ROUTELEDGER: ROUTING-AWARE ANSWER SELECTION WITH TASK-GATED VISUAL VERIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Mixture-of-experts vision-language models expose input-dependent routing signals, but directly intervening on expert routes is unreliable and can replace correct answers with plausible errors. We propose RouteLedger, a certificate-based residual selection framework that uses routing evidence conservatively. RouteLedger records expert assignments under text-only, visual, and full multimodal views, and converts their differences into route-aware candidate proposals. A beam-first selector accepts a route residual only when it receives sufficient independent support and passes explicit fidelity constraints. We further introduce SafeSwitch-Margin, a task-family-gated CLIP-L/336 verifier that evaluates a fixed non-route candidate pool and preserves the Beam-based answer when no residual is certified. On 45,857 examples from 12 VQA-style datasets, RouteLedger obtains 23,419 correct answers, improving Beam by 277 predictions and significantly outperforming the verifier-free selector. A matched Beam+Anchor system with the same verifier reaches 23,397 correct answers, leaving a small and non-significant residual for routing; visual verification therefore explains most of the gain. These results support MoE routing as supplementary evidence within a conservative candidate recipe.

1 INTRODUCTION

Large vision–language models (LVLMs) have become general interfaces for visual question answering, document understanding, optical character recognition, and multimodal reasoning. Visual instruction tuning enables open-ended image–text interaction Liu et al. (2023), while Qwen-VL strengthens visual grounding and text-rich image understanding Bai et al. (2023). Sparse mixture-of-experts (MoE) models increase capacity through conditional expert activation Fedus et al. (2022), and MoE-LLaVA extends this mechanism to multimodal instruction following Lin et al. (2026). Nevertheless, LVLMs still produce fluent but visually unsupported answers Liu et al. (2024a). POPE exposes object-existence hallucination Li et al. (2023), whereas MME reveals related failures in attributes, counting, spatial perception, and multimodal reasoning Fu et al. (2026).

Existing inference-time methods improve visual grounding by modifying decoding or internal representations. Visual Contrastive Decoding contrasts predictions from original and distorted images Leng et al. (2024). OPERA suppresses over-trusted attention patterns through retrospection Huang et al. (2024), while PAI increases the contribution of visual evidence during generation Liu et al. (2024b). Representation-level methods instead apply latent steering Liu et al. (2025), remove hallucination-related subspaces Yang et al. (2024), or predict input-conditioned intervention directions Cheng et al. (2026). These approaches can recover missed visual evidence, but direct intervention is not uniformly safe: the same modification may correct one answer and rewrite another correct prediction into a plausible error.

Candidate aggregation and external verification provide a complementary strategy. Self-consistency selects answers supported across multiple generation paths Wang et al. (2023), while Woodpecker repairs claims that are unsupported by external evidence Yin et al. (2024). CLIP-Guided Decoding uses image–text compatibility as an additional grounding signal Deng et al. (2024), building on the transferable visual–text representation of CLIP Radford et al. (2021). However, generic voting ignores candidate provenance: several answers produced by related route interventions may form a correlated majority rather than independent evidence. A global visual verifier is also unreliable

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.

across heterogeneous tasks, especially for charts, numerical reasoning, knowledge-intensive QA, and synthetic benchmarks. The key question is therefore when a route-derived alternative is sufficiently supported to replace a strong default answer.

To address this question, we propose RouteLedger, a routing-aware framework that formulates answer correction as risk-constrained residual selection. RouteLedger records MoE gate behavior under text-only, visual, and full multimodal prompt views and converts route differences into controlled candidate proposals. A beam-first selector accepts a route-derived answer only when it receives independent candidate support and passes explicit fidelity checks for attributes, numbers, OCR strings, dates, units, and answer form. We further introduce SafeSwitch-Margin, a task-family-gated CLIP-L/336 margin verifier that arbitrates over a fixed non-route candidate pool when routing evidence is insufficient. Otherwise, the method abstains and preserves the Beam-based answer. Controlled experiments across 12 VQA-style datasets show that selective visual verification supplies most of the aggregate improvement, while routing evidence provides a smaller residual benefit within the conservative selector.

Contributions. The main contributions of this paper are:

- We introduce a paired modality routing ledger that records MoE expert behavior across text-only, visual, and multimodal views and converts routing differences into controlled answer proposals.
- We formulate answer correction as beam-preserving, risk-constrained residual selection with independent support and explicit fidelity constraints.
- We introduce a task-family-gated visual–text verifier and provide matched-verifier, shuffled-unit, and paired statistical analyses that separate the contributions of routing evidence and visual arbitration.

2 Related Work

2.1 Large Vision–Language Models

Large vision–language models (LVLMs) typically connect a pretrained visual encoder with a large language model for multimodal understanding and generation. CLIP learns transferable image–text representations through contrastive pretraining Radford et al. (2021). LLaVA introduces visual instruction tuning by projecting image features into the language model Liu et al. (2023), while Qwen-VL extends this paradigm to visual grounding and text-rich image understanding Bai et al. (2023).

Mixture-of-experts (MoE) models improve computational efficiency by routing each token to a small subset of experts. Switch Transformers demonstrate the scalability of sparse expert activation Fedus et al. (2022), and MoE-LLaVA applies conditional expert routing to multimodal instruction following Lin et al. (2026). RouteLedger treats routing behavior as an observable multimodal signal: it compares expert paths across text-only, visual, and multimodal inputs and uses route changes to construct answer candidates.

Recent work also studies routing as a behavioral signal or an inference-time control mechanism. Bandarkar et al. (2026) analyze how routing changes across languages, while Fayyaz et al. (2026) steer an MoE model by deactivating experts. These studies make clear that observing or changing expert activation is not, by itself, the novelty of RouteLedger. Our focus is the answer-level decision rule: paired modality views provide proposals, and a conservative selector requires support and fidelity checks before allowing a proposal to replace Beam.

2.2 Hallucination Mitigation in LVLMs

Existing hallucination mitigation methods can be broadly divided into decoding-based and representation-based approaches. Decoding methods adjust generation at inference time without modifying model parameters. Visual Contrastive Decoding contrasts predictions from clean and distorted images Leng et al. (2024). OPERA penalizes over-trusted attention patterns through retrospection allocation Huang et al. (2024), while PAI increases the contribution of visual evidence during generation Liu et al. (2024b).

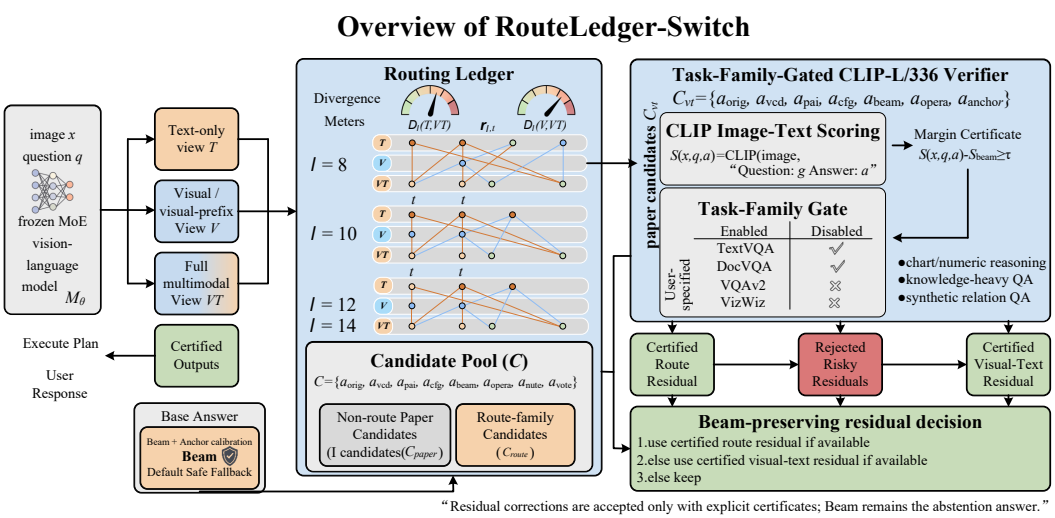


Figure 1: Overview of RouteLedger. Text-only, visual-prefix, and full multimodal views form a paired routing ledger. Route interventions provide residual candidates, while SafeSwitch-Margin verifies a fixed non-route candidate pool. Certified route residuals have first priority, followed by certified visual–text residuals; otherwise, the Beam-based answer is preserved.

