QUOTIENT-SET EVIDENCE FUSION FOR ROBUST MULTIMODAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable reasoning over multiple images and document pages requires combining distinct facts, yet repeated content can disproportionately influence predictions even when it adds no information. We make multimodal reasoning resistant to artificial repetition while preserving the complementary and conflicting evidence needed to answer a question. We introduce Quotient-Set Evidence Fusion (QSEF), which identifies redundancy relative to the question and adjusts each evidence unit's attention weight according to its estimated repetition count. QSEF corrects attention before pooling independently encoded inputs into a compact evidence representation, separating the relevance of information from its frequency while retaining every input. Across eligible subsets of six multi-image benchmarks, including MuirBench and BLINK, and three vision–language backbones, QSEF improves accuracy-based scores by 1.4–2.7 percentage points over the strongest compared baseline in all 18 comparisons. Our analysis establishes exact repetition invariance for correctly identified copies with matching representations and bounds representation drift under imperfect redundancy estimates, clarifying when robust multimodal evidence aggregation is achievable.

1 INTRODUCTION

Multimodal large language models increasingly answer questions whose evidence is distributed across multiple images or document pages. Multi-image instruction tuning has expanded their ability to compare and combine visual information (Jiang et al., 2024), while document retrieval extends their reach to large collections (Chen et al., 2025). These applications require reliable evidence aggregation: a model must integrate complementary facts while remaining insensitive to changes in presentation that add no information.

Evidence count and evidence content are different quantities. Copying a document page adds no independent evidence, while still changing how a model uses its contents. SoFA mitigates image-position bias (Tian et al., 2025), and delimiter token scaling reduces cross-image information leakage (Lee et al., 2026). These advances improve access to and separation of images; repeated information remains unnormalized. In text retrieval-augmented generation, duplication can alter observable evidence-use behavior even when the final answer remains unchanged (Jain & Vedam, 2026).

Suppressing repetition must preserve complementary information. Text-guided visual selection identifies relevant regions (Zeng et al., 2025); relevance alone leaves interchangeability unresolved. Two product cards may repeat a price yet report different delivery dates: either suffices for the price query, whereas comparing dates requires both. Permutation-invariant attention handles presentation order (Lee et al., 2019), yet uneven repetition can change its output. Max pooling ignores exact copies (Qi et al., 2017); complementary details can still be suppressed during aggregation. The challenge is to remain stable when information is repeated and responsive when question-relevant information changes.

We ask whether multimodal reasoning can resist artificial repetition while preserving complementary and conflicting evidence. Our scope covers questions whose answers are independent of presentation order and copy count, with task-relevant metadata retained. We introduce *Quotient-Set Evidence Fusion* (QSEF), which allocates influence according to distinct question-relevant information. It estimates redundancy relative to the question and corrects the influence of repeated content

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.

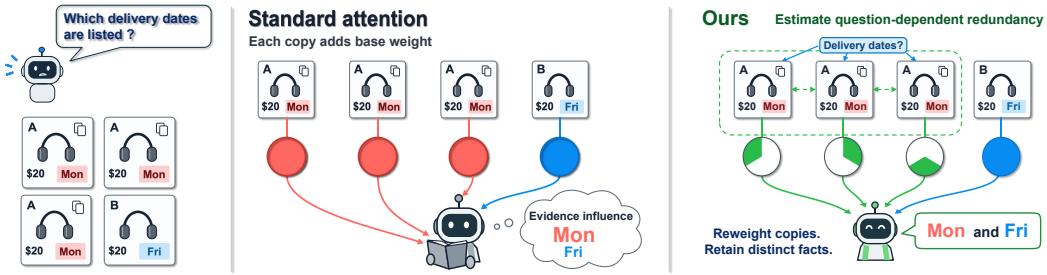


Figure 1: **QSEF corrects repeated evidence while retaining distinct facts.** Standard attention can amplify "Mon" simply because it appears three times. QSEF estimates redundancy relative to the question and reweights copies while retaining the complementary "Fri" evidence. The circles show base weights for exact copies with ideal redundancy estimates, before relevance-based attention.

during fusion, while retaining distinct evidence (Figure 1). A compact representation of the combined evidence supports answer generation.

Our contributions are threefold:

- **Question-dependent evidence redundancy.** Our preliminary study separates repetition effects from input size: concentrating copies on one unit doubles the median prediction-distribution drift of balanced copying under an equal input budget. Question-switch diagnostics further motivate evaluating stability together with complementary-information use.
- **Evidence fusion with explicit robustness conditions.** We develop QSEF and establish exact duplication invariance under ideal equivalence relations and matching within-class representations. A sharp conditional bound characterizes how redundancy-estimation errors affect fused representations; further analysis identifies failure under extreme repetition.
- **Consistent gains across tasks and backbones.** On eligible subsets of six multi-image benchmarks and three backbones, QSEF outperforms the strongest compared baseline by 1.4–2.7 percentage points. Controlled tests show improved complementary-information use and fact sensitivity, with additional transfer to document question answering under shared retrieval outputs.

2 Related Work

Robust multi-image reasoning. Multi-image instruction tuning, exemplified by Mantis (Jiang et al., 2024), improves visual reasoning while leaving evidence-presentation invariance unspecified. Subsequent interventions address position bias through SoFA's attention interpolation (Tian et al., 2025), and cross-image interference through delimiter scaling (DTS) (Lee et al., 2026), contrastive decoding (FOCUS) (Park et al., 2025), and layer-wise attention masking (Das et al., 2026). These mechanisms regulate interactions or decoding while leaving information-class multiplicity unnormalized. CUE-R (Jain & Vedam, 2026) finds that duplicating text evidence can alter retrieval-use behavior even when answers remain unchanged. QSEF targets artificial repetition during multimodal fusion, for questions whose answers are insensitive to presentation order and copies.

Set aggregation and sampling correction. Deep Sets (Zaheer et al., 2017) formalizes permutation-invariant pooling; Set Transformer (Lee et al., 2019) extends it with attention-based interactions and learned-query pooling (PMA). Sum, mean, and softmax pooling generally remain sensitive to unequal repetition, whereas PointNet-style max pooling (Qi et al., 2017) already ignores exact copies. PointConv (Wu et al., 2019) also uses inverse-density weighting to compensate for nonuniform sampling. QSEF builds on these principles by estimating question-dependent evidence equivalence and correcting attention for multiplicity. Its exact duplication invariance requires correct equivalence grouping and identical within-class keys and values.

Visual compression and evidence mass. Visual bottlenecks evolved from BLIP-2's learned queries (Li et al., 2023) to instruction-aware extraction in InstructBLIP (Dai et al., 2023). LLaVA-PruMerge (Shang et al., 2025) prunes and merges visual tokens, while AVAM (Zeng et al., 2025) uses text-guided region selection for multi-image QA. Relevance and compressibility leave inter-

changeability unresolved. ToMe's proportional attention (Bolya et al., 2023) adds log token size to preserve merged attention mass. QSEF subtracts log multiplicity to counter artificial copies while retaining distinct complementary or conflicting evidence. Its fixed prefix controls decoder context length, with per-unit encoding and pairwise relation costs still growing with the evidence set.

3 METHOD

3.1 OVERVIEW AND TASK SETTING

Quotient-Set Evidence Fusion (QSEF) adapts a pretrained vision–language model to evidence sets whose answers should be invariant to presentation order and artificial copies. Standard occurrence-level attention overweights repeated units; QSEF estimates each unit's question-dependent multiplicity, subtracts its logarithm from the attention logits, and pools independently encoded units into a fixed-length evidence prefix (Figure 2).

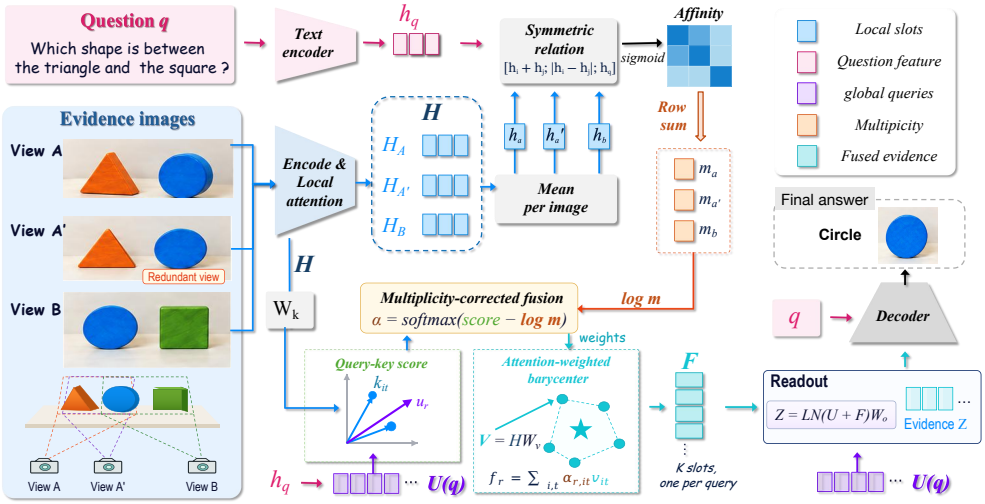


Figure 2: **Overview of Quotient-Set Evidence Fusion (QSEF).** Evidence units are independently encoded into local slots H_i. Given h_q, symmetric affinities yield multiplicities $m_i = \sum_j A_{ij}$; subtracting $\log m_i$ corrects redundant mass before softmax. The K question-conditioned queries $U(q)$ pool local values into F, and $Z = \mathrm{LN}(U + F)W_o$ supplies a fixed-length evidence prefix. The example combines redundant views A, A' with complementary view B to identify the circle between the triangle and square.

Let q be a question and $E = ((e_i, M_i))_{i=1}^{n}$, $n \geq 1$, a presentation of an evidence multiset. Each visual unit e_i carries semantic metadata M_i; local structure, reading order, and source attributes remain attached, while the presentation index is nonsemantic. We target tasks invariant to reordering and artificial copies; questions about indices, order, input cardinality, or independent-observation counts are outside this setting.

Two units are equivalent for q when $I_q(e_i, M_i) = I_q(e_j, M_j)$, where I_q is their question-relevant information, including task-relevant source attributes. This partitions units into classes $\mathcal{C}_1, \ldots, \mathcal{C}_C$. Equivalence means information-level substitutability: matching answers alone is insufficient, and partially overlapping evidence remains distinct. QSEF assigns each class one unit of base mass while attention learns relevance; because classes are latent, it estimates their multiplicities from pairwise affinities.

3.2 INDEPENDENT EVIDENCE ENCODING

A shared frozen visual encoder processes each unit independently. Its features and metadata are linearly mapped into $X_i \in \mathbb{R}^{L_i \times d}$, where L_i is the local-token count; positions restart within each unit, so the outer presentation does not affect X_i. A question-only pass through the frozen text

backbone and a learned projection produces $h_q \in \mathbb{R}^d$. Appendix A.2 specifies feature and metadata construction.

A learned bank $S \in \mathbb{R}^{s\times d}$ supplies s local queries, conditioned on the question as $Q_{\mathrm{loc}} = S + \mathbf{1}_s(W_{\mathrm{loc}}h_q)^\top$. Each unit is resampled independently:

$$H_i = \mathrm{softmax}_{\mathrm{row}}\left(\frac{Q_{\mathrm{loc}}(X_iW_k^{\mathrm{loc}})^\top}{\sqrt{d}}\right)X_iW_v^{\mathrm{loc}}, \qquad h_i = \frac{1}{s}H_i^\top\mathbf{1}_s. \tag{1}$$

Here $H_i \in \mathbb{R}^{s\times d}$ contains local evidence slots and $h_i \in \mathbb{R}^d$ summarizes the unit for relation estimation. Every unit produces s slots, preventing local token count from adding a multiplicity factor; the frozen encoders retain local geometry and text order before resampling.

3.3 Multiplicity-Corrected Evidence Fusion

Estimating multiplicity. For an ideal equivalence matrix $A^\star_{ij} = \mathbf{1}[(e_i, M_i) \sim_q (e_j, M_j)]$, the row sum is the class size of unit i. Its reciprocal gives every class unit total weight and uniquely minimizes the squared weight norm under that constraint (Appendix D.1).

QSEF estimates this matrix with symmetric pair features $\xi_{ij} = [h_i + h_j;\, |h_i - h_j|;\, h_q] \in \mathbb{R}^{3d}$ and a shared scalar network g_ψ:

$$z_{ij} = g_\psi(\xi_{ij}), \qquad A_{ij} = \begin{cases} 1, & i = j, \\ \sigma(z_{ij}), & i \neq j, \end{cases} \qquad m_i = \sum_{j=1}^{n} A_{ij}. \tag{2}$$

The sigmoid bounds off-diagonal affinities, while the fixed diagonal gives $1 \le m_i \le n$, so m_i^{-1} needs no offset. Symmetric features give $A_{ij} = A_{ji}$; relation supervision promotes transitive class assignments and distinguishes interchangeable from complementary evidence (Section 3.5).

Correcting attention mass. A learned bank $U_0 \in \mathbb{R}^{K\times d}$ defines K global queries, $U = U_0 + \mathbf{1}_K(W_uh_q)^\top$. Let $u_r = U_{r,:}^\top$ and define $k_{it} = W_k^\top H_{i,t,:}^\top$, $v_{it} = W_v^\top H_{i,t,:}^\top$. Subtracting $\log m_i$ from every logit for unit i yields

$$f_r = \frac{\displaystyle\sum_{i=1}^{n} m_i^{-1} \sum_{t=1}^{s} \exp(u_r^\top k_{it}/\sqrt{d})\, v_{it}}{\displaystyle\sum_{i=1}^{n} m_i^{-1} \sum_{t=1}^{s} \exp(u_r^\top k_{it}/\sqrt{d})}, \qquad r = 1, \dots, K. \tag{3}$$

Dot products determine relevance, while m_i^{-1} adjusts the base mass assigned to repeated information. The fixed-query pooling structure of set attention (Lee et al., 2019) is retained; every local value participates, with no representative selection.

Stacking $f_r^\top$ as the rows of $F \in \mathbb{R}^{K\times d}$ gives the language-model evidence prefix

$$Z = \mathrm{LN}(U + F)W_o \in \mathbb{R}^{K\times d_{\mathrm{LM}}}, \qquad W_o \in \mathbb{R}^{d\times d_{\mathrm{LM}}}. \tag{4}$$

Layer normalization is applied independently to each row. The decoder receives $[Z; \mathrm{Emb}(q)]$ and generates $p_\theta(y_\tau \mid y_{<\tau}, Z, q)$ autoregressively. Evidence occupies fixed positions $1, \dots, K$, so decoder length and positions are independent of n.

3.4 Invariance and Its Conditions

Presentation order. For any permutation π, deterministic per-unit preprocessing gives $Z(q, E_\pi) = Z(q, E)$. The local slot blocks, both axes of A, and the multiplicities permute together, leaving the sums in Equation (3) and each next-token distribution unchanged for a fixed answer prefix. Appendix D.2 gives the proof.

Artificial duplication. Suppose $A = A^\star$ and all members of class $\mathcal{C}_c$ share keys $k^\star_{ct}$ and values $v^\star_{ct}$. Then $m_i = |\mathcal{C}_c|$ for $i \in \mathcal{C}_c$, and Equation (3) reduces to

$$f^\star_r = \frac{\sum_{c=1}^{C}\sum_{t=1}^{s} \exp(u_r^\top k^\star_{ct}/\sqrt{d})\, v^\star_{ct}}{\sum_{c=1}^{C}\sum_{t=1}^{s} \exp(u_r^\top k^\star_{ct}/\sqrt{d})}. \tag{5}$$

Class cardinalities cancel, so adding copies leaves the evidence prefix and conditional output distributions unchanged. This requires recovered $A^\star$ and shared representations within each class; Appendix D.3 gives the cancellation proof, and Subsection 3.6 quantifies multiplicity-estimation error with extensions in Appendix D.

3.5 Learning and Computational Cost

We mix eligible Mantis-Instruct examples (Jiang et al., 2024) and controlled evidence cards from ChartQA tables (Masry et al., 2022) with equal probability. Answer labels supervise generation; relation labels use verified copies and known question-relevant field maps. Natural examples contribute only verified positive pairs; other pairs are unlabeled. Appendix A.1 specifies labels and source-disjoint construction.

For each example, we retain the original input with probability $1/2$. Otherwise, we select one unit uniformly and append a uniformly sampled number of copies from $\{1, 2, 4\}$, preserving its semantic metadata. The resulting input $\widetilde{E}$ supplies both losses, and all original units remain present. Let $Z = Z(q, \widetilde{E})$, let $y^\star$ be the target answer of length T_y, and let $\mathcal{P}$ contain the labeled pairs (i, j) with $i < j$ in $\widetilde{E}$. We minimize

$$\begin{aligned} \mathcal{L}_{\text{ans}} &= -\frac{1}{T_y}\sum_{\tau=1}^{T_y} \log p_\theta(y^\star_\tau \mid y^\star_{<\tau}, Z, q), \\ \mathcal{L}_{\text{eq}} &= \frac{1}{|\mathcal{P}|}\sum_{(i,j)\in\mathcal{P}} \big[\text{softplus}(z_{ij}) - r_{ij} z_{ij}\big], \\ \mathcal{L} &= \mathbb{E}_{\mathcal{D},\mathcal{A}}[\mathcal{L}_{\text{ans}} + \mathcal{L}_{\text{eq}}]. \end{aligned} \tag{6}$$

Here θ collects trainable parameters, $r_{ij} \in \{0, 1\}$ is the equivalence label, $\mathcal{D}$ the equal source mixture, and $\mathcal{A}$ the copy augmentation. The relation term is logit-space binary cross-entropy with $\text{softplus}(z) = \log(1 + e^z)$ and is zero when $\mathcal{P}$ is empty. Both losses are mean-normalized with unit coefficient. QSEF and rank-16 language LoRA adapters (Hu et al., 2021) are trainable, pretrained weights remain frozen, and training uses one answer-decoding pass per example. Appendices A.3 and A.4 give the architecture, augmentation, and input budgets.

With relation-head width h, pair scoring costs $O(n^2 dh)$ and fusion costs $O(Knsd + nsd^2 + d^2 + Kdd_{\text{LM}})$; local resampling adds $O(\sum_i L_i d^2 + s\sum_i L_i d + d^2)$. Frozen feature extraction and input projections are detailed in Appendix A.5. The decoder always receives K evidence tokens, while visual encoding and pair scoring grow with n.

3.6 Robustness to Repetition

The following result quantifies how multiplicity-estimation errors affect QSEF's repetition stability. Fix a question and a deterministic model. Consider two presentations with count vectors $\nu, \nu' \in \mathbb{N}_+^C$ for the same evidence classes. All classes remain present, and corresponding keys and values agree within each class and across presentations. For query r, let V_{rc} be the relevance-weighted mean value of class c's prototype slots, and set $D = (\sum_{r=1}^{K} \max_{c,j} \|V_{rc} - V_{rj}\|_2^2)^{1/2}$, the aggregate class-value diameter.

Theorem 1 (Stability under multiplicity-estimation error)**.** *If every occurrence in both presentations satisfies $|m_i/\nu_{c(i)} - 1| \le \epsilon$ for some $0 \le \epsilon < 1$, with each presentation using its own counts and estimated multiplicities, then $\|F(\nu') - F(\nu)\|_{\text{F}} \le D\epsilon$. Thus ideal multiplicities give exact invariance, and the coefficient one is optimal over the admissible fusion-module family.*

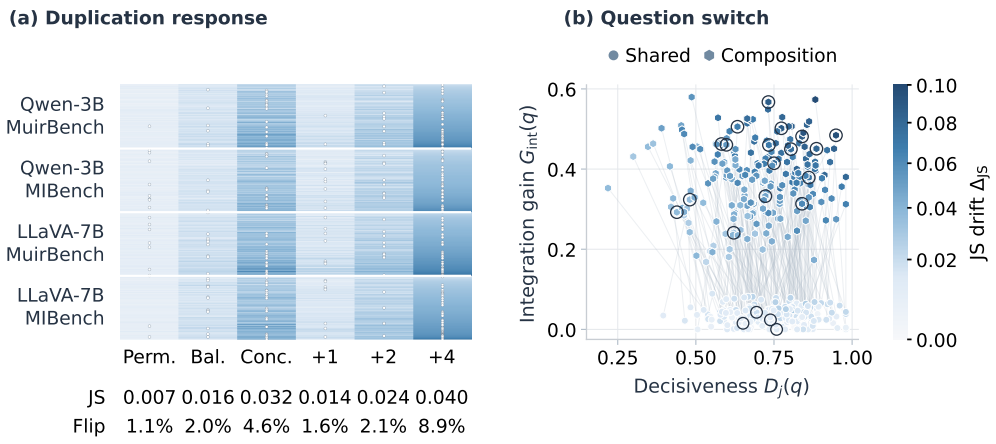


Figure 3: **Evidence repetition changes predictions in a question-dependent way. (a)** Responses to permutation, balanced and concentrated duplication, and $+1/2/4$ targeted copies; summary rows give median drift and flip rate. **(b)** Each line connects the same card pair under shared-field (circle) and composition (hexagon) questions. Colors encode Δ_{JS}; white dots in (a) and outer rings in (b) mark answer flips. See Appendix C.1 for plotting details.

Proof sketch. Grouping Equation (3) by class makes the effective mass an average of inverse relative multiplicities, lying in $[(1+\epsilon)^{-1}, (1-\epsilon)^{-1}]$. Hence the logarithmic mass ratios between presentations have oscillation at most $2\log((1+\epsilon)/(1-\epsilon))$. The sharp exponential-tilting inequality bounds each query's class-weight total variation by ϵ. Multiplying by the corresponding class-value diameter and summing the squared row bounds proves the claim. A realizable two-class construction attains equality. Appendix D.9 gives the full proof, and Appendix D.12 gives the construction. □

4 Experiments

We organize the experiments around four questions: when repeated evidence changes predictions (Q1), whether QSEF improves task quality while preserving evidence (Q2), whether the correction follows its mechanism and stability analysis (Q3), and where robustness gives way under extreme repetition (Q4).

