ROUTE EVIDENCE BEFORE REVISING: CHANNEL-SELECTIVE VERIFICATION FOR LOW-CALL MULTIMODAL REPAIR

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-hoc multimodal correction is a one-edit decision: boxes, counts, geometry, and temporal tracks may support different answers, so a verifier must repair errors without overwriting correct drafts. Selective Channel Verification (SCV) manages this trade-off with a utility–risk router that invokes a grounding, counting, geometry, or temporal critic on only its aligned cache record. On identical GQA draft/cache states pooled over five seed-paired runs, SCV repairs 3,411 initially wrong drafts and damages 1,578 correct ones, compared with 3,204 repairs and 1,685 damaged drafts for a union-trained, capacity-matched verifier. Spatial accuracy is 58.2±0.3 versus 56.6±0.3, a 1.6-point difference (95% CI: 1.1, 2.1). A second intervention sends all four records to the unchanged channel-trained critics; spatial accuracy falls from 58.4 to 55.4 and over-correction rises from 3.8% to 4.4%. NLVR2 shows the same spatial and draft-preservation pattern, and replacing the backbone with Qwen2.5-VL-7B-Instruct yields 62.0±0.3 versus 60.8±0.3 spatial accuracy. An outcome-blind proposal-disagreement score places the largest difference—2.5±0.6 points—among examples with competing channel-local proposals. At approximately 1.4 verifier calls per question, SCV improves the repair–damage operating point and establishes evidence access as an architectural choice in multimodal verification.

1 INTRODUCTION

A spatial question can leave a post-hoc verifier with several individually meaningful but action-incompatible views of the same scene: grounding boxes nominate entities, count summaries aggregate detections, geometry records encode relations, and temporal tracks describe motion. The verifier, however, may have only one opportunity to replace the draft answer. Concatenating every record therefore changes more than how much evidence is visible; it changes which cues compete inside the keep-or-revise decision. On GQA, this distinction is consequential: giving a called channel-trained critic all four records instead of its aligned record lowers spatial accuracy by 3.0 points and raises over-correction by 0.6 points at the same approximately 1.4-call operating point.

Selective Channel Verification (SCV) makes evidence choice explicit. A lightweight router predicts each critic's correction utility and collateral-edit risk, orders four channel-specialized critics, and stops when no intervention is expected to help. Each call receives the question, draft, confidence, and exactly one aligned record; a common gate accepts at most one valid revision. This design uses specialization not only to choose computation, as in mixture-of-experts routing, but also to control the evidence interface at the moment an answer can be overwritten.

The central comparison evaluates complete repair behavior rather than accuracy alone. Over five paired GQA seeds, SCV exceeds a union-trained, capacity-matched verifier by 1.6 spatial points (95% CI: 1.1, 2.1), repairs 207 more initially wrong drafts, and damages 107 fewer initially correct drafts. The same spatial ordering appears on NLVR2 (+1.1 points) and after replacing the frozen backbone with Qwen2.5-VL-7B-Instruct (+1.2 points). Moreover, the SCV advantage rises from 0.6±0.4 points among low-disagreement proposals to 2.5±0.6 points among high-disagreement proposals, linking the endpoint difference to the examples where evidence selection matters most.

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.

Two all-channel variants assign these results precise experimental roles. Union-restored sends the canonical four-channel serialization to the existing channel-trained critics without changing their weights or router order; it measures the response of the deployed selective scaffold to input expansion. Competitive-Union trains the same 8.2M post-hoc capacity directly on that serialization; it measures whether interface-specific adaptation closes the observed gap inside the common four-slot architecture. Both start from identical frozen drafts and caches, use the same output contract and one-revision gate, and realize nearly identical calls. The first comparison isolates interface sensitivity in a fixed model, while the second supplies the practical learned endpoint against which SCV is evaluated.

This after-draft access decision is usually coupled to other changes in the correction procedure. Iterative refinement changes the revision sequence (Madaan et al., 2023); self-consistency changes sampling and aggregation (Wang et al., 2023); DeGF regenerates visual evidence during decoding (Zhang et al., 2025); Sherlock learns trajectory-level correction (Ding & Zhang, 2025); REVERSE backtracks during generation (Wu et al., 2025); and visually grounded reflection reacquires image regions across rounds (Zhang et al., 2026). These systems establish effective end-to-end procedures, whereas SCV targets the specific architectural decision that remains after the draft, structured evidence, call controller, output space, and revision rule have already been fixed.

SCV's technical contribution is the joint control of critic selection and evidence exposure: utility–risk routing decides which specialist may act, while the aligned record constrains what can influence that action. Evaluating this design against both fixed-critic input expansion and union-specific learning separates interface sensitivity from adaptation without changing the draft/cache state or revision contract. GQA and NLVR2 seed-paired experiments, a backbone transfer, taxonomy controls, and proposal-disagreement strata then connect the accuracy advantage to its operational source—additional repairs with fewer collateral edits, concentrated where channel-local proposals compete. Evidence access is therefore an architectural variable to report alongside accuracy and verifier calls.

2 Related Work and Positioning

Multimodal correction. LURE reconstructs descriptions from hallucination-related factors (Zhou et al., 2024), Woodpecker stages image checks before correction (Yin et al., 2024), and ReCoVERR asks follow-up visual questions while filtering evidence for reliability and relevance (Srinivasan et al., 2024). Self-refinement, retrospective resampling, and visually grounded reflection instead alter the generation trajectory or reacquire visual evidence. Debate and judge systems vary interaction depth (Du et al., 2024; Chan et al., 2024). SCV intervenes at a different boundary: a frozen draft and structured cache already exist, and the remaining decision is which cached record a bounded post-hoc critic may use before the single revision gate.