Representation-based methods directly modify hidden states or hallucination-related feature directions. VTI applies latent-space steering vectors Liu et al. (2025), Nullu removes a hallucination-related subspace through projection Yang et al. (2024), and RFI predicts input-specific intervention vectors using rectified flow Cheng et al. (2026). Although these methods can recover visually grounded outputs, an intervention may also replace a correct answer with a plausible error.

Candidate aggregation provides a complementary solution. Self-consistency selects answers supported across multiple generation paths Wang et al. (2023), and CLIP-Guided Decoding uses image–text compatibility as an external grounding signal Deng et al. (2024). However, existing selectors often ignore candidate provenance and apply the same criterion across heterogeneous tasks. RouteLedger instead treats route interventions only as proposals, preserves Beam as the fallback, and accepts a correction only when it receives independent support and passes explicit fidelity checks. Its visual verifier is further restricted to pre-specified task families and non-route candidates, preventing route proposals from verifying themselves.

3 Method

RouteLedger formulates answer correction as certificate-based residual selection. Here, "certificate" denotes a deterministic heuristic acceptance predicate, not a formal correctness or risk guarantee. As shown in Figure 1, paired modality views expose MoE routing differences and generate route candidates. The selector accepts a non-base answer only when it passes a route certificate or the task-gated SafeSwitch-Margin visual–text certificate; otherwise, it preserves the Beam-based answer.

3.1 Paired Modality Routing Ledger

Let x denote the image, q the question, and M_θ the frozen MoE vision–language model. We construct text-only, visual-prefix, and full multimodal views, denoted by T, V, and VT, respectively. At layer ℓ and token position t, the gate produces

$$r^m_{\ell,t} = \mathrm{TopK}\left(g^m_{\ell,t}\right), \qquad m \in \{T, V, VT\}, \tag{1}$$

where $g^m_{\ell,t}$ is the expert-logit vector for view m, and $r^m_{\ell,t}$ is its selected expert set.

The paired routing ledger is

$$\mathcal{R}(x,q) = \left\{ \left(\ell, t, g^T_{\ell,t}, g^V_{\ell,t}, g^{VT}_{\ell,t}, r^T_{\ell,t}, r^V_{\ell,t}, r^{VT}_{\ell,t} \right) \right\}. \tag{2}$$

Here, $\mathcal{R}(x,q)$ stores aligned gate logits and expert assignments for image x and question q; each aligned layer–token record is a routing unit. Because the three views have different modality prefixes, we align their shared text suffix by token index and discard modality-only prefix positions from paired comparisons. The default capture plan uses the middle layers $\{8, 10, 12, 14\}$, while the additional layers recorded for the audit are reported in the experimental setup.

The route divergence between views A and B is

$$D_\ell(A,B) = \frac{1}{|\mathcal{T}_\ell|} \sum_{t \in \mathcal{T}_\ell} \mathbb{I}\left[r^A_{\ell,t} \neq r^B_{\ell,t} \right], \tag{3}$$

where $\mathcal{T}_\ell$ is the captured token set at layer ℓ, $|\mathcal{T}_\ell|$ is its size, and $\mathbb{I}[\cdot]$ is the indicator function.

3.2 Route Candidate Proposals

For selected expert e, a route proposal modifies the gate logit as

$$\widetilde{g}_{\ell,t,e} = g_{\ell,t,e} + \alpha_{\ell,t,e}, \tag{4}$$

where $g_{\ell,t,e}$ is the original logit, $\alpha_{\ell,t,e}$ is the sparse intervention bias, and $\widetilde{g}_{\ell,t,e}$ is the modified logit.

Let $\mathcal{C}_{\text{route}}$ be the route-family candidate set and $n(a)$ the normalized form of answer a. The route consensus is

$$\widehat{a}_{\text{gate}} = \arg\max_{a} \sum_{c \in \mathcal{C}_{\text{route}}} \mathbb{I}\left[n(c) = n(a) \right], \tag{5}$$

where $\widehat{a}_{\text{gate}}$ is retained only when at least two route-family candidates agree.

3.3 Certificate-Based Residual Selection

Figure 2 summarizes candidate evidence, fidelity constraints, and the final residual hierarchy.

The candidate pool is

$$\begin{aligned} \mathcal{C} &= \mathcal{C}_{\text{paper}} \cup \mathcal{C}_{\text{route}}, \\ \mathcal{C}_{\text{paper}} &= \{a_{\text{orig}}, a_{\text{vcd}}, a_{\text{pai}}, a_{\text{cfg}}, a_{\text{beam}}, a_{\text{opera}}, a_{\text{anchor}}\}, \\ \mathcal{C}_{\text{route}} &= \{a_{\text{route}}, a_{\text{vote}}\}. \end{aligned} \tag{6}$$

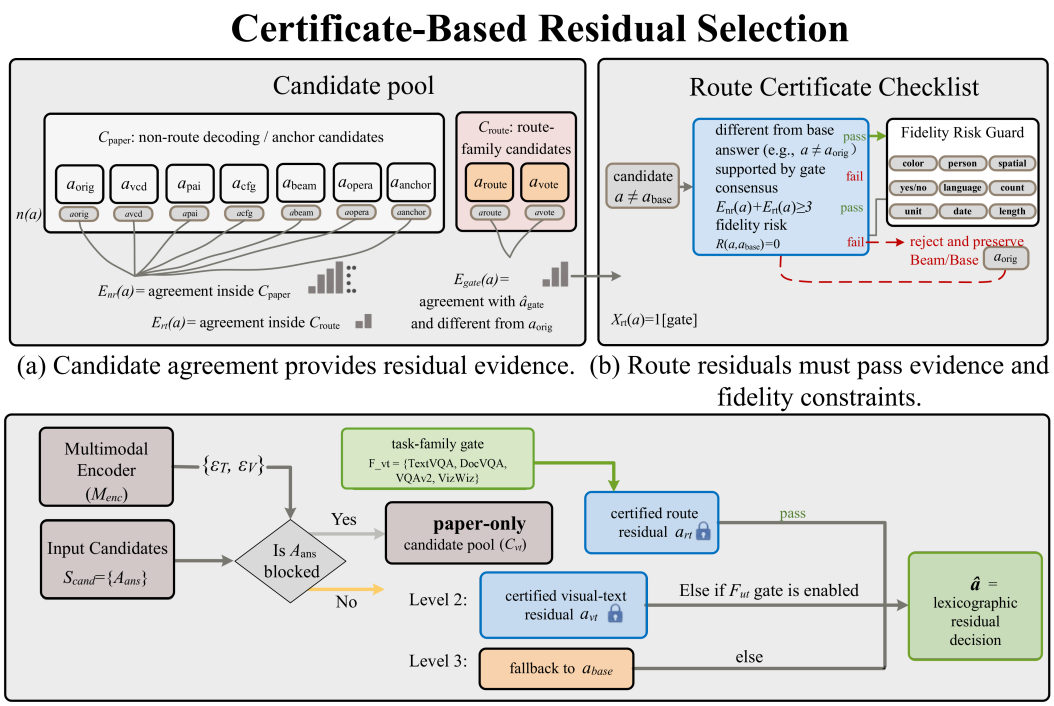
Here, $\mathcal{C}$ is the complete pool, $\mathcal{C}_{\text{paper}}$ contains non-route decoding and anchor candidates, and $\mathcal{C}_{\text{route}}$ contains route-derived candidates.