4.1 Experimental Setup

Benchmarks and metrics. We evaluate order- and multiplicity-insensitive subsets of MuirBench (Wang et al., 2025), MIBench (Liu et al., 2024), MMIU (Meng et al., 2025), MIRB (Zhao et al., 2024), Mantis-Eval (Jiang et al., 2024), and BLINK (Fu et al., 2024), plus the official Document Haystacks tests (Chen et al., 2025). We report native task scores, Qwen/MuirBench permutation and duplication diagnostics, and controlled-card tests of complementary information, fact sensitivity, and class-mass conservation.

Comparison protocol. We compare QSEF with native and LoRA-tuned (Hu et al., 2021) readers, SoFA (Tian et al., 2025), DTS (Lee et al., 2026), FOCUS (Park et al., 2025), attention masking (Das et al., 2026), and pooling/deduplication controls. Trainable methods share data and updates within each backbone, fusion controls share the prefix budget, and inference costs are measured on Qwen2.5-VL-3B.

Backbones and implementation. We use Qwen2.5-VL-3B (Bai et al., 2025), LLaVA-OneVision-7B (Li et al., 2025), and InternVL3-8B (Zhu et al., 2025). QSEF uses rank-16 LoRA and $K = 64$ prefix tokens on four NVIDIA A100 80 GB GPUs. Appendix B gives the full protocol.

4.2 Q1. When Does Repeated Evidence Change Predictions?

We audit Qwen2.5-VL-3B-Instruct and LLaVA-OneVision-7B on 400 eligible examples from each of MuirBench and MIBench (1,600 model–example pairs), measuring Jensen–Shannon drift Δ_{JS}

Table 1: **Multi-image task performance on eligible subsets.** Scores are percentages (↑). C-Acc. denotes MIBench circular accuracy; MMIU uses the retained-task macro-average. The retained item counts for each backbone and benchmark are reported in Appendix B.5. Trainable rows show three-seed means; QSEF additionally reports ± sample SD. Bold and underline mark the best and second-best scores within each backbone, including ties; shading identifies QSEF.

Backbone	Method	MuirBench Acc.	MIBench C-Acc.	MMIU Macro	MIRB Acc.	Mantis-Eval Acc.	BLINK Acc.
Qwen2.5-VL-3B-Instruct (Bai et al., 2025)	Native	41.2	39.8	44.6	56.1	59.8	46.7
	LoRA (Hu et al., 2021)	44.6	42.7	47.5	57.8	62.2	48.6
	LoRA + SoFA (Tian et al., 2025)	45.8	43.4	48.0	58.6	63.1	49.5
	LoRA + DTS (Lee et al., 2026)	47.7	43.3	48.5	59.0	<u>64.6</u>	49.8
	LoRA + FOCUS (Park et al., 2025)	45.5	43.1	47.9	58.4	64.1	49.2
	LoRA + attn. mask (Das et al., 2026)	47.0	44.2	49.2	59.8	63.7	50.7
	PMA (Lee et al., 2019)	<u>48.5</u>	<u>45.5</u>	<u>49.9</u>	<u>60.2</u>	64.4	<u>51.1</u>
	QSEF (ours)	**51.2** ± 0.2	**47.6** ± 0.4	**51.5** ± 0.2	**61.8** ± 0.3	**66.3** ± 0.3	**53.2** ± 0.3
LLaVA-OneVision-Qwen2-7B-OV (Li et al., 2025)	Native	41.5	41.8	44.9	47.4	62.1	50.3
	LoRA (Hu et al., 2021)	43.9	44.0	46.6	49.1	63.5	51.4
	LoRA + SoFA (Tian et al., 2025)	45.2	44.8	47.2	49.9	64.6	52.2
	LoRA + DTS (Lee et al., 2026)	44.4	44.5	47.0	49.6	<u>65.3</u>	52.0
	LoRA + FOCUS (Park et al., 2025)	44.9	44.6	47.1	49.7	64.9	51.9
	LoRA + attn. mask (Das et al., 2026)	48.2	45.9	47.5	51.8	64.6	53.0
	PMA (Lee et al., 2019)	<u>48.8</u>	<u>46.8</u>	<u>48.6</u>	<u>52.1</u>	65.1	<u>53.4</u>
	QSEF (ours)	**51.0** ± 0.3	**48.9** ± 0.5	**50.3** ± 0.3	**53.8** ± 0.3	**66.7** ± 0.3	**55.1** ± 0.4
InternVL3-8B (Zhu et al., 2025)	Native	44.8	45.9	51.7	52.5	67.2	53.8
	LoRA (Hu et al., 2021)	46.7	48.1	53.2	54.0	68.1	55.0
	LoRA + SoFA (Tian et al., 2025)	47.6	48.9	53.8	54.7	68.6	55.7
	LoRA + DTS (Lee et al., 2026)	46.9	48.6	53.7	54.4	69.1	55.5
	LoRA + FOCUS (Park et al., 2025)	47.1	48.5	53.6	54.6	68.4	55.4
	LoRA + attn. mask (Das et al., 2026)	48.5	49.5	54.4	55.5	68.9	56.4
	PMA (Lee et al., 2019)	<u>49.2</u>	<u>50.2</u>	<u>55.0</u>	<u>55.9</u>	<u>69.4</u>	<u>56.9</u>
	QSEF (ours)	**51.4** ± 0.2	**52.3** ± 0.2	**56.4** ± 0.3	**57.4** ± 0.4	**71.0** ± 0.3	**58.4** ± 0.5

and answer-flip rate. Appendices B.6 and C.1 give the protocol and complete results. **Multiplicity matters beyond input size.** We compare balanced copying (one copy of every unit) with concentrated copying (the same budget assigned to one unit); both inputs contain $2n$ units. Concentrated duplication doubles median drift (0.032 vs. 0.016), while permutation alone yields 0.007 (Figure 3a). Targeted copies from one to four raise median drift from 0.014 to 0.040 and flip rate from 1.6% to 8.9%.

Redundancy depends on the question. On 220 held-out card pairs with shared and distinct fields, four targeted copies produce median drift 0.014 for shared-field questions and 0.059 for composition questions, with flip rates 1.8% and 7.7%. Figure 3b relates these responses to single-unit decisiveness $D_j(q)$ and the gain from combining evidence $G_{\text{int}}(q)$, motivating question-dependent redundancy control.

Robustness follows question-relevant information classes: equal-length presentations can exert different influence, and a locally decisive unit can remain complementary to another.

4.3 Q2. Does QSEF Improve Task Quality While Preserving Useful Evidence?

Task accuracy across backbones. QSEF achieves the highest score in all 18 backbone–benchmark comparisons in Table 1, improving over the strongest baseline by 1.4–2.7 percentage points across all three backbones, including the same-budget PMA fusion baseline.

Robustness and evidence preservation. On Qwen2.5-VL-3B, QSEF retains 51.2% accuracy under permutation and 50.8% under duplication, a 0.4-point clean-accuracy drop (Table 2). Its duplicate-induced flip rate is 0.7% versus 5.0% for PMA. QSEF also gives the highest complementary-card gain and fact-edit accuracy, exceeding PMA with a relation head by 3.2 and 5.2 points.

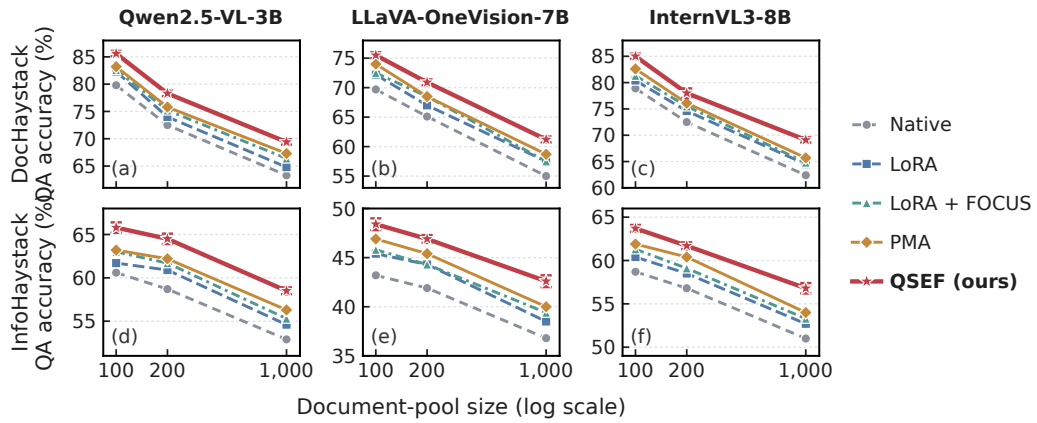


Figure 4: **Document transfer.** QA accuracy on DocHaystack (top) and InfoHaystack (bottom), with shared frozen V-RAG retrieval (Chen et al., 2025). Columns correspond to backbones. Trainable methods show three-seed means; only QSEF displays ± 1 sample SD. Each panel uses its own y-axis range. Pool sizes use a logarithmic axis; lines connect measured conditions.

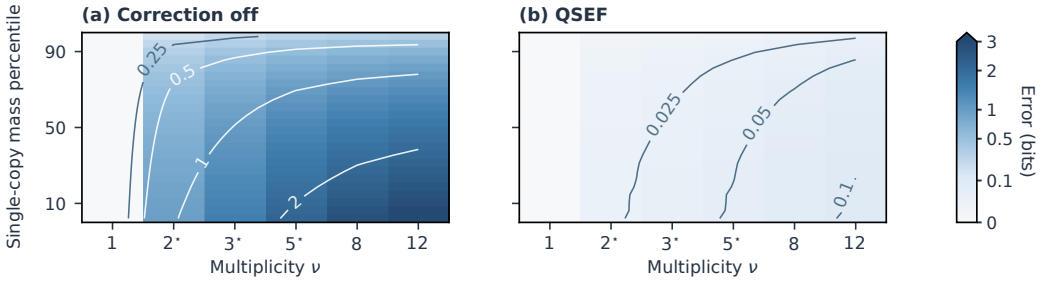


Figure 5: **Class-mass conservation.** Bin-median error on a shared nonlinear scale. Stars mark training multiplicities; Appendix C.3 details smoothing and statistics.

Transfer to document evidence. QSEF achieves the highest QA accuracy in all 18 document conditions under common cached retrieval outputs (Figure 4); at 1,000 documents, gains over PMA range from 2.1 to 3.4 percentage points.

4.4 Q3. Does the Correction Behave as Its Mechanism and Stability Analysis Predict?

We test whether all copies of an evidence unit retain the same total attention as a single copy. On 800 held-out two-card trials, we vary one card's multiplicity over $\nu \in \{1, 2, 3, 5, 8, 12\}$ while fixing the complementary card, question, and answer. Using one Qwen2.5-VL-3B checkpoint, we compare correction on versus off ($m_i = 1$), holding learned parameters and query–key scores fixed.

For each of 64 global queries, we sum attention over all copies and measure the absolute log-ratio error relative to $\nu = 1$, in bits. Zero means unchanged mass, and one bit means a twofold change. Figure 5 groups observations into 24 bins by single-copy attention mass.

At $\nu = 12$, correction reduces median error from 1.710 to 0.080 bits (95.3%) and the 90th percentile from 2.846 to 0.127 bits. Error stays low across bins at $\nu \in \{8, 12\}$, beyond training multiplicities. The median nevertheless rises from 0.016 bits at $\nu = 2$ to 0.080 at $\nu = 12$, supporting approximate mass conservation with residual drift (Appendix C.3).

Appendix C.4 also finds that fusion drift tracks the product of multiplicity-estimation error and the spread of class values ($\rho_s = 0.919$). Since the class-value spread varies across copy strengths, this is an empirical diagnostic of the scaling in Theorem 1. Its exact guarantee applies when class keys and values remain fixed across presentations.

Table 2: **Robustness and information preservation.** All methods use Qwen2.5-VL-3B. Entries show three-seed means; QSEF additionally reports ± sample SD. Accuracy and flip rates are percentages; G_{comp} is in percentage points and D_{JS} in 10^{-3} nats. Bold and underline mark the best and second-best scores per column, including ties.

		MuirBench subset					Controlled cards	
Category	Method	Clean ↑	R_π ↑	R_{dup} ↑	Flip ↓	D_{JS} ↓	G_{comp} ↑	A_{edit} ↑
Native-token readers with matched-data LoRA	LoRA (Hu et al., 2021)	44.6	40.9	41.7	7.8	11.8	14.7	48.4
	+ SoFA (Tian et al., 2025)	45.8	43.6	42.9	6.5	9.6	15.4	49.2
	+ DTS (Lee et al., 2026)	47.7	43.8	45.1	5.3	7.8	15.8	50.6
	+ FOCUS (Park et al., 2025)	45.5	42.8	43.4	4.8	7.1	15.2	49.5
	+ attn. mask (Das et al., 2026)	47.0	45.2	45.8	3.6	4.9	16.8	52.2
Fixed-prefix fusion ($K = 64$), with the same decoder LoRA	Deep Sets (sum) (Zaheer et al., 2017)	45.0	45.0	41.8	7.3	10.9	13.5	46.5
	Deep Sets (mean) (Zaheer et al., 2017)	46.7	46.7	45.0	4.2	5.7	15.6	49.9
	Max pooling (Qi et al., 2017)	44.8	44.8	44.8	**0.0**	**0.0**	11.8	44.1
	PMA (Lee et al., 2019)	48.5	48.5	45.7	5.0	6.3	17.7	55.9
	PMA (Lee et al., 2019) + relation head	<u>49.1</u>	<u>49.1</u>	46.4	4.7	5.8	<u>19.0</u>	<u>57.8</u>
	Exact-copy + PMA (Lee et al., 2019)	48.4	48.4	<u>48.4</u>	**0.0**	**0.0**	17.6	55.7
	Similarity + PMA (Lee et al., 2019)	47.9	47.9	47.9	**0.0**	**0.0**	13.8	51.0
	QSEF (ours)	**51.2**±0.2	**51.2**±0.3	**50.8**±0.3	<u>0.7</u>±0.1	<u>0.8</u>±0.1	**22.2**±0.4	**63.0**±0.5

4.5 Q4. Where Does Robustness Give Way Under Extreme Repetition?

We stress-test QSEF with Qwen2.5-VL-3B on two complementary evidence classes, varying one class over $N \in \{1, 2, 4, 8, 16, 32, 64\}$ occurrences while keeping the other as a necessary singleton. We analyze 5,292 records from 252 held-out questions and three checkpoints, stratified by clean cross-class affinity $\bar{A}_{\text{cross}}$. Failure is a three-percentage-point QA-accuracy drop from $N = 1$; Appendix C.5 gives the full aggregation and bin summaries.

For $N \leq 4$, every affinity-bin mean has an accuracy drop below 0.14 points and $J < 0.34$, and the four lowest-affinity bins remain below the failure threshold through $N = 64$. At higher affinity, the interpolated boundary is $N \approx 16.4$ at bin center 0.205 and $N \approx 9.7$ at 0.285. At $N = 64$, the 0.285 bin reaches a 12.72-point drop, with $J = 7.20$ and normalized fusion drift 0.82 (Appendix D.11).

Repetition robustness is a budgeted property whose usable range is set by the interaction between copy count and cross-class affinity. Monitoring redundancy geometry can therefore reveal risk earlier than counting inputs alone.

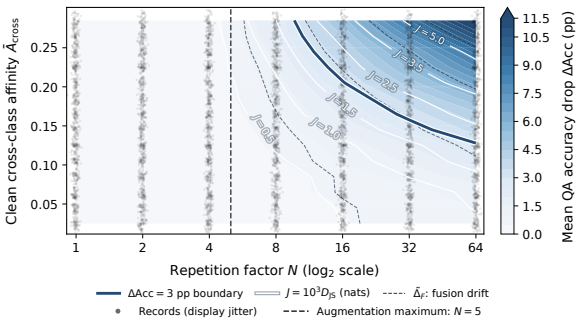


Figure 6: **Extreme-repetition stress test.** Color shows mean QA-accuracy drop; the blue curve marks the interpolated three-point failure boundary. White contours show $J = 10^3 D_{\text{JS}}$, while dashed curves show normalized fusion drift. The vertical dashed line marks the five-occurrence augmentation limit.

5 Conclusion

QSEF corrects attention mass using question-dependent redundancy. It guarantees exact invariance for correctly identified copies with matching representations and bounds fusion drift under bounded multiplicity-estimation error. Experiments across six multi-image benchmarks and three backbones show improved accuracy. Controlled tests demonstrate greater copy stability and sensitivity to complementary and edited facts. Evaluation should capture both properties, and future work should preserve them as retrieved evidence and provenance evolve.

AI USE STATEMENT

We used generative AI tools to assist with checking mathematical proofs, supporting literature searches and related-work synthesis, analyzing and interpreting experimental results, and drafting and editing the manuscript. The authors take full responsibility for the final content of this submission.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. URL https://arxiv.org/abs/2502.13923.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=JroZRaRw7Eu.

Jun Chen, Dannong Xu, Junjie Fei, Chun-Mei Feng, and Mohamed Elhoseiny. Document haystacks: Vision-language reasoning over piles of 1000+ documents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24817–24826, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Chen_Document_Haystacks__Vision-Language_Reasoning_Over_Piles_of_1000_Documents_CVPR_2025_paper.html.

Samuel N. Cohen and Eliana Fausti. Hyperbolic contractivity and the Hilbert metric on probability measures. *arXiv preprint arXiv:2309.02413*, 2023. URL https://arxiv.org/abs/2309.02413.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/9a6a435e75419a836fe47ab6793623e6-Abstract-Conference.html.

Anurag Das, Adrian Bulat, Alberto Baldrati, Ioannis Maniadis Metaxas, Bernt Schiele, Georgios Tzimiropoulos, and Brais Martinez. More images, more problems? a controlled analysis of VLM failure modes. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 7423–7442, 2026. doi: 10.18653/v1/2026.findings-acl.366. URL https://aclanthology.org/2026.findings-acl.366/.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. VLMEvalKit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 11198–11201. Association for Computing Machinery, 2024. doi: 10.1145/3664647.3685520. URL https://doi.org/10.1145/3664647.3685520.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pp. 148–166, 2024. doi: 10.1007/978-3-031-73337-6_9. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/03356.pdf.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. URL https://arxiv.org/abs/2106.09685.

Siddharth Jain and Venkat Narayan Vedam. CUE-R: Beyond the final answer in retrieval-augmented generation. *arXiv preprint arXiv:2604.05467*, 2026. URL https://arxiv.org/abs/2604.05467.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. MANTIS: Interleaved multi-image instruction tuning. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2405.01483.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set Transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of ICML*, volume 97 of *PMLR*, pp. 3744–3753, 2019. URL https://proceedings.mlr.press/v97/lee19d.html.

Minyoung Lee, Yeji Park, Dongjun Hwang, Yejin Kim, Seong Joon Oh, and Junsuk Choe. Enhancing multi-image understanding through delimiter token scaling. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=7QFf05KrOm.

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

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023. URL https://proceedings.mlr.press/v202/li23q.html.

Haowei Liu, Xi Zhang, Haiyang Xu, Yaya Shi, Chaoya Jiang, Ming Yan, Ji Zhang, Fei Huang, Chunfeng Yuan, Bing Li, and Weiming Hu. MIBench: Evaluating multimodal large language models over multiple images. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 22417–22428, 2024. doi: 10.18653/v1/2024.emnlp-main.1250. URL https://aclanthology.org/2024.emnlp-main.1250/.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of ACL*, pp. 2263–2279, 2022. URL https://aclanthology.org/2022.findings-acl.177/.

Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Tianshuo Yang, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. MMIU: Multimodal multi-image understanding for evaluating large vision-language models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/5f92032278b1c70946e0b753068de51e-Abstract-Conference.html.

Yeji Park, Minyoung Lee, Sanghyuk Chun, and Junsuk Choe. Mitigating cross-image information leakage in multi-image understanding with large vision-language models, 2025. URL https://arxiv.org/abs/2508.13744. Version 2 revised June 2026.

Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *Proceedings of CVPR*, pp. 652–660, 2017. URL https://arxiv.org/abs/1612.00593.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22857–22867, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Shang_LLaVA-PruMerge_Adaptive_Token_Reduction_for_Efficient_Large_Multimodal_Models_ICCV_2025_paper.html.

Xinyu Tian, Shu Zou, Zhaoyuan Yang, and Jing Zhang. Identifying and mitigating position bias of multi-image vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10599–10609, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/

Tian_Identifying_and_Mitigating_Position_Bias_of_Multi-image_Vision-Language_Models_CVPR_2025_paper.html.

Fei Wang, Xingyu Fu, James Y. Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, Tianyi Yan, Wenjie Mo, Hsiang-Hui Liu, Pan Lu, Chunyuan Li, Chaowei Xiao, Kai-Wei Chang, Dan Roth, Sheng Zhang, Hoifung Poon, and Muhao Chen. MuirBench: A comprehensive benchmark for robust multi-image understanding. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/9cf6139382f98623d08cc595622f3fb1-Abstract-Conference.html.

Wenxuan Wu, Zhongang Qi, and Li Fuxin. PointConv: Deep convolutional networks on 3D point clouds. In *Proceedings of CVPR*, pp. 9621–9630, 2019. URL https://arxiv.org/abs/1811.07246.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep Sets. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/f22e4747da1aa27e363d86d40ff442fe-Abstract.html.

Kang Zeng, Guojin Zhong, Jintao Cheng, Jin Yuan, and Zhiyong Li. AVAM: a universal training-free adaptive visual anchoring embedded into multimodal large language model for multi-image question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2292–2302, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Zeng_AVAM_a_Universal_Training-free_Adaptive_Visual_Anchoring_Embedded_into_Multimodal_ICCV_2025_paper.html.