Evidence selection and modular computation. Long-context studies show that irrelevant images and plausible semantic distractions can enter VLM reasoning (Sharma et al., 2024; Sun et al., 2026); compositional and hallucination diagnostics expose failures to bind answers to the relevant visual content (Thrush et al., 2022; Yuksekgonul et al., 2023; Li et al., 2023; Kamath et al., 2023). These works perturb raw multimodal context or diagnose the base model. Mixture-of-experts and modular reasoning systems route latent computation during generation (Shazeer et al., 2017; Fedus et al., 2022; Andreas et al., 2016; Gupta & Kembhavi, 2023). SCV instead routes an observable, typed evidence interface after generation and measures the downstream correction and collateral-edit consequences. This placement permits the selective and all-channel systems to reuse the same draft, normalized cache records, call controller, output contract, and gate.

Appendix Table 8 makes the boundary operational. GQA and NLVR2 carry the paired access tests, while RefCOCOg and NExT-QA provide grounding and temporal operating-point context (Hudson & Manning, 2019; Suhr et al., 2019; Mao et al., 2016; Xiao et al., 2021).

3 Problem Formulation

Let a frozen VLM f_θ map query q and image or video I to a draft answer $\hat{y}$, confidence c, and evidence trace e. A deterministic extractor combines e with detector outputs into a shared cache

$$\mathcal{E} = \{\mathcal{E}_g, \mathcal{E}_c, \mathcal{E}_r, \mathcal{E}_t\}, \tag{1}$$

containing grounding candidates, count summaries, geometry tuples, and temporal tracklets. Every compared verifier starts from the same $(q, \hat{y}, c, \mathcal{E})$ state, uses at most $B = 4$ critic invocations, emits KEEP or REVISE, and passes a common deterministic gate that accepts at most one correction.

The extractor first maps normalized object mentions in q and e to GroundingDINO detections, producing labeled confidence–box rows for $\mathcal{E}_g$. It aggregates detections with the same normalized label for $\mathcal{E}_c$, derives pairwise center offsets, overlap, and discrete relations for $\mathcal{E}_r$, and links label-matched detections across the shared sampled-frame schedule to form ordered tracklets for $\mathcal{E}_t$. The scalar c is the frozen VLM's draft confidence and is copied unchanged to every verifier. Extraction precedes method selection, so all compared interfaces receive the same cached rows, including the same missing detections and track breaks.

The intervention variable is an access map $A(k)$. Under selective access, critic k receives only its corresponding channel, $A_{\text{sel}}(k) = \mathcal{E}_k$. Under union access, it receives the fixed serialization $A_{\text{union}}(k) = \mathcal{E}_g \| \mathcal{E}_c \| \mathcal{E}_r \| \mathcal{E}_t$. For evaluation slice S, we measure the union-minus-selective access effect

$$\Delta_{\text{fixed}}(S) = \big(\text{Acc}_{\text{union-restored}}(S) - \text{Acc}_{\text{SCV}}(S),\ \text{OC}_{\text{union}} - \text{OC}_{\text{sel}}\big). \tag{2}$$

This fixed-model quantity measures the deployed response to expanding a channel-trained critic's interface, including its response to the larger serialization. Union-restored estimates Δ_{fixed} while holding the draft, cache, verifier parameters, router scores and ordering, call cap, output normalization, and acceptance rule fixed. Because a KEEP decision advances to the next ranked critic, output-dependent continuation produces the small observed difference between 1.40 and 1.41 mean calls.

A second quantity compares separately optimized endpoints:

$$\Delta_{\text{adapt}}(S) = \big(\text{Acc}_{\text{Competitive-Union}}(S) - \text{Acc}_{\text{SCV}}(S),\ \text{OC}_{\text{Competitive-Union}} - \text{OC}_{\text{SCV}}\big). \tag{3}$$

Competitive-Union is trained directly for A_{union} at the same total post-hoc capacity and call target. Thus Δ_{fixed} is the controlled interface intervention, while Δ_{adapt} tests whether union-specific learning produces a stronger deployed endpoint.

For each critic $k \in \{g, c, r, t\}$, the router r_ψ uses $(q, \hat{y}, c, e)$ to predict two quantities: expected correction utility $\hat{u}_k$ and the risk of over-correction $\hat{o}_k$. Here e denotes the frozen generation-side evidence trace consumed by both cache extraction and routing; it is recorded once and never recomputed for an access variant:

$$r_\psi(q, \hat{y}, c, e) = \{(\hat{u}_k, \hat{o}_k)\}_k. \tag{4}$$

It also predicts $p_{\texttt{stop}} = P(\text{no critic will help})$. If the stop threshold is not met, critics are ordered by

$$\pi = \text{argsort}_k\big(\hat{u}_k - \gamma \hat{o}_k\big), \tag{5}$$

with ties resolved in the fixed order grounding, counting, geometry, temporal. The controller returns the draft when $p_{\texttt{stop}} > \tau_s$ and otherwise follows π until a valid revision is accepted or $B = 4$ is exhausted. We use $\tau_s = 0.65$ and $\gamma = 1.5$, selected on validation and then frozen.

Each critic receives `question`, `draft`, `confidence`, a channel tag, and evidence rows. Grounding rows are $(\texttt{object}, \texttt{score}, x_1, y_1, x_2, y_2)$; count rows are $(\texttt{object}, \texttt{count}, \texttt{mean score})$; geometry rows contain subject and object identifiers, a relation label, normalized center offsets, and overlap; temporal rows contain track identifier, frame index, box, label, and order indicators. Coordinates lie in $[0, 1]$, scores and counts use canonical decimal and integer formats, rows use a deterministic confidence–identifier order, and an absent channel is serialized as an explicit `EMPTY` record. Per-channel caps and truncation are fixed on validation and applied before access assignment, making each SCV record identical to its segment inside the union serialization.