The preserved base answer is

$$a_{\text{base}} = \begin{cases} a_{\text{anchor}}, & \chi_{\text{anchor}} = 1, \\ a_{\text{beam}}, & \text{otherwise}, \end{cases} \tag{7}$$

where χ_{anchor} indicates a high-support yes/no or single-choice anchor answer.

For candidate a, the evidence scores are



(a) Candidate agreement provides residual evidence. (b) Route residuals must pass evidence and fidelity constraints.

(c) Switch verifies paper-only candidates and preserves Beam when uncertified.

Figure 2: Certificate-based residual selection. Candidate agreement provides non-route, route-family, and gate-consensus evidence. Route residuals must differ from the base answer, receive sufficient support, and pass the fidelity-risk guard. If no route residual is certified, SafeSwitch-Margin verifies the fixed paper-only candidate pool; otherwise, the base answer is retained.

$$
\begin{aligned}
E_{\mathrm{nr}}(a) &= \sum_{c \in \mathcal{C}_{\mathrm{paper}}} \mathbb{I}[n(c) = n(a)], \\
E_{\mathrm{rt}}(a) &= \sum_{c \in \mathcal{C}_{\mathrm{route}}} \mathbb{I}[n(c) = n(a)], \\
E_{\mathrm{gate}}(a) &= \mathbb{I}[n(a) = n(\widehat{a}_{\mathrm{gate}})]\mathbb{I}[n(\widehat{a}_{\mathrm{gate}}) \neq n(a_{\mathrm{orig}})].
\end{aligned}
\tag{8}
$$

Here, $E_{\mathrm{nr}}(a)$ counts non-route support, $E_{\mathrm{rt}}(a)$ counts route-family support, and $E_{\mathrm{gate}}(a)$ indicates agreement with a route consensus that differs from the original answer.

Let $R(a, a_{\mathrm{base}}) \in \{0, 1\}$ denote the combined fidelity risk over attributes, numbers, OCR strings, dates, units, and answer length. A route candidate is certified by

$$
\begin{aligned}
\chi_{\mathrm{rt}}(a) = \mathbb{I}[n(a) \neq n(a_{\mathrm{base}})]\mathbb{I}[E_{\mathrm{gate}}(a) = 1] \\
\cdot \mathbb{I}[E_{\mathrm{nr}}(a) + E_{\mathrm{rt}}(a) \geq 3]\mathbb{I}[R(a, a_{\mathrm{base}}) = 0].
\end{aligned}
\tag{9}
$$

Here, $\chi_{\mathrm{rt}}(a) = 1$ means that a differs from the base, matches the route consensus, has at least three supporting votes, and introduces no detected fidelity risk.

The certified route set is

$$
\mathcal{A}_{\mathrm{rt}} = \{a \in \mathcal{C} : \chi_{\mathrm{rt}}(a) = 1\}. \tag{10}
$$

The route-selected answer is

$$a_{\mathrm{rt}} = \begin{cases} \arg\max_{a \in \mathcal{A}_{\mathrm{rt}}} \left(E_{\mathrm{gate}}(a), E_{\mathrm{nr}}(a) + E_{\mathrm{rt}}(a)\right), & \mathcal{A}_{\mathrm{rt}} \neq \varnothing, \\ a_{\mathrm{base}}, & \mathcal{A}_{\mathrm{rt}} = \varnothing. \end{cases} \tag{11}$$

Here, a_{rt} is the best certified route residual, selected lexicographically by gate agreement and total support.

3.4 Task-Family-Gated SafeSwitch-Margin

The verifier pool excludes route candidates:

$$\mathcal{C}_{\mathrm{vt}} = \mathcal{C}_{\mathrm{paper}}, \tag{12}$$

where $\mathcal{C}_{\mathrm{vt}}$ is deduplicated by normalized answer string.

For candidate a, the frozen CLIP-L/336 verifier computes

$$S(x, q, a) = \gamma \left\langle \frac{\phi_I(x)}{\|\phi_I(x)\|_2}, \frac{\phi_T(t(q,a))}{\|\phi_T(t(q,a))\|_2} \right\rangle, \tag{13}$$

where ϕ_I and ϕ_T are the CLIP image and text encoders, γ is the learned logit scale, and

$$t(q, a) = \texttt{Question: } q \ \ \texttt{Answer: } a \tag{14}$$

is the candidate text template.

The enabled task families are

$$\mathcal{F}_{\mathrm{vt}} = \{\mathrm{TextVQA}, \mathrm{DocVQA}, \mathrm{VQAv2}, \mathrm{VizWiz}\}, \tag{15}$$

where $\mathcal{F}_{\mathrm{vt}}$ is fixed before evaluation.

The visual–text certificate is

$$\begin{aligned} \chi_{\mathrm{vt}}(a) = \mathbb{I}[d \in \mathcal{F}_{\mathrm{vt}}]\, \mathbb{I}[n(a) \neq n(a_{\mathrm{beam}})] \\ \times \mathbb{I}[S(x, q, a) - S(x, q, a_{\mathrm{beam}}) \geq \tau], \end{aligned} \tag{16}$$

where d is the dataset family and τ is the globally calibrated margin.

The certified visual–text set and its best candidate are

$$\begin{aligned} \mathcal{A}_{\mathrm{vt}} &= \{a \in \mathcal{C}_{\mathrm{vt}} : \chi_{\mathrm{vt}}(a) = 1\}, \\ a^{\star}_{\mathrm{vt}} &= \arg\max_{a \in \mathcal{A}_{\mathrm{vt}}} S(x, q, a). \end{aligned} \tag{17}$$

Here, $\mathcal{A}_{\mathrm{vt}}$ is the certified verifier set and $a^{\star}_{\mathrm{vt}}$ is its highest-scoring candidate.

3.5 Final Decision

The final answer is

$$\widehat{a} = \begin{cases} a_{\mathrm{rt}}, & n(a_{\mathrm{rt}}) \neq n(a_{\mathrm{base}}), \\ a^{\star}_{\mathrm{vt}}, & \mathcal{A}_{\mathrm{vt}} \neq \varnothing, \\ a_{\mathrm{base}}, & \text{otherwise.} \end{cases} \tag{18}$$

Here, $\widehat{a}$ is the deployed answer: certified route residuals have first priority, certified visual–text residuals have second priority, and a_{base} is the fallback.

Appendix Algorithm 1 gives the complete executable workflow.

4 EXPERIMENTS

We evaluate whether RouteLedger contributes useful routing evidence beyond strong decoding and answer-consensus baselines. All methods share the same samples, cached model outputs, answer normalizer, and scorer; only candidate generation and final selection differ. We therefore focus on paired aggregate evidence rather than per-dataset dominance.

4.1 EXPERIMENTAL SETUP

Model and routing traces. We use MoE-LLaVA-Qwen-1.8B-4e (Lin et al., 2026) and record gate decisions at layers $0, 2, \ldots, 22$; layers $8, 10, 12, 14$ form the default middle-layer capture plan. We generate each candidate answer once and compare selectors offline, so every comparison is paired at the sample level.