Bingchen Zhao, Yongshuo Zong, Letian Zhang, and Timothy Hospedales. Benchmarking multi-image understanding in vision and language models: Perception, knowledge, reasoning, and multi-hop reasoning, 2024. URL https://arxiv.org/abs/2406.12742.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Yue Cao, Yangzhou Liu, Weiye Xu, Hao Li, Jiahao Wang, Han Lv, Dengnian Chen, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. URL https://arxiv.org/abs/2504.10479.

APPENDIX CONTENTS

A Technical Details **14**
A.1 Question-Dependent Relation Labels 14
A.2 Local Features, Metadata, and Question Encoding 14
A.3 Fixed Architecture and Trainable Parameters 15
A.4 Copy Augmentation and Input Budgets 15
A.5 Inference and Computational Accounting 15

B Experimental Details **17**
B.1 Benchmark Scope, Splits, and Native Scoring 17
B.2 Perturbations and Complementary-Information Diagnostics 17
B.3 Comparison Groups and Fairness Controls 18
B.4 Backbones, Optimization, and Reproducibility 19
B.5 Reading the Main Results 21
B.6 Preliminary Study Protocol 22

C Additional Experimental Results **24**
C.1 Preliminary Study Results 24
C.2 Targeted Ablations and Diagnostics 24
C.3 Information-Class Mass Conservation: Protocol and Results 26
C.4 Empirical Scaling Consistent with the Stability Envelope 28
C.5 Extreme-Repetition Stress Test 30

D Proofs **31**
D.1 Equal Weights Within an Equivalence Class 31
D.2 Proof of Presentation-Order Invariance 31
D.3 Proof of Ideal Duplication Invariance 31
D.4 Formal Setup and General Stability Theorem 32
D.5 A Sharp Exponential-Tilting Inequality 33
D.6 From Probability Mass to Vector Displacement 34
D.7 Proof of Theorem 2 34
D.8 A Geometry-Sensitive Refinement 35
D.9 Degree Envelopes and Relation-Error Amplification 36
D.10 Proof of Sharpness and the Unbounded-Repetition Limitation 38
D.11 An Exact Repetition Budget in the Two-Class Model 39
D.12 Sharpness of the Symmetric Degree-Error Bound 40
D.13 Explicit Dependence on Counts and Conditional Certificates 40
D.14 Direct Stability for Nonidentical Within-Class Representations 42
D.15 A Complementary Bound from Certified Prototype Errors 44
D.16 Application to the Answer Cross-Entropy Loss 46

A TECHNICAL DETAILS

A.1 QUESTION-DEPENDENT RELATION LABELS

Controlled evidence cards. We construct auxiliary cards from the underlying tables in ChartQA (Masry et al., 2022). Each field identifier includes its entity, row, column, and task-relevant source attributes, so equal numeric values attached to different facts remain distinct. A card displays a nonempty subset of the source table together with the labels needed to interpret its values. Questions are generated from executable programs using three operations: lookup, summation, and subtraction. The program identifies the required fields $S(q)$ and computes the target answer from their values. All required fields must occur in the union of the cards; examples failing this condition are excluded.

Let $\mathcal{F}_i$ be the field–value map visible on card i, and define $\mathcal{J}_i(q) = \{(f, \mathcal{F}_i(f)) : f \in S(q) \cap \mathrm{dom}(\mathcal{F}_i)\}$. For pairs with both maps nonempty, the relation label is

$$r_{ij} = \begin{cases} 1, & \mathcal{J}_i(q) = \mathcal{J}_j(q), \\ 0, & \mathcal{J}_i(q) \neq \mathcal{J}_j(q). \end{cases} \tag{7}$$

Pairs involving an empty relevant map are excluded from the controlled relation loss. Equal maps provide positives even when irrelevant fields and rendering styles differ. Distinct nonempty maps provide negatives, including partial overlap and complementary field sets. This construction makes the relation question-dependent: two cards with the same answer-relevant fields can be positive for one question and negative for another. Labels are derived from visible facts, never from equality of final answers. These are newly constructed auxiliary annotations, not native ChartQA relation labels.

Natural evidence. We retain Mantis-Instruct examples (Jiang et al., 2024) whose questions identify semantic entities and whose targets are unchanged by duplicating an input unit. Examples requiring presentation indices, inter-unit sequence positions, evidence cardinality, and independent-observation counts are excluded. Positive relation labels require a verified copy with matching semantic metadata. Lossless re-encoding qualifies only when decoded RGB arrays and task-relevant metadata agree. Pixel equality with different provenance is insufficient when provenance changes the task semantics. Compression, resizing, cropping, and OCR extraction qualify only when decoded RGB arrays and task-relevant metadata agree. All remaining natural pairs are unlabeled, including visually distinct pairs.

Splits and sampling. We split by source before generating derived questions, cards, and copies. All material from the same source table remains in one split. Natural examples sharing a source image are grouped together before splitting; connected components under this relation prevent a shared image from crossing split boundaries. The controlled development and test partitions also use held-out question and rendering templates. Rendering style is sampled independently of field content and pair labels. Copy flags and generation identifiers serve only to construct supervision and never enter the model. During training, each example selects the natural source with probability $1/2$ and the controlled source with probability $1/2$, then samples uniformly from the retained examples of that source.

A.2 LOCAL FEATURES, METADATA, AND QUESTION ENCODING

Our primary backbone is Qwen2.5-VL-3B-Instruct (Bai et al., 2025); Appendix B.4 specifies two additional backbone instantiations. For unit i, let $V_i \in \mathbb{R}^{L_i^v \times d_v}$ denote the output of its frozen visual encoder and pretrained visual merger or projector. Image preprocessing is applied independently with a fixed per-unit configuration. It never allocates a shared image-token budget according to presentation order. All visual positions are local to the unit.

Metadata are serialized as field-name–value pairs in a fixed field order and encoded by the frozen text backbone, with language adapters disabled. Its token states form $T_i \in \mathbb{R}^{L_i^m \times d_{\mathrm{LM}}}$. The serialization includes semantic attributes supplied with the task, such as source identity and observation time when relevant. It excludes the unit's presentation index and augmentation bookkeeping. Empty metadata contribute zero tokens. Shared learned projections $P_v \in \mathbb{R}^{d_v \times d}$ and $P_m \in \mathbb{R}^{d_{\mathrm{LM}} \times d}$ produce $X_i = [V_i P_v; T_i P_m]$, with $L_i = L_i^v + L_i^m$.

The question is encoded separately by the same frozen text backbone with language adapters disabled. If $t_q \in \mathbb{R}^{d_{\mathrm{LM}}}$ is the last question-token hidden state, the learned projection $W_q \in \mathbb{R}^{d \times d_{\mathrm{LM}}}$ gives $h_q = W_q t_q$. No evidence is present in this pass. The question and metadata passes share pretrained weights with the answer decoder. Their frozen features can be cached; trainable projections and resampler outputs are recomputed after parameter updates.

At answer generation, the model receives continuous embeddings $[Z; \mathrm{Emb}(q)]$ through its language input interface. The evidence slots use a fixed sequence of K positions; question and answer positions follow this prefix. The language position sequence is independent of the original image count and geometry. All visual and metadata information reaches the answer decoder through Z.

A.3 Fixed Architecture and Trainable Parameters

The fusion width is $d = 512$, with $s = 16$ local slots per unit, $K = 64$ output evidence slots, and relation-head width $h = 64$. The resampler and fusion each use one attention operation as specified in Equations (1) and (3). The relation network is $g_\psi(\xi) = a^\top \mathrm{GELU}(W_g \xi + b_g) + b_0$, where $W_g \in \mathbb{R}^{h \times 3d}$, $a, b_g \in \mathbb{R}^h$, and $b_0 \in \mathbb{R}$. We evaluate each off-diagonal pair once and mirror its score to construct the symmetric affinity matrix. Diagonal entries are fixed at one and excluded from relation supervision.

Language adaptation uses rank-16 LoRA (Hu et al., 2021) on the query and value projections in every decoder self-attention layer, with scaling parameter $\alpha = 16$ and dropout zero. For a column-vector weight matrix $W_0 \in \mathbb{R}^{d_{\mathrm{out}} \times d_{\mathrm{in}}}$, the adapted weight is $W_0 + (\alpha/16) D_{\mathrm{up}} D_{\mathrm{down}}$, where $D_{\mathrm{up}} \in \mathbb{R}^{d_{\mathrm{out}} \times 16}$ and $D_{\mathrm{down}} \in \mathbb{R}^{16 \times d_{\mathrm{in}}}$. The up projection is initialized to zero and the down projection uses Kaiming initialization. These adapters are active only for answer decoding.

Trainable parameters comprise P_v, P_m, W_q, the local query bank and projections, the relation network, the global query bank and projections, the readout layer normalization and W_o, and the language LoRA matrices. All pretrained weights remain frozen. Gradients from the answer loss pass through the frozen decoder operations to the evidence prefix and QSEF parameters; frozen weights still permit this gradient flow. The answer and relation losses jointly update the relation estimator. The affinity matrix is used directly, without discretization.

A.4 Copy Augmentation and Input Budgets

The original input contains at most 12 units. Natural examples contain at least two units, and controlled examples contain at least one. With probability $1/2$ the original input is retained. With the remaining probability, a unit is chosen uniformly and receives $c \in \{1, 2, 4\}$ additional copies, with each count equally likely. The resulting input contains at most 16 units. Copies preserve decoded visual content and semantic metadata, and all original units remain present.

Pair labels and their mask $\mathcal{P}$ are constructed for the sampled input. For controlled cards, copies inherit their source field maps and Equation (7) is applied to the resulting pairs. For natural evidence, source–copy and copy–copy pairs with verified equality receive positive labels; other pairs are included only when their equality is independently verified. We use every verified pair with $i < j$ and average the relation loss over those pairs. When no valid pair is present, the relation loss is zero.

One sampled input supplies Equation (6). Its fixed K-slot prefix leaves decoder context length unchanged under copying. Eligible examples retain their full question and target within the backbone context. Per-unit visual preprocessing is fixed before augmentation and applied identically to copies.

A.5 Inference and Computational Accounting

Inference follows the same sequence: independently encode the units and question, resample H_i, compute the symmetric affinities and row sums, fuse local slots, and decode from Z. Attention is evaluated using stable softmax after subtracting $\log m_i$ from each corresponding logit. Since $A_{ii} = 1$, no degree offset is added. Stochastic preprocessing and dropout are disabled. A single unit gives $m_1 = 1$. Empty inputs lie outside the method's domain and are rejected; nonfinite activations are reported as numerical failures.

The pair network costs $O(n^2dh)$, the fusion dot products and weighted sums cost $O(Knsd)$, and the key and value projections cost $O(nsd^2)$. Conditioning the global queries costs $O(d^2+Kd)$. The final readout costs $O(Kdd_{\mathrm{LM}})$. Local resampling costs $O(\sum_i L_i d^2 + s\sum_i L_i d + d^2)$, including shared local query construction. Feature projections add $O(\sum_i L_i^v d_v d + \sum_i L_i^m d_{\mathrm{LM}} d + d_{\mathrm{LM}} d)$ for the visual, metadata, and question features. Frozen visual encoding, metadata encoding, question encoding, and answer decoding contribute their own backbone costs. All backbone costs are included for uncached inference.

A dense implementation stores $O(n^2)$ affinities, $O(Kns)$ fusion attention entries, and $O(nsd + Kd)$ local and global slot representations, in addition to local-resampling activations and backbone memory. Fixing K controls the decoder's evidence-token budget, while total computation still grows with the evidence count.

B EXPERIMENTAL DETAILS

This appendix describes benchmark selection, evaluation controls, and implementation settings. The architecture and single-pass objective follow Appendices A.3 and A.4.

B.1 BENCHMARK SCOPE, SPLITS, AND NATIVE SCORING

Admissible evidence sets. The natural-image evaluations require that permuting units or inserting a verified artificial copy leaves the reference answer unchanged. We first apply a task-category allowlist, then screen questions, answer options, and metadata. We exclude presentation-index references, inter-unit temporal-order questions, input-cardinality questions, and tasks where repeated independent observations change the answer. Semantic identifiers already present in the task move with their units; benchmark identifiers and wording remain unchanged. Multi-image inputs contain 2–12 original units, and every retained unit stays intact. We manually adjudicate borderline cases before inspecting model outputs. Item identifiers, source groups, exclusion rules, and retained counts define each evaluation subset. All reported scores use these defined subsets.

Natural multi-image benchmarks. MuirBench (Wang et al., 2025) supplies 12 task categories and paired answerable/unanswerable examples. We use multiple-choice accuracy and retain source pairing when constructing splits. For MIBench (Liu et al., 2024), we retain eligible multi-image instruction and knowledge-seeking examples; its interleaved in-context demonstrations are outside the evidence-set interface. The score is circular accuracy: an item is correct only if all cyclic option placements are answered correctly. MMIU (Meng et al., 2025) uses the unweighted macro-average of accuracy over retained non-temporal tasks. MIRB (Zhao et al., 2024) retains its official answer normalization, scoring, and category aggregation for choice-based and free-form questions. We use native accuracy on eligible Mantis-Eval (Jiang et al., 2024) and multi-image BLINK (Fu et al., 2024) items. We retain native benchmark aggregation and report each metric separately.

Document retrieval and fusion. Clean document transfer uses the complete CVPR 2025 official test sets of DocHaystack and InfoHaystack (Chen et al., 2025): 109 and 155 unique QA questions, respectively, at each of the 100-, 200-, and 1,000-document pool sizes. The denominators remain fixed across seeds, retrieved documents, and repeated inference. No test questions are used for development. All readers share the same frozen V-RAG retrieval pipeline: the released CLIP/SigLIP/OpenCLIP ensemble retrieves 60 candidates, the frozen LLaVA-OneVision relevance filter screens them, and the top five retained documents are cached for every reader. If fewer are returned, all available documents are used; empty retrievals are scored as failures within the full test denominator. QA accuracy uses the official judge prompt and GPT-4o-mini judge, with an identical judge snapshot across methods. This evaluates reader-side fusion of shared retrieved evidence. Multi-image checkpoints transfer without document-specific fine-tuning.

Development isolation and contamination control. Official training, validation, and evaluation splits take precedence. If a released labeled evaluation split has no separate development split, a deterministic source-grouped 10% holdout is removed before tuning; every method is evaluated on the same remaining 90%. We explicitly name that reduced split. All sources already inspected in the preliminary study are reserved for development or pilot analysis and excluded from final held-out comparisons; the split manifest records their original benchmark identifiers. The 10% rule applies when no such development pool exists. Sources shared across items, answerable/unanswerable pairs, and all derived copies stay in the same partition. Training components sharing source identifiers or decoded-image hashes with any evaluation or development item are removed. Near-match screening and manual inspection address resized or re-encoded overlaps. Benchmark labels are used only for development selection or final scoring, never as training targets. We use official evaluators, with a pinned VLMEvalKit version where supported (Duan et al., 2024); prompts, parsing rules, and any judge calls are identical across methods.

B.2 PERTURBATIONS AND COMPLEMENTARY-INFORMATION DIAGNOSTICS

Same-input and permutation controls. The final-model robustness and mechanism diagnostics use Qwen2.5-VL-3B on the same eligible MuirBench subset as Table 2, together with the source- and template-disjoint controlled cards described below. Three-backbone coverage applies to the clean multi-image and document-transfer tables. Each MuirBench item is evaluated in its original

presentation, in an identical repeated invocation, and under five distinct nonidentity permutations sampled with seed 42, or all such permutations when fewer exist. Questions, option order, local preprocessing, and semantic metadata remain fixed. We report per-item worst-presentation accuracy R_π; the identical-input repeat provides a measured Flip and distribution-drift reference. The supplementary controls compare the matched-data attention-mask reader, PMA with a relation head, and QSEF (Table 7).

Duplication and length controls. A uniformly sampled unit receives $c \in \{1, 2, 4\}$ additional lossless copies, preserving decoded pixels and semantic metadata and retaining every original unit. Copies are appended, giving at most 16 units. We report R_{dup} separately for each c; Table 2 additionally reports the equally weighted means of R_{dup}, Flip, and token-level divergence. A fixed-length control contrasts four copies of one unit with four copies distributed as evenly as possible across units; ties in allocation use a fixed seeded draw shared by all methods. Both inputs contain $n + 4$ units. This control uses the subset on which the two allocations have exactly the same total processed visual-token count under the shared Qwen processor, before any fixed-prefix fusion. Its retained count is denoted N_{len}; the full MuirBench scores continue to use the complete eligible subset. Each allocation is compared with its own unchanged original input.

Task quality and distribution drift. Let $s_b(y, y^\star)$ be benchmark b's native item score, and let $a(\cdot)$ be its normalized answer. For a perturbation family $\mathcal{T}$, we use

$$\mathrm{Flip}_{\mathcal{T}} = \mathbb{E}_{E,T}\, \mathbf{1}[a(\hat{y}(E)) \neq a(\hat{y}(T(E)))], \tag{8}$$

$$\mathrm{RobustScore}_{\mathcal{T}} = \mathbb{E}_{E} \min_{T \in \mathcal{T} \cup \{\mathrm{id}\}} s_b(\hat{y}(T(E)), y^\star). \tag{9}$$

Here R_π uses the sampled permutations, whereas $R_{\mathrm{dup}}(c)$ uses the original input and its single appended-copy presentation at count c. The fixed-length control applies the same definition separately to each allocation. To detect drift hidden by greedy decoding, we compare full-vocabulary next-token distributions at identical teacher-forced reference-answer prefixes:

$$D_{\mathrm{JS}}(E, T(E)) = \frac{1}{T_y} \sum_{\tau=1}^{T_y} \mathrm{JS}(p_\theta(\cdot \mid y^\star_{<\tau}, E, q), p_\theta(\cdot \mid y^\star_{<\tau}, T(E), q)) \,. \tag{10}$$

Natural logarithms give scores in $[0, \log 2]$; tables display $10^3 D_{\mathrm{JS}}$. Reference-prefix comparisons stay within one backbone and tokenizer. These final-model diagnostics use token-level divergence; the preliminary study retains its separately defined candidate-answer divergence. These diagnostics use the existing inference path and add no consistency loss or second training branch.

Information preservation. On source- and template-disjoint ChartQA-derived cards (Masry et al., 2022), executable field maps define question-dependent equivalence as in Appendix A.1. We measure full-evidence accuracy A_{full} on the complementary-card partition, the gain G_{comp} when a withheld necessary card is restored, and joint correctness A_{edit} before and after a targeted fact edit with a recomputed answer. Edited facts change consistently in every affected copy. The complementary and edit-pair counts are N_{comp} and N_{edit}; a question pair is the evaluation unit for A_{edit}.

Relation AUROC/AUPRC use only pairs with field-map labels. False-positive rates are reported separately for complementary and partially overlapping pairs, with a threshold fixed on development data at 80% positive-pair recall. Question-switch pairs reuse the same cards for one shared-field question and one composition question with opposite relation labels. Switch accuracy is the fraction of these pairs whose two relation decisions are both correct under that fixed threshold. Each seed uses its own development-selected threshold without test adjustment. The labeled-pair, positive-pair, complementary-negative, partial-overlap-negative, and switch-pair counts accompany Table 6. Stability is interpreted alongside answer quality and fact sensitivity.

B.3 Comparison Groups and Fairness Controls

Native multi-image readers. Each of the three backbone checkpoints is evaluated directly and after matched-data LoRA supervised fine-tuning (Hu et al., 2021). SoFA (Tian et al., 2025), DTS (Lee et al., 2026), and FOCUS (Park et al., 2025) are applied to the matched-data LoRA checkpoints, as labeled in the result tables. SoFA adjusts inter-image attention; DTS rescales image-delimiter representations; FOCUS combines image-focused predictions with a noise-reference prediction. These

methods span CVPR 2025 (SoFA), ICLR 2026 (DTS), and Findings of ACL 2026 (attention masking); FOCUS is a preprint. The choices follow comparisons and backbone families used in their experimental tables. The masking intervention of Das et al. (2026) is trained from the same multi-image checkpoint as the other matched-data readers, restricting visual queries to their own image in the last half of decoder layers while preserving text attention. It is labeled a matched-data reimplementation: the original work uses a different initialization stage, training mixture, and rank-128 adaptation. This comparison attributes the measured change to the isolated intervention.

Controlled fusion baselines. All fusion controls use the same independent encoder, local resampler, $K = 64$ output prefix, decoder LoRA, and answer training. Deep Sets-style sum and mean pooling (Zaheer et al., 2017) and PointNet-style coordinatewise max pooling (Qi et al., 2017) pool a shared unit-wise projection of the local summaries and map the pooled vector to K output slots. They adapt these pooling operators to the fixed-prefix interface. Ordinary pooling by multihead attention (PMA) from Set Transformer (Lee et al., 2019) is instantiated here with the same single attention operation as QSEF and $m_i = 1$. A capacity-matched PMA control retains the relation head and its auxiliary labels while omitting the multiplicity correction, separating correction from extra supervision.

Exact-copy deduplication removes units only when decoded pixels and semantic metadata match, then uses PMA. Question-blind similarity deduplication forms connected components of a thresholded cosine-similarity graph over frozen mean visual features, with identical metadata required, and averages local slots within each component before PMA. This avoids choosing representatives by presentation position. Both are explicitly defined controls built on Set Transformer pooling (Lee et al., 2019). QSEF ablations remove relation supervision or copy augmentation one at a time. A question-blind relation ablation replaces its inputs by question-independent frozen per-unit summaries with learned projections, while retaining question-conditioned answer fusion; its relation estimator therefore remains question-blind. The Qwen cost comparison includes each fusion operator under the same interface.