The critic emits KEEP or REVISE and one candidate. For GQA, NLVR2, and NExT-QA, the candidate is a lower-cased, whitespace- and punctuation-normalized answer string; for RefCOCOg it is an `xyxy` tuple on $[0, 1]$. The gate accepts the first REVISE candidate in router order that differs from the normalized draft and satisfies the task contract: a nonempty answer string for QA or ordered in-range coordinates for RefCOCOg. Invalid candidates are treated as KEEP; exhausting the ranked list preserves the draft. No verifier receives raw pixels, video, benchmark slice IDs, or perturbation metadata.

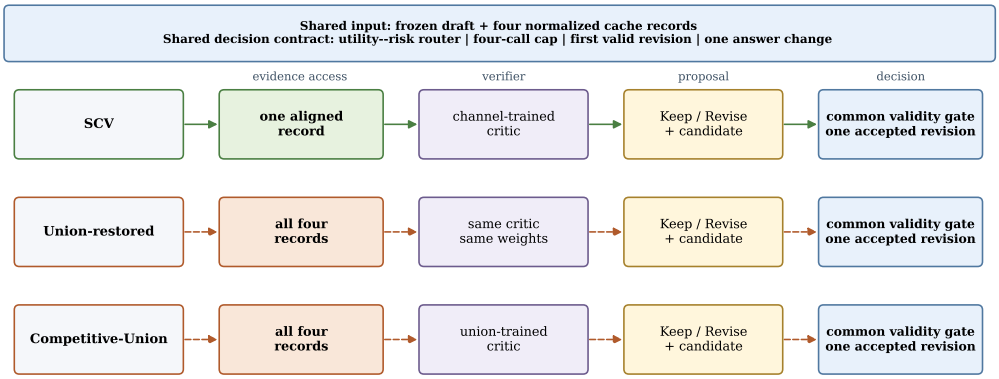


Figure 1: Evidence access after draft generation. SCV routes one aligned record to a channel-trained critic; Union-restored gives the same critic all four records; Competitive-Union trains equal post-hoc capacity for the all-channel serialization. Every row shares the draft/cache state, router objective, call and revision caps, output contract, and gate.

4 Selective Channel Verification

Selective Channel Verification (SCV) combines a router, four channel-local critics, and an explicit stop action. The frozen base-model pass and cache construction execute once. The router prioritizes expected correction utility net of collateral-edit risk; each critic then receives one aligned record, and the gate accepts the first valid proposed revision in router order. Union-restored preserves all learned SCV parameters but changes A_{sel} to A_{union}. Competitive-Union retains the four-slot controller, output heads, gate, and 8.2M total budget while training every critic and its input projection from scratch on the canonical union serialization.

Figure 1 separates the two inferential roles. The SCV–Union-restored pair changes only the serialized input presented to fixed channel-trained critics. The SCV–Competitive-Union pair compares systems optimized for their respective interfaces at equal total post-hoc capacity. Cache extraction cost is shared and excluded from verifier-call accounting; none of the paths reopens raw pixels or video.

4.1 Architecture and Inference

Algorithm 1 SCV inference

Require: Query q, image/video I, frozen VLM f_θ, access map A, budget $B = 4$

$(\hat{y}, c, e) \leftarrow f_\theta(q, I)$
$\mathcal{E} \leftarrow \textsc{ExtractAndNormalize}(e, I)$
$\{(\hat{u}_k, \hat{o}_k)\}_k, p_{\texttt{stop}} \leftarrow r_\psi(q, \hat{y}, c, e)$
if $p_{\texttt{stop}} > \tau_s$ **then**
  return $\hat{y}$
end if
$\pi \leftarrow \text{argsort}_k(\hat{u}_k - \gamma \hat{o}_k)$
for $b = 1, \ldots, \min(B, 4)$ **do**
  $k \leftarrow \pi_b$
  $(d, \hat{y}') \leftarrow v_k(A(k), q, \hat{y}, c)$ {one counted call}
  if $d =$ Revise and Valid$(\hat{y}')$ and $\hat{y}' \neq \hat{y}$ **then**
    return $\hat{y}'$
  end if
end for
return $\hat{y}$

The router is a 2-layer transformer (256-d, 4 heads) over tokenized query and draft representations, the scalar confidence embedding, and generation-trace features. Each typed critic is a 2-head cross-attention module (128-d) over the deterministic row serialization defined in Section 3. The critic output heads jointly predict the binary KEEP/REVISE decision and the task-contract candidate consumed by the common gate. The router, four critics, input projections, and output heads contain 8.2M learned parameters in total. Single-Verify applies one critic block once to the union serialization. Union-restored uses the SCV router and critic weights unchanged. Competitive-Union uses the same four critic slots and controller, but all four slots read the union serialization and all post-hoc parameters are trained for that interface; retaining module identities keeps the ordered call policy and total parameter allocation comparable.

4.2 TRAINING AND TAXONOMY CONTROLS

Training examples receive channel labels through a fixed precedence rule. Benchmark metadata supplies a cue when explicit; otherwise cache-derived grounding, count, geometry, or tracklet signals nominate a channel. For unresolved cases, one channel-local perturbation—a swapped relation, mismatched grounding candidate, count offset, or temporal permutation—identifies the cue that changes the verifier target. The router minimizes

$$\mathcal{L}_{\text{router}} = \mathcal{L}_{\text{utility}} + \lambda_1 \mathcal{L}_{\text{overcorrect}} + \lambda_2 \mathcal{L}_{\text{stop}}, \tag{6}$$

where the terms respectively provide correction-utility predictions, penalties for over-correction, and supervision of the stop decision.

