair. The feature is the hidden state of the final prompt token, taken from the output of a specified decoder layer. Each probe standardizes features within the training fold and fits an ℓ_2-regularized logistic regression with $C = 1$. Accuracy is averaged over five stratified folds; held-out examples are not used to fit the standardization or classifier.

The early reference is decoder layer 5 for every model. The other two reference layers are selected using selection and later open-ended depth estimates available for the probe experiment. Table 17 gives the actual layer index beside each accuracy. These are layer *outputs*, while removal depths elsewhere refer to layer *inputs*; the later probe reference can also differ from the open-ended depth in the comparison with and without the GAS.

At the selection reference, probe accuracy is 0.720–0.980, with a median of 0.943. It matches or exceeds clean yes/no accuracy in 19 of 22 pairs. Moreover, the median gain over layer 5 is 19 pp, while the median absolute difference from the later reference is below two points. Thus, proposal correctness is often linearly accessible in intermediate text representations. This readout establishes accessibility under the verification prompt, complementing the direct interventions on open-ended answers provided by Transfer and Freeze.

Table 17: **Linear-probe accuracy at three reference layers.** Parentheses give zero-based decoder-layer output indices. Accuracy is the mean over five held-out folds. Each pair has 300 examples; clean yes/no accuracy is the model's response accuracy on that probe input set.

TextVQA Model	Early	Selection ref.	Later ref.	Clean yes/no
LLaVA-OV-Qwen2-0.5B	0.500 (5)	0.827 (13)	0.797 (23)	0.563
LLaVA-OV-Qwen2-7B	0.753 (5)	0.943 (15)	0.953 (27)	0.927
LLaVA-OV-1.5-4B	0.783 (5)	0.970 (16)	0.973 (35)	0.910
LLaVA-OV-1.5-8B	0.777 (5)	0.927 (14)	0.973 (34)	0.930
Qwen2-VL-2B	0.787 (5)	0.943 (10)	0.950 (27)	0.907
Qwen2-VL-7B	0.837 (5)	0.977 (15)	0.947 (27)	0.923
Qwen2.5-VL-3B	0.747 (5)	0.947 (18)	0.967 (33)	0.943
Qwen2.5-VL-7B	0.863 (5)	0.980 (15)	0.977 (27)	0.913
InternVL3-1B	0.673 (5)	0.863 (12)	0.933 (23)	0.897
InternVL3-8B	0.820 (5)	0.970 (17)	0.973 (27)	0.897
InternVL3-14B	0.653 (5)	0.943 (22)	0.967 (46)	0.893
GQA Model	Early	Selection ref.	Later ref.	Clean yes/no
LLaVA-OV-Qwen2-0.5B	0.617 (5)	0.813 (13)	0.843 (18)	0.620
LLaVA-OV-Qwen2-7B	0.763 (5)	0.957 (15)	0.950 (22)	0.847
LLaVA-OV-1.5-4B	0.710 (5)	0.857 (10)	0.967 (30)	0.823
LLaVA-OV-1.5-8B	0.713 (5)	0.943 (14)	0.970 (25)	0.877
Qwen2-VL-2B	0.790 (5)	0.920 (12)	0.913 (24)	0.673
Qwen2-VL-7B	0.743 (5)	0.940 (14)	0.957 (23)	0.667
Qwen2.5-VL-3B	0.697 (5)	0.950 (18)	0.973 (32)	0.763
Qwen2.5-VL-7B	0.783 (5)	0.960 (16)	0.967 (27)	0.787
InternVL3-1B	0.703 (5)	0.720 (7)	0.917 (21)	0.797
InternVL3-8B	0.803 (5)	0.957 (17)	0.973 (22)	0.813
InternVL3-14B	0.667 (5)	0.973 (26)	0.973 (34)	0.873

D CANDIDATE SET SIZE AND ANSWER DIVERSITY

D.1 CANDIDATE SET SIZE AND OUTPUT FORMAT

To test Conjecture 1 within selection, the factorial experiment varies candidate count separately from the required output format. We evaluate $N \in \{2, 4, 8\}$ with either an option-label or answer-text response on LLaVA-OV-1.5-8B, Qwen2.5-VL-7B, and InternVL3-8B. For each model and task, all six conditions use the same 300 images, questions, target answers, checkpoint, and evaluated layers. Candidate sets are nested. At each N, both formats present the same candidates in the same order, and the correct candidate's position is balanced. This gives 36 conditions across the three models and two tasks.

For answer-text output, we ignore case and collapse whitespace, then match the complete response to a candidate. A match to any candidate counts as format compliance; a match to the correct candidate counts as a correct answer. Punctuation is preserved. Option-label output is scored against the correct candidate position.

Table 18: **Candidate count and response format: all 36 conditions.** Each condition uses 300 examples. Top: accuracy and answer-text format compliance, as proportions. Bottom: first-crossing and sustained-crossing layer indices. Sustained crossing requires all later evaluated layers to remain within tolerance.