Training and inference budgets. Within a backbone, the main trainable comparisons share source examples, answer targets, sample order, copy augmentation, LoRA rank, optimizer schedule, and update count. The no-augmentation ablation changes only its named intervention. Relation supervision is used by models with a relation head, except in the explicitly named no-supervision ablation; auxiliary-head PMA controls for its contribution. Fusion comparisons additionally fix s and K, except in the one-factor sensitivity analysis. Native readers retain their native visual-token interfaces, and FOCUS incurs $n+1$ branches per prediction step. The primary-backbone cost table reports task score and measured cost separately while preserving each interface's native compute. We report original published scores separately from these restricted-subset measurements.

B.4 Backbones, Optimization, and Reproducibility

Checkpoint and preprocessing alignment. The checkpoints are Qwen2.5-VL-3B-Instruct (Bai et al., 2025), LLaVA-OneVision-Qwen2-7B-OV (Li et al., 2025), and InternVL3-8B (Zhu et al., 2025). Their families overlap with the DTS and FOCUS comparisons, while LLaVA-OneVision also supports the attention-masking comparison. Qwen2.5-VL is the primary development backbone; the other two test transfer of the fusion design. Each uses its released tokenizer, visual encoder, and pretrained visual projector or merger. Within a backbone, a fixed independent per-unit processor configuration is shared by all methods and remains unchanged under permutation or copying. The per-unit limits are 256–1,280 merged visual tokens for Qwen2.5-VL (pixel bounds $256 \cdot 28^2$ and $1280 \cdot 28^2$), at most four 448×448 tiles plus the standard thumbnail for InternVL3, and the released `anyres_max_9` preprocessing/merging policy for LLaVA-OneVision. These settings are common within each backbone; token counts vary across model families. There is no image-count-dependent downsampling or position-dependent image splitting. Processor settings, resulting token counts, checkpoint revisions, and attention backend are recorded before evaluation. We admit only examples whose complete questions and evidence fit every compared reader in that backbone group, using the same retained items for every compared method. All methods use greedy decoding, a 256-token answer limit, and identical benchmark prompts. Outputs that reach the limit are scored as generated under the same evaluator.

Architecture and objective. We retain $d = 512$, $s = 16$, $K = 64$, relation width 64, $A_{ii} = 1$, and correction $-\log m_i$, with no degree offset. Language LoRA targets query and value projections in every decoder layer, with rank 16, scaling 16, and zero dropout. The frozen question and metadata passes disable these adapters. Only answer decoding activates them. The decoder receives visual information exclusively through the fused prefix. Both mean-normalized losses have coefficient one, and each example supplies one sampled input and one answer pass. Pairs lacking verified equality are masked and contribute no negative labels. Copy augmentation follows Appendix A.4 exactly.

Data and optimization. After source grouping and contamination removal, the training pool contains up to 25,000 eligible Mantis-Instruct examples (Jiang et al., 2024) and up to 25,000 controlled examples. Training samples the two pools with probability $1/2$ each, independently of their retained sizes. Resource calibration uses at most 10,000 unique development examples. Optimization uses 5,000 AdamW updates with $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$, weight decay 0.01 on non-bias/non-normalization parameters, gradient clipping at norm 1, linear warmup for 150 updates, and cosine decay. The initial learning rates are 10^{-4} for new fusion parameters and 2×10^{-5} for LoRA. Effective batch size is 32: one example per GPU, four GPUs, and eight accumulation steps. Training uses BF16, activation checkpointing, and optimizer-state sharding on $4\times$ NVIDIA A100 80 GB. Resource-driven configuration changes apply to all matched comparisons and are fixed before evaluation; the cost diagnostic uses the single-GPU configuration specified below.

Development selection and ablations. Hyperparameters are selected on development data only. Baseline selection maximizes mean development task score within a backbone. We allow at most 12 development trials per inference intervention: SoFA uses $\sigma \in \{0, 0.25, 0.5, 0.75, 1\}$ with an every-two-layer schedule; DTS considers scale factors $\{1, 2, 4, 8\}$ over the first $\{1, 2, 4\}$ layers; FOCUS includes its reported $(\lambda, \alpha) = (0.3, 0.4)$ setting plus 11 seeded trials over its stated search ranges. The unmodified reader is always an available fallback. Similarity deduplication uses thresholds $\{0.8, 0.9, 0.95, 0.99\}$ selected by task score. A QSEF checkpoint is selected by development task score, breaking ties by smaller duplicate-induced divergence. Checkpoints are evaluated every 500 updates; the candidate schedule and selection rule are shared across trainable comparisons and seeds. Mechanism ablations and one-factor sensitivity use Qwen2.5-VL-3B, the fixed MuirBench subset, and the same controlled-card partitions. The default is $(K, s) = (64, 16)$; the four additional configurations are $(32, 16)$, $(128, 16)$, $(64, 8)$, and $(64, 32)$. Together with the three mechanism ablations, they define seven additional training configurations, each following the same three-seed protocol.

Uncertainty, cost, and reporting. Trainable models use seeds 42, 43, and 44. For item-level accuracy, let c_s be the integer number of correctly scored items in seed s and N the fixed number of evaluated items. We compute

$$a_s = 100\frac{c_s}{N}, \qquad \bar{a} = \frac{1}{3}\sum_{s=1}^{3} a_s,$$
$$\mathrm{SD} = \sqrt{\frac{\sum_{s=1}^{3}(a_s - \bar{a})^2}{3-1}}. \tag{11}$$

Each seed's score is computed at full precision; the mean and sample standard deviation ($\mathrm{ddof} = 1$) are calculated from the unrounded scores and rounded only once to one decimal place, using round-to-nearest with ties to even. Native deterministic accuracies are computed directly from integer correct counts and fixed denominators, and rounded once. For task/category aggregates and other metrics, we first compute each seed's benchmark-defined score at full precision, then apply the same mean, sample-SD, and final-rounding procedure; aggregation uses the benchmark-defined score with seed-level averaging. All rounding occurs once at the final reporting step. In result tables, sample SD is displayed only for the complete QSEF configuration; other trainable rows show means.

Paired uncertainty is restricted to the QSEF–PMA comparisons for multi-image and document-transfer accuracy (Table 8). We use 2,000 paired resamples of source groups and percentile 95% intervals for QSEF minus PMA, recomputing the benchmark-native score for each seed before averaging the three seed scores. Copies, paired questions, and all seed predictions remain attached to the same source group; seeds do not multiply the number of independent test questions.

The cost diagnostic uses Qwen2.5-VL-3B on 200 fixed MuirBench queries shared by all methods, after 20 warmup queries. Uncached end-to-end timing uses batch size one on a single A100 80 GB,

including frozen image, metadata and question encoding, relation estimation where present, all FOCUS branches, and answer generation. We report median and 95th-percentile latency, peak allocated GPU memory, mean processed visual-token count, mean language-model evidence-token count, and trainable parameter count. A FOCUS latency includes all branches; its evidence-token column sums the branch input counts. For fixed-prefix methods, the language-model evidence count is $K = 64$; upstream encoding and fusion costs remain included.

B.5 Reading the Main Results

In Tables 1 and 2, SoFA, DTS, and FOCUS operate on the matched-data LoRA reader. Attention masking is a matched-data reimplementation. PMA and QSEF share the 64-token prefix, training data, and updates. Native-token and fixed-prefix readers have different inference costs, as detailed in Appendix B.3.

Table 2 uses the eligible MuirBench subset for its natural-image metrics. Clean denotes unperturbed accuracy. R_π is the per-item worst-presentation accuracy over the sampled permutations and the original input. For each copy count $c \in \{1, 2, 4\}$, R_{dup} compares the original input with the appended-copy input; the three scores are averaged equally. Flip and D_{JS} use the same appended-copy interventions and average over the same counts. D_{JS} is the conditional next-token divergence in Appendix B.2, reported in 10^{-3} nats.

Retained counts for Table 1. To keep the main table compact, the retained item counts are reported here in the column order (MuirBench, MIBench, MMIU, MIRB, Mantis-Eval, BLINK). They are $(1476, 2418, 1108, 714, 194, 1052)$ for Qwen2.5-VL-3B-Instruct, $(1452, 2356, 1086, 702, 190, 1024)$ for LLaVA-OneVision-Qwen2-7B-OV, and $(1464, 2397, 1096, 708, 192, 1041)$ for InternVL3-8B.

For controlled cards, G_{comp} is the change in accuracy when a necessary complementary card is restored. A_{edit} is the fraction of pairs for which both the original answer and the recomputed answer after a fact edit are correct. These columns use source- and template-disjoint cards, separately from the MuirBench items.

The fixed-prefix controls adapt Deep Sets pooling (Zaheer et al., 2017), PointNet-style coordinate-wise maxima (Qi et al., 2017), and Set Transformer PMA (Lee et al., 2019), as specified in Appendix B.3. PMA with a relation head retains the auxiliary supervision but omits multiplicity correction. In the main tables, shading identifies QSEF; boldface and underline denote the best and second-best displayed means within each backbone or metric column, including ties. Native rows use one deterministic greedy-decoding evaluation of the fixed checkpoint, without seed averaging. For Mantis-Eval, each retained question contributes once: accuracy is $100\, N_{\mathrm{correct}}/N$. The Native counts are 116/194 for Qwen2.5-VL-3B, 118/190 for LLaVA-OneVision-7B, and 129/192 for InternVL3-8B. Other benchmarks retain their specified scoring and aggregation rules. Trainable rows report means across seeds 42, 43, and 44; only the complete QSEF configuration additionally displays $\pm$ sample standard deviation ($\mathrm{ddof} = 1$). All statistics use full-precision calculation and a single final rounding as specified in Equation (11). Method citations identify the underlying approaches; their table entries are measurements under our matched protocol. Differences discussed in the main text are computed from the displayed means and expressed in percentage points.

Interpreting the comparisons. The eligible item counts differ across backbones, and Table 1 reports method comparisons within each backbone. QSEF's gains over the strongest competing row range from 1.6–2.7 points for Qwen2.5-VL-3B, 1.4–2.2 for LLaVA-OneVision-7B, and 1.4–2.2 for InternVL3-8B. The benchmark-specific metrics are reported separately.

Table 2 exposes a stability–utility distinction: exact-copy and similarity deduplication both have zero displayed duplicate drift, yet their fact-edit accuracies are 55.7% and 51.0%, compared with 63.0% for QSEF. QSEF retains a small nonzero duplicate-induced drift, consistent with its estimated soft relation. The PMA-with-relation-head comparison controls for auxiliary supervision when assessing multiplicity correction.

Document-transfer figure. Figure 4 compares Native, LoRA (Hu et al., 2021), LoRA+FOCUS (Park et al., 2025), PMA (Lee et al., 2019), and QSEF. The columns use Qwen2.5-VL-3B-Instruct (Bai et al., 2025), LLaVA-OneVision-Qwen2-7B-OV (Li et al., 2025), and InternVL3-8B (Zhu et al., 2025), with independently scaled y-axes that include all methods and dis-

played error bars. Native uses a single deterministic evaluation; trainable methods show three-seed means. QSEF's error bars represent sample SD across seeds. Trained readers reuse their multi-image checkpoints without document-specific tuning; retrieved evidence is shared within each backbone and pool-size condition. Every condition uses the complete official test set: 109 unique questions for DocHaystack and 155 for InfoHaystack. Pool size changes the retrieval candidate collection; the question counts remain fixed across pool sizes, seeds, and repeated inference. Document transfer evaluates clean QA; final-model perturbation diagnostics remain restricted to Qwen/MuirBench.

B.6 Preliminary Study Protocol

B.6.1 Evaluation Units and Eligibility

The natural-example audit uses unmodified Qwen2.5-VL-3B-Instruct and LLaVA-OneVision-7B, without retraining, relation estimation, deduplication, or the proposed fusion module. Both models evaluate the same 400 examples from each benchmark. Thus, the audit contains 800 distinct benchmark examples and 1,600 model–example pairs; the latter are repeated model evaluations, not 1,600 distinct examples. Each model/dataset block in Figure 3a contains 400 rows.

Eligibility requires that artificial copies leave the target answer unchanged. We exclude questions that depend on image indices, evidence cardinality, temporal order, or presentation order, as well as inputs that would undergo context truncation after intervention. Every original unit remains present. The retained inputs contain two to six units (median four) in both benchmarks. Their counts for $n = 2, 3, 4, 5, 6$ are $(85, 80, 78, 74, 83)$ for MuirBench and $(76, 80, 77, 79, 88)$ for MIBench. Sample identifiers and unit counts match across models within each benchmark.

B.6.2 Interventions and Response Measures

For an input $E = \{e_1, \ldots, e_n\}$, all transformations preserve its underlying information. *Permutation* changes the outer presentation order, with the protocol averaging over up to five non-identity permutations. *Balanced* adds one copy of each unit. With the same budget $B = n$, *Concentrated* adds all copies to one uniformly selected target e_j. Both duplication conditions contain $2n$ units, but differ in their allocation of occurrences. This control matches the number of evidence units; tokenized sequence lengths can still differ. The *Target* $+c$ conditions add $c \in \{1, 2, 4\}$ exact copies of the same target and contain $n + c$ units. Copy positions are balanced across the sequence. Consequently, the retained inputs contain four to twelve units in the Balanced/Concentrated comparison and at most ten units under Target $+4$.

Let P_E denote the distribution over candidate answers $\mathcal{Y}$ obtained by normalizing their log-likelihoods for a fixed question q. For a transformation T, define

$$\begin{aligned} \Delta_{\mathrm{JS}}(E, T) &= D_{\mathrm{JS}}\big(P_E \,\|\, P_{T(E)}\big)\,, \\ \mathrm{Flip}(E, T) &= \mathbf{1}\Big[\arg\max_y P_E(y) \neq \arg\max_y P_{T(E)}(y)\Big]\,. \end{aligned} \tag{12}$$

Here $D_{\mathrm{JS}}(P\|Q) = \frac{1}{2}D_{\mathrm{KL}}(P\|M) + \frac{1}{2}D_{\mathrm{KL}}(Q\|M)$ with $M = (P + Q)/2$. Drift measures a change in the full answer distribution, whereas a flip records a change in its highest-probability answer. Accuracy is evaluated separately because a flip can correct an initially wrong answer or spoil a correct one. All comparisons use the original input as the reference for that same example and question.

B.6.3 Question-Switch Pairs and Diagnostic Axes

The controlled audit contains 220 held-out two-card pairs. A pair displays (s, a) on one card and (s, b) on the other, where s is shared and a, b are distinct fields. A *shared-field* question requires only s, making the cards substitutable for the requested information. A *composition* question requires an operation involving both a and b, making the cards complementary. The images remain identical between these two questions; the required information changes. Each pair contributes one response under each question type, giving 440 records. The reported drift and flip indicators use Target $+4$. This diagnostic uses the original model and card pair, with no learned relation classifier or proposed fusion mechanism.

The horizontal axis measures how sharply the duplicated unit alone favors an answer:

$$D_j(q) = 1 - \frac{H(P_{\{e_j\}})}{\log |\mathcal{Y}|}, \tag{13}$$

where H is entropy and all distributions condition on the same question. For at least two answer options, $D_j \in [0, 1]$: a uniform distribution gives zero, and a point mass gives one. High decisiveness can accompany an incorrect preferred answer.

The vertical axis measures the positive gain in probability assigned to the correct answer when all evidence is available:

$$G_{\text{int}}(q) = \left[P_E(y^\star) - \max_i P_{\{e_i\}}(y^\star) \right]_+, \tag{14}$$

where $y^\star$ is the ground-truth answer and $[x]_+ = \max(x, 0)$. The single-unit distributions are evaluated with each unit presented alone. Thus $G_{\text{int}} \in [0, 1]$, and a positive value indicates that joint evidence supports the correct answer more strongly than the best individual unit. Both axes describe responses to the original evidence, while color describes the subsequent duplication response.

C Additional Experimental Results

C.1 Preliminary Study Results

The natural audit contains 9,600 records: six conditions for each of 1,600 model–example pairs. We pool observations across the four equally sized blocks and compute the median drift and fraction of flips directly. Controlled questions are summarized separately over the same 220 pairs (Table 3).

Table 3: Aggregate duplication-audit results. N counts model–example pairs for natural inputs and card pairs for each question type. The last column gives count (percentage). Controlled questions use Target $+4$.

Audit	Condition	N	Median Δ_{JS}	Flips (%)
Natural	Permutation	1,600	0.0075	18 (1.12)
	Balanced	1,600	0.0160	32 (2.00)
	Concentrated	1,600	0.0319	74 (4.62)
	Target $+1$	1,600	0.0139	26 (1.62)
	Target $+2$	1,600	0.0236	33 (2.06)
	Target $+4$	1,600	0.0403	142 (8.88)
Controlled	Shared field	220	0.0140	4 (1.82)
	Composition	220	0.0590	17 (7.73)

In Figure 3a, rows within each block are sorted by increasing Target $+4$ drift, breaking ties by sample identifier; this order is shared across columns. Perm., Bal., Conc., and $+c$ denote Permutation, Balanced, Concentrated, and Target $+c$. The JS and Flip rows report pooled median drift and flip percentage. In panel (b), each segment links the shared-field and composition questions for one card pair. White dots in (a) and outer rings in (b) mark individual flips.

The shared color scale spans 0–0.10 and includes every observed value without clipping. Power normalization with exponent 0.72 makes small drifts visible; ticks retain the original JS units. Before plotting, we check finite values, complete conditions, unique sample/condition records, and matching sample identifiers across models and question pairs.

C.2 Targeted Ablations and Diagnostics

The mechanism, sensitivity, relation, robustness, and cost diagnostics use Qwen2.5-VL-3B. Clean-task confidence intervals cover all three backbones. Shared reference entries repeat the corresponding main-result measurements. Trainable performance entries show means over seeds 42/43/44. Only the complete QSEF configuration displays $\pm$ sample SD ($\mathrm{ddof} = 1$); all statistics are rounded once from full precision.

Mechanism and capacity. Table 4 isolates relation supervision, copy augmentation, and question conditioning of the relation estimator. PMA with a relation head supplies the no-multiplicity-correction control. Table 5 varies one slot budget at a time. Both use the same 1,476 MuirBench questions and card partitions as Table 2. The complementary-card count is $N_{\mathrm{comp}} = 800$ and the fact-edit pair count is $N_{\mathrm{edit}} = 600$; neither count is multiplied by the three seeds.

Table 4: **Mechanism ablations.** Qwen2.5-VL-3B with $K = 64$, $s = 16$. Scores are percentages, except G_{comp} in percentage points. Shading identifies the complete QSEF method.

	MuirBench			Controlled cards	
Method	Clean ↑	R_{dup} ↑	Flip ↓	G_{comp} ↑	A_{edit} ↑
PMA (Lee et al., 2019)	48.5	45.7	5.0	17.7	55.9
PMA + relation head (Lee et al., 2019)	49.1	46.4	4.7	19.0	57.8
QSEF w/o relation supervision	49.8	48.3	2.6	19.6	58.9
QSEF w/o copy augmentation	50.6	48.9	2.3	21.5	61.7
QSEF with question-blind relations	50.1	49.3	1.5	18.3	56.9
QSEF (ours)	51.2 $\pm$ 0.2	50.8 $\pm$ 0.3	0.7 $\pm$ 0.1	22.2 $\pm$ 0.4	63.0 $\pm$ 0.5

Question-dependent relations. Table 6 evaluates relation heads and full-evidence answer quality. Switch denotes joint correctness of the two opposite-label relation decisions for the same card pair.

Table 5: **Slot-budget sensitivity.** Qwen2.5-VL-3B; all rows use QSEF. Only the indicated slot budget changes. Scores are percentages, except G_{comp} in percentage points.

Configuration	K	s	Clean ↑	R_{dup} ↑	G_{comp} ↑	A_{edit} ↑
QSEF: fewer global slots	32	16	50.4	50.0	20.7	60.8
QSEF: more global slots	128	16	51.3	50.9	22.4	63.2
QSEF: fewer local slots	64	8	50.7	50.2	21.0	61.4
QSEF: more local slots	64	32	51.1	50.7	22.1	62.8
QSEF (default)	64	16	51.2 ± 0.2	50.8 ± 0.3	22.2 ± 0.4	63.0 ± 0.5

All relation rates use a development-selected threshold; AUROC and AUPRC are threshold-free. The labeled-pair and positive-pair counts are $N_{\text{rel}} = 12{,}480$ and $N_{+} = 4{,}160$. Complementary-negative and partial-overlap-negative counts are $N_{\text{rel,comp}} = 2{,}880$ and $N_{\text{rel,part}} = 2{,}560$; the question-switch count is $N_{\text{switch}} = 640$. A_{full} uses the N_{comp} complementary-card questions above.

Table 6: **Question-dependent relation estimation and answer quality.** Qwen2.5-VL-3B on controlled cards. All entries, including AUROC/AUPRC, use the 0–100 scale.

	Relation ranking		False positives		Question / answer	
Method	AUROC ↑	AUPRC ↑	Comp. ↓	Partial ↓	Switch ↑	A_{full} ↑
PMA + relation head (Lee et al., 2019)	93.1	91.0	10.8	16.2	84.6	70.9
QSEF w/o relation supervision	86.2	82.8	18.9	26.1	67.4	71.6
QSEF with question-blind relations	84.7	81.5	22.7	31.4	0.0	69.2
QSEF (ours)	96.0 ± 0.4	94.4 ± 0.5	6.4 ± 0.8	10.9 ± 0.9	91.2 ± 1.0	75.4 ± 0.5

Multiplicity and matched length. Table 7 separates identical-input drift, permutation robustness, and copy strength on the full MuirBench subset. Its second panel uses $N_{\text{len}} = 436$ questions with equal total processed visual-token counts in the two $n + 4$ allocations. The mean augmented visual-token count in that subset is 3,728. "Balanced" distributes the four copies as evenly as possible; "concentrated" assigns all four to one unit. Every metric compares the indicated input with the same question's original evidence.