Critics train independently on incorrect drafts, correct drafts, and channel-local perturbations. A critic-specific positive requires one dominant cue; multi-cue examples supervise router utility without forcing a critic identity, and no-cue examples train KEEP/`stop`. Competitive-Union uses the same draft labels, correction targets, router losses, minibatch order, training budget, and validation selection rule, replacing only the critic inputs and their training distribution with the union serialization. All validation-selected thresholds, loss weights, record caps, and truncation rules are frozen before final evaluation. Evaluation excludes gold slices, gold boxes, and perturbation metadata. Table 6 probes taxonomy dependence with a held-out mixed-mode subset, training without perturbation-derived labels, and a channel-shuffling stress test that preserves cache values and calls while disrupting the learned channel interface.

5 EVALUATION

5.1 PROTOCOL AND COMPARISON CONTRACT

We evaluate GQA (Hudson & Manning, 2019), NLVR2 (Suhr et al., 2019), RefCOCOg (Mao et al., 2016), and NExT-QA (Xiao et al., 2021). Exact-match accuracy is used for the three QA tasks and Acc@0.5 for RefCOCOg. The primary access comparisons use the official GQA test-dev set with 12,578 questions and the public NLVR2 development set with 6,982 examples. GQA benchmark operation metadata identifies 3,742 questions whose functional programs contain spatial relation, position, or counting operations; the count-operation subset has 1,136 questions and overlaps the broader spatial family. The NLVR2 spatial manifest contains 2,106 examples tagged before training for count, arrangement, or cross-image correspondence. RefCOCOg uses its standard test split. The NExT-QA temporal subset contains event-order and trajectory-continuation questions. Manifests and all overlaps are frozen before model fitting and applied identically to every method.

The primary stack uses frozen LLaVA-1.6-13B (Liu et al., 2024a); the transfer stack uses frozen Qwen2.5-VL-7B-Instruct (Bai et al., 2025). GroundingDINO (Liu et al., 2024b) supplies boxes; counts aggregate detector outputs; geometry records normalized pairwise relations; temporal records contain sampled-frame tracklets and order indicators. Decoding and extraction execute once per example, and all post-hoc methods reuse the same normalized cache. Thus calls/question measures verifier work, while end-to-end runtime also includes the common draft and extraction pass.

New trained comparisons use five paired seeds. A pair shares draft/cache examples, split, initialization seed, minibatch order, normalization, gate, and revision cap across access variants. Tables report mean $\pm$ standard deviation over seeds. Confidence intervals come from 10,000 hierarchical paired bootstrap replicates: each replicate samples the five seed indices with replacement, samples example

Table 1: Repair contract. SCV, Union-restored, and Competitive-Union share frozen draft/cache records, a four-call cap and approximately 1.4 realized calls/question, output normalization, gate, and one-revision cap.

Method	Verifier access	Parameters	Calls/q	Revisions
Single-Verify	union cache, one generic block	one block	1.0	at most 1
All-Critics	one channel per critic	8.2M	4.0	at most 1
SCV	routed channel, `stop` enabled	8.2M	1.4	at most 1
Union-restored	SCV scaffold, union cache per call	8.2M	1.4	at most 1
Competitive-Union	union-trained capacity-matched verifier	8.2M	1.4	at most 1

Table 2: Fixed-model GQA input expansion. Union-restored presents the four-channel serialization to the channel-trained SCV critics; both rows use 1.4 calls/question after rounding.

Variant	Overall	Spatial	Count	Calls/q	Over-corr.%
SCV	66.8	58.4	62.1	1.4	3.8
− Spatial critic	65.1	53.2	62.0	1.2	3.5
− Counting critic	66.1	58.1	57.4	1.3	3.6
− Router (round-robin)	65.8	56.9	60.8	2.8	5.1
− Stop option	66.4	57.8	61.6	2.1	4.8
Union-restored	65.9	55.4	60.3	1.4	4.4

indices with replacement inside each selected seed, preserves method pairing, and recomputes the equal-seed-weighted difference. Training uses 8×A100 80GB GPUs; the router requires about 3 hours and each critic about 2 hours. Router accuracy is correct routing among examples with one hindsight target; true correction is fixed initially-wrong drafts divided by all initially-wrong drafts; over-correction is damaged initially-correct drafts divided by all initially-correct drafts; stop accuracy is a correct keep decision divided by examples whose draft should be retained.

Single-Verify reads `question`, `draft`, `confidence`, and the tagged `BOX`, `COUNT`, `GEOM`, and `TEMP` channels through one critic block. Union-restored applies the same serialization to every called critic without retraining, measuring fixed-model input expansion inside the SCV scaffold. Competitive-Union addresses interface adaptation: its 8.2M post-hoc parameters are trained end-to-end for union input under the same router objective, call cap, normalization, and keep/revise gate. Together they prevent a fixed-model distribution shift from standing in for the stronger practical comparison.

The results follow the decision path of the method: sensitivity to evidence exposure, adaptation to the all-channel interface, benchmark and backbone replication, dependence on channel semantics, and localization by proposal disagreement. The final cross-task table places SCV's accuracy/call operating point in broader task context.

5.2 Sensitivity to Evidence Exposure

Table 2 yields $\Delta_{\text{fixed}}(\text{spatial}) = (-3.0, +0.6)$: expanding the trained critics' input loses 3.0 spatial-accuracy points while raising collateral edits by 0.6 points. It also loses 1.8 points on the nested count subset. The component ablations locate the selective gains: removing the spatial critic erases the spatial advantage, whereas removing the counting critic primarily affects count questions. Round-robin spends twice as many calls and remains below SCV; disabling `stop` raises over-correction. The simultaneous loss of repairs and increase in damage identifies all-channel input sensitivity in the fixed SCV scaffold.