			Accuracy			Format compliance		
Model	Task	Output	N=2	4	8	N=2	4	8
LLaVA-OV-1.5-8B	TextVQA	Option label	0.983	0.967	0.947	—	—	—
LLaVA-OV-1.5-8B	TextVQA	Answer text	0.923	0.923	0.920	0.943	0.953	0.953
Qwen2.5-VL-7B	TextVQA	Option label	0.977	0.957	0.953	—	—	—
Qwen2.5-VL-7B	TextVQA	Answer text	0.907	0.913	0.917	0.920	0.930	0.940
InternVL3-8B	TextVQA	Option label	0.980	0.963	0.950	—	—	—
InternVL3-8B	TextVQA	Answer text	0.930	0.920	0.927	0.937	0.937	0.950
LLaVA-OV-1.5-8B	GQA	Option label	0.940	0.923	0.877	—	—	—
LLaVA-OV-1.5-8B	GQA	Answer text	0.827	0.907	0.870	0.837	0.957	0.963
Qwen2.5-VL-7B	GQA	Option label	0.957	0.910	0.887	—	—	—
Qwen2.5-VL-7B	GQA	Answer text	0.923	0.883	0.847	0.953	0.957	0.977
InternVL3-8B	GQA	Option label	0.960	0.937	0.867	—	—	—
InternVL3-8B	GQA	Answer text	0.940	0.910	0.843	0.977	0.970	0.973

			First crossing			Sustained crossing		
Model	Task	Output	N=2	4	8	N=2	4	8
LLaVA-OV-1.5-8B	TextVQA	Option label	14	16	16	14	16	16
LLaVA-OV-1.5-8B	TextVQA	Answer text	14	14	16	14	14	16
Qwen2.5-VL-7B	TextVQA	Option label	16	16	16	16	16	16
Qwen2.5-VL-7B	TextVQA	Answer text	16	16	16	16	16	16
InternVL3-8B	TextVQA	Option label	16	16	16	16	16	16
InternVL3-8B	TextVQA	Answer text	12	16	16	12	16	16
LLaVA-OV-1.5-8B	GQA	Option label	0	16	18	6	16	18
LLaVA-OV-1.5-8B	GQA	Answer text	4	20	16	22	20	16
Qwen2.5-VL-7B	GQA	Option label	16	16	18	16	16	24
Qwen2.5-VL-7B	GQA	Answer text	16	16	16	28	16	16
InternVL3-8B	GQA	Option label	18	18	18	18	18	18
InternVL3-8B	GQA	Answer text	16	16	18	16	16	18

The first-crossing depth is nondecreasing with candidate count in 11 of the 12 combinations of model, task, and format. Comparing $N = 8$ with $N = 2$, depth increases in seven combinations and is equal in five. However, dataset-level thresholds can remain equal even when individual examples change because accuracy is aggregated and only a set of layers is evaluated. We therefore also report the paired per-example analysis used in the main text.

For example i, candidate count N, and output format o, the stable depth $s_i(N, o)$ is the earliest evaluated layer from which the answer remains correct at every later evaluated layer. A paired comparison includes examples with observed stable depths under both conditions. We report the number retained in each comparison, so the resulting mean describes this paired subset. Confidence intervals use 10,000 bootstrap resamples of examples with seed 20260813.

Table 19: **Increasing the candidate set from two to eight.** Mean $\Delta_N = [s_i(8, o) - s_i(2, o)]/L$ on examples with both stable depths observed. Positive values indicate later dependence with more candidates. Intervals use 10,000 paired example bootstrap resamples.

Model	Task	Output	paired n	mean Δ_N	95% CI
LLaVA-OV-1.5-8B	TextVQA	Option label	284	0.0812	[0.0646, 0.0980]
LLaVA-OV-1.5-8B	TextVQA	Answer text	272	0.0231	[0.0010, 0.0451]
Qwen2.5-VL-7B	TextVQA	Option label	284	0.1281	[0.1019, 0.1547]
Qwen2.5-VL-7B	TextVQA	Answer text	271	0.0563	[0.0271, 0.0850]
InternVL3-8B	TextVQA	Option label	284	0.1084	[0.0850, 0.1315]
InternVL3-8B	TextVQA	Answer text	273	0.0634	[0.0406, 0.0862]
LLaVA-OV-1.5-8B	GQA	Option label	256	0.0646	[0.0409, 0.0888]
LLaVA-OV-1.5-8B	GQA	Answer text	227	0.0460	[0.0142, 0.0776]
Qwen2.5-VL-7B	GQA	Option label	265	0.1407	[0.1050, 0.1759]
Qwen2.5-VL-7B	GQA	Answer text	247	0.0882	[0.0564, 0.1209]
InternVL3-8B	GQA	Option label	259	0.0718	[0.0393, 0.1041]
InternVL3-8B	GQA	Answer text	251	0.1017	[0.0731, 0.1319]

The mean normalized change from two to eight candidates is positive in all twelve combinations, with all twelve 95% intervals above zero (Table 19). The median of these mean changes is 0.0765, with values from 0.0231 to 0.1407. Importantly, increasing candidate count extends visual dependence under both response formats, supporting the result on the number of candidates in Section 5.1.

Table 20: **Changing option-label output to answer-text output.** Mean $\Delta_{\text{form}} = [s_i(N, \text{text}) - s_i(N, \text{label})]/L$ on examples with both stable depths observed. Positive values indicate later dependence for answer text. Intervals use 10,000 paired example bootstrap resamples.