Table 7: **Copy strength and length controls.** Qwen2.5-VL-3B. Accuracy and Flip use percentages; JS columns show $10^3 D_{\text{JS}}$. Panels use distinct, explicitly identified evaluation sets.

(a) Full MuirBench subset: $N = 1{,}476$

	Identical-input repeat		Permutation	Appended-copy R_{dup} ↑		
Method	Flip ↓	JS ↓	R_{π} ↑	$c = 1$	$c = 2$	$c = 4$
LoRA + attn. mask (Das et al., 2026)	0.0	0.0	45.2	46.4	45.9	45.1
PMA + relation head (Lee et al., 2019)	0.0	0.0	49.1	47.5	46.5	45.2
QSEF (ours)	0.0 ± 0.0	0.0 ± 0.0	51.2 ± 0.3	51.0 ± 0.3	50.8 ± 0.3	50.6 ± 0.3

(b) Equal-unit and equal-token subset: $N = 436$

	Balanced: $n + 4$ units			Concentrated: $n + 4$ units		
Method	Robust ↑	Flip ↓	JS ↓	Robust ↑	Flip ↓	JS ↓
LoRA + attn. mask (Das et al., 2026)	46.1	2.8	3.7	44.7	4.9	6.2
PMA + relation head (Lee et al., 2019)	48.4	2.5	3.1	45.5	5.2	6.5
QSEF (ours)	51.0 ± 0.3	0.4 ± 0.1	0.5 ± 0.1	50.6 ± 0.3	0.9 ± 0.2	1.0 ± 0.2

Paired task differences. Table 8 reports QSEF minus the same-backbone PMA baseline on identical questions. Each entry is the difference in percentage points followed by its 95% source-group paired bootstrap interval. Point differences are computed from unrounded seed scores; displayed one-decimal scores are not bootstrap inputs. The original benchmark scoring and the complete document-test denominators remain unchanged.

Table 8: **Paired clean-task differences: QSEF (ours) minus PMA (Lee et al., 2019).** Entries are Δ [95% CI] in percentage points; each interval uses 2,000 source-group resamples.

(a) Eligible multi-image benchmarks

Backbone	MuirBench	MIBench	MMIU	MIRB	Mantis-Eval	BLINK
Qwen2.5-VL-3B (Bai et al., 2025)	2.7 [1.7, 3.7]	2.1 [1.4, 2.9]	1.6 [0.6, 2.6]	1.6 [0.2, 3.0]	1.9 [-1.0, 4.8]	2.1 [1.0, 3.3]
LLaVA-OneVision-7B (Li et al., 2025)	2.2 [1.3, 3.2]	2.1 [1.3, 2.9]	1.7 [0.7, 2.7]	1.7 [0.3, 3.1]	1.6 [-1.3, 4.6]	1.7 [0.6, 2.9]
InternVL3-8B (Zhu et al., 2025)	2.2 [1.3, 3.1]	2.1 [1.3, 2.9]	1.4 [0.4, 2.4]	1.5 [0.2, 2.8]	1.6 [-1.2, 4.4]	1.5 [0.4, 2.7]

(b) Complete official document test sets

	DocHaystack ($N = 109$)			InfoHaystack ($N = 155$)		
Backbone	100	200	1,000	100	200	1,000
Qwen2.5-VL-3B (Bai et al., 2025)	2.4 [-0.5, 5.4]	2.4 [-0.4, 5.5]	2.1 [-0.8, 5.0]	2.6 [0.2, 5.1]	2.4 [-0.1, 4.8]	2.2 [-0.3, 4.8]
LLaVA-OneVision-7B (Li et al., 2025)	1.5 [-1.4, 4.5]	2.4 [-0.5, 5.4]	2.4 [-0.4, 5.5]	1.5 [-0.9, 3.9]	1.5 [-0.9, 3.9]	2.6 [0.1, 5.2]
InternVL3-8B (Zhu et al., 2025)	2.4 [-0.5, 5.3]	1.8 [-1.0, 4.8]	3.4 [0.4, 6.4]	1.7 [-0.6, 4.2]	1.3 [-1.1, 3.7]	2.8 [0.2, 5.4]

Measured cost. Table 9 uses the same 200 MuirBench queries for every Qwen method. Enc. tokens counts the processed visual tokens before fusion; LM tokens counts the evidence tokens consumed by the language model, summing across FOCUS branches. Both token columns report means over queries. Latency percentiles summarize queries, not training seeds. For each trainable method, timing uses its seed-42 checkpoint with fixed generation settings; parameters are in millions and memory is peak allocated GPU memory in GiB. The other seed checkpoints contribute to the performance statistics, not to the latency sample size. Native has no trainable parameters; fixed-prefix methods use 64 evidence tokens.

Table 9: **Uncached end-to-end cost.** Qwen2.5-VL-3B on one A100 80 GB, batch size one, 20 warmup queries and 200 measured queries. The accompanying task scores are in Tables 1 and 2.

	Mean token count		Trainable	Latency (s) ↓		Peak GPU
Method	Enc.	LM	params (M)	Median	p95	memory (GiB)
Native	1,936	1,936	0	2.16	3.28	15.6
LoRA (Hu et al., 2021)	1,936	1,936	3.8	2.18	3.31	15.7
LoRA + SoFA (Tian et al., 2025)	1,936	1,936	3.8	2.24	3.39	15.8
LoRA + DTS (Lee et al., 2026)	1,936	1,936	3.8	2.21	3.36	15.8
LoRA + FOCUS (Park et al., 2025)	1,936	10,418	3.8	10.61	15.74	17.2
LoRA + attn. mask (Das et al., 2026)	1,936	1,936	3.8	2.27	3.45	16.0
Deep Sets (sum) (Zaheer et al., 2017)	1,936	64	8.9	1.59	2.44	11.8
Deep Sets (mean) (Zaheer et al., 2017)	1,936	64	8.9	1.60	2.45	11.8
Max pooling (Qi et al., 2017)	1,936	64	8.9	1.60	2.46	11.8
PMA (Lee et al., 2019)	1,936	64	9.3	1.67	2.55	12.1
PMA + relation head (Lee et al., 2019)	1,936	64	9.4	1.74	2.66	12.3
Exact-copy + PMA (Lee et al., 2019)	1,936	64	9.3	1.69	2.58	12.1
Similarity + PMA (Lee et al., 2019)	1,936	64	9.3	1.75	2.69	12.4
QSEF (ours)	1,936	64	9.4	1.76	2.68	12.3

C.3 Information-Class Mass Conservation: Protocol and Results

Controlled duplication. We use one Qwen2.5-VL-3B QSEF checkpoint ($K = 64$, $s = 16$) on 800 source- and template-disjoint two-card trials. Both cards contain question-relevant information. The target card $C^\star$ appears $\nu \in \{1, 2, 3, 5, 8, 12\}$ times; its content and semantic metadata, the complementary card, question, and answer remain fixed. Training adds $1, 2$, or 4 copies (total multiplicities

$2, 3, 5$); 8 and 12 test extrapolation. The largest presentation contains 13 units, within the 16-unit budget.

Intervention and metric. For query r, local slot t, and occurrence i, let $a_{rit} = u_r^\top k_{it}/\sqrt{d}$. The paired intervention holds query–key scores and learned parameters fixed:

$$\omega_{rit}^{\mathrm{QSEF}} = \frac{\exp(a_{rit} - \log m_i)}{\sum_{j,t'} \exp(a_{rjt'} - \log m_j)}, \qquad \omega_{rit}^{\mathrm{off}} = \frac{\exp(a_{rit})}{\sum_{j,t'} \exp(a_{rjt'})}. \tag{15}$$

Setting $m_i \equiv 1$ disables correction at inference; this changes inference only and trains no separate PMA baseline. For each intervention, define

$$M_r^{(\nu)}(C^\star) = \sum_{i \in C^\star} \sum_{t=1}^{s} \omega_{rit}^{(\nu)}, \qquad E_{\mathrm{mass},r}^{(\nu)} = \left| \log_2 \frac{M_r^{(\nu)}(C^\star)}{M_r^{(1)}(C^\star)} \right|. \tag{16}$$

Each intervention uses its own single-copy reference; these coincide here. Zero error indicates conservation; one bit indicates a factor-of-two change in either direction. The main text abbreviates this metric as $E_r^{(\nu)}$.

Aggregation and visualization. Table 10 summarizes 51,200 unsmoothed question–query observations per condition (800×64). Queries and repeated presentations derive from the same 800 question-level trials and seeds. The 95.3% reduction uses the unrounded overall medians, $100(1 - 0.0799574/1.7098951)$.

Figure 5 uses 24 fixed single-copy mass-percentile bins (2,133–2,134 observations each). Baseline attention mass serves as a model-internal stratifier. Displayed bin medians are Gaussian-smoothed along this axis only ($\sigma = 0.55$ bins; nearest-edge handling); contours use the same matrix. Both panels share a power-normalized color scale (exponent 0.43; 0–3 bits with an upper extension). Multiplicities are discrete, equally spaced columns. Reported statistics use no smoothing.

Table 10: **Class-mass conservation error (bits).** Unsmoothed median and 90th percentile over 51,200 observations per condition, rounded to three decimal places.

	Correction off		QSEF	
Multiplicity ν	Median	90th percentile	Median	90th percentile
1	0.000	0.000	0.000	0.000
2	0.687	0.915	0.016	0.027
3	1.014	1.419	0.025	0.042
5	1.344	2.008	0.039	0.064
8	1.568	2.488	0.057	0.092
12	1.710	2.846	0.080	0.127

Interpretation. Each bin's median error increases with multiplicity in both interventions. Without correction, distortion decreases with baseline mass. At $\nu = 12$, QSEF's unsmoothed bin medians span 0.021–0.110 bits, versus 0.299–3.234 without correction. Residual drift is compatible with imperfect soft affinities. The intervention supports approximate conservation under controlled duplication. Table 2 supplies separate downstream evidence.

C.4 Empirical Scaling Consistent with the Stability Envelope

We test whether the quantities motivated by Theorem 1 empirically track fusion displacement on 800 held-out complementary-card questions, three QSEF checkpoints, and $c \in \{1, 2, 4\}$ appended copies, yielding 7,200 observations. We compare per-query fusion displacement $\Delta_F = \|F(E^{(c)}) - F(E)\|_{\mathrm{F}}/\sqrt{K}$ with the recorded product $B_{\mathrm{T1}} = \bar{D}\widehat{\epsilon}$, where $\bar{D} = D/\sqrt{K}$ is the normalized class-value diameter and $\widehat{\epsilon}$ is the maximum relative multiplicity error across the pair.

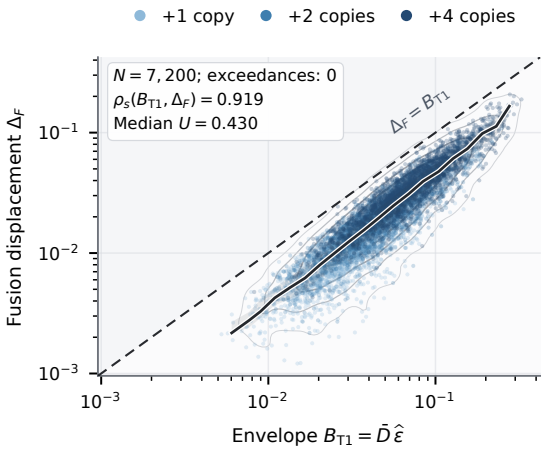


Figure 7: **Empirical scaling of the stability envelope.** Fusion displacement versus the recorded product (log axes). Dashed line: equality; solid curve and band: binned median and interquartile range. Gray contours show observation density.

Across all 7,200 observations, the recorded product exceeds the observed displacement (Figure 7). The product has a stronger Spearman association with displacement ($\rho_s = 0.919$) than either $\widehat{\epsilon}$ (0.658) or $\bar{D}$ (0.674). Utilization $U = \Delta_F/B_{\mathrm{T1}}$ has median 0.430 and 90th percentile 0.592. Its median rises from 0.370 to 0.430 and 0.492 for $c = 1, 2, 4$, respectively; this is an aggregate trend across copy strengths.

The recorded product tracks displacement strongly and remains conservative on this sample. Since the logged $\bar{D}$ varies across copy strengths, this appendix presents B_{T1} as an empirical scaling result that complements the theorem's exact guarantee under its fixed-representation premise. The definitions and diagnostics below give the complete protocol and geometry analysis.

The data establish B_{T1} as a strong empirical diagnostic of fusion drift: it tracks displacement well and remains conservative here. The theorem supplies the corresponding exact guarantee under its fixed-representation geometry conditions.

Protocol and quantities. We evaluate 800 held-out complementary-card questions with three QSEF checkpoints (seeds 42, 43, and 44), appending $c \in \{1, 2, 4\}$ copies of a necessary evidence unit. Each copy strength has 2,400 observations. The endpoint is pre-readout fusion displacement. For a presentation pair $(E, E^{(c)})$, define

$$\widehat{\epsilon} = \max_{\eta \in \{E, E^{(c)}\}} \max_{i \in \eta} \left| \frac{m_i(\eta)}{\nu_{c(i)}(\eta)} - 1 \right|, \qquad m_i(\eta) = \sum_j A_{ij}(\eta), \tag{17}$$

Class counts follow the controlled-card field maps; $c(i)$ indexes classes and c counts appended copies. Using the relevance-weighted class values V_{rj} in Equation (22), the theorem defines $D = (\sum_r \max_{j,\ell} \|V_{rj} - V_{r\ell}\|_2^2)^{1/2}$. The normalized quantities are

$$\bar{D} = \frac{D}{\sqrt{K}}, \qquad \Delta_F = \frac{\|F(E^{(c)}) - F(E)\|_{\mathrm{F}}}{\sqrt{K}}, \qquad B_{\mathrm{T1}} = \bar{D}\widehat{\epsilon}, \qquad U = \frac{\Delta_F}{B_{\mathrm{T1}}}. \tag{18}$$

All recorded $B_{\mathrm{T1}} > 0$ and $0 < \widehat{\epsilon} < 0.119$. The stored products and ratios agree with the recorded factors. Under the theorem's fixed-representation premise, Theorem 1 guarantees $\Delta_F \le B_{\mathrm{T1}}$ when $\widehat{\epsilon} < 1$. The analysis below evaluates the corresponding empirical scaling and reports the representation geometry.

Stratified results. Pooled Spearman correlations with Δ_F are 0.91855, 0.65774, and 0.67405 for B_{T1}, $\widehat{\epsilon}$, and $\bar{D}$, respectively. The product's correlations are 0.91605, 0.92156, and 0.91773 across seeds, and 0.87210, 0.90204, and 0.91651 within $c = 1, 2, 4$. In each copy condition, the product also exceeds both individual factors in rank correlation. Thus, the association persists within copy

Table 11: **Statistics by appended copies.** Each row pools 2,400 observations. All columns are medians except the final column, which reports the 90th percentile of utilization. Values use unsmoothed records.

Copies c	$\widehat{\epsilon}$	$\bar{D}$	B_{T1}	Δ_F	U	$P_{90}(U)$
1	0.0374	0.9176	0.0338	0.0120	0.3702	0.5188
2	0.0523	0.9192	0.0479	0.0200	0.4296	0.5800
4	0.0750	0.9187	0.0675	0.0324	0.4922	0.6438

strengths. Across all 7,200 observations, no observed displacement exceeds the recorded product at relative tolerance 10^{-10}; pooled utilization has median 0.42988, 90th percentile 0.59191, and maximum 0.94260. Utilization strictly increases across all three strengths in 994/2,400 question–checkpoint groups (41.42%).

Visualization. All observations are plotted on logarithmic axes. The curve and band show medians and interquartile ranges within 20 logarithmically spaced envelope bins (5–918 observations each). Gray contours use a 94×94 log-space histogram, Gaussian smoothing ($\sigma = 2.15$ bins), and the 55th, 72nd, 84th, and 92nd percentiles of positive smoothed density. Reported statistics use unsmoothed records.

Geometry consistency and interpretation. Fixed class keys and values imply constant $\bar{D}$ for a question–checkpoint pair under pure copies. The logged $\bar{D}$ varies in all 2,400 groups: the relative range $(\max_c \bar{D}_c - \min_c \bar{D}_c)/\operatorname{mean}_c \bar{D}_c$ has median 6.33% and maximum 19.54%. These records show that B_{T1} functions as an empirical scaling diagnostic motivated by the theorem. It tracks displacement without exceedances in this sample, while the theorem's exact guarantee applies to representations satisfying its fixed-representation geometry conditions.

C.5 Extreme-Repetition Stress Test

Protocol and endpoints. Each of 252 held-out base questions requires two evidence classes. Presentations E_N contain N occurrences of the dominant class and one necessary singleton, with facts unchanged. We use $N \in \{1, 2, 4, 8, 16, 32, 64\}$ and checkpoints 42, 43, and 44. Fourteen affinity bins have centers $0.025, 0.045, \ldots, 0.285$, with 18 base questions per bin. Each bin–N cell averages 54 records (18 questions $\times$ three checkpoints); there are 5,292 records in total. Affinity is measured on E_1:

$$\bar{A}_{\text{cross}} = \frac{\sum_{i \in C_{\text{dom}}} \sum_{j \in C_{\text{rare}}} A_{ij}}{|C_{\text{dom}}|\,|C_{\text{rare}}|}, \qquad \Delta\text{Acc}(N) = \text{Acc}(E_1) - \text{Acc}(E_N). \tag{19}$$

The other endpoints are $J = 10^3 D_{\text{JS}}(E_1, E_N)$, using the token-level definition in Appendix B.2, and $\widetilde{\Delta}_F = \|F(E_N) - F(E_1)\|_{\text{F}}/D$, with the class-value diameter D from Theorem 1. All three recorded changes are zero at $N = 1$.

Aggregation and scope. The export contains percentage-valued accuracies per base-question/checkpoint record; binary correctness outcomes and the trial counts underlying these percentages are unavailable. We average the supplied values directly. Repeated presentations and checkpoints remain grouped by question, so the summaries use question-level records and direct averaging. Affinity bins are stratified by observed relation scores. The normalized fusion endpoint is reported as logged; the underlying class-value tensors are unavailable for reconstructing D.

Table 12: **Affinity-stratified stress test.** Cell means from unsmoothed records. N_3 interpolates the three-point crossing linearly in $\log_2 N$; a dash indicates no crossing through $N = 64$. J and $\widetilde{\Delta}_F$ are shown at $N = 64$.

	ΔAcc (percentage points)						
$\bar{A}_{\text{cross}}$	$N = 8$	$N = 16$	$N = 32$	$N = 64$	J	$\widetilde{\Delta}_F$	N_3
0.025	0.25	0.48	0.63	0.94	0.94	0.159	–
0.045	0.26	0.49	0.81	1.01	1.02	0.164	–
0.065	0.28	0.68	0.84	1.24	1.09	0.167	–
0.085	0.34	0.67	1.25	1.70	1.39	0.199	–
0.105	0.45	1.07	1.58	2.45	1.72	0.241	–
0.125	0.60	1.21	1.99	2.88	1.94	0.270	–
0.145	0.72	1.64	2.63	3.71	2.36	0.310	40.6
0.165	0.91	2.08	3.21	4.34	2.76	0.346	28.1
0.185	1.11	2.43	3.78	5.45	3.30	0.408	21.4
0.205	1.27	2.94	4.74	6.69	4.01	0.477	16.4
0.225	1.47	3.43	5.58	7.93	4.66	0.545	13.7
0.245	1.71	3.99	6.31	9.24	5.33	0.622	11.8
0.265	2.00	4.51	7.66	10.87	6.23	0.720	10.5
0.285	2.15	5.22	9.02	12.72	7.20	0.824	9.7

Visualization and interpretation. Figure 6 interpolates the 14×7 grid in $(\log_2 N, \bar{A}_{\text{cross}})$ without smoothing cell means. The supplied jitter coordinates affect only point placement. The color scale extends above 11.5 points. The augmentation protocol in Appendix A.4 appends up to four copies, so its maximum induced multiplicity is $N = 5$; $N = 8$ is the first tested value beyond this range. The three-point criterion defines the operational boundary. The pattern supports a finite-range association between larger imbalance, residual affinity, and degradation.

D PROOFS

This appendix proves Theorem 1 through a sharper class-mass stability theorem and develops its extensions. Appendix D.8 preserves the actual class-value geometry, Appendices D.11 and D.13 make repetition budgets explicit, and Appendices D.14 and D.15 treat nonidentical representations. The direct comparison in Theorem 12 is zero for identical presentations and avoids counting a shared prototype error twice.

Scope and verification. The bound controls pre-readout fusion under the stated representation conditions; sharpness concerns the admissible model family, while downstream risk depends on the fixed decoder. Relation training determines the estimated ϵ, and decoder answers can remain stable when fusion displacement is nonzero. A target-domain degree envelope requires known class membership together with verified affinity/count bounds, and the risk bound additionally requires a useful downstream Lipschitz constant. Identical count laws give zero in the coupling bound, while separate domain envelopes may lose this cancellation. The answer cross-entropy is covered by Appendix D.16; unchanged greedy answers require a decision margin. The sharper class-mass bound, conditional risk bounds, count-dependent certificates, and extensions to nonidentical representations are developed below. These guarantees concern multiplicity shift; finite-sample learning and new or missing evidence classes require separate analysis.

D.1 EQUAL WEIGHTS WITHIN AN EQUIVALENCE CLASS

Let $\mathcal{C}_1, \dots, \mathcal{C}_C$ partition the evidence units. Assign real weights w_i subject to $\sum_{i \in \mathcal{C}_c} w_i = 1$ for every class. By Cauchy–Schwarz, $1 = (\sum_{i \in \mathcal{C}_c} w_i)^2 \le |\mathcal{C}_c| \sum_{i \in \mathcal{C}_c} w_i^2$. Equality holds exactly when all weights in that class are equal. Summing over classes proves that $w_i = 1/|\mathcal{C}_c|$ for $i \in \mathcal{C}_c$ uniquely minimizes $\sum_i w_i^2$. These weights are positive and assign each class unit mass. Since $\sum_j A^\star_{ij} = |\mathcal{C}_c|$, inverse row sums recover precisely this measure.