5.3 Adaptation to the All-Channel Interface

Union-specific training recovers 1.1 of the 2.7 spatial points lost by fixed-model expansion, showing that adaptation matters, while SCV remains ahead on every reported GQA outcome. Against Union-restored, paired SCV differences are +0.9 (0.6, 1.2) overall, +2.7 (2.2, 3.2) spatial, +1.5 (1.0, 2.0)

Table 3: Five-seed GQA access test (mean ± standard deviation). Competitive-Union is trained for union input with the same 8.2M post-hoc parameter budget.

Method	Overall	Spatial	Count	Calls/q	Over-corr.%
SCV	66.7±0.2	58.2±0.3	61.9±0.4	1.41±0.02	3.9±0.1
Union-restored	65.8±0.2	55.5±0.3	60.4±0.4	1.40±0.02	4.5±0.1
Competitive-Union	66.2±0.2	56.6±0.3	61.0±0.4	1.39±0.02	4.2±0.1

count, and -0.6 (-0.8, -0.4) over-correction points. Against Competitive-Union, the corresponding differences are +0.5 (0.2, 0.8), +1.6 (1.1, 2.1), +0.9 (0.4, 1.4), and -0.3 (-0.5, -0.1). Parentheses are hierarchical paired 95% confidence intervals.

Pooled over the five paired GQA runs, fixed/initially-wrong counts are SCV 3,411/22,775, Union-restored 3,068/22,775, and Competitive-Union 3,204/22,775. Damaged/initially-correct counts are respectively 1,578/40,115, 1,805/40,115, and 1,685/40,115. The common denominators verify that the accuracy gain is produced by more repairs and fewer damaged correct drafts on identical examples.

5.4 Replication on NLVR2

Table 4: Five-seed NLVR2 access replication (mean ± standard deviation).

Method	Overall	Spatial	Calls/q	Over-corr.%
SCV	81.0±0.2	74.2±0.3	1.41±0.02	4.2±0.1
Union-restored	80.5±0.2	72.1±0.3	1.41±0.02	4.8±0.2
Competitive-Union	80.8±0.2	73.1±0.3	1.40±0.02	4.5±0.1

NLVR2 reproduces the spatial and collateral-damage pattern. SCV exceeds Union-restored by +0.5 (0.2, 0.8) overall and +2.1 (1.6, 2.6) spatial points while reducing over-correction by 0.6 (0.3, 0.9) points. Against Competitive-Union, the corresponding differences are +0.2 (0.0, 0.4) overall, +1.1 (0.7, 1.5) spatial, and 0.3 (0.1, 0.5) over-correction points. The predeclared spatial and damage intervals remain positive.

Pooled fixed/initially-wrong counts are SCV 2,057/7,540, Union-restored 2,046/7,540, and Competitive-Union 2,069/7,540. Damaged/initially-correct counts are 1,149/27,370, 1,313/27,370, and 1,232/27,370. The repair–damage decomposition explains the small overall difference: Competitive-Union makes 12 more repairs, whereas SCV preserves 83 more initially correct drafts.

5.5 Backbone Transfer

Table 5: Five-seed GQA replication with frozen Qwen2.5-VL-7B-Instruct (mean ± standard deviation). Only post-hoc components are retrained.

Method	Overall	Spatial	Calls/q	Over-corr.%
No-Verify	68.9±0.2	57.1±0.3	0.0	—
Single-Verify	70.0±0.2	59.1±0.4	1.0	4.4±0.1
SCV	71.2±0.2	62.0±0.3	1.36±0.02	3.6±0.1
Union-restored	70.4±0.2	59.8±0.3	1.36±0.02	4.2±0.1
Competitive-Union	70.8±0.2	60.8±0.3	1.35±0.02	3.9±0.1

Absolute accuracy rises with Qwen2.5-VL, while the interface ordering persists. SCV exceeds Competitive-Union by 0.4 overall and 1.2 spatial points and reduces over-correction by 0.3 points at near-identical realized calls (1.36 versus 1.35). Pooled fixed/initially-wrong and damaged/initially-correct counts are Single-Verify 2,599/19,560 and 1,907/43,330; SCV 3,006/19,560 and 1,560/43,330; Union-restored 2,757/19,560 and 1,820/43,330; Competitive-Union 2,886/19,560 and 1,690/43,330.

5.6 Semantic Alignment Controls

Table 6: GQA taxonomy controls across five paired seeds. The held-out mixed-mode subset contains 864 examples with multiple independently detected cues.

Method	Overall	Spatial	Mixed	Calls/q	Over-corr.%
SCV	66.7±0.2	58.2±0.3	60.7±0.4	1.41±0.02	3.9±0.1
Union-restored	65.8±0.2	55.5±0.3	59.3±0.5	1.40±0.02	4.5±0.1
Competitive-Union	66.2±0.2	56.6±0.3	59.9±0.4	1.39±0.02	4.2±0.1
SCV, no perturb labels	66.5±0.2	57.7±0.4	60.3±0.4	1.42±0.02	4.0±0.1
SCV, shuffled channels	65.9±0.2	55.9±0.4	59.0±0.5	1.40±0.02	4.6±0.2

SCV retains a 0.8-point mixed-mode advantage over Competitive-Union while cues coexist. Removing perturbation-derived labels preserves 1.1 of the 1.6 spatial points separating SCV from Competitive-Union, so perturbation supervision is not necessary for most of the gain. Shuffling channel identities removes the slice-local advantage and raises over-correction. The shuffle disrupts both semantic alignment and the critic's familiar input distribution; together with label removal, this pattern demonstrates that the learned channel identity is consequential.