Model	Task	N	paired n	mean Δ_{form}	95% CI
LLaVA-OV-1.5-8B	TextVQA	2	277	0.0556	[0.0373, 0.0744]
LLaVA-OV-1.5-8B	TextVQA	4	276	-0.0012	[-0.0193, 0.0167]
LLaVA-OV-1.5-8B	TextVQA	8	275	0.0024	[-0.0127, 0.0180]
Qwen2.5-VL-7B	TextVQA	2	271	0.0818	[0.0524, 0.1122]
Qwen2.5-VL-7B	TextVQA	4	272	0.0636	[0.0364, 0.0918]
Qwen2.5-VL-7B	TextVQA	8	274	0.0125	[-0.0136, 0.0391]
InternVL3-8B	TextVQA	2	279	0.0127	[-0.0123, 0.0381]
InternVL3-8B	TextVQA	4	276	-0.0060	[-0.0272, 0.0155]
InternVL3-8B	TextVQA	8	276	-0.0305	[-0.0497, -0.0114]
LLaVA-OV-1.5-8B	GQA	2	238	0.0766	[0.0502, 0.1039]
LLaVA-OV-1.5-8B	GQA	4	262	0.0683	[0.0422, 0.0952]
LLaVA-OV-1.5-8B	GQA	8	248	0.0665	[0.0426, 0.0923]
Qwen2.5-VL-7B	GQA	2	274	0.0184	[-0.0132, 0.0503]
Qwen2.5-VL-7B	GQA	4	258	0.0159	[-0.0148, 0.0476]
Qwen2.5-VL-7B	GQA	8	244	-0.0280	[-0.0590, 0.0035]
InternVL3-8B	GQA	2	278	0.0117	[-0.0134, 0.0378]
InternVL3-8B	GQA	4	269	0.0467	[0.0198, 0.0755]
InternVL3-8B	GQA	8	239	0.0254	[-0.0006, 0.0517]

By contrast, changing the output format has a smaller and less consistent effect (Table 20). Across eighteen comparisons, the mean changes from label to answer text range from -0.0305 to 0.0818, with a median of 0.0171. Seven intervals are above zero, one is below zero, and ten include zero. These results identify candidate count as the more consistent factor within this experiment.

Scoring sensitivity. Clean answer-text format compliance ranges from 0.837 to 0.977. To assess whether the result depends on exact candidate matching, we also score responses by the open-ended substring matching rule, which accepts a correct target answer within a longer response. Under this more permissive rule, the change from two to eight candidates is positive in all six answer-text conditions, with a median normalized change of 0.0819 and all intervals above zero. The median format difference is 0.0124; five intervals are positive, two negative, and eleven include zero. The qualitative comparison between candidate count and output format therefore persists.

Dataset-level summary. Ordinary least squares on the 36 first-crossing depths gives

$$\bar{\ell}_{\mathrm{dep}} = 0.4158 + 0.0522 \log_2 N - 0.0097 \mathbf{1}_{\mathrm{text}} - 0.0020 \log_2 N \, \mathbf{1}_{\mathrm{text}}. \tag{8}$$

Here $\mathbf{1}_{\mathrm{text}}$ indicates answer-text output. This is a descriptive summary; the paired stable-depth comparisons above provide the main result on the number of candidates.

For the likelihood and factorial tables, thresholds are evaluated from integer correct answer counts. At $n = 300$, a 0.02 drop is exactly six answers, so the inclusive test is $50(c_{\mathrm{clean}} - c_k) \leq n$.

D.2 Answer entropy and the included model-task pairs

The cross-benchmark analysis complements the intervention on the number of candidates with a descriptive measure of answer diversity. For answer frequencies $p(z)$, $H = -\sum_z p(z) \log_2 p(z)$ is measured in bits. Specifically, a higher value means that the distribution covers more answers or is less concentrated. The answer corpora used to estimate these values are larger than the 124–300-example intervention sets. Their sizes and entropy values are listed in Table 21.

Table 21: **Answer distribution summaries used for entropy.** Counts refer to the answer corpus used for each task. BLINK uses a uniform two-label reference.

Task	Answers	Unique	H (bits)	Basis
POPE	3,000	2	1.0000	Answer frequencies
BLINK-spatial	—	2	1.0000	Uniform reference
BLINK-depth	—	2	1.0000	Uniform reference
MMBench	4,329	4	1.9960	Answer frequencies
ScienceQA	4,241	5	1.7214	Answer frequencies
GQA	12,578	722	6.5363	Answer frequencies
TextVQA	8,652	6,049	11.7746	Answer frequencies
DocVQA	2,674	2,272	10.9662	Answer frequencies

GQA uses the benchmark's balanced test-development answer corpus. TextVQA uses five training shards, totaling 8,652 answers; the DocVQA corpus contains 2,674 answers. TextVQA and DocVQA use the most frequent annotated answer per example, lowercased with surrounding whitespace removed. For the fixed-choice MMBench and ScienceQA tasks, we instead use validation answer label frequencies. The binary reference uses $H = 1$: BLINK is assigned a uniform two-label distribution, while POPE's balanced yes/no distribution has the same entropy. These choices summarize answer diversity at the task level, separately from the sample used to measure dependency depth.

Inclusion rules. Two of the 88 benchmark pairs have clean accuracy below 0.20 and are excluded from all pooled regressions, leaving 86. For the main visualization and regression across the eight task means, the additional rule is $A^{\mathrm{clean}} - 2^{-H} \geq 0.10$, leaving 81 pairs. Here 2^{-H} is an entropy-based reference value. It equals uniform chance for a balanced fixed-choice task; for a nonuniform or open-ended answer distribution it is a screening reference, not the accuracy of a particular guessing model. The 81-pair analysis includes two combinations with $\ell_{\mathrm{dep}} = L$.

A sensitivity analysis restricts the 86-pair set to the 84 combinations with $\ell_{\mathrm{dep}} < L$. The image-dependence analysis below uses weights on all 86 pairs; restricting it to pairs with $w = 1$ leaves 73. Each analysis reports its corresponding set of pairs.

Image-dependence weighting. We compare clean accuracy with a no-image control A^B. These controls include both randomized projected visual features and replaced input images, with benchmark-specific prompts and answer matching. The image intervention is therefore not uniform across this analysis. It is distinct from the Gaussian-image controls on the same examples in Appendix B. The weight is

$$w = \mathrm{clip}\left(\frac{A^{\mathrm{clean}} - A^B - 0.05}{0.10}, 0, 1\right). \tag{9}$$

The thresholds 0.05 and 0.15 define a gradual weighting rule. There are 73 pairs with $w = 1$, eleven with $0 < w < 1$, and two with $w = 0$. The zero-weight pairs are LLaVA-OV-Qwen2-0.5B and Qwen2.5-VL-3B on BLINK-depth, whose clean accuracies are near the binary reference and show no advantage over their no-image controls. Weighted least squares minimizes $\sum_j w_j(\bar{\ell}_j - \hat{\ell}_j)^2$.

The main task mean fit has $R^2 = 0.936$. The 86-pair unweighted fit gives $R^2 = 0.693$; with image-dependence weighting, this rises to $R^2 = 0.815$; restricting to the 73 pairs with $w = 1$ gives $R^2 = 0.865$. Thus the association persists under these different inclusion rules. Its interpretation remains descriptive because the benchmarks differ in visual content as well as answer requirements.

D.3 Alternative predictors

Table 22: **Alternative predictors on the same 86 pairs.** Unweighted ordinary least squares gives R^2 for each predictor alone, with answer entropy H, and the increment over H alone.

Predictor	R^2 alone	With H	Increment
Answer entropy H	0.693	—	—
No-image accuracy	0.520	0.693	+0.000
log(parameters)	0.045	0.723	+0.029
Hidden size	0.031	0.713	+0.020
Model family (five families)	0.014	0.701	+0.008
Clean accuracy	0.000	0.709	+0.015
Model depth L	0.000	0.694	+0.001

To contextualize the entropy association, Table 22 compares entropy with no-image accuracy, model size, hidden size, family, clean accuracy, and decoder depth on the same 86 pairs. Model family uses five categories, matching the model list in Appendix A. No-image accuracy is strongly correlated with entropy, and adding it to entropy changes R^2 by less than 0.001.

Answer length. Answer diversity and response length describe different properties of a task. We therefore compare entropy with the logarithm of mean answer length, measured in whitespace-separated words in the ground-truth answers. The task means are 1.00 for POPE, both BLINK tasks, and ScienceQA; 1.80 for MMBench; 1.06 for GQA; 1.39 for TextVQA; and 2.42 for DocVQA. This analysis uses the 84 combinations with $\ell_{\mathrm{dep}} < L$. Without weighting, length alone has $R^2 = 0.226$, entropy has 0.677, and their joint fit has 0.677. With the weights in Eq. (9), the corresponding values are 0.210, 0.804, and 0.807. Importantly, adding entropy to length increases the weighted R^2 by 0.597, whereas adding length to entropy increases it by 0.003. Thus mean answer length explains little additional variation once entropy is included at the task level.

D.4 Robustness of the entropy regression

The eight task means used in the main text satisfy the descriptive fit

$$\bar{\ell}_{\mathrm{dep}} = 0.038H + 0.522, \qquad R^2 = 0.936, \quad n_{\mathrm{tasks}} = 8. \tag{10}$$

Because entropy has one value per task, this regression has eight distinct task-level observations. To assess robustness at that level, Table 23 uses the alternative 84-pair set with $\ell_{\mathrm{dep}} < L$. It reports a task mean fit, pooled fits after leaving out one task at a time, and a bootstrap that resamples whole tasks.

Table 23: **Entropy regression for the 84 combinations with $\ell_{\text{dep}} < L$.** The task mean fit uses eight means. Leave-one-task-out fits use the remaining model-task pairs; cluster-bootstrap intervals use 5,000 resamples of whole tasks. These are sensitivity analyses with a different inclusion rule from the main 81-pair task mean fit.

Analysis	slope	R^2	note
Pooled settings (n=84)	0.043	0.68	8 tasks
Task mean (n=8)	0.043	0.94	8 task means
Leave-one-task-out	0.041 to 0.045	n/a	TextVQA most influential
Task cluster bootstrap	$[0.037, 0.058]$	$[0.23, 0.90]$	resamples tasks

Importantly, all leave-one-task-out slopes are positive. On these 84 pairs, entropy alone gives $R^2 = 0.68$, no-image accuracy gives 0.50, and their joint fit gives 0.68. The two predictors are correlated ($r = -0.86$), so the joint fit should be interpreted with that overlap in mind.

Within fixed-choice tasks, the association is weak ($R^2 = 0.004$, slope 0.021, three distinct entropy values). By contrast, within the three open-ended tasks, the pooled association is stronger ($R^2 = 0.73$, slope 0.027), with three task-level entropy values. These results support entropy as a summary of broad benchmark differences. Alongside this descriptive association, the experiments on the same examples and on the number of candidates provide controlled evidence for the effect of the GAS.