D.2 PROOF OF PRESENTATION-ORDER INVARIANCE

Fix a question and a permutation matrix $P \in \{0,1\}^{n \times n}$. Under the permuted presentation, each visual unit retains its metadata, local positions, and preprocessing parameters. Its feature matrix X_i, local slots H_i, and summary h_i are therefore unchanged; only the outer sequence is permuted. The question vector and global queries also remain unchanged because their computation does not depend on the evidence presentation.

Shared pair scoring gives $A' = PAP^\top$. Writing $m = A\mathbf{1}_n$, we obtain $m' = A'\mathbf{1}_n = Pm$, so each multiplicity remains attached to its original unit. For a fixed query, the numerator and denominator in Equation (3) are sums over all unit–slot pairs. Reindexing the outer sum leaves both quantities unchanged. Hence $f'_r = f_r$ for every r, followed by $F' = F$ and $Z' = Z$ through Equation (4).

Identical decoder inputs imply identical next-token distributions for each fixed prefix $y_{<\tau}$, hence identical joint answer distributions and greedy outputs with fixed tie-breaking. The result holds in exact arithmetic for every affinity matrix produced by Equation (2). Floating-point reduction order can introduce numerical differences.

D.3 PROOF OF IDEAL DUPLICATION INVARIANCE

Assume the conditions stated in Section 3.4. For $i \in \mathcal{C}_c$, the degree is $m_i = |\mathcal{C}_c|$ and the corresponding keys and values equal $k^\star_{ct}, v^\star_{ct}$. The contribution of class c to the numerator of Equation (3) is therefore $|\mathcal{C}_c|\,|\mathcal{C}_c|^{-1} \sum_t \exp(u_r^\top k^\star_{ct}/\sqrt{d}) v^\star_{ct}$. The same cancellation occurs in the denominator. Summing these class contributions yields Equation (5).

Appending artificial copies changes only class cardinalities, leaving the class prototypes, global queries, fused slots, and decoder inputs unchanged. Literal copies have identical representations under deterministic independent encoding with matching visual content and semantic metadata. For semantic equivalence beyond literal copies, representation equality remains an explicit assumption.

D.4 Formal Setup and General Stability Theorem

Multiplicity-shift model. Let $\xi = (q, \mathcal{Q}, y) \sim \mu$, where $\mathcal{Q}$ contains $C \geq 1$ evidence classes and their fixed semantic metadata and representations. A presentation is specified by $\nu = (\nu_1, \ldots, \nu_C) \in \mathbb{N}_+^C$. Class c has ν_c occurrences, each with the same corresponding keys $k_{ct}^\star$ and values $v_{ct}^\star$, $t \in \{1, \ldots, s\}$. This holds for literal copies under deterministic independent encoding; representation equality for semantic equivalence must be verified separately. All classes remain present, and ν changes neither source identity nor the number of genuinely independent observations.

The source and target domains share μ but use arbitrary conditional multiplicity laws $\kappa_{\mathrm{s}}(\cdot \mid \xi)$ and $\kappa_{\mathrm{t}}(\cdot \mid \xi)$. Their supports need not overlap. All spaces and maps are measurable, with standard Borel underlying spaces. Write $F_\xi(\nu)$ for the K fused slots before the readout. The loss $\ell_\xi(F)$ includes the readout and decoder. Assume, on the relevant representations,

$$0 \leq \ell_\xi(F) \leq B, \qquad |\ell_\xi(F) - \ell_\xi(G)| \leq L(\xi)\|F - G\|_{\mathrm{F}}, \tag{20}$$

where $B \in (0, \infty]$, $0 \leq L(\xi) < \infty$, and both population risks are finite. Taking $B = \infty$ allows unbounded losses.

$$R_a = \mathbb{E}_{\xi\sim\mu}\mathbb{E}_{\nu\sim\kappa_a(\cdot|\xi)}[\ell_\xi(F_\xi(\nu))], \qquad a \in \{\mathrm{s}, \mathrm{t}\}. \tag{21}$$

Effective class masses. Let $\mathcal{C}_c(\nu)$ denote the occurrences of class c. All prototype vectors and raw logits are finite. The actual learned affinities determine $m_i(\nu) = \sum_j A_{ij}(\nu)$. Since $A_{ii} = 1$ and $A_{ij} \in [0, 1]$, all multiplicities are positive. For each query r, define

$$\begin{aligned} a_{rct}^\star &= u_r(q)^\top k_{ct}^\star/\sqrt{d}, \qquad S_{rc} = \sum_{t=1}^{s} e^{a_{rct}^\star}, \\ V_{rc} &= S_{rc}^{-1}\sum_{t=1}^{s} e^{a_{rct}^\star} v_{ct}^\star, \quad b_c(\nu) = \sum_{i\in\mathcal{C}_c(\nu)} m_i(\nu)^{-1}, \\ \pi_{rc}(\nu) &= \frac{S_{rc}b_c(\nu)}{\sum_{j=1}^{C} S_{rj}b_j(\nu)}, \qquad D_r = \max_{c,j}\|V_{rc} - V_{rj}\|_2. \end{aligned} \tag{22}$$

Dependence on ξ is suppressed. In particular, $D_r = 0$ when $C = 1$. For two presentations of the same ξ, put

$$\begin{aligned} h_c(\nu, \nu') &= \log\frac{b_c(\nu')}{b_c(\nu)}, & \Omega_\xi(\nu, \nu') &= \max_c h_c - \min_c h_c, \\ T_r(\nu, \nu') &= \frac{1}{2}\sum_c |\pi_{rc}(\nu') - \pi_{rc}(\nu)|, & D_\xi &= \left(\sum_r D_r^2\right)^{1/2}, \\ J_\xi(\nu, \nu') &= \left(\sum_r D_r^2 T_r(\nu, \nu')^2\right)^{1/2}. \end{aligned} \tag{23}$$

Thus Ω_ξ measures the change in relative class masses; a common rescaling of all class masses contributes zero.

Theorem 2 (Sharp fusion stability and a consequent risk bound)**.** *Under the preceding assumptions, every pair of presentations satisfies*

$$\|F_\xi(\nu') - F_\xi(\nu)\|_{\mathrm{F}} \leq J_\xi(\nu, \nu') \leq D_\xi \tanh\left(\frac{\Omega_\xi(\nu, \nu')}{4}\right). \tag{24}$$

Let Γ be the set of joint laws of (ξ, ν, ν') with marginals $\mu(d\xi)\kappa_{\mathrm{s}}(d\nu \mid \xi)$ and $\mu(d\xi)\kappa_{\mathrm{t}}(d\nu' \mid \xi)$. Then

$$\begin{aligned} |R_{\mathrm{t}} - R_{\mathrm{s}}| &\leq \inf_{\gamma\in\Gamma} \mathbb{E}_\gamma \min\{B, L(\xi)J_\xi(\nu, \nu')\} \\ &\leq \inf_{\gamma\in\Gamma} \mathbb{E}_\gamma \min\left\{B, L(\xi)D_\xi \tanh\left(\frac{\Omega_\xi(\nu, \nu')}{4}\right)\right\}. \end{aligned} \tag{25}$$

In particular, if $m_i(\nu) = \nu_{c(i)}$ on both domain supports, then $R_{\mathrm{t}} = R_{\mathrm{s}}$ for arbitrary multiplicity laws.

D.5 A Sharp Exponential-Tilting Inequality

For a finite vector z, write $\operatorname{osc}(z) = \max_j z_j - \min_j z_j$. For probability vectors p, q, use the convention $\mathrm{TV}(p,q) = \frac{1}{2}\sum_j |p_j - q_j|$. The following lemma is the finite-dimensional sharp relation between Hilbert distance and total variation; see Cohen & Fausti (2023). We provide an elementary proof, including the equality conditions.

Lemma 3 (Sharp exponential tilting)**.** *Let $p_j > 0$, $\sum_j p_j = 1$, and $h \in \mathbb{R}^M$. Define*

$$q_j = \frac{p_j e^{h_j}}{\sum_k p_k e^{h_k}}. \tag{26}$$

Then

$$\mathrm{TV}(p,q) \le \tanh\left(\frac{\operatorname{osc}(h)}{4}\right). \tag{27}$$

For every prescribed positive oscillation, equality is possible with $M = 2$.

Proof. Let Y take value e^{h_j} with probability p_j, and put $a = \min_j e^{h_j}$, $b = \max_j e^{h_j}$, and $m = \mathbb{E}Y$. If $a = b$, the tilt is constant, $p = q$, and the result follows. Assume $0 < a < b$. Convexity of $x \mapsto |x - m|$ gives the pointwise chord inequality

$$\begin{aligned} |Y - m| &= \left| \frac{b - Y}{b - a}(a - m) + \frac{Y - a}{b - a}(b - m) \right| \\ &\le \frac{b - Y}{b - a}|a - m| + \frac{Y - a}{b - a}|b - m| \\ &= \frac{(b - Y)(m - a) + (Y - a)(b - m)}{b - a}. \end{aligned} \tag{28}$$

Taking expectations and using $\mathbb{E}Y = m$ yields

$$\begin{aligned} \mathrm{TV}(p,q) &= \frac{1}{2}\sum_j p_j \left| \frac{e^{h_j}}{m} - 1 \right| \\ &= \frac{\mathbb{E}|Y - m|}{2m} \\ &\le \frac{(b - m)(m - a) + (m - a)(b - m)}{2m(b - a)} \\ &= \frac{(b - m)(m - a)}{m(b - a)} \\ &= \frac{a + b - m - ab/m}{b - a} \\ &\le \frac{a + b - 2\sqrt{ab}}{b - a} \\ &= \frac{(\sqrt{b} - \sqrt{a})^2}{(\sqrt{b} - \sqrt{a})(\sqrt{b} + \sqrt{a})} \\ &= \frac{\sqrt{b} - \sqrt{a}}{\sqrt{b} + \sqrt{a}} \\ &= \tanh\left(\frac{1}{4}\log\frac{b}{a}\right) = \tanh\left(\frac{\operatorname{osc}(h)}{4}\right). \end{aligned} \tag{29}$$

The only nonalgebraic steps are the chord inequality and $m + ab/m \ge 2\sqrt{ab}$, which is equivalent to $(\sqrt{m} - \sqrt{ab/m})^2 \ge 0$.

To attain equality, let Y take only the values a, b, with

$$\mathbb{P}(Y = a) = \frac{\sqrt{b}}{\sqrt{a} + \sqrt{b}}, \qquad \mathbb{P}(Y = b) = \frac{\sqrt{a}}{\sqrt{a} + \sqrt{b}}. \tag{30}$$

Then $m = \sqrt{ab}$, and the chord inequality is an equality at both endpoints. Choosing $a = 1$ and $b = e^{\omega}$ realizes any $\operatorname{osc}(h) = \omega > 0$. □

D.6 From Probability Mass to Vector Displacement

Lemma 4 (Diameter contraction)**.** *Let $z_1, \dots, z_M$ belong to a normed vector space, and let $D = \max_{i,j} \|z_i - z_j\|$. For any probability vectors p, q,*

$$\left\| \sum_j (q_j - p_j) z_j \right\| \le D \, \mathrm{TV}(p, q). \tag{31}$$

The coefficient D is sharp.

Proof. Write $d_j = q_j - p_j$, $d_j^+ = \max\{d_j, 0\}$, and $d_j^- = \max\{-d_j, 0\}$. Since $\sum_j d_j = 0$,

$$\sum_j d_j^+ = \sum_j d_j^- = \frac{1}{2} \sum_j |d_j| = \mathrm{TV}(p, q) =: \tau. \tag{32}$$

For $\tau = 0$, the assertion is immediate. Otherwise, define $\alpha_j = d_j^+ / \tau$ and $\beta_j = d_j^- / \tau$. Both vectors are probability distributions. Therefore

$$\begin{aligned} \left\| \sum_j (q_j - p_j) z_j \right\| &= \tau \left\| \sum_i \alpha_i z_i - \sum_j \beta_j z_j \right\| \\ &= \tau \left\| \sum_{i,j} \alpha_i \beta_j (z_i - z_j) \right\| \\ &\le \tau \sum_{i,j} \alpha_i \beta_j \|z_i - z_j\| \\ &\le \tau D \sum_{i,j} \alpha_i \beta_j = \tau D. \end{aligned} \tag{33}$$

For two points at distance D, every nonzero signed mass transfer between the points attains equality. □

D.7 Proof of Theorem 2

Proof. Fix ξ throughout the pointwise argument.

Step 1: exact reduction to class attention. Because corresponding keys and values agree within each class, the QSEF fusion formula can be regrouped without approximation:

$$\begin{aligned} f_r(\nu) &= \frac{\sum_c \sum_{i \in \mathcal{C}_c(\nu)} m_i^{-1} \sum_t e^{a^\star_{rct}} v^\star_{ct}}{\sum_c \sum_{i \in \mathcal{C}_c(\nu)} m_i^{-1} \sum_t e^{a^\star_{rct}}} \\ &= \frac{\sum_c b_c(\nu) S_{rc} V_{rc}}{\sum_c b_c(\nu) S_{rc}} \\ &= \sum_c \pi_{rc}(\nu) V_{rc}. \end{aligned} \tag{34}$$

In particular, the vectors being averaged are independent of multiplicity; only their class probabilities change.

Step 2: identify the exact exponential tilt. For a second presentation ν', direct cancellation gives

$$\begin{aligned} \pi_{rc}(\nu') &= \frac{S_{rc} b_c(\nu) e^{h_c}}{\sum_j S_{rj} b_j(\nu) e^{h_j}} \\ &= \frac{\pi_{rc}(\nu) e^{h_c}}{\sum_j \pi_{rj}(\nu) e^{h_j}}. \end{aligned} \tag{35}$$

Thus the normalizing constants disappear from the oscillation. Lemmas 3 and 4 imply

$$\begin{aligned}\|f_r(\nu') - f_r(\nu)\|_2 &= \left\| \sum_c [\pi_{rc}(\nu') - \pi_{rc}(\nu)] V_{rc} \right\|_2 \\ &\le D_r T_r(\nu, \nu') \\ &= \frac{D_r}{2} \sum_c \pi_{rc}(\nu) \left| \frac{e^{h_c}}{\sum_j \pi_{rj}(\nu) e^{h_j}} - 1 \right| \\ &\le D_r \tanh\left(\frac{\operatorname{osc}(h)}{4} \right) \\ &= D_r \tanh\left(\frac{\Omega_\xi(\nu, \nu')}{4} \right).\end{aligned} \tag{36}$$

Consequently,

$$\begin{aligned}\|F_\xi(\nu') - F_\xi(\nu)\|_{\mathrm{F}}^2 &= \sum_r \|f_r(\nu') - f_r(\nu)\|_2^2 \\ &\le \sum_r D_r^2 T_r(\nu, \nu')^2 \\ &\le \sum_r D_r^2 \tanh^2\left(\frac{\Omega_\xi(\nu, \nu')}{4} \right) \\ &= D_\xi^2 \tanh^2\left(\frac{\Omega_\xi(\nu, \nu')}{4} \right).\end{aligned} \tag{37}$$

Taking square roots proves (24).

Step 3: couple the domains before taking expectations. The admissible coupling set is nonempty: conditional independence gives the joint law

$$\gamma_0(d\xi, d\nu, d\nu') = \mu(d\xi) \kappa_{\mathrm{s}}(d\nu \mid \xi) \kappa_{\mathrm{t}}(d\nu' \mid \xi). \tag{38}$$

For any $\gamma \in \Gamma$, its marginal identities and the finiteness of the risks justify

$$\begin{aligned}|R_{\mathrm{t}} - R_{\mathrm{s}}| &= |\mathbb{E}_\gamma[\ell_\xi(F_\xi(\nu')) - \ell_\xi(F_\xi(\nu))]| \\ &\le \mathbb{E}_\gamma |\ell_\xi(F_\xi(\nu')) - \ell_\xi(F_\xi(\nu))| \\ &\le \mathbb{E}_\gamma \min\{B, L(\xi) \|F_\xi(\nu') - F_\xi(\nu)\|_{\mathrm{F}}\} \\ &\le \mathbb{E}_\gamma \min\{B, L(\xi) J_\xi(\nu, \nu')\} \\ &\le \mathbb{E}_\gamma \min\left\{ B, L(\xi) D_\xi \tanh\left(\frac{\Omega_\xi(\nu, \nu')}{4} \right) \right\}.\end{aligned} \tag{39}$$

Taking infima over the same nonempty coupling set proves (25). No exchange of an infimum and an expectation, and no attainment of an optimal coupling, is required.

Step 4: exact invariance as a special case. If $m_i(\nu) = \nu_{c(i)}$, then

$$b_c(\nu) = \sum_{i \in \mathcal{C}_c(\nu)} \frac{1}{\nu_c} = 1, \qquad \pi_{rc}(\nu) = \frac{S_{rc}}{\sum_j S_{rj}}. \tag{40}$$

These probabilities are independent of ν. Hence the fused representation and the loss are equal pointwise across presentations, and their expectations under arbitrary multiplicity laws agree. □

D.8 A Geometry-Sensitive Refinement

The diameter bound deliberately summarizes all class values by one number. When their actual geometry is available, a finite transport problem gives a certificate at least as strong as the total-variation bound.

Proposition 5 (Transport refinement)**.** *For the exact-copy setting, let Λ_r be the matrices with nonnegative entries, row sums $\pi_r(\nu)$, and column sums $\pi_r(\nu')$. Define*

$$W_r = \min_{\lambda \in \Lambda_r} \sum_{c,j} \lambda_{cj} \|V_{rj} - V_{rc}\|_2, \qquad H_\xi = \left(\sum_r W_r^2 \right)^{1/2}. \tag{41}$$

Then J_ξ can be replaced by H_ξ in both conclusions of Theorem 2, since

$$\|F_\xi(\nu') - F_\xi(\nu)\|_{\mathrm{F}} \le H_\xi \le J_\xi \le D_\xi \tanh(\Omega_\xi/4). \tag{42}$$

Proof. Fix r, abbreviate $p = \pi_r(\nu)$, $q = \pi_r(\nu')$, and put $\tau = \mathrm{TV}(p, q)$. The finite coupling polytope is nonempty (it contains $pq^\top$) and compact; its continuous cost therefore attains its minimum. For any admissible coupling,

$$\begin{aligned} \left\| \sum_j q_j V_{rj} - \sum_c p_c V_{rc} \right\|_2 &= \left\| \sum_{c,j} \lambda_{cj} (V_{rj} - V_{rc}) \right\|_2 \\ &\le \sum_{c,j} \lambda_{cj} \|V_{rj} - V_{rc}\|_2. \end{aligned} \tag{43}$$

Taking the minimum bounds the row displacement by W_r. For $\tau = 0$, the diagonal coupling gives $W_r = 0$. If $\tau > 0$, write $x_+ = \max\{x, 0\}$ and construct

$$\lambda^0_{cj} = \mathbf{1}[c = j] \min\{p_c, q_c\} + \frac{(p_c - q_c)_+ (q_j - p_j)_+}{\tau}. \tag{44}$$

Both residual masses sum to τ, so this matrix has the required row and column sums. Its diagonal part has zero cost, and hence

$$\begin{aligned} W_r &\le \sum_{c,j} \lambda^0_{cj} \|V_{rj} - V_{rc}\|_2 \\ &= \frac{1}{\tau} \sum_{c,j} (p_c - q_c)_+ (q_j - p_j)_+ \|V_{rj} - V_{rc}\|_2 \\ &\le \frac{D_r}{\tau} \left(\sum_c (p_c - q_c)_+ \right) \left(\sum_j (q_j - p_j)_+ \right) = D_r \tau. \end{aligned} \tag{45}$$

Square and sum the nonnegative row inequalities to obtain (42). For the risk conclusion, substitute this pointwise bound in the first three lines of (39) and take the infimum over Γ. The finite transportation value is measurable in its finite-dimensional inputs, so the expectation is well defined. This refinement preserves class geometry but cannot improve the worst-case D_ξ, Ω_ξ-only envelope: the two-point construction already attains it. □

D.9 Degree Envelopes and Relation-Error Amplification

Corollary 6 (Conditional degree-error certificate)**.** *Suppose that, for μ-almost every ξ, every presentation in domain a satisfies*

$$0 < \underline{d}_a(\xi) \le \frac{m_i(\nu)}{\nu_{c(i)}} \le \overline{d}_a(\xi) < \infty, \qquad \chi_a(\xi) = \log \frac{\overline{d}_a(\xi)}{\underline{d}_a(\xi)}. \tag{46}$$

Then

$$|R_{\mathrm{t}} - R_{\mathrm{s}}| \le \mathbb{E}_\mu \min \left\{ B, L(\xi) D_\xi \tanh\left(\frac{\chi_{\mathrm{s}}(\xi) + \chi_{\mathrm{t}}(\xi)}{4} \right) \right\}. \tag{47}$$

For the common symmetric condition $|m_i/\nu_{c(i)} - 1| \le \epsilon$ with $0 \le \epsilon < 1$, every source–target pair also satisfies $\|F_\xi(\nu') - F_\xi(\nu)\|_{\mathrm{F}} \le D_\xi \epsilon$, and the risk bound simplifies to

$$|R_{\mathrm{t}} - R_{\mathrm{s}}| \le \mathbb{E}_\mu \min\{B, L(\xi) D_\xi \epsilon\}. \tag{48}$$