A 384-example audit of each official StableLM and Phi2 MoE-LLaVA variant yields weak and heterogeneous effects. We therefore report these results only as a appendix portability check.

Datasets. We evaluate 12 VQA-style datasets spanning charts, documents, OCR-heavy images, open visual QA, and multimodal exams: ChartQA (Masry et al., 2022), DocVQA (Mathew et al., 2021), TextVQA (Singh et al., 2019), VQAv2 (Goyal et al., 2017), GQA (Hudson & Manning, 2019), ScienceQA (Lu et al., 2022), OK/A-OKVQA (Marino et al., 2019; Schwenk et al., 2022), InfographicVQA (Mathew et al., 2022), VizWiz (Gurari et al., 2018), MME (Fu et al., 2026), MMBench (Liu et al., 2024c), and OCRBench (Liu et al., 2024d). Large splits are capped at 5,000 examples and smaller splits are used in full, yielding 45,857 scored examples.

Each method is scored with the same deterministic normalized-answer evaluator and the same sample set. The dataset rows report the fraction of correct sample-level decisions; "All" is the micro-average (the sum of correct decisions divided by 45,857), rather than an unweighted mean over datasets. Table 6 gives the per-dataset sample counts used for this aggregate.

Compared methods. **Original** denotes the sampled answer. **VCD** (Leng et al., 2024), **PAI** and **PAI-CFG** (Liu et al., 2024b), and **OPERA** (Huang et al., 2024) are reproduced in the common framework, and **Beam** is the strongest standard decoder. **Beam+Anchor** is a non-route consensus control that accepts a high-support anchor only for yes/no or single-choice outputs. **RouteLedger** first applies beam-first anchor and route-consensus checks, then invokes the task-gated visual-text verifier only if the guarded selector would retain Beam. The verifier considers Beam and the non-route paper/decoding/anchor pool. Oracle coverage is diagnostic and never used by the deployable selector.

Metrics and statistical tests. We report normalized answer accuracy, paired correct-count deltas, paired bootstrap intervals, exact two-sided sign tests, and a train1024 robustness audit. The paired tests and matched ablations define the scope of our claims; aggregate gains do not establish per-dataset state-of-the-art performance.

Inference accounting. For one uncached example, the fixed candidate recipe produces six non-route LVLM outputs (Original, VCD, PAI, PAI-CFG, Beam, and OPERA), two route-intervention outputs, and three additional forward traces for the text-only, visual-prefix, and multimodal views. Anchor and consensus decisions are deterministic; the verifier is one batched CLIP call over the deduplicated candidate pool. These are pass counts for the offline implementation, while wall-clock latency and memory depend on batching and hardware.

4.2 MAIN EXPERIMENTS

Table 1 summarizes the central result. RouteLedger answers 23,419 of 45,857 examples correctly (51.07%), compared with 23,142 (50.47%) for Beam, 23,178 (50.54%) for Beam+Anchor, and 23,225 (50.65%) for the verifier-free selector. This corresponds to 277 additional correct answers over Beam and 194 over the verifier-free selector. The improvement is aggregate: RouteLedger helps on several OCR-heavy and open-VQA datasets, whereas VCD remains stronger on MME. As shown by Table 3, the verifier accounts for most of the gain, with routing providing a smaller residual contribution.

Dataset	Original	VCD	PAI	PAI-CFG	Beam	OPERA	Beam+Anchor	RouteLedger
ChartQA	9.7	9.2	9.6	9.5	**10.5**	**10.5**	**10.5**	**10.5**
DocVQA	9.6	**10.3**	10.2	9.8	10.2	10.2	**10.3**	**10.3**
TextVQA	46.1	47.5	47.7	47.1	47.7	47.2	47.8	**49.9**
VQAv2	72.1	71.6	72.2	72.4	72.2	72.2	72.3	**72.5**
GQA	66.4	65.8	65.9	66.3	66.9	66.4	66.8	**67.0**
ScienceQA	62.4	60.9	62.1	61.9	**62.6**	62.0	62.5	**62.6**
OK/A-OKVQA	63.3	62.1	63.0	63.1	64.6	64.5	64.6	**64.8**
InfoVQA	9.1	9.1	8.9	8.8	**9.4**	**9.4**	**9.4**	**9.4**
VizWiz	46.8	47.0	48.2	47.7	47.3	46.9	47.3	**49.6**
MME	66.6	**67.4**	67.0	67.2	66.7	66.7	67.1	67.1
MMBench	71.8	71.4	72.0	72.1	71.7	71.8	72.1	**72.2**
OCRBench	27.8	26.5	28.3	28.0	**28.9**	28.6	**28.9**	**28.9**
All	49.9	49.7	50.2	50.1	50.5	50.3	50.5	**51.1**

Table 1: Normalized answer accuracy (%) on 12 VQA-style datasets. Bold marks the best displayed non-oracle result, including ties. Oracle coverage is reported only in the appendix.

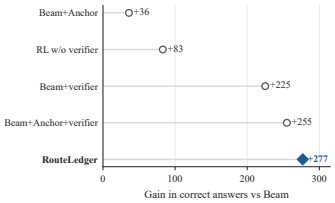


Figure 3: Correct-count gains over Beam on 45,857 examples.

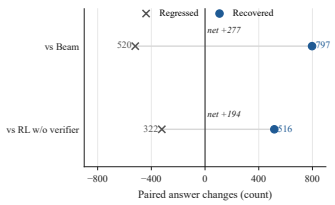


Figure 4: Paired recoveries and regressions against Beam and verifier-free RL.

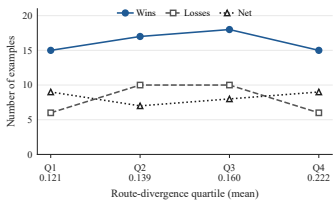


Figure 5: Verifier-free RL versus Beam+Anchor by route-divergence quartile.

Reference	Δ correct	Wins/Losses
Beam	+277	797/520
95% CI: [+0.0045,+0.0076]; sign $p = 2.29\times10^{-14}$		
RouteLedger w/o verifier	+194	516/322
95% CI: [+0.0030,+0.0055]; sign $p = 2.14\times10^{-11}$		

Table 2: Paired evidence on 45,857 examples. CI is the 95% paired bootstrap interval for the accuracy delta.

The paired comparisons in Table 2 support the aggregate improvement, with 797 wins versus 520 losses against Beam and 516 wins versus 322 losses against the verifier-free selector. On train1024, however, fusion adds 23 correct answers over the verifier-free selector but only 11 over Beam, and the latter difference is not significant. The complete evaluation therefore supports the final selector, while independent robustness evidence over Beam remains limited.

The appendix paper-pool oracle reaches 25,708 correct answers, leaving 2,289 recoverable cases beyond the final selector. We exclude this oracle from the main table because it is diagnostic rather than deployable. The larger cached-pool oracle in the appendix is likewise not directly comparable.

Figures 3 and 4 summarize the component and paired results; the matched-verifier comparison keeps the routing contribution separate from the larger verifier gain. The quartile plot in Figure 5 shows that raw route divergence is not a reliable effect-strength proxy.