E Error recovery and practical implications

E.1 Two-stage inference with the model-generated GAS

To evaluate two-stage inference with the GAS under practical conditions, we construct candidate pools without ground-truth answers. LLaVA-OV-1.5-8B, Qwen2.5-VL-7B, and InternVL3-8B each serve as the target model on the same 300 TextVQA and 300 GQA examples. Each model first generates one greedy answer and seven sampled answers under the original open-ended prompt. Sampling uses temperature 0.7, top-p 0.9, and a maximum of 16 new tokens.

For each target, the primary pool takes four answers from each of the other two models: the greedy answer and the first three sampled answers. The comparison takes eight answers from the target itself. This matches the number of raw answers generated, while the source models and their computation costs can differ. Ground-truth answers are used only for evaluation, after generation, pool construction, ordering, and selection.

Candidates are deduplicated after case folding and whitespace normalization. We alternate between the two source models and keep at most eight distinct answers. An answer generated by another model is retained even when its text matches the target's original answer; the target's original answer is not separately added to the primary pool.

For selection, candidates are shuffled deterministically for presentation. Each candidate is assigned an option label (A, B, and so on). Direct selection compares the model's next-token logits for these labels and returns the candidate associated with the highest-scoring label. A single candidate is returned directly; an empty pool falls back to the original answer. No primary pool is empty. Both generated and selected answers use the same substring matching rule against the example's reference answer.

Metrics and full results. *Coverage* is the fraction of original errors for which the pool contains a correct answer. *Recovery* is the fraction of original errors corrected by the complete procedure. *Reverse errors* are originally correct answers made incorrect. *Net accuracy gain* therefore equals the number recovered minus reverse errors, divided by all predictions. The recovery denominator is 624 original errors, while accuracy uses all 1,800 predictions.

The primary procedure recovers 148 errors and introduces 84 reverse errors. Accuracy therefore rises from $1176/1800$ to $1240/1800$ (65.33% to 68.89%), a gain of 3.56 pp. By comparison, candidates from the target model recover 76 errors and introduce 58 reverse errors. Moreover, recovery

Table 24: **Model-generated candidate selection on six model-task pairs.** Each pair uses 300 examples. Top: original errors, recovery with a supplied correct candidate, and model-generated candidate coverage and recovery. Bottom: selection accuracy with candidates from other models and reverse errors. Rates are percentages; pools are constructed without ground truth.

Task	Target model	Errors	Supplied	Coverage	Other recovery	Own recovery
TextVQA	LLaVA-OV-1.5-8B	99	94.9	35.4	27.3	11.1
TextVQA	Qwen2.5-VL-7B	84	91.7	25.0	20.2	11.9
TextVQA	InternVL3-8B	101	94.1	38.6	26.7	8.9
GQA	LLaVA-OV-1.5-8B	110	85.5	46.4	22.7	15.5
GQA	Qwen2.5-VL-7B	118	89.0	51.7	21.2	15.3
GQA	InternVL3-8B	112	91.1	50.0	24.1	9.8
Pooled	—	624	90.9	42.1	23.7	12.2

Task	Target model	Original accuracy	Reverse errors	Final accuracy
TextVQA	LLaVA-OV-1.5-8B	67.0	5	74.3
TextVQA	Qwen2.5-VL-7B	72.0	14	73.0
TextVQA	InternVL3-8B	66.3	7	73.0
GQA	LLaVA-OV-1.5-8B	63.3	16	66.3
GQA	Qwen2.5-VL-7B	60.7	15	64.0
GQA	InternVL3-8B	62.7	27	62.7
Pooled	—	65.3	84	68.9

with candidates from other models exceeds recovery with candidates from the same model in all six pairs; final accuracy improves over original generation in five pairs and ties in one.

The 95% intervals use 10,000 bootstrap resamples with seed 20260904. Within each task, we resample its 300 examples and keep the three target models' predictions for an example together. This accounts for shared questions across models. The recovery interval is $[19.7, 28.1]\%$, and the net gain interval is $[2.11, 5.06]$ pp. Compared with candidates from the same model, recovery increases by 11.5 points ($[7.8, 15.4]$), and final accuracy by 2.56 points ($[1.17, 4.00]$).

Table 25: **Candidate generation budget.** All rows use direct selection. Coverage and recovery are percentages of the same 624 original errors; net gain uses all 1,800 predictions. The cap of eight distinct candidates is applied after deduplication.

Candidate source	Raw answers	Distinct answers (mean)	Coverage	Recovery	Reverse	Net gain (pp)
Other models, 1 answer each	2	1.28	34.1	24.0	96	+3.00
Other models, 4 answers each	8	1.73	42.1	23.7	84	+3.56
Other models, 8 answers each	16	2.00	47.6	23.6	90	+3.17
Target model, 8 answers	8	1.52	24.2	12.2	58	+1.00

Candidate count and selection method. Using one, four, or eight answers from each other model gives two, eight, or sixteen raw generations before deduplication; the distinct answer cap remains eight. Although coverage increases with this budget, recovery remains near 24% (Table 25). Thus, additional correct candidates do not automatically translate into greater recovery. We therefore compare selection methods on identical pools to examine the contribution of selection while holding candidate generation fixed.

On the primary pool, we compare direct selection with frequency voting, mean per-token candidate likelihood under the original prompt, and answer-text generation with candidates in the prompt (Table 1). Recovery rates and counts of reverse errors are 22.9% and 125 for frequency voting, 19.6% and 128 for likelihood, 19.9% and 82 for answer-text generation, and 23.7% and 84 for direct selection. The generated answer matches a candidate in 97.6% of cases.