Proof of Corollary 6. For $i \in \mathcal{C}_c(\nu)$, let $d_i = m_i/\nu_c$. For a presentation from domain a, inversion and averaging give

$$\begin{aligned} \overline{d}_a^{-1} &\le \frac{1}{\nu_c} \sum_{i \in \mathcal{C}_c(\nu)} d_i^{-1} \\ &= \sum_{i \in \mathcal{C}_c(\nu)} m_i^{-1} = b_c(\nu) \le \underline{d}_a^{-1}. \end{aligned} \tag{49}$$

Since logarithms preserve order,

$$-\log \overline{d}_a \le \log b_c(\nu) \le -\log \underline{d}_a, \qquad \operatorname{osc}_c \log b_c(\nu) \le \chi_a. \tag{50}$$

For a source–target pair, the elementary oscillation inequality yields

$$\begin{aligned} \Omega_\xi(\nu, \nu') &= \max_{c,j} \left[\log b_c(\nu') - \log b_c(\nu) - \log b_j(\nu') + \log b_j(\nu)\right] \\ &\le \max_{c,j} [\log b_c(\nu') - \log b_j(\nu')] + \max_{c,j} [\log b_j(\nu) - \log b_c(\nu)] \\ &= \operatorname{osc} \log b(\nu') + \operatorname{osc} \log b(\nu) \\ &\le \chi_{\mathrm{t}} + \chi_{\mathrm{s}}. \end{aligned} \tag{51}$$

The last bound depends only on ξ, so every coupling in Theorem 2 has the same upper envelope, proving (47). Under the symmetric assumption with $0 \le \epsilon < 1$,

$$\begin{aligned} \tanh\left(\frac{\chi_{\mathrm{s}} + \chi_{\mathrm{t}}}{4}\right) &\le \tanh\left(\frac{1}{2} \log \frac{1+\epsilon}{1-\epsilon}\right) \\ &= \frac{(1+\epsilon)/(1-\epsilon) - 1}{(1+\epsilon)/(1-\epsilon) + 1} = \epsilon, \end{aligned} \tag{52}$$

which proves (48). Applying the pointwise fusion bound in (24) before taking expectations also gives $\|F_\xi(\nu') - F_\xi(\nu)\|_{\mathrm{F}} \le D_\xi \epsilon$, as asserted in Corollary 6. □

Proof of Theorem 1. Fix the question and class prototypes, so that D in the main text is D_ξ in (23). For each $\eta \in \{\nu, \nu'\}$, the assumed relative-degree interval and (49) give $b_c(\eta) \in [(1+\epsilon)^{-1}, (1-\epsilon)^{-1}]$. Applying (51) and the monotonicity of $\tanh$, followed by (52), yields

$$\begin{aligned} \|F(\nu') - F(\nu)\|_{\mathrm{F}} &\le \left(\sum_r D_r^2 T_r(\nu, \nu')^2\right)^{1/2} \\ &\le D \tanh\left(\frac{\Omega_\xi(\nu, \nu')}{4}\right) \\ &\le D \tanh\left(\frac{1}{2} \log \frac{1+\epsilon}{1-\epsilon}\right) = D\epsilon. \end{aligned} \tag{53}$$

The first two inequalities are (24), proved in Appendix D.7; the pointwise proof applies across losses and count laws. When $m_i(\eta) = \eta_{c(i)}$, take $\epsilon = 0$ to obtain $F(\nu') = F(\nu)$. With the question fixed, the deterministic readout and decoder therefore have identical conditional next-token distributions for any common answer prefix. For every $0 < \epsilon < 1$ and $D > 0$, the construction in Proposition 10, including its proof, satisfies these assumptions and attains $D\epsilon$ exactly. Thus no smaller universal coefficient is valid. At $D = 0$ or $\epsilon = 0$, the zero bound is already exact. □

Proposition 7 (Signed affinity errors and rare-class amplification)**.** *For a fixed presentation, define the normalized within-class deficit and cross-class leakage by*

$$\eta_i^- = \frac{1}{\nu_{c(i)}} \sum_{j \in \mathcal{C}_{c(i)}} (1 - A_{ij}), \qquad \eta_i^+ = \frac{1}{\nu_{c(i)}} \sum_{j \notin \mathcal{C}_{c(i)}} A_{ij}. \tag{54}$$

Then $m_i/\nu_{c(i)} = 1 - \eta_i^- + \eta_i^+$ exactly. If $1 - A_{ij} \le \alpha$ within classes off the diagonal and $A_{ij} \le \beta$ across classes, with $\alpha, \beta \in [0, 1]$, then

$$1 - \alpha\left(1 - \frac{1}{\nu_{c(i)}}\right) \le \frac{m_i}{\nu_{c(i)}} \le 1 + \beta\left(\frac{n}{\nu_{c(i)}} - 1\right). \tag{55}$$

Consequently, if an entire domain obeys $n \le n_{\max}$ and $\nu_{\min} \le \nu_c \le \nu_{\max}$, valid uniform envelopes are

$$\underline{d} = 1 - \alpha(1 - \nu_{\max}^{-1}) > 0, \qquad \overline{d} = 1 + \beta(n_{\max}/\nu_{\min} - 1). \tag{56}$$

Proof. Splitting the degree into its two types of edges gives

$$\begin{aligned} \frac{m_i}{\nu_{c(i)}} &= \frac{1}{\nu_{c(i)}} \left(\sum_{j \in \mathcal{C}_{c(i)}} A_{ij} + \sum_{j \notin \mathcal{C}_{c(i)}} A_{ij} \right) \\ &= \frac{1}{\nu_{c(i)}} \left(\nu_{c(i)} - \sum_{j \in \mathcal{C}_{c(i)}} (1 - A_{ij}) + \sum_{j \notin \mathcal{C}_{c(i)}} A_{ij} \right) \\ &= 1 - \eta_i^- + \eta_i^+. \end{aligned} \tag{57}$$

Because the diagonal deficit is zero and there are respectively $\nu_{c(i)} - 1$ and $n - \nu_{c(i)}$ eligible off-diagonal edges,

$$\begin{aligned} 0 \le \eta_i^- &\le \frac{\nu_{c(i)} - 1}{\nu_{c(i)}} \alpha \le (1 - \nu_{\max}^{-1})\alpha, \\ 0 \le \eta_i^+ &\le \frac{n - \nu_{c(i)}}{\nu_{c(i)}} \beta \le (n_{\max}/\nu_{\min} - 1)\beta, \\ 1 - (1 - \nu_{\max}^{-1})\alpha \le 1 - \eta_i^- &\le 1 - \eta_i^- + \eta_i^+ \le 1 + \eta_i^+ \le 1 + (n_{\max}/\nu_{\min} - 1)\beta. \end{aligned} \tag{58}$$

The lower endpoint is positive since it is at least $\nu_{\max}^{-1}$. The row-specific inequalities follow before the last uniform relaxations. Notice that retaining $1 - \eta_i^- + \eta_i^+$ preserves cancellation between the two errors; replacing it by a sum of absolute errors can be unnecessarily loose. □

D.10 Proof of Sharpness and the Unbounded-Repetition Limitation

Proposition 8 (Attainability and the limit of soft correction)**.** *The envelope $D_\xi \tanh(\Omega_\xi/4)$ cannot be uniformly reduced using only D_ξ and Ω_ξ. Its equality cases are realizable by a two-class QSEF fusion model with one fixed shared symmetric scorer and finite off-diagonal sigmoid logits. The coefficient of $D_\xi \epsilon$ in Corollary 6 is also optimal under its stated information. Moreover, every fixed positive cross-class affinity admits a two-class construction whose fusion drift approaches the full class-value diameter as the multiplicity imbalance grows without bound.*

Proof of Proposition 8. Take $C = 2$, $K = s = 1$, equal raw relevance logits, and scalar values $v_1 = 0, v_2 = D > 0$. Fix $\rho, \beta \in (0, 1)$. Set every within-class off-diagonal affinity to ρ, every cross-class affinity to β, and every diagonal entry to one. Both off-diagonal values have finite sigmoid logits, so no exact zero or one is needed. The construction uses one shared symmetric scorer: choose scalar unit summaries $h_i \in \{0, D\}$, set the key projection to zero, and use

$$g_\psi(h_i + h_j, |h_i - h_j|, h_q) = \mathrm{logit}(\rho) + [\mathrm{logit}(\beta) - \mathrm{logit}(\rho)] \frac{|h_i - h_j|}{D}, \tag{59}$$

where $\mathrm{logit}(p) = \log(p/(1-p))$. This affine function of a permitted pair feature can also be represented by a GELU head using $\mathrm{GELU}(x) - \mathrm{GELU}(-x) = x$.

Compare $\nu = (N, 1)$ and $\nu' = (1, N)$, where $N > 1$. The two effective class-mass vectors are (a_N, b_N) and (b_N, a_N), where

$$a_N = \frac{N}{1 + (N-1)\rho + \beta}, \qquad b_N = \frac{1}{1 + N\beta}, \qquad r_N = \frac{a_N}{b_N} = \frac{N(1 + N\beta)}{1 + (N-1)\rho + \beta}. \tag{60}$$

The ratio exceeds one, since

$$\begin{aligned} N(1 + N\beta) - [1 + (N-1)\rho + \beta] &= (N-1)(1-\rho) + \beta(N^2 - 1) \\ &= (N-1)[1 - \rho + \beta(N+1)] > 0. \end{aligned} \tag{61}$$

Consequently $h = (-\log r_N, \log r_N)$, $\Omega = 2\log r_N$, and direct calculation gives

$$\begin{aligned} |f(\nu') - f(\nu)| &= \left| D\frac{a_N}{a_N + b_N} - D\frac{b_N}{a_N + b_N} \right| \\ &= D\frac{a_N - b_N}{a_N + b_N} \\ &= D\frac{r_N - 1}{r_N + 1} \\ &= D\tanh\left(\frac{1}{2}\log r_N\right) = D\tanh\left(\frac{\Omega}{4}\right). \end{aligned} \tag{62}$$

Both the total-variation and diameter bounds are therefore attained simultaneously by QSEF fusion.

Every prescribed $\Omega > 0$ can be attained. To see this, put $r = e^{\Omega/2} > 1$, choose an integer $N > 1$ with $N(N+1)/2 > r$, and set $\rho = 1 - \beta$. The ratio becomes the continuous function

$$r_N(\beta) = \frac{N(1 + N\beta)}{N + (2 - N)\beta}, \qquad r_N(0) = 1, \qquad r_N(1) = \frac{N(N+1)}{2}. \tag{63}$$

Its denominator is positive on $[0, 1]$. The intermediate value theorem supplies $\beta \in (0, 1)$ with $r_N(\beta) = r$, and then $\rho \in (0, 1)$ as required. The case $\Omega = 0$ is attained by identical presentations.

Finally, for any fixed $\rho, \beta \in (0, 1)$,

$$\lim_{N\to\infty} \frac{r_N}{N} = \frac{\beta}{\rho} > 0, \qquad \lim_{N\to\infty} |f(\nu') - f(\nu)| = D. \tag{64}$$

In particular, one may fix $\rho = 1 - \beta$ and let β be arbitrarily small: the supremum displacement over all N is still D. Thus small fixed affinity errors alone cannot yield a vanishing guarantee uniformly over unbounded multiplicities. This establishes sharpness over admissible fusion instances at the fusion-module level. The example embeds into any fixed $K, s, d \geq 1$ by repeating the same local slot s times, taking identical queries, and embedding scalar values along one coordinate. Then both the Frobenius displacement and D_ξ are multiplied by $\sqrt{K}$. The resulting equality concerns the fusion module; a fixed downstream decoder can introduce additional loss. In particular, one-dimensional layer normalization is constant, so the scalar fusion equality alone does not transfer automatically to the readout. □

D.11 An Exact Repetition Budget in the Two-Class Model

Proposition 9 (Exact tolerance threshold)**.** *In the preceding two-class model with $D = 1$, compare $(N, 1)$ with the balanced presentation $(1, 1)$. Let $p_N = f(N, 1)$ and $\delta = 1 - \rho + \beta$. For $\rho, \beta \in (0, 1)$,*

$$\begin{aligned} p_N &= \frac{\rho N + \delta}{\beta N^2 + (1 + \rho)N + \delta}, \\ \Delta_N := \frac{1}{2} - p_N &= \frac{(N - 1)[\beta(N + 1) + 1 - \rho]}{2[\beta N^2 + (1 + \rho)N + \delta]}. \end{aligned} \tag{65}$$

The drift increases with N and converges to $1/2$. For $\tau \in (0, 1/2)$, define $c_\tau = 1 - \rho - 2\tau(1 + \rho)$. Then the exact admissible integer budget is

$$\Delta_N \leq \tau \quad \Longleftrightarrow \quad N \leq \frac{-c_\tau + \sqrt{c_\tau^2 + 4\beta(1 - 4\tau^2)\delta}}{2\beta(1 - 2\tau)}. \tag{66}$$

Proof. The singleton carries value one, so $p_N = b_N/(a_N + b_N)$ with the masses in (60). Clearing denominators gives (65); at $N = 1$, both masses agree and $p_1 = 1/2$. Extending the rational expression to real $N \geq 1$, differentiation and cancellation yield

$$\begin{aligned} p'_N &= \frac{\rho[\beta N^2 + (1 + \rho)N + \delta] - (\rho N + \delta)[2\beta N + 1 + \rho]}{[\beta N^2 + (1 + \rho)N + \delta]^2} \\ &= -\frac{\rho\beta N^2 + 2\delta\beta N + \delta}{[\beta N^2 + (1 + \rho)N + \delta]^2} < 0. \end{aligned} \tag{67}$$

Since the numerator of p_N is linear and its denominator is quadratic with positive leading coefficient, $p_N \to 0$. For the budget, all denominators are positive and

$$\begin{aligned}\Delta_N \le \tau &\iff \beta N^2 + (1-\rho)N - \delta \le 2\tau[\beta N^2 + (1+\rho)N + \delta] \\ &\iff \beta(1-2\tau)N^2 + c_\tau N - (1+2\tau)\delta \le 0 \\ &\iff N \le \frac{-c_\tau + \sqrt{c_\tau^2 + 4\beta(1-2\tau)(1+2\tau)\delta}}{2\beta(1-2\tau)}.\end{aligned} \tag{68}$$

The last equivalence uses $N \ge 1$: the quadratic has positive leading coefficient and negative constant term, hence exactly one negative root and one positive root. This proves the claim. □

D.12 Sharpness of the Symmetric Degree-Error Bound

Proposition 10 (Attainability at every degree-error level)**.** *For every $0 < \epsilon < 1$ and $D > 0$, there is a fixed QSEF fusion module and two presentations with $C = 2$, $K = s = 1$, and class-value diameter D, such that every relative degree in both presentations belongs to $[1-\epsilon, 1+\epsilon]$ and*

$$|f(\nu') - f(\nu)| = D\epsilon. \tag{69}$$

All off-diagonal affinities lie strictly between zero and one and are generated by one shared symmetric scorer. Consequently, the coefficient of $D\epsilon$ in the fusion specialization of Corollary 6 cannot be uniformly reduced.

Proof. Since $N(N-1)/(N^2+1) \to 1$, choose an integer $N > 1$ and set

$$\epsilon < \frac{N(N-1)}{N^2+1}, \qquad \beta = \frac{\epsilon}{N}, \qquad \rho = 1 - \frac{\epsilon(N^2+1)}{N(N-1)}. \tag{70}$$

These choices ensure $0 < \rho, \beta < 1$. Use the two-class construction of Proposition 8: values $0, D$, zero raw relevance logits, within-class off-diagonal affinity ρ, and cross-class affinity β. The shared scorer in (59) realizes these affinities with finite parameters. Compare $\nu = (N, 1)$ and $\nu' = (1, N)$. For the repeated class and the singleton, respectively, the relative degrees are exactly

$$\begin{aligned}d_{\mathrm{rep}} &= \frac{1 + (N-1)\rho + \beta}{N} \\ &= \frac{N - \epsilon(N^2+1)/N + \epsilon/N}{N} = 1 - \epsilon, \\ d_{\mathrm{single}} &= 1 + N\beta = 1 + \epsilon.\end{aligned} \tag{71}$$

The degrees exchange classes between the two presentations. Their class-mass vectors are therefore (a, b) and (b, a), where $a = (1-\epsilon)^{-1}$ and $b = (1+\epsilon)^{-1}$. Hence

$$\begin{aligned}f(\nu) = D\frac{b}{a+b} = \frac{D(1-\epsilon)}{2}, \qquad f(\nu') = D\frac{a}{a+b} = \frac{D(1+\epsilon)}{2}, \\ |f(\nu') - f(\nu)| = D\frac{a-b}{a+b} = D\frac{(1+\epsilon)-(1-\epsilon)}{(1+\epsilon)+(1-\epsilon)} = D\epsilon.\end{aligned} \tag{72}$$

Thus both the inverse-degree envelope and the fusion bound are simultaneously attained. The scalar construction embeds into any larger feature dimension along a fixed unit vector; duplicating identical local slots preserves the calculation, and using identical pooling queries makes all K rows attain their bounds, yielding $\|F(\nu') - F(\nu)\|_{\mathrm{F}} = D_\xi \epsilon$. This sharpness statement concerns the fusion-module family with the specified local representations. A particular frozen encoder, readout, decoder, or population loss requires additional conditions to realize the same equality. □

D.13 Explicit Dependence on Counts and Conditional Certificates

The next result makes the multiplicity dependence explicit. It requires equality of the relation-head summaries within each class, in addition to equality of corresponding keys and values. Literal copies under the deterministic independent encoder satisfy both conditions. The scorer and question are fixed throughout.

Proposition 11 (Certificates on bounded multiplicity sets)**.** *Let $\rho_c \in [0,1]$ denote the scorer's affinity between distinct occurrences of class c, and let $\beta_{cj} = \beta_{jc} \in [0,1]$ denote its affinity between classes $c \neq j$. Every occurrence of class c has the same degree, and its relative degree and effective class mass satisfy*

$$\begin{aligned} m_c(\nu) &= 1 + (\nu_c - 1)\rho_c + \sum_{j \neq c} \nu_j \beta_{cj}, \\ d_c(\nu) := \frac{m_c(\nu)}{\nu_c} &= \rho_c + \frac{1 - \rho_c + \sum_{j \neq c} \nu_j \beta_{cj}}{\nu_c}, \qquad b_c(\nu) = d_c(\nu)^{-1}. \end{aligned} \tag{73}$$

In particular, the exact pairwise oscillation is

$$\Omega(\nu, \nu') = \operatorname{osc}_c \log \frac{d_c(\nu')}{d_c(\nu)}. \tag{74}$$

Let $\mathcal{V}_{\mathrm{s}}, \mathcal{V}_{\mathrm{t}} \subset \mathbb{N}_+^C$ be nonempty finite sets of admissible counts. Define

$$\begin{aligned} Q^-_{a,cj} &= \min_{\nu \in \mathcal{V}_a} \frac{d_c(\nu)}{d_j(\nu)}, \qquad Q^+_{a,cj} = \max_{\nu \in \mathcal{V}_a} \frac{d_c(\nu)}{d_j(\nu)}, \\ M &= \max_{c,j} \log \frac{Q^+_{\mathrm{t},cj}}{Q^-_{\mathrm{s},cj}}. \end{aligned} \tag{75}$$

Then M is the exact largest oscillation over these count sets, and the fusion representation has the uniform conditional certificate

$$\begin{aligned} \max_{\nu \in \mathcal{V}_{\mathrm{s}},\, \nu' \in \mathcal{V}_{\mathrm{t}}} \Omega(\nu, \nu') &= M, \\ \max_{\nu \in \mathcal{V}_{\mathrm{s}},\, \nu' \in \mathcal{V}_{\mathrm{t}}} \|F(\nu') - F(\nu)\|_{\mathrm{F}} &\leq D_\xi \tanh(M/4). \end{aligned} \tag{76}$$

The first equality is exact for the specified scorer and count sets; the second bound need not be attained for their particular value vectors and relevance logits.

Proof. Identical summaries give the same scorer input to every within-class pair, and likewise to every pair from two given classes. Counting the diagonal, the $\nu_c - 1$ other within-class occurrences, and all cross-class occurrences proves the first identity in (73). Division by ν_c gives the second, while summing the ν_c identical reciprocal degrees gives $b_c = \nu_c / m_c = d_c^{-1}$. All d_c are strictly positive because $m_c \geq 1$. Thus logarithms below are well defined. Since oscillation is unchanged by negation,

$$\begin{aligned} \Omega(\nu, \nu') &= \operatorname{osc}_c \log \frac{b_c(\nu')}{b_c(\nu)} \\ &= \operatorname{osc}_c \log \frac{d_c(\nu')}{d_c(\nu)} \\ &= \max_{c,j} \left[\log \frac{d_c(\nu')}{d_j(\nu')} - \log \frac{d_c(\nu)}{d_j(\nu)} \right] \\ &\leq \max_{c,j} \left[\log Q^+_{\mathrm{t},cj} - \log Q^-_{\mathrm{s},cj} \right] = M. \end{aligned} \tag{77}$$

All maxima and minima are over nonempty finite sets. Choose an ordered class pair (c, j) attaining the last maximum, a target presentation attaining $Q^+_{\mathrm{t},cj}$, and a source presentation attaining $Q^-_{\mathrm{s},cj}$. For this pair of presentations, the corresponding term in the preceding maximum equals M, so its oscillation is at least M. The upper bound just proved gives equality. Terms with $c = j$ equal zero, so $M \geq 0$ also when $C = 1$. Applying Theorem 2 and monotonicity of $\tanh$ proves the fusion certificate. □

Closed-form bounds for count boxes. For positive integer endpoints satisfying $\ell_{a,c} \leq \nu_c \leq u_{a,c}$, define

$$\begin{aligned} d^-_{a,c} &= \rho_c + \frac{1 - \rho_c + \sum_{j \neq c} \ell_{a,j} \beta_{cj}}{u_{a,c}}, \\ d^+_{a,c} &= \rho_c + \frac{1 - \rho_c + \sum_{j \neq c} u_{a,j} \beta_{cj}}{\ell_{a,c}}. \end{aligned} \tag{78}$$

These are the exact individual extrema of d_c over the full box: the nonnegative numerator in (73) increases with the other counts, and division decreases with the class's own count. The two opposite box corners attain the stated endpoints. For any admissible subsets of these boxes, it follows that

$$d_{a,c}^{-} \le d_c(\nu) \le d_{a,c}^{+}, \qquad \frac{d_{a,c}^{-}}{d_{a,j}^{+}} \le \frac{d_c(\nu)}{d_j(\nu)} \le \frac{d_{a,c}^{+}}{d_{a,j}^{-}},$$
$$M \le \begin{cases} 0, & C = 1, \\ \max\left\{0, \max\limits_{c \ne j} \log \dfrac{d_{\mathrm{t},c}^{+} d_{\mathrm{s},j}^{+}}{d_{\mathrm{t},j}^{-} d_{\mathrm{s},c}^{-}}\right\}, & C \ge 2. \end{cases} \tag{79}$$

The ratio certificate retains dependencies between degrees of different classes; the box relaxation can discard them. Similarly, the direct pairwise expression retains cancellation: if $d_c(\nu') = \lambda d_c(\nu)$ for one common $\lambda > 0$, then $\Omega(\nu, \nu') = 0$, even if the degrees in each presentation are far from one. These certificates rely on the specified classes, affinities, and admissible counts. A small average relation-training loss alone leaves them undetermined, and deployment requires the corresponding latent classes to be observable. The maximum over count pairs also discards distributional coupling: it may be positive when the two count laws coincide, whereas the diagonal coupling in Theorem 2 gives zero risk difference in that case.