5.7 Proposal Disagreement Localizes the Gap

For each example, let d_k be critic k's keep/revise decision and a_k its candidate when revising. We compute

$$C = \binom{4}{2}^{-1} \sum_{i<j} \left[\frac{1}{2} \mathbb{1}(d_i \neq d_j) + \frac{1}{2} \mathbb{1}(d_i = d_j = \textsc{Revise}) \big(1 - s(a_i, a_j)\big) \right], \tag{7}$$

where s is token-F1 between normalized GQA candidates. Validation-set terciles define low, medium, and high conflict, with threshold ties kept in the same bin; thresholds are frozen before outcome analysis. The score is computed from model proposals without outcome labels. Figure 2 reports SCV minus Competitive-Union spatial accuracy by conflict level. Critic outputs and bin membership are seed-specific; supports are shown for one seed, whereas deltas average paired estimates across five seeds.

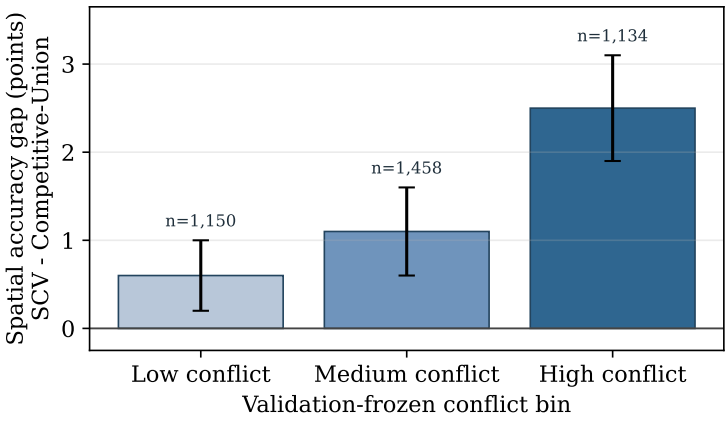


Figure 2: The SCV–Competitive-Union gap increases with proposal conflict. Spatial-accuracy differences are 0.6±0.4, 1.1±0.5, and 2.5±0.6 points for low, medium, and high conflict (paired 95% intervals). Single-seed supports are 1,150, 1,458, and 1,134; effect estimates average five paired seeds. SCV-minus-Competitive-Union over-correction differences are -0.1 ± 0.2, -0.3 ± 0.2, and -0.8 ± 0.3 points.

The gradient links the endpoint difference to proposal competition. From low to high conflict, SCV's spatial advantage rises from 0.6 ± 0.4 to 2.5 ± 0.6 points, while its over-correction difference moves from -0.1 ± 0.2 to -0.8 ± 0.3 points. Low-conflict examples leave little separation between interfaces; high-conflict examples are where explicit selection most improves both spatial accuracy and draft preservation.

5.8 Operating Points Across Four Tasks

Table 7: Operating points across four tasks. Calls count method-native verifier or generation invocations; paired access results for GQA and NLVR2 appear in Tables 3 and 4.

Method	Calls	GQA		NLVR2		RefCOCOg	NExT-QA	
		Overall	Spatial	Overall	Spatial	Acc@0.5	Overall	Temporal
No-Verify	0.0	63.8	51.2	78.4	68.1	71.3	62.1	48.7
Self-Consistency	4.0	65.1	53.8	79.7	70.2	72.1	63.4	50.9
Self-Refine	2.0	64.9	52.6	79.1	69.4	72.4	63.0	50.2
Single-Verify	1.0	65.4	54.1	79.9	70.8	73.2	63.7	51.4
All-Critics	4.0	66.2	56.8	80.6	73.1	74.8	64.8	53.2
SCV	1.4	66.8	58.4	81.1	74.3	75.6	65.2	54.8

SCV improves over Single-Verify by 4.3 points on GQA spatial, 3.5 on NLVR2 spatial, 2.4 Acc@0.5 on RefCOCOg, and 3.4 on NExT-QA temporal; All-Critics remains below SCV on every structured slice. Self-Refine and Self-Consistency also remain below SCV. The paired GQA and NLVR2 tables provide the access comparisons, while RefCOCOg and NExT-QA show that the same router-and-critic design reaches favorable grounding and temporal operating points.

6 Discussion

SCV changes the quality of revision, not only its frequency. Across five GQA runs, it repairs 207 more wrong drafts than Competitive-Union while damaging 107 fewer correct drafts. On NLVR2, Competitive-Union makes 12 additional repairs, but SCV preserves 83 additional correct drafts; spatial accuracy and over-correction therefore favor SCV even when aggregate accuracy is close. Union-specific training recovers part of the fixed-model input-expansion loss, and the remaining ordering persists under label removal, NLVR2 replication, and Qwen2.5-VL transfer. The near-tie at low proposal disagreement and larger high-disagreement gap further identify the cases in which explicit evidence selection changes the revision outcome.

The conclusions apply to detector-derived shared caches, a single accepted correction, and the shared four-slot architecture evaluated here. Union-restored includes the complete response to a longer, position-shifted serialization, while the conflict strata localize rather than causally decompose that response; channel shuffling likewise tests alignment sensitivity by disrupting the learned interface. RefCOCOg and NExT-QA provide task-level operating points, with the paired access evidence supplied by GQA and NLVR2. Within this design space, a robust comparison reports realized calls together with repairs and damaged correct drafts, because aggregate accuracy alone hides the central preservation benefit.

7 Conclusion

Routing evidence before revision turns heterogeneous cache records into a stronger low-call correction interface. SCV combines utility–risk routing, channel-local critics, and a one-revision gate to correct more wrong drafts while preserving more correct ones than the all-channel endpoints evaluated at comparable capacity and calls. The spatial advantage repeats across five paired seeds, NLVR2, and Qwen2.5-VL and is largest when channel-local proposals disagree. Multimodal verification should therefore expose its evidence-access structure and evaluate correction yield, collateral edits, and calls as a joint operating point.