4.3 Ablation Studies

Table 3 separates routing from verification. Applying the same verifier to Beam and Beam+Anchor already yields 23,367 and 23,397 correct answers, respectively. RouteLedger reaches 23,419, a significant gain of 52 over Beam+verifier ($p = 0.0188$) but a non-significant gain of 22 over Beam+Anchor+verifier ($p = 0.0903$). The verifier therefore explains most of the 6.04 additional

(a) Aggregate components			
Method	Correct	Acc. (%)	Δ Beam
Beam	23142	50.47	–
Beam+Anchor	23178	50.54	+36
RL w/o VT	23225	50.65	+83
Beam+VT	23367	50.96	+225
Beam+Anchor+VT	23397	51.02	+255
RL+VT	**23419**	**51.07**	**+277**
(b) RL+VT matched comparisons			
Reference	Δ correct	95% CI	Sign p
Beam+VT	+52	[.0002,.0020]	.0188
Beam+Anchor+VT	+22	[-.00004,.00102]	.0903

Table 3: Component evidence on 45,857 examples. RL is RouteLedger and VT the same frozen verifier. Panel (b) gives paired 95% bootstrap intervals.

correct answers per 1,000 examples over Beam; routing contributes only 0.48 additional answers per 1,000 over the strongest matched control. RouteLedger should thus be interpreted as a routing-aware conservative selector with selective visual-text arbitration, not as a standalone gate corrector or global reranker.

Panel (b) makes the matched-control boundary explicit: all compared systems use the same task-family gate, candidate pool, Beam-current guard, and verifier margin. The $+52$ residual over Beam+VT has a paired interval above zero, whereas the $+22$ residual over Beam+Anchor+VT has an interval that includes zero. The former supports a narrow routing-aware contribution; the stronger anchor-matched comparison does not support an independent routing gain at conventional significance levels.

Calibration and mechanism audits. We freeze the single verifier margin at $\tau = 1.0$ on train1024 before the paper-evaluation run. The complete threshold grid and mechanism audits are in Appendix A.5 and Appendix A.8. Across the grid, development accuracy varies by only 11 correct answers while the verifier selection rate falls from 1,293 to 373; the train1024 comparison with Beam is not significant.

The unit-shuffle control changes route-branch accuracy by only 0.08 points, while the final verifier recovers 516 and regresses 322 predictions (Appendix A.8). These diagnostics support a dominant visual-verifier contribution with a smaller routing residual; the appendix candidate-dependency audit likewise finds no strong dependence on route candidates.

Figure 5 further constrains the mechanism claim. Across the four divergence quartiles, verifier-free RouteLedger has similar small paired gains over Beam+Anchor: $+9$, $+7$, $+8$, and $+9$ correct answers. Thus, higher route divergence does not predict a larger benefit on the traced subset. This boundary result motivates certificate-based residual use rather than direct route transfer.

Taken together, the ablations support a narrow conclusion: the pre-specified verifier supplies most of the aggregate improvement, while RouteLedger adds a small residual over the same verifier applied to Beam. Its residual over Beam+Anchor+verifier is not statistically significant. Unit alignment and route candidates remain useful audit and construction signals, but neither is established as the sole necessary mechanism.

5 Conclusion

We proposed RouteLedger, which combines paired modality routing traces, controlled route proposals, beam-preserving residual selection, explicit fidelity guards, and a task-family-gated visual–text verifier. Across 12 VQA-style datasets, the procedure improves strong decoding and consensus baselines without unrestricted route intervention. Matched controls show that the verifier supplies most of the aggregate gain, while aligned routing contributes a smaller residual and does not yield a significant advantage over the anchor-matched verifier. Calibration supports a fixed operating point; portability beyond the primary Qwen backbone and production cost benefits remain open.

AI USE STATEMENT

Generative AI tools were used solely to assist with and polish the writing of this manuscript. The authors reviewed and revised all AI-assisted content and take full responsibility for the final manuscript.

REPRODUCIBILITY STATEMENT

The paper specifies the MoE routing ledger, candidate construction, beam-first selector, route-support and fidelity guards, task-family gates, CLIP-L/336 visual–text verifier, datasets, thresholds, and evaluation metrics. The appendix records the frozen selector specification, candidate pools, dataset sizes, verifier audits, robustness checks, and computational accounting. This repository contains the manuscript source, figures, tables, and references; it does not include the complete inference runner, model checkpoints, cached routing traces, or all external model assets. Full replication requires those artifacts, the named model versions, the public or licensed datasets, and the decoding settings reported in the paper.

REFERENCES

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A versatile vision–language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Lalit Bandarkar, Chao Yang, Mahan Fayyaz, Junjie Hu, and Nanyun Peng. Multilingual routing in mixture-of-experts. In *The Fourteenth International Conference on Learning Representations*, 2026.

Junyu Cheng, Zhibiao Liang, Yidong Chen, and Shuangyin Li. RFI: Rectified flow intervention for mitigating object hallucination in large vision–language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 3255–3263, 2026.

Ailin Deng, Zhirui Chen, and Bryan Hooi. Seeing is believing: Mitigating hallucination in large vision–language models via CLIP-guided decoding. *arXiv preprint arXiv:2402.15300*, 2024.

Mahan Fayyaz, Amir Hossein Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Ryan A. Rossi, Trung Bui, Hinrich Schuetze, and Nanyun Peng. Steering moe llms via expert (de)activation. In *The Fourteenth International Conference on Learning Representations*, 2026.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Advances in Neural Information Processing Systems*, 38, 2026.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3608–3617, 2018.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13418–13427, 2024.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13872–13882, 2024.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023.

Bin Lin, Zhenyu Tang, Yang Ye, Jinfa Huang, Junwu Zhang, Yatian Pang, Peng Jin, Munan Ning, Jiebo Luo, and Li Yuan. Moe-llava: Mixture of experts for large vision-language models. *IEEE Transactions on Multimedia*, 2026.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision–language models. *arXiv preprint arXiv:2402.00253*, 2024a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Sheng Liu, Haotian Ye, and James Zou. Reducing hallucinations in large vision-language models via latent space steering. In *ICLR*. OpenReview.net, 2025.

Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *European Conference on Computer Vision*, pp. 125–140. Springer, 2024b.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024c.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024d.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in neural information processing systems*, 35:2507–2521, 2022.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pp. 3195–3204, 2019.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq R. Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *ACL (Findings)*, volume ACL 2022 of *Findings of ACL*, pp. 2263–2279. Association for Computational Linguistics, 2022.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2200–2209, 2021.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1697–1706, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pp. 146–162. Springer, 2022.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8317–8326, 2019.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

Le Yang, Ziwei Zheng, Boxu Chen, Zhengyu Zhao, Chenhao Lin, and Chao Shen. Nullu: Mitigating object hallucinations in large vision–language models via halluspace projection. *arXiv preprint arXiv:2412.13817*, 2024.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: hallucination correction for multimodal large language models. *Sci. China Inf. Sci.*, 67(12), 2024.

A Appendix

A.1 Final Selector Specification

The main paper gives the high-level inference procedure. This appendix records the exact deployed selector in audit-table form. The selector is beam-first, then anchor-safe, then gate-safe, and finally visual-text-safe for pre-specified task families. It does not use gold labels or the oracle upper bound.

Stage	Acceptance condition	Output if accepted
Beam fallback	Always available	Beam answer
Anchor override	Anchor support ≥ 3 and answer is yes/no or single-choice	Anchor answer
Gate candidate	Gate ensemble selected a non-original answer	Continue checks
External support	At least 3 non-gate candidates match the gate answer	Continue checks
Attribute guard	No protected visual-attribute rewrite is detected	Gate answer
Visual-text task gate	Route selector kept Beam and dataset is in $\mathcal{F}_{\mathrm{vt}}$	Continue checks
Visual-text candidate pool	Paper/decoding/anchor candidates plus Beam; route/gate candidates excluded	Continue checks
Visual-text margin	Top verifier candidate beats Beam by $\tau = 1.0$	Verifier answer
Otherwise	Any check fails	Current selector answer

Table 4: Frozen final selector used for the paper-evaluation results.