D.14 Direct Stability for Nonidentical Within-Class Representations

We now allow representations to vary within a semantic class and across presentations. Fix a question q and two finite presentations E, E'. Their units are partitioned into the same C semantic identities, and every class is nonempty in both presentations. The numbers of units may differ. All queries u_r are shared because they depend only on q; all keys, values, and positive degrees are finite. In particular, the QSEF construction gives $m_i \ge 1$. No correspondence between individual units in E and E' is needed.

For $i \in E$, write $a_{rit}(E) = u_r^{\top} k_{it}(E)/\sqrt{d}$ and let $\mathcal{C}_c(E)$ denote the units in class c. Define the unnormalized class masses, class probabilities, and class barycenters by

$$W_{rc}(E) = \sum_{i \in \mathcal{C}_c(E)} \sum_{t=1}^{s} \frac{e^{a_{rit}(E)}}{m_i(E)}, \qquad P_{rc}(E) = \frac{W_{rc}(E)}{\sum_{j=1}^{C} W_{rj}(E)},$$
$$V_{rc}(E) = \frac{1}{W_{rc}(E)} \sum_{i \in \mathcal{C}_c(E)} \sum_{t=1}^{s} \frac{e^{a_{rit}(E)}}{m_i(E)} v_{it}(E). \tag{80}$$

Every $W_{rc}(E)$ is finite and strictly positive. Use primes for quantities evaluated at E', and abbreviate

$$d_{rc} = \|V'_{rc} - V_{rc}\|_2, \qquad T_r = \tfrac{1}{2} \sum_c |P'_{rc} - P_{rc}|,$$
$$D_r(E) = \max_{c,j} \|V_{rc} - V_{rj}\|_2, \quad \Omega_r = \mathrm{osc}_c \log \frac{W'_{rc}}{W_{rc}}. \tag{81}$$

Theorem 12 (Direct pairwise stability)**.** *Under the preceding conditions, define*

$$\mathcal{B}_r(E, E') = \min\left\{ \sum_c P'_{rc} d_{rc} + D_r(E) T_r,\ \sum_c P_{rc} d_{rc} + D_r(E') T_r \right\},$$
$$\widehat{\mathcal{B}}_r(E, E') = \min\left\{ \sum_c P'_{rc} d_{rc} + D_r(E) \tanh(\Omega_r/4),\ \sum_c P_{rc} d_{rc} + D_r(E') \tanh(\Omega_r/4) \right\}. \tag{82}$$

Then

$$\|F(E') - F(E)\|_{\mathrm{F}} \le \mathcal{B}(E, E') := \left(\sum_r \mathcal{B}_r(E, E')^2 \right)^{1/2} \le \left(\sum_r \widehat{\mathcal{B}}_r(E, E')^2 \right)^{1/2}. \tag{83}$$

In particular, $\mathcal{B}(E, E) = 0$. When corresponding keys and values are identical within each class and across presentations, the bound reduces to the class-weight bound in Theorem 2.

For a distributional statement, let ξ specify the question, semantic class identities, and target answer, with a common law μ. Let $\kappa_{\mathrm{s}}(dE \mid \xi)$ and $\kappa_{\mathrm{t}}(dE' \mid \xi)$ be conditional laws of complete presentations satisfying the preceding conditions. These laws may change both multiplicity and the within-class representations. Assume the kernels and all quantities above are measurable on standard Borel spaces. Suppose the two risks are finite and the fixed loss, including the downstream readout and decoder, obeys $0 \le \ell_\xi \le B$ and $|\ell_\xi(F) - \ell_\xi(G)| \le L(\xi)\|F - G\|_{\mathrm{F}}$, where $B \in (0, \infty]$ and $0 \le L(\xi) < \infty$. Let Γ_E contain joint laws of (ξ, E, E') with marginals $\mu(d\xi)\kappa_{\mathrm{s}}(dE \mid \xi)$ and $\mu(d\xi)\kappa_{\mathrm{t}}(dE' \mid \xi)$. Then

$$|R_{\mathrm{t}} - R_{\mathrm{s}}| \le \inf_{\gamma \in \Gamma_E} \mathbb{E}_\gamma \min\{B, L(\xi)\mathcal{B}(E, E')\}. \tag{84}$$

Proof. Fix the question and the pair of presentations throughout the pointwise argument.

Step 1: exact aggregation without identical representations. Grouping the finite sums in the fusion equation yields

$$\begin{aligned} f_r(E) &= \frac{\sum_c \sum_{i \in \mathcal{C}_c(E)} \sum_t e^{a_{rit}(E)} v_{it}(E)/m_i(E)}{\sum_c \sum_{i \in \mathcal{C}_c(E)} \sum_t e^{a_{rit}(E)}/m_i(E)} \\ &= \frac{\sum_c W_{rc}(E) V_{rc}(E)}{\sum_j W_{rj}(E)} = \sum_c P_{rc}(E) V_{rc}(E). \end{aligned} \tag{85}$$

The same identity holds for E'. Unlike the exact-copy setting, both the class probabilities and the class barycenters can change.

Step 2: compare the two presentations directly. Add and subtract $\sum_c P'_{rc} V_{rc}$ and then apply the triangle inequality and Lemma 4:

$$\begin{aligned} \|f_r(E') - f_r(E)\|_2 &= \left\| \sum_c P'_{rc}(V'_{rc} - V_{rc}) + \sum_c (P'_{rc} - P_{rc}) V_{rc} \right\|_2 \\ &\le \left\| \sum_c P'_{rc}(V'_{rc} - V_{rc}) \right\|_2 + \left\| \sum_c (P'_{rc} - P_{rc}) V_{rc} \right\|_2 \\ &\le \sum_c P'_{rc} \|V'_{rc} - V_{rc}\|_2 + D_r(E)\, \mathrm{TV}(P'_r, P_r) \\ &= \sum_c P'_{rc} d_{rc} + D_r(E) T_r. \end{aligned} \tag{86}$$

Alternatively, add and subtract $\sum_c P_{rc} V'_{rc}$:

$$\begin{aligned} \|f_r(E') - f_r(E)\|_2 &= \left\| \sum_c P_{rc}(V'_{rc} - V_{rc}) + \sum_c (P'_{rc} - P_{rc}) V'_{rc} \right\|_2 \\ &\le \sum_c P_{rc} \|V'_{rc} - V_{rc}\|_2 + \left\| \sum_c (P'_{rc} - P_{rc}) V'_{rc} \right\|_2 \\ &\le \sum_c P_{rc} d_{rc} + D_r(E') T_r. \end{aligned} \tag{87}$$

Both inequalities hold simultaneously, so their minimum is also an upper bound. This retains cancellation of any representation errors shared by the two presentations; no intermediate prototype is introduced.

Step 3: control the class-probability change. Put $h_{rc} = \log(W'_{rc}/W_{rc})$. Since all masses are positive,

$$\begin{aligned} P'_{rc} &= \frac{W_{rc} e^{h_{rc}}}{\sum_j W_{rj} e^{h_{rj}}} = \frac{P_{rc} e^{h_{rc}}}{\sum_j P_{rj} e^{h_{rj}}}, \\ T_r &= \frac{1}{2} \sum_c P_{rc} \left| \frac{e^{h_{rc}}}{\sum_j P_{rj} e^{h_{rj}}} - 1 \right| \le \tanh\left(\frac{\mathrm{osc}_c\, h_{rc}}{4} \right) = \tanh(\Omega_r/4), \end{aligned} \tag{88}$$

where the inequality is Lemma 3. All coefficients multiplying T_r are nonnegative. Thus $\mathcal{B}_r \le \widehat{\mathcal{B}}_r$, and

$$\begin{aligned}\|F(E') - F(E)\|_{\mathrm{F}}^2 &= \sum_r \|f_r(E') - f_r(E)\|_2^2 \\ &\le \sum_r \mathcal{B}_r(E, E')^2 \le \sum_r \widehat{\mathcal{B}}_r(E, E')^2.\end{aligned} \tag{89}$$

Taking square roots proves (83).

Step 4: verify the limiting cases. If $E' = E$, then $d_{rc} = 0$, $T_r = 0$, and $\Omega_r = 0$, so both row budgets vanish exactly. For exact copies of fixed class representations, let $S_{rc} = \sum_t e^{u_r^\top k_{ct}^\star/\sqrt{d}}$ and $b_c(E) = \sum_{i \in \mathcal{C}_c(E)} m_i(E)^{-1}$. Then

$$\begin{aligned}W_{rc}(E) = S_{rc} b_c(E), \quad V_{rc}(E) = S_{rc}^{-1} \sum_t e^{u_r^\top k_{ct}^\star/\sqrt{d}} v_{ct}^\star, \\ d_{rc} = 0, \qquad\qquad \Omega_r = \operatorname{osc}_c \log \frac{b_c(E')}{b_c(E)}.\end{aligned} \tag{90}$$

Consequently $\mathcal{B}_r = D_r T_r$, which is precisely the class-weight bound in the main theorem.

Step 5: transfer the pairwise bound to risk. Conditional independence supplies an admissible coupling $\mu(d\xi)\kappa_{\mathrm{s}}(dE \mid \xi)\kappa_{\mathrm{t}}(dE' \mid \xi)$, so Γ_E is nonempty. For every $\gamma \in \Gamma_E$, finite risks and the marginal constraints give

$$\begin{aligned}|R_{\mathrm{t}} - R_{\mathrm{s}}| &= |\mathbb{E}_\gamma[\ell_\xi(F(E')) - \ell_\xi(F(E))]| \\ &\le \mathbb{E}_\gamma |\ell_\xi(F(E')) - \ell_\xi(F(E))| \\ &\le \mathbb{E}_\gamma \min\{B, L(\xi)\|F(E') - F(E)\|_{\mathrm{F}}\} \\ &\le \mathbb{E}_\gamma \min\{B, L(\xi)\mathcal{B}(E, E')\}.\end{aligned} \tag{91}$$

The assertion follows by taking the infimum over Γ_E. No optimal coupling is assumed to exist. □

The theorem provides a direct comparison for presentations whose semantic equivalence is accompanied by controlled d_{rc} and Ω_r. Those quantities must be measured for the presentations under study. The diameter and triangle inequalities can be strict, so the complete near-duplicate and downstream-risk bound depends on the observed geometry.

D.15 A Complementary Bound from Certified Prototype Errors

When a reference representation is available, an error certificate provides another computable bound. Fix one prototype key and value $k_{ct}^\star, v_{ct}^\star$ for each class and local slot, and let $F^\star$ be attention over exactly one copy of these prototype slots, with the same queries as the actual presentation.

Proposition 13 (Oscillation-sensitive prototype certificate)**.** *For a presentation E, let $\nu_i = |\mathcal{C}_{c(i)}(E)|$, $d_i = m_i/\nu_i$, $d_- = \min_i d_i > 0$, and $d_+ = \max_i d_i$. Suppose, for every i, t, r,*

$$\left|\frac{u_r^\top (k_{it} - k_{c(i)t}^\star)}{\sqrt{d}}\right| \le \epsilon_{k,r}, \qquad \|v_{it} - v_{c(i)t}^\star\|_2 \le \epsilon_v, \tag{92}$$

where the error tolerances are finite and nonnegative. Let $D_\star = \max_{c,t,j,l} \|v_{ct}^\star - v_{jl}^\star\|_2$. Then

$$\|F(E) - F^\star\|_{\mathrm{F}} \le \mathcal{E}(E) := \left\{\sum_r \left[\epsilon_v + D_\star \tanh\left(\frac{2\epsilon_{k,r} + \log(d_+/d_-)}{4}\right)\right]^2\right\}^{1/2}. \tag{93}$$

Proof. Fix a query r and index all occurrence slots by (i, t). Introduce the corrected logits and reference occurrence logits

$$\begin{aligned}\alpha_{it} &= u_r^\top k_{it}/\sqrt{d} - \log m_i, \\ \alpha_{it}^\star &= u_r^\top k_{c(i)t}^\star/\sqrt{d} - \log \nu_i, \\ \zeta_{it} &= \alpha_{it} - \alpha_{it}^\star = \frac{u_r^\top (k_{it} - k_{c(i)t}^\star)}{\sqrt{d}} - \log d_i.\end{aligned} \tag{94}$$

Let $p = \mathrm{softmax}(\alpha^\star)$ and $\widehat{p} = \mathrm{softmax}(\alpha)$. Summing the reference occurrence masses within a class cancels the number of occurrences:

$$\begin{aligned} \sum_{i \in \mathcal{C}_c(E)} e^{\alpha^\star_{it}} &= \sum_{i \in \mathcal{C}_c(E)} \nu_c^{-1} e^{u_r^\top k^\star_{ct}/\sqrt{d}} = e^{u_r^\top k^\star_{ct}/\sqrt{d}}, \\ \sum_{i,t} p_{it} v^\star_{c(i)t} &= \frac{\sum_{c,t} e^{u_r^\top k^\star_{ct}/\sqrt{d}} v^\star_{ct}}{\sum_{c,t} e^{u_r^\top k^\star_{ct}/\sqrt{d}}} = f_r^\star. \end{aligned} \tag{95}$$

The assumptions and monotonicity of the logarithm imply

$$\begin{aligned} -\epsilon_{k,r} - \log d_+ \le \zeta_{it} &\le \epsilon_{k,r} - \log d_-, \\ \mathrm{osc}_{i,t}\, \zeta_{it} &\le (\epsilon_{k,r} - \log d_-) - (-\epsilon_{k,r} - \log d_+) \\ &= 2\epsilon_{k,r} + \log(d_+/d_-). \end{aligned} \tag{96}$$

Since $\widehat{p}$ is the exponential tilt of p by ζ, Lemmas 3 and 4 give

$$\begin{aligned} \|f_r(E) - f_r^\star\|_2 &= \left\| \sum_{i,t} \widehat{p}_{it}(v_{it} - v^\star_{c(i)t}) + \sum_{i,t} (\widehat{p}_{it} - p_{it}) v^\star_{c(i)t} \right\|_2 \\ &\le \sum_{i,t} \widehat{p}_{it} \|v_{it} - v^\star_{c(i)t}\|_2 + D_\star\, \mathrm{TV}(\widehat{p}, p) \\ &\le \epsilon_v \sum_{i,t} \widehat{p}_{it} + D_\star \tanh\!\left(\frac{\mathrm{osc}_{i,t}\, \zeta_{it}}{4} \right) \\ &\le \epsilon_v + D_\star \tanh\!\left(\frac{2\epsilon_{k,r} + \log(d_+/d_-)}{4} \right). \end{aligned} \tag{97}$$

The first term uses the actual attention weights, whose sum is one; the second uses the diameter of the reference values, unchanged by their replication. Finally,

$$\begin{aligned} \|F(E) - F^\star\|_{\mathrm{F}}^2 &= \sum_r \|f_r(E) - f_r^\star\|_2^2 \\ &\le \sum_r \left[\epsilon_v + D_\star \tanh\!\left(\frac{2\epsilon_{k,r} + \log(d_+/d_-)}{4} \right) \right]^2 = \mathcal{E}(E)^2. \end{aligned} \tag{98}$$

Taking square roots proves the proposition. □

If $|d_i - 1| \le \epsilon_m < 1$ and $\|v^\star_{ct}\|_2 \le V_{\max}$, the certificate satisfies

$$\begin{aligned} \log \frac{d_+}{d_-} \le \log \frac{1+\epsilon_m}{1-\epsilon_m} &= \log(1-\epsilon_m^2) - 2\log(1-\epsilon_m) \le -2\log(1-\epsilon_m), \\ D_\star \le \max_{c,t,j,l} (\|v^\star_{ct}\|_2 + \|v^\star_{jl}\|_2) &\le 2V_{\max}. \end{aligned} \tag{99}$$

The oscillation form therefore retains more information than a uniform absolute corrected-logit error combined with a value-norm bound. In particular, common shifts of all corrected logits make no contribution to their oscillation.

Corollary 14 (Combining direct and prototype information)**.** *If both presentations satisfy Proposition 13 relative to the same prototype collection, with their respective error tolerances, then*

$$\|F(E') - F(E)\|_{\mathrm{F}} \le \min\{\mathcal{B}(E, E'), \mathcal{E}(E) + \mathcal{E}(E')\}. \tag{100}$$

Under the distributional assumptions of Theorem 12, if the prototype collection is a measurable function of ξ shared by both domains and these certificates hold almost surely for their presentations, then

$$|R_{\mathrm{t}} - R_{\mathrm{s}}| \le \inf_{\gamma \in \Gamma_E} \mathbb{E}_\gamma \min\{B, L(\xi)\mathcal{B}(E, E'), L(\xi)[\mathcal{E}(E) + \mathcal{E}(E')]\}. \tag{101}$$

Proof. The shared prototype collection implies a common $F^\star$. The triangle inequality and the two certificates give

$$\begin{aligned}\|F(E') - F(E)\|_{\mathrm{F}} &\le \|F(E') - F^\star\|_{\mathrm{F}} + \|F(E) - F^\star\|_{\mathrm{F}} \\ &\le \mathcal{E}(E') + \mathcal{E}(E).\end{aligned} \tag{102}$$

Combining this with (83) proves (100). Substitution into the loss bound in (91), followed by the same infimum over Γ_E, proves (101). □

The prototype certificate is complementary: its sum can remain positive even when $E' = E$. The direct term restores exact self-comparison and is never worsened by retaining the optional prototype certificate through the minimum above.

D.16 Application to the Answer Cross-Entropy Loss

Lemma 15 (A sufficient decoder condition)**.** *For a fixed question and teacher-forced answer prefix, let $g_{\xi,\tau}(F)$ be the finite-vocabulary logit vector at answer position τ. Suppose $0 \le G_\tau(\xi) < \infty$ and*

$$\|g_{\xi,\tau}(F) - g_{\xi,\tau}(G)\|_2 \le G_\tau(\xi)\|F - G\|_{\mathrm{F}}. \tag{103}$$

Then the mean answer cross-entropy satisfies the loss Lipschitz condition with

$$L(\xi) = \frac{\sqrt{2}}{T_y}\sum_{\tau=1}^{T_y} G_\tau(\xi). \tag{104}$$

Thus Theorem 2 applies with $B = \infty$ whenever the two risks are finite.

Proof. For a target token y, set $c_y(z) = \log\sum_j e^{z_j} - z_y$ and $p = \mathrm{softmax}(z)$. Its gradient is $p - e_y$, so

$$\begin{aligned}\|\nabla c_y(z)\|_2^2 &= (1-p_y)^2 + \sum_{j\ne y} p_j^2 \\ &\le (1-p_y)^2 + \left(\sum_{j\ne y} p_j\right)^2 \\ &= 2(1-p_y)^2 \le 2.\end{aligned} \tag{105}$$

Integrating the gradient along the line segment from z to z' and applying Cauchy–Schwarz gives

$$\begin{aligned}|c_y(z') - c_y(z)| &= \left|\int_0^1 \langle \nabla c_y(z + t(z'-z)), z' - z\rangle\, dt\right| \\ &\le \int_0^1 \|\nabla c_y(z + t(z'-z))\|_2 \|z' - z\|_2\, dt \\ &\le \sqrt{2}\|z' - z\|_2.\end{aligned} \tag{106}$$

Therefore the teacher-forced mean loss obeys

$$\begin{aligned}|\ell_\xi(F) - \ell_\xi(G)| &\le \frac{1}{T_y}\sum_{\tau=1}^{T_y} |c_{y_\tau}(g_{\xi,\tau}(F)) - c_{y_\tau}(g_{\xi,\tau}(G))| \\ &\le \frac{\sqrt{2}}{T_y}\sum_{\tau=1}^{T_y} \|g_{\xi,\tau}(F) - g_{\xi,\tau}(G)\|_2 \\ &\le \frac{\sqrt{2}}{T_y}\sum_{\tau=1}^{T_y} G_\tau(\xi)\|F - G\|_{\mathrm{F}} \\ &= L(\xi)\|F - G\|_{\mathrm{F}}.\end{aligned} \tag{107}$$

□

The logit Lipschitz constants are assumptions on the fixed downstream map, not consequences of accurate relation prediction. No bounded cross-entropy assumption is made. Discrete greedy accuracy is not a Lipschitz loss and requires additional decision-margin assumptions.