8 AI USE DISCLOSURE

Generative language models assisted with manuscript organization and English-language revision, including this submission-readiness audit. This preparation pass did not create, modify, or validate experimental artifacts.

AI USE STATEMENT

During manuscript preparation, generative AI was used to refine wording, readability, and presentation. The authors reviewed the revisions and retain responsibility for the technical content and final text.

REFERENCES

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 39–48, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/Andreas_Neural_Module_Networks_CVPR_2016_paper.html.

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. *arXiv preprint arXiv:2502.13923*, 2025. URL https://arxiv.org/abs/2502.13923.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better LLM-based evaluators through multi-agent debate. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=FQepisCUWu.

Yi Ding and Ruqi Zhang. Sherlock: Self-correcting reasoning in vision-language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3403. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/93397b489222f1e66a862f795fc32b96-Abstract-Conference.html.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11733–11763. PMLR, 2024. URL https://proceedings.mlr.press/v235/du24e.html.

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. URL https://www.jmlr.org/papers/v23/21-0998.html.

Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14953–14962, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Gupta_Visual_Programming_Compositional_Visual_Reasoning_Without_Training_CVPR_2023_paper.html.

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. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Hudson_GQA_A_New_Dataset_for_Real-World_Visual_Reasoning_and_Compositional_CVPR_2019_paper.html.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9161–9175, 2023. doi: 10.18653/v1/2023.emnlp-main.568. URL https://aclanthology.org/2023.emnlp-main.568/.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne 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. doi: 10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20/.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge, January 2024a. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *Computer Vision – ECCV 2024*, Lecture Notes in Computer Science, pp. 38–55. Springer, 2024b. doi: 10.1007/978-3-031-72970-6_3. URL https://doi.org/10.1007/978-3-031-72970-6_3.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2019. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91edff07232fb1b55a505a9e9f6c0ff3-Abstract-Conference.html.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11–20, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/Mao_Generation_and_Comprehension_CVPR_2016_paper.html.

Aditya Sharma, Michael Saxon, and William Yang Wang. Losing visual needles in image haystacks: Vision language models are easily distracted in short and long contexts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 5429–5451. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.312. URL https://aclanthology.org/2024.findings-emnlp.312/.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=B1ckMDqlg.

Tejas Srinivasan, Jack Hessel, Tanmay Gupta, Bill Yuchen Lin, Yejin Choi, Jesse Thomason, and Khyathi Chandu. Selective "selective prediction": Reducing unnecessary abstention in vision-language reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12935–12948. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.767. URL https://aclanthology.org/2024.findings-acl.767/.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6418–6428, 2019. doi: 10.18653/v1/P19-1644. URL https://aclanthology.org/P19-1644/.

Yizheng Sun, Mochuan Zhan, Yanan Ma, Jia Tong See, Yifan Wang, Ziyi Wang, Hao Li, Yang Cui, Wenhao Cai, Jingyu Sun, Chenghua Lin, Riza Batista-Navarro, and Jingyuan Sun. Are reasoning vision-language models robust to semantic visual distractions? *arXiv preprint arXiv:2606.08894*, 2026. URL https://arxiv.org/abs/2606.08894.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5238–5248, 2022. doi: 10.1109/CVPR52688.2022.00517. URL https://doi.org/10.1109/CVPR52688.2022.00517.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed 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. URL https://openreview.net/forum?id=1PL1NIMMrw.

Tsung-Han Wu, Heekyung Lee, Jiaxin Ge, Joseph E. Gonzalez, Trevor Darrell, and David M. Chan. Generate, but verify: Reducing hallucination in vision-language models with retrospective resampling. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2206. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/5eff08bd064f0cdd92182cdf6fd06b99-Abstract-Conference.html.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. NExT-QA: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9777–9786, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Xiao_NExT-QA_Next_Phase_of_Question-Answering_to_Explaining_Temporal_Actions_CVPR_2021_paper.html.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12):220105, 2024. doi: 10.1007/s11432-024-4251-x. URL https://doi.org/10.1007/s11432-024-4251-x.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=KRLUvxh8uaX.

Ce Zhang, Zifu Wan, Zhehan Kan, Martin Q. Ma, Simon Stepputtis, Deva Ramanan, Russ Salakhutdinov, Louis-Philippe Morency, Katia P. Sycara, and Yaqi Xie. Self-correcting decoding with generative feedback for mitigating hallucinations in large vision-language models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=tTBXePRKSx.

Haoyu Zhang, Yuwei Wu, Pengxiang Li, Xintong Zhang, Zhi Gao, Rui Gao, Mingyang Gao, Che Sun, and Yunde Jia. Bridging modality disconnect in self-reflection via closed-loop visually grounded verification. *arXiv preprint arXiv:2602.18746*, 2026. doi: 10.48550/arXiv.2602.18746. URL https://arxiv.org/abs/2602.18746.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/fc625e831361cfcc82cb74224fdc66cb-Paper-Conference.pdf.

A Method Positioning

The table distinguishes post-hoc cache access from methods that reacquire raw images or alter the generation trajectory. The access intervention is defined only for the SCV–Union-restored pair.

Table 8: Comparison dimensions for multimodal correction. "Capacity ctrl." denotes equal total post-hoc parameters; "access int." denotes a fixed-model comparison that changes serialized verifier input while retaining the remaining repair contract.