The non-gate support pool contains the original sampled answer, VCD, PAI, PAI-CFG, Beam, OPERA, anchor-family vote, route boost, and route vote. This pool is fixed before evaluation. The protected visual-attribute groups are colors, person terms, spatial terms, yes/no polarity, and language names. The visual-text verifier is enabled only for TextVQA, DocVQA, VQAv2, and VizWiz. Its final candidate pool is restricted to the non-route paper/decoding/anchor candidates plus Beam, following the visual-text candidate-dependency audit below.

A.2 Visual-Text Verifier Details

Item	Frozen setting
Checkpoint	`openai/clip-vit-large-patch14-336` via local `clip-vit-large-patch14-336` snapshot
Loader	HuggingFace `CLIPModel` and `AutoProcessor`, `local_files_only=True`
Image input	RGB image from the prepared paper-evaluation sample cache
Text input	`Question: {question} Answer: {candidate text}`
Text length	First 250 characters before processor tokenization; processor truncation enabled
Candidate pool	Paper/decoding/anchor candidates plus Beam; RouteLedger route/gate candidates excluded
Candidate preprocessing	Deduplicate candidates by normalized answer string
Score	CLIP `logits_per_image`; margin is top-candidate logit minus Beam-candidate logit
Margin threshold	Single global threshold $\tau = 1.0$ selected on train1024 from grid $\{0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0\}$
Additional guard	Verifier can fire only when the current RouteLedger answer is still Beam
Enabled task families	TextVQA, DocVQA, VQAv2, VizWiz

Table 5: Implementation details for the visual-text verifier.

The task-family gate is not selected by taking the best per-dataset final accuracy. It is a metadata-level compatibility rule: enable CLIP-style question-answer image scoring for open-ended natural-image, scene-text, or document VQA where the candidate answer is expected to be visually grounded; disable it for chart/numeric reasoning, synthetic relation QA, external knowledge QA, and benchmark-suite tasks. This rule includes DocVQA even though its final net gain is zero, and it excludes visually related but mismatched families such as GQA and OCRBench. The threshold is selected on train1024 before the paper-evaluation run, so no per-dataset or test-set accuracy tuning is used.

Algorithm 1 Overall workflow for RouteLedger.

Require: Image x, question q, dataset family d, frozen model M_θ, non-route generators, route interventions, enabled families $\mathcal{F}_{\mathrm{vt}}$, verifier margin τ, answer normalizer $n(\cdot)$, and fidelity test $R(\cdot,\cdot)$.

Ensure: Final answer $\widehat{a}$.

Generate the non-route candidate set $\mathcal{C}_{\mathrm{paper}}$, including a_{beam} and a_{anchor}.
Execute the text-only, visual-prefix, and multimodal views and construct the paired routing ledger $\mathcal{R}(x,q)$.
Generate $\mathcal{C}_{\mathrm{route}}$ from the selected routing interventions.
Set $a_{\mathrm{base}} \leftarrow a_{\mathrm{anchor}}$ if $\chi_{\mathrm{anchor}} = 1$; otherwise set $a_{\mathrm{base}} \leftarrow a_{\mathrm{beam}}$.
Obtain route consensus $\widehat{a}_{\mathrm{gate}}$, then compute non-route support, route support, gate agreement, and fidelity risk.
Construct $\mathcal{A}_{\mathrm{rt}} \leftarrow \{a : \chi_{\mathrm{rt}}(a) = 1\}$.
if $\mathcal{A}_{\mathrm{rt}} \neq \varnothing$ **then**
    return its candidate with the strongest gate agreement and total support.
end if
if $d \in \mathcal{F}_{\mathrm{vt}}$ **then**
    Deduplicate the non-route verifier pool: $\mathcal{C}_{\mathrm{vt}} \leftarrow \mathcal{C}_{\mathrm{paper}}$.
    Score every $a \in \mathcal{C}_{\mathrm{vt}}$ with the frozen visual–text verifier $S(x,q,a)$.
    Construct $\mathcal{A}_{\mathrm{vt}} \leftarrow \{a : n(a) \neq n(a_{\mathrm{beam}}),\ S(x,q,a) - S(x,q,a_{\mathrm{beam}}) \geq \tau\}$.
    if $\mathcal{A}_{\mathrm{vt}} \neq \varnothing$ **then**
        return $\arg\max_{a \in \mathcal{A}_{\mathrm{vt}}} S(x,q,a)$.
    end if
end if
return a_{base}.

A.3 Paper-Evaluation Dataset Sizes

Dataset	N	RouteLedger correct	Accuracy (%)
ChartQA	1920	202	10.5
DocVQA	5000	516	10.3
TextVQA	5000	2497	49.9
VQAv2	5000	3627	72.5
GQA	5000	3349	67.0
ScienceQA	4114	2575	62.6
OK/A-OKVQA	5000	3239	64.8
InfoVQA	2801	264	9.4
VizWiz	4319	2143	49.6
MME	2374	1592	67.1
MMBench	4329	3126	72.2
OCRBench	1000	289	28.9
All	45857	23419	51.1

Table 6: Paper-evaluation split sizes and final RouteLedger counts. Large evaluation splits are capped at 5,000 examples; smaller splits are used in full.

A.4 Robustness Audit

The train1024 audit is not the headline benchmark; it is a smaller robustness check over cached paired outputs. It supports the same component-level interpretation as the paper-evaluation set: the task-family-gated verifier is positive over the verifier-free RouteLedger selector. It does not establish strong independent robustness over Beam, because the +11 Beam comparison is not significant on this subset.

Reference	Split	Δ correct	CI / Result	Sign p
Beam	paper eval	+277	[+0.0045,+0.0076]	2.29×10^{-14}
RouteLedger w/o verifier	paper eval	+194	[+0.0030,+0.0055]	2.14×10^{-11}
Beam	train1024	+11	not significant	–
RouteLedger w/o verifier	train1024	+23	positive	–

Table 7: Paired robustness audit on cached outputs. Paper-evaluation rows report 95% paired bootstrap intervals; train1024 rows summarize the development audit used for threshold selection.

τ	Correct	Δ vs Beam	Δ vs w/o verifier	Selected VT	Wins/Losses vs Beam	Sign p vs Beam
0.5	3901	+0	+12	1293	129/129	1.000
1.0	**3912**	**+11**	**+23**	1077	124/113	0.516
1.5	3908	+7	+19	892	113/106	0.685
2.0	3911	+10	+22	737	107/97	0.529
2.5	3910	+9	+21	614	94/85	0.550
3.0	3906	+5	+17	518	85/80	0.756
3.5	3910	+9	+21	438	84/75	0.526
4.0	3904	+3	+15	373	76/73	0.870

Table 8: Train1024 development grid for the final CLIP-L/336 paper-only verifier. The chosen threshold $\tau = 1.0$ maximizes development correct count in this grid and is frozen before the paper-evaluation run.

A.5 Development Threshold Grid

Table 8 is a development-only calibration audit, not a paper-evaluation sweep. It explains why the reported L/336 verifier uses $\tau = 1.0$, which maximizes correct count among the evaluated development thresholds. The development evidence remains weak against Beam, so the paper treats the threshold choice as calibration rather than as an independent robustness claim.

A.6 Per-Dataset Deltas

Dataset	vs Beam	vs RouteLedger w/o verifier
ChartQA	+0	+0
DocVQA	+4	+0
TextVQA	+113	+99
VQAv2	+17	+1
GQA	+5	+0
ScienceQA	+0	+0
OK/A-OKVQA	+10	+0
InfoVQA	+0	+0
VizWiz	+99	+94
MME	+8	+0
MMBench	+21	+0
OCRBench	+0	+0
All	+277	+194

Table 9: Per-dataset paired deltas for the final selector. Deltas are correct counts relative to the reference on the same samples.

