Hidden Clones: Exposing and Fixing Family Bias in Vision-Language Model Ensembles

Anonymous CVPR submission

Paper ID *****

Abstract

Ensembling Vision-Language Models (VLMs) maximizes benchmark accuracy, yet models from the same architectural family *share correlated errors that standard voting ignores. We study 17 VLMs from 8 families on VQAv2, TextVQA, and GQA. Family-correlated errors reduce effective ensemble dimensionality to 2.5–3.6 independent voters and create a* Misleading *tier (1.5–6.5% of questions) where calibrated voting collapses to 0% despite the best model being correct. We propose three family-aware methods:* ***Hierarchical Family Voting*** *(HFV) recovers +18–26 pp on the Misleading tier;* ***QualRCCV****, a training-free method, is the first to beat calibrated voting on all three benchmarks ($p<0.05$);* ***Learned Candidate Scoring*** *(LCS) achieves $+0.68\%$ VQAv2, $+0.61\%$ TextVQA, $+2.45\%$ GQA—all significant—and is the only learned method that never degrades any benchmark.*

1. Introduction

Combining predictions from multiple models is the default strategy for maximizing VQA accuracy [8, 17]. Condorcet's jury theorem [3] guarantees that majority voting improves with more independent, better-than-random voters. In practice, however, state-of-the-art VLM ensembles draw from a small number of *architectural families*—Qwen2.5-VL, InternVL, Molmo, LLaVA, etc.—where models within a family share training data, architecture, and pre-training methodology. This creates a hidden structure: **within-family errors are strongly correlated**, violating the independence assumption.

We present the first multi-benchmark study of this family structure across 17 VLMs from 8 families on VQAv2 (N=20,001), TextVQA (N=5,000), and GQA (N=12,578). Our contributions: (1) A **multi-benchmark analysis** showing that eigenvalue structure reduces 17 models to only 2–4 effective voters, with a *Misleading tier* where calibrated voting achieves 0% despite the best model being correct. (2) **Hierarchical Family Voting** (HFV), a training-free method that recovers the Misleading tier by +18–26 pp. (3) **QualRCCV**, the first training-free method to beat calibrated voting on all three benchmarks ($p<0.05$). (4) **Learned Candidate Scoring** (LCS), achieving the largest gains ($+2.45\%$ GQA) and the only learned method that never degrades any benchmark.

2. Related Work

Condorcet's jury theorem [3] establishes that majority voting improves with more independent, better-than-random voters. Extensions to correlated voters [1, 11] predict degradation when errors are positively correlated, while the bias–variance–covariance decomposition [2, 14] formalizes how ensemble error depends on individual accuracy and pairwise diversity. Kuncheva & Whitaker [10] survey diversity measures and find no single measure reliably predicts performance.

Recent LLM ensemble work includes LLM-Blender [7], which trains a ranking model to select the best response, and More-Agents-Is-All-You-Need [12], which shows scaling agents improves reasoning via