Method	Phase	Raw image	Structured-cache access
Self-Refine	after	yes	none
Self-Consistency	during	yes	none
DeGF	during	yes	regenerated image
Sherlock	train/after	yes	reasoning trajectory
REVERSE	during	yes	confidence tokens
MIRROR	iterative	yes	marked image regions
SCV / Union-restored	after	no	same typed / union cache

Method	Calls / rounds	Revision freedom	Capacity ctrl.	Access int.
Self-Refine	fixed rounds	full answer	no	no
Self-Consistency	samples	resample and vote	no	no
DeGF	recursive	token decoding	no	no
Sherlock	multi-step	erroneous suffix	no	no
REVERSE	adaptive	backtrack and resample	no	no
MIRROR	repeated	region-guided answer	no	no
SCV / Union-restored	1.4 expected	one answer	yes	yes

B Additional Diagnostics

Table 9: SCV diagnostics. Router accuracy conditions on examples with one hindsight target; true correction and over-correction use the event denominators defined in Section 5.1.

Metric	GQA	NLVR2	RefCOCOg	NExT-QA
Router accuracy (%)	71.2	66.8	69.4	66.2
Over-correction rate (%)	3.8	4.2	3.1	4.0
True correction rate (%)	15.0	27.7	22.7	14.7
Stop accuracy (%)	82.1	79.4	83.7	78.2

Stop accuracy remains near 80% while over-correction remains near 4%, showing that the router's gains arise from selective revision rather than indiscriminate rewriting. True correction divides repaired drafts by all initially wrong drafts, consistent with the pooled GQA and NLVR2 counts in Sections 5.3 and 5.4. NLVR2 has the highest true-correction rate, while RefCOCOg combines the next-highest correction rate with the lowest over-correction.

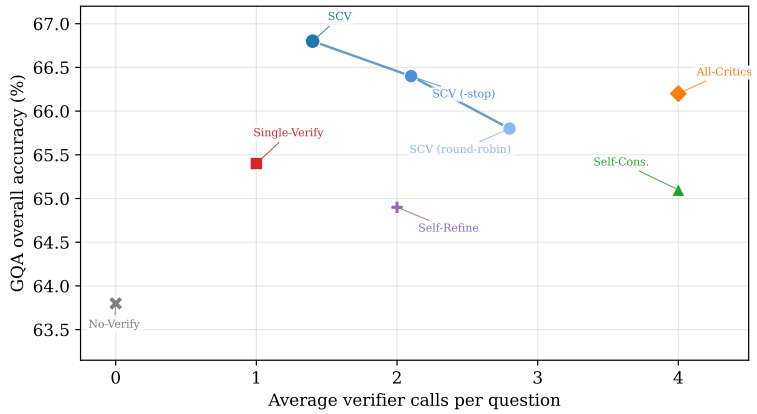


Figure 3: GQA operating points under shared-cache verifier-call accounting. Every point corresponds to a reported table row; the plot does not interpolate. Generation rounds and verifier calls are method-native units.

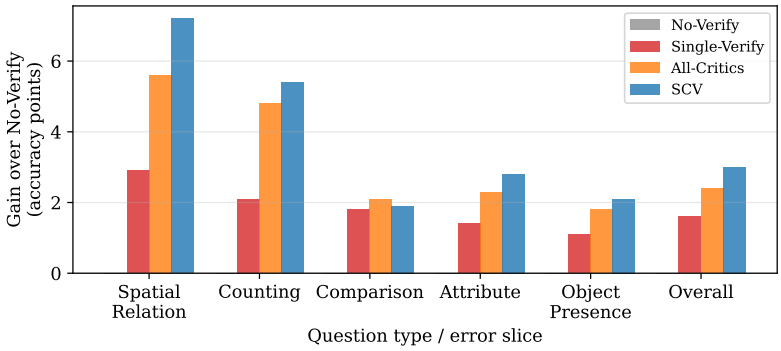


Figure 4: GQA gains over No-Verify by frozen question type. Counting is nested within, and therefore overlaps, the broader spatial family; plotted supports are not disjoint. SCV's advantage is largest on spatial relation and counting, while comparison, attribute, and object-presence gains are smaller.

C Experimental Design Summary

Table 10 maps each result to its shared quantities and interpretation. This evidence map keeps the fixed-model intervention, interface adaptation, transfer, and heterogeneity analyses aligned with the decisions they evaluate.

Table 10: Experimental questions, controlled comparisons, and supported interpretations.

Question	Comparison	Shared quantities	Supported interpretation
Fixed-model expansion	SCV vs. Union-restored	weights, router order, records, cap, gate	deployed response to a larger serialization (Tables 2, 3)
Interface adaptation	SCV vs. Competitive-Union	8.2M capacity, training budget, call target, gate	selective and union endpoints after interface-specific training (Table 3)
Benchmark transfer	three access variants on NLVR2	schema, metrics, pairing protocol	spatial and collateral-damage replication (Table 4)
Backbone transfer	three access variants with Qwen2.5-VL	manifests, cache schema, post-hoc training	persistence after changing the frozen VLM (Table 5)
Taxonomy sensitivity	label removal and channel shuffle	examples, call cap, gate	robustness to label source and sensitivity to interface alignment (Table 6)
Conflict heterogeneity	frozen score strata	paired predictions and thresholds	association between proposal disagreement and endpoint gap (Figure 2)

Across these comparisons, the protected object is the one-revision decision: every access-controlled row begins with the same draft and cache and is evaluated on the same examples. This design makes the repair–damage decomposition directly comparable even when the trained interface changes. All access variants use the deterministic channel order `BOX`, `COUNT`, `GEOM`, `TEMP`; per-channel normalization and truncation occur before access assignment, so the SCV record is byte-for-byte the corresponding segment of the all-channel serialization. Validation fixes $\tau_s = 0.65$, $\gamma = 1.5$, all cache caps, truncation rules, and training-selection choices before evaluation.