Table 9 is included to make the claim boundary explicit. RouteLedger is positive in aggregate, with the largest verifier gains concentrated in TextVQA and VizWiz. The table also keeps the limitation visible: several datasets receive no additional benefit from the visual-text verifier, and MME remains stronger under VCD in the main table.

Quantity	Count
Guarded RouteLedger w/o verifier correct	23225
Full RouteLedger correct	23419
Selected visual-text verifier answer	5843
Verifier recovered w/o-verifier mistakes	516
Verifier regressed w/o-verifier correct answers	322
Verifier net delta over w/o verifier	+194

Table 10: Visual-text verifier audit for the final selector. Counts are over all 45,857 paper-evaluation examples.

A.7 Verifier Selection Audit

The verifier branch is selective but no longer as sparse as the pre-verifier override. It improves the frozen RouteLedger selector when restricted to the visual task families where it is reliable and to the non-route paper-candidate pool where the verifier is most stable.

A.8 Mechanism and Selection Audits

(a) Unit alignment

Variant	Acc. (%)	Selection rate (%)	Δ vs aligned
Aligned units	50.88	17.5	–
Shuffled units	50.80	17.1	-0.08 pp

(b) Final verifier selection

Quantity	Count
Guarded RouteLedger w/o verifier correct	23225
Selected visual-text verifier answer	5843
Verifier recovered / regressed	516 / 322
Verifier net gain over w/o verifier	+194
Full RouteLedger correct	23419

Table 11: Mechanism and selection audits. Panel (a) shuffles aligned routing units while preserving the recipe, selector, and fallback. Panel (b) audits the final verifier against the guarded selector.

The unit-alignment control changes the route-branch accuracy by only 0.08 points and the selection rate by 0.4 points. The final verifier is applied to 5,843 examples, recovering 516 and regressing 322 relative to the guarded selector.

A.9 Qualitative Selector Audit

Figure 6 illustrates the residual hierarchy without expanding the main-paper claim. Routing evidence can recover a supported candidate when Beam and the matched Beam+Anchor control fail; visual evidence can recover a different paper-pool candidate; and the absence of a certificate can preserve a correct Beam answer. The included failure makes explicit that the task-family gate and fixed margin reduce, but do not eliminate, visually plausible regressions.

A.10 Non-Route Verifier Ablation

Tables 12 and 13 show that the paper-only visual verifier is the dominant source of the final aggregate improvement. A non-RouteLedger Beam+Anchor+verifier baseline reaches 23,397/45,857, close to the final 23,419/45,857 result. RouteLedger adds a small +52 gain over Beam+verifier, but only +22 over Beam+Anchor+verifier; the latter comparison is not significant at conventional thresholds. The claim must therefore be framed as routing-aware selection plus task-gated visual verification rather than as route candidates being the main source of gain.

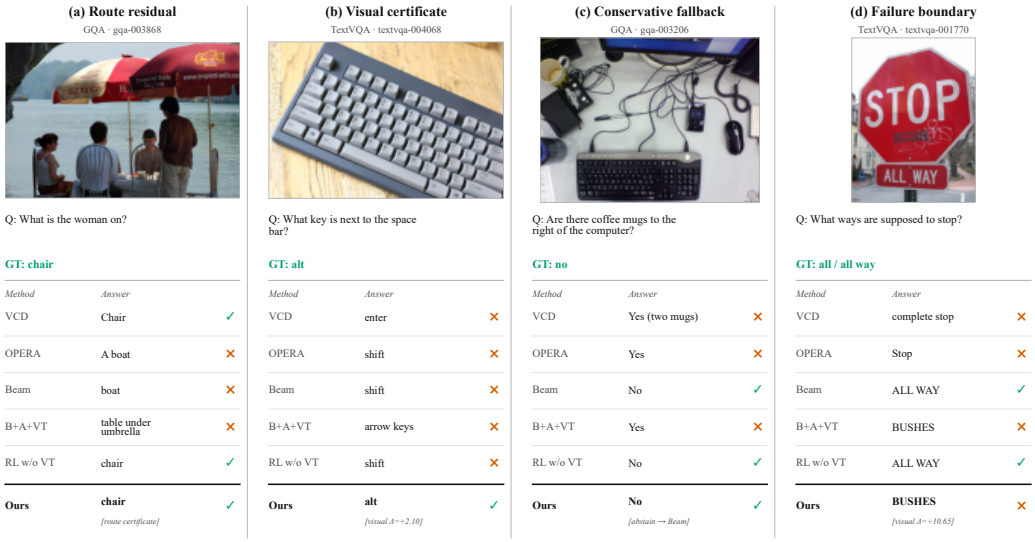


Figure 6: Qualitative audit on the formal `paper_eval5k` cache. Answer-bearing spans are displayed for readability, while correctness is computed from the complete cached outputs. From left to right: a route-certified residual correction, a SafeSwitch-Margin visual-certificate correction, a conservative abstention to Beam, and a failure boundary. The final example shows that a large CLIP-L/336 margin can still favor salient but question-irrelevant text ("BUSHES") over the correct Beam answer ("ALL WAY"). One visually legible example is selected per outcome class; the panels do not represent class frequencies.

Method	Correct	Acc. (%)	Δ vs Beam	Key comparison
Beam	23142	50.47	–	–
Beam+Anchor	23178	50.54	+36	–
RouteLedger w/o verifier	23225	50.65	+83	–
Beam + same verifier	23367	50.96	+225	+225 vs Beam
Beam+Anchor + same verifier	23397	51.02	+255	+219 vs Beam+Anchor
RouteLedger + same verifier	23419	51.07	+277	+52 vs Beam+verifier

Table 12: Applying the same paper-only task-gated verifier to non-RouteLedger bases. All verifier variants use the same candidate pool, task-family gate, Beam-current guard, and margin threshold.

Comparison	Δ correct	CI	Sign p
Beam+verifier vs Beam	+225	[+0.0036,+0.0062]	6.75×10^{-14}
Beam+Anchor+verifier vs Beam+Anchor	+219	[+0.0035,+0.0060]	2.58×10^{-13}
RouteLedger+verifier vs Beam+verifier	+52	[+0.0002,+0.0020]	0.0188
RouteLedger+verifier vs Beam+Anchor+verifier	+22	[-0.00004,+0.00102]	0.0903

Table 13: Paired comparisons for Table 12. CI denotes 95% paired bootstrap intervals for the accuracy delta.

Variant	Base	Correct	Acc. (%)	Δ vs Beam	Δ vs RL w/o VT	Sign p vs RL w/o VT
No verifier	Beam	22767	53.90	–	-80	2.68×10^{-4}
No verifier	RouteLedger	22847	54.09	+80	–	–
Full pool verifier	Beam	22851	54.10	+84	+4	0.948
Paper-only verifier	Beam	22862	54.13	+95	+15	0.757
Route-only verifier	Beam	22859	54.12	+92	+12	0.743
Shuffled-route verifier	Beam	22060	52.23	-707	-787	1.58×10^{-49}
Full pool verifier	RouteLedger	22905	54.23	+138	+58	0.158
Paper-only verifier	RouteLedger	22919	54.26	+152	+72	0.0739
Route-only verifier	RouteLedger	22920	54.26	+153	+73	0.00635
Shuffled-route verifier	RouteLedger	22188	52.53	-579	-659	1.15×10^{-42}

Table 14: Final L/336 candidate-dependency audit for the visual-text verifier at the frozen threshold $\tau = 1.0$. The audit covers 42,239 examples for which prepared images are available; 3,617 DocVQA rows are missing from the prepared image cache and one MME image is truncated.