After deduplication, 973 of 1,800 primary pools (54.1%) contain one answer. Consequently, the target chooses among alternatives in the remaining 827 predictions. Of the 148 recovered errors, 109 (73.6%) come from these multiple-candidate cases. Crucially, direct selection improves accuracy over frequency voting by 2.56 points overall and 5.56 points on those 827 predictions. These comparisons establish the complementary contributions of candidate generation and selection: other models supply useful answers, and the selection procedure affects accuracy when alternatives remain.

Visual input during selection. To measure the image contribution after candidate generation, we reuse the primary pools from other models and hold the question, candidate order, labels, and selection prompt fixed. Only the selector's image is replaced, using either a uniform gray image with RGB values $(128, 128, 128)$ or an image from another example in the same task. Replacement images have the original dimensions. Donor images are sampled with seed 20260926 after excluding images with identical pixels. The 973 single-candidate cases return the same answer in all conditions. Using the same references and matching rule as above, accuracy across all 1,800 predictions is 68.89% with the original image, 64.44% with the gray image, and 62.94% with another example's image. The original image therefore contributes to selection accuracy even after the candidate answers have been generated.

Table 26: **Including the target model's original answer.** All rows use 300 examples per task and target. Recovery is a percentage of original errors; reverse errors are counts; net gain uses all 1,800 predictions.

Candidate source	Predictions	Recovery	Reverse	Net gain (pp)
Other models, 1 answer each	1800	24.0	96	+3.00
Other models, 1 answer each + target answer	1800	16.0	54	+2.56
Other models, 4 answers each	1800	23.7	84	+3.56
Other models, 4 answers each + target answer	1800	17.1	62	+2.50

Including the target's original answer. Retaining the target's original answer introduces a trade-off. Adding it to the pool reduces both recovery and reverse errors in the 300-example evaluations (Table 26): it preserves some original successes but also makes some original failures harder to correct.

Table 27: **Additional ablations on 100 examples per task and target.** Recovery is a percentage of each row's original errors; net gain uses all predictions in that row. Single-source rows evaluate different target-model sets, as specified below.

Candidate source	Method	Predictions	Recovery	Reverse	Net gain (pp)
Candidate retention					
Other models, 4 each, retain matches	Answer text	600	23.4	31	+2.67
Other models, 4 each, exclude matches	Answer text	600	29.9	62	-0.33
Other models, 8 each, retain matches	Answer text	600	22.4	29	+2.67
Other models, 8 each, exclude matches	Answer text	600	29.4	65	-1.00
Candidate source					
Only LLaVA-OV-1.5-8B	Direct	400	28.0	25	+3.00
Only Qwen2.5-VL-7B	Direct	400	32.4	22	+5.75
Only InternVL3-8B	Direct	400	21.4	31	-0.75
Selection prompt					
Other models, 1 answer each	Direct	600	29.4	32	+4.50
Other models, 1 answer each	Reconsider	600	29.9	32	+4.67
Other models, 4 answers each	Direct	600	31.3	25	+6.33
Other models, 4 answers each	Reconsider	600	33.3	33	+5.67

Additional ablations on smaller subsets. These ablations use 100 examples per task and target, separately from the 300-example primary evaluation. First, excluding answers from other models

that match the target's greedy answer increases recovery but reduces net accuracy because some removed answers are correct. Second, a single source is evaluated on the other two targets, giving 400 predictions per source; comparisons with two sources have 600. Their aggregate accuracies therefore compare different target sets as well as different sources. Finally, asking the model to reconsider has no consistent advantage over direct selection.

E.2 Recovery when the correct answer is supplied

To place the model-generated candidate results in context, the diagnostic with a supplied correct candidate measures recovery when a correct candidate is guaranteed. It compares open-ended answering with two-choice selection on the same image and question, including the correct answer among the choices. This diagnostic quantifies the opportunity that practical candidate generation and selection seek to realize.

Let $c_i^I \in \{0, 1\}$ indicate a correct answer under setting I. The selection recovery rate is

$$\mathrm{SR} = \frac{\sum_i \mathbf{1}[c_i^{\mathrm{open}} = 0 \wedge c_i^{\mathrm{sel}} = 1]}{\sum_i \mathbf{1}[c_i^{\mathrm{open}} = 0]}. \tag{11}$$

The evaluation across the six pairs in the main text corrects 567 of 624 errors, or 90.9%. To assess the breadth of this result, the extended experiment below uses 1,000 examples per pair across eleven models and three tasks, with its own error sets.

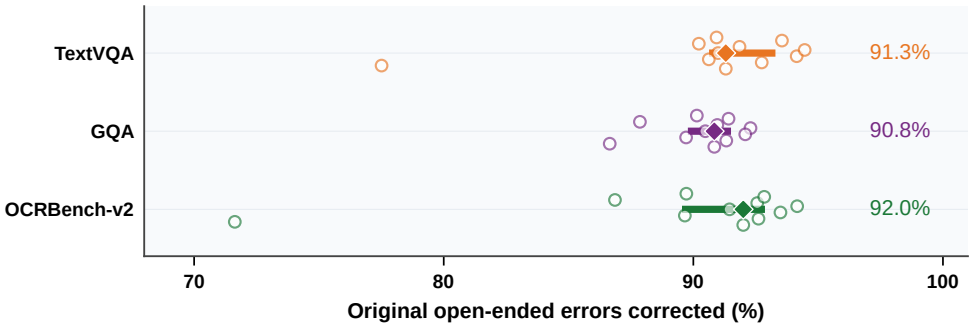

Figure 12: **Error recovery with a correct candidate.** Recovery among original open-ended errors for eleven models per task (1,000 examples per model). Diamonds and intervals summarize the model results as task medians and interquartile ranges.

Across this larger evaluation, median recovery is 0.91 on TextVQA, 0.91 on GQA, and 0.92 on OCRBench-v2 (Figure 12; Table 28). The noise-image column reports open-ended accuracy. The image control on the same selection examples appears in Table 8.

Table 28: **Recovery with a supplied correct candidate: full results.** Each pair uses the same 1,000 examples in both answering settings. Noise-image accuracy uses the open-ended prompt. Recovery is the fraction of original open-ended errors corrected by two-choice selection (Eq. (11)). All values are proportions.

TextVQA				
Model	Open-ended	Two-choice	Noise image	Recovery
LLaVA-OV-Qwen2-0.5B	0.38	0.84	0.03	0.78
LLaVA-OV-Qwen2-7B	0.57	0.98	0.04	0.94
LLaVA-OV-1.5-4B	0.65	0.97	0.05	0.91
LLaVA-OV-1.5-8B	0.65	0.97	0.06	0.91
Qwen2-VL-2B	0.64	0.97	0.04	0.92
Qwen2-VL-7B	0.68	0.97	0.04	0.90
Qwen2.5-VL-3B	0.66	0.98	0.05	0.94
Qwen2.5-VL-7B	0.69	0.98	0.07	0.94
InternVL3-1B	0.53	0.95	0.05	0.91
InternVL3-8B	0.59	0.97	0.05	0.93
InternVL3-14B	0.59	0.96	0.07	0.91
Median	**0.64**	**0.97**	**0.05**	**0.91**

GQA				
Model	Open-ended	Two-choice	Noise image	Recovery
LLaVA-OV-Qwen2-0.5B	0.50	0.90	0.27	0.87
LLaVA-OV-Qwen2-7B	0.62	0.96	0.31	0.92
LLaVA-OV-1.5-4B	0.62	0.95	0.32	0.90
LLaVA-OV-1.5-8B	0.64	0.96	0.31	0.91
Qwen2-VL-2B	0.60	0.95	0.30	0.90
Qwen2-VL-7B	0.62	0.96	0.32	0.91
Qwen2.5-VL-3B	0.61	0.96	0.33	0.92
Qwen2.5-VL-7B	0.61	0.94	0.38	0.88
InternVL3-1B	0.56	0.94	0.29	0.90
InternVL3-8B	0.62	0.96	0.31	0.91
InternVL3-14B	0.64	0.96	0.34	0.91
Median	**0.62**	**0.96**	**0.31**	**0.91**

OCRBench-v2				
Model	Open-ended	Two-choice	Noise image	Recovery
LLaVA-OV-Qwen2-0.5B	0.21	0.75	0.07	0.72
LLaVA-OV-Qwen2-7B	0.39	0.95	0.12	0.93
LLaVA-OV-1.5-4B	0.50	0.95	0.15	0.92
LLaVA-OV-1.5-8B	0.50	0.94	0.15	0.90
Qwen2-VL-2B	0.45	0.96	0.11	0.93
Qwen2-VL-7B	0.50	0.96	0.15	0.94
Qwen2.5-VL-3B	0.49	0.93	0.15	0.90
Qwen2.5-VL-7B	0.54	0.96	0.16	0.93
InternVL3-1B	0.35	0.90	0.10	0.87
InternVL3-8B	0.46	0.94	0.14	0.93
InternVL3-14B	0.46	0.95	0.17	0.91
Median	**0.46**	**0.95**	**0.15**	**0.92**

E.3 Recovery with candidates included in both prompts

To examine output requirements when candidates are already provided, we compare answer-text and option-label responses with both candidate answers included in the prompt. Each condition uses 300 examples per pair. The comparison changes candidate presentation and the required response, so it measures their combined effect. Recovery uses errors from the answer-text condition with candidates in the prompt.

Table 29: **Recovery with candidates included in both prompts.** Each pair uses 300 examples. Corrected and new errors count the two directions between answer-text and label outputs. Recovery uses answer-text errors, with 95% Wilson intervals. Noise is answer-text accuracy with a noise image; all rates are proportions.

TextVQA

Model	Answer text	Label	Corrected	New errors	Recovery [95% CI]	Noise
LLaVA-OV-Qwen2-0.5B	0.397	0.823	140	12	0.773 [0.707,0.828]	0.233
LLaVA-OV-Qwen2-7B	0.817	0.987	53	2	0.964 [0.877,0.990]	0.453
LLaVA-OV-1.5-4B	0.843	0.983	44	2	0.936 [0.828,0.978]	0.490
LLaVA-OV-1.5-8B	0.897	0.973	23	0	0.742 [0.568,0.863]	0.733
Qwen2-VL-2B	0.873	0.980	35	3	0.921 [0.792,0.973]	0.540
Qwen2-VL-7B	0.837	0.983	44	0	0.898 [0.782,0.956]	0.587
Qwen2.5-VL-3B	0.827	0.977	46	1	0.885 [0.770,0.946]	0.717
Qwen2.5-VL-7B	0.717	0.983	80	0	0.941 [0.870,0.975]	0.547
InternVL3-1B	0.843	0.947	35	4	0.745 [0.605,0.848]	0.653
InternVL3-8B	0.820	0.973	50	4	0.926 [0.825,0.971]	0.577
InternVL3-14B	0.950	0.970	9	3	0.600 [0.357,0.802]	0.807

GQA

Model	Answer text	Label	Corrected	New errors	Recovery [95% CI]	Noise
LLaVA-OV-Qwen2-0.5B	0.570	0.917	111	7	0.861 [0.790,0.910]	0.527
LLaVA-OV-Qwen2-7B	0.910	0.967	22	5	0.815 [0.633,0.918]	0.773
LLaVA-OV-1.5-4B	0.927	0.957	13	4	0.591 [0.387,0.767]	0.780
LLaVA-OV-1.5-8B	0.940	0.957	11	6	0.611 [0.386,0.797]	0.863
Qwen2-VL-2B	0.900	0.973	24	2	0.800 [0.627,0.905]	0.720
Qwen2-VL-7B	0.900	0.980	26	2	0.867 [0.703,0.947]	0.800
Qwen2.5-VL-3B	0.917	0.967	18	3	0.720 [0.524,0.857]	0.897
Qwen2.5-VL-7B	0.877	0.963	29	3	0.784 [0.628,0.886]	0.870
InternVL3-1B	0.880	0.963	29	4	0.806 [0.650,0.902]	0.790
InternVL3-8B	0.947	0.963	8	3	0.500 [0.280,0.720]	0.783
InternVL3-14B	0.960	0.970	6	3	0.500 [0.254,0.746]	0.887

Option-label selection corrects many answer-text errors, although it can also introduce new errors. Accordingly, Table 29 reports both directions and Wilson intervals for the recovery proportions. This control complements the factorial experiment, where candidate order is held fixed across output formats.

E.4 APPLYING AN MMBENCH CUTOFF TO OTHER TASKS

If visual dependency depth depends on the GAS, a layer chosen on one benchmark may not preserve accuracy on another. To test this implication, we use the MMBench-selected cutoff to zero visual states from layer 19 on LLaVA-OV-1.5-8B and layer 15 on Qwen2.5-VL-7B. Table 30 reports accuracy divided by the clean baseline and includes FastV with a 50% visual token budget for reference.

Table 30: **Accuracy under an MMBench-selected cutoff.** Visual states are zeroed from layer 19 on LLaVA-OV-1.5-8B and layer 15 on Qwen2.5-VL-7B. Values are fractions of clean accuracy. FastV retains 50% of visual tokens and is included as an accuracy reference.

Task	Model	MMBench cutoff	FastV, 50% budget
MMBench	Qwen2.5-VL-7B	1.01	0.99
POPE	Qwen2.5-VL-7B	0.99	0.99
GQA	Qwen2.5-VL-7B	0.96	0.98
TextVQA	Qwen2.5-VL-7B	0.14	0.86
TextVQA	LLaVA-OV-1.5-8B	0.71	1.00
DocVQA	Qwen2.5-VL-7B	0.06	0.92
DocVQA	LLaVA-OV-1.5-8B	0.77	0.97

The cutoff retains less accuracy on TextVQA and DocVQA than on the fixed-choice settings. Moreover, the loss varies across models: retained accuracy on TextVQA is 0.71 for LLaVA-OV-1.5-8B and 0.14 for Qwen2.5-VL-7B; the DocVQA values are 0.77 and 0.06. Consistent with this pattern, the cutoff lies before the open-ended dependency depth ℓ_{open} in each case. Thus, a cutoff's performance depends on both the task and model, motivating evaluation under the intended answering setting. This diagnostic measures accuracy under removal; it does not measure runtime.

F REPRODUCIBILITY AND INTERPRETATION

Evaluation settings. Each intervention gain uses the clean baseline from the same experiment. Different sample sets and prompts can produce different clean accuracies. The 300-example TextVQA/GQA comparison of answer requirements, 300-example candidate experiment, and 1,000-example extended recovery evaluation therefore use separate evaluation settings.

Layer indices and missing values. Removal, Transfer, and Freeze act at decoder-layer inputs. By contrast, probe features are layer outputs, with actual indices in Table 17. The figures and tables mark low clean baselines as LB and unreported conditions as —.

Uncertainty and data access. Bootstrap units follow the comparison: paired examples for replacement and stable-depth analyses, model-task pairs for the stated cross-pair Transfer and Freeze intervals, and whole tasks for the entropy robustness check. Candidate selection intervals resample examples within each task while keeping the three target models' predictions for each example together. The accompanying numerical data and scripts reproduce the appendix figures and regression checks without model checkpoints. Full model evaluations additionally require the benchmark images and checkpoints.

Interpretation. The controlled comparisons establish that the GAS affects visual dependency depth. Crucially, Transfer shows that intermediate visual states contribute to open-ended answers under the original prompt, while Freeze demonstrates the benefit of continued access without further state updates. Together, these interventions support the Answer Bottleneck in the tested decoder-only models on short answer, single-image tasks. The diagnostic with a supplied correct candidate then quantifies the opportunity for error recovery, and model-generated candidates improve accuracy without ground-truth access during candidate construction or selection. Finally, comparisons on identical candidate pools show that the selection procedure affects accuracy even when the candidate answers are unchanged.