A.11 Visual-Text Candidate Dependency Audit

Table 14 is the final candidate-dependency audit under the L/336 verifier used by SafeSwitch-Margin. It does not support a strong claim that the verifier gain requires route candidates: paper-only and route-only verifier pools are effectively tied on correct count, and the full route-plus-paper pool is lower than either restricted pool. At the same time, route-only candidates do contain signal, while shuffled-route candidates are strongly harmful. The final selector therefore uses the pre-specified paper-only verifier pool and treats route evidence as part of the guarded selector rather than as a required verifier input.

Variant	Correct	Acc. (%)	Δ vs Beam	Δ vs w/o verifier	Selected VT	Sign p vs w/o verifier
Beam	8550	44.26	–	-39	0	5.31×10^{-5}
RouteLedger w/o verifier	8589	44.46	+39	–	0	1
Full candidate pool	8703	45.05	+153	+114	3842	5.07×10^{-7}
Paper-only candidates	8707	45.07	+157	+118	3703	1.75×10^{-7}
Route-only candidates	8615	44.59	+65	+26	1433	0.0768
Paper + shuffled-route candidates	8519	44.10	-31	-70	4636	0.00906

Table 15: Earlier B/32 candidate-dependency diagnostic for the visual-text verifier on the four enabled task families: TextVQA, DocVQA, VQAv2, and VizWiz. This audit uses the earlier CLIP-B/32 verifier configuration to compare candidate-pool choices while keeping the text template, margin rule, and beam-first guard fixed. Deltas are paired correct-count changes on 19,319 examples.

Dataset	RouteLedger w/o verifier	Full	Paper-only	Route-only	Paper + shuffled-route
DocVQA	516	516	516	514	439
TextVQA	2398	2448	2451	2398	2409
VQAv2	3626	3636	3635	3628	3584
VizWiz	2049	2103	2105	2075	2087

Table 16: Dataset breakdown for Table 15.

Table 15 is the earlier B/32 diagnostic that motivated freezing the paper-only verifier pool before the final L/336 run. It reaches the same conservative boundary as the final audit: route candidates contain signal, shuffled route candidates are unsafe, and the verifier branch should be interpreted as visual-text arbitration over a guarded candidate portfolio rather than as evidence that route candidates are necessary for the verifier gain.

A.12 Route-Divergence Stratification

Table 17 is a mechanism-boundary audit. RouteLedger is positive in every quartile, but the highest-divergence quartile does not dominate the lower-divergence quartiles. This supports the paper's cautious claim: route divergence is useful for auditing and candidate construction, but it is not yet a strong standalone risk score that explains the final gain.

Stratum	N	Mean div.	Beam+Anchor	RouteLedger no VT	Δ	Wins/Losses	Changed	Precision
Q1 low	3001	0.1207	1878	1887	+9	15/6	531	0.714
Q2	3000	0.1393	1696	1703	+7	17/10	481	0.630
Q3	3001	0.1603	1381	1389	+8	18/10	415	0.643
Q4 high	3000	0.2219	1051	1060	+9	15/6	316	0.714

Table 17: Route-divergence stratification on 12,002 traced examples before the final visual-text verifier. The table tests whether route disagreement itself cleanly predicts useful RouteLedger switches.

A.13 Oracle Upper Bound and Remaining Headroom

Oracle pool	Correct	Gap to RouteLedger	Candidate-pool definition
Paper-pool oracle	25708	2289	Restricted candidate pool used for the paper oracle row
Full cached-candidate oracle	26334	2915	Broader diagnostic cache over all paper-eval portfolio candidates

Table 18: Oracle upper bounds under two different candidate pools. Both oracles are diagnostic only and use gold correctness to test whether at least one cached candidate is correct.

Table 18 separates two numbers that can otherwise look inconsistent. The paper-pool oracle, 25,708/45,857, is the upper bound paired with the main paper table. The larger 26,334/45,857 full cached-candidate oracle comes from a separate headroom audit over the broader cached portfolio, including additional external and route-family candidates. It is useful for diagnosing unused candidate-selection headroom, but it is not the oracle used for the main paper-pool gap. Neither oracle is an achievable method result.

A.14 Cross-Backbone Audit

Backbone	N	Original	Copy-route	Δ	Heuristic-route	Δ
MoE-LLaVA-StableLM-1.6B-4e	384	193	192	-1	191	-2
MoE-LLaVA-Phi2-2.7B-4e	384	196	198	+2	196	0
MoE-LLaVA-OpenChat-7B-4e	3072	1492	1484	-8	1474	-18

Table 19: Cross-backbone audit on official MoE-LLaVA variants. StableLM and Phi2 use 32 prepared examples from each paper dataset; OpenChat-7B uses 256 examples per dataset. Deltas are correct-count changes relative to the same backbone's original output.

Table 19 checks whether the implemented RouteLedger pipeline transfers beyond the Qwen backbone used in the main experiments. The audit uses the same prepared 12-dataset sample cache, the same local CLIP image tower, model-specific conversation templates, and middle-layer MoE gate capture plans for the StableLM, Phi2, and OpenChat official MoE-LLaVA checkpoints. The result is best interpreted as an engineering portability check: route extraction, generation, and intervention execute successfully on additional backbones, but the accuracy effect is small and inconsistent. StableLM is slightly negative under both route controls, Phi2 shows only a small positive copy-route delta, and the larger OpenChat-7B audit is negative. We therefore keep the cross-backbone result out of the main claim and treat broad backbone generality as future work.

A.15 Computational Accounting

Table 20 reports pass counts rather than a hardware-specific latency claim. The candidate outputs, routing traces, and verifier scores can be cached for the paired comparisons used here; a deployment study would still need wall-clock and memory measurements on a fixed serving stack.

A.16 Claim Boundary

The paper supports three claims. First, RouteLedger is the strongest aggregate comparable method in the 12-dataset table. Second, pre-specified task-family-gated visual-text arbitration is the dominant source of the final gain and must be treated as a core component rather than an auxiliary add-on. Third, RouteLedger's routing-aware selector adds a small significant residual gain over the same

Component	Extra work per example	Main resource	Cache-able
Non-route candidates	Six decoding outputs: Original, VCD, PAI, PAI-CFG, Beam, OPERA	LVLM decoding	Yes
Route traces	Three text/visual/full forward traces at the captured layers	LVLM forward traces	Yes
Route candidates	Two fixed route-intervention decoding outputs	LVLM decoding	Yes
CLIP-L/336 verifier	One batched image-text call over deduplicated candidates	CLIP forward pass	Yes
Final selector	Deterministic table lookup and normalization	CPU only	Yes

Table 20: Pass accounting for one uncached example in the evaluated recipe. The paper reports offline selection over cached candidate outputs; this table states which components add model passes and whether their outputs can be cached. Wall-clock latency depends on batching and hardware.

verifier applied to Beam, while the residual over Beam+Anchor with the same verifier is weak. The shuffled-unit negative control, the verifier candidate-dependency audit, and the non-route verifier ablation all limit the mechanism claim: aligned routing units and route candidates are consistent with the method, but the current evidence does not prove that either is the dominant source of the final gain.

The paper does not claim that standalone gate correction beats all baselines, that broad text-to-vision route transfer is always beneficial, that global visual-text reranking is reliable, that aligned unit evidence is necessary by itself, that the final verifier gain requires route candidates, or that RouteLedger is state of the art on every individual dataset. These exclusions are part of the method design: the final result depends on a conservative beam-first fallback, recipe-matched candidate construction, preservation fallback, fixed route-support guards, and task-family-gated paper-only visual-text fidelity arbitration, with routing evidence contributing a residual selector gain rather than the majority of the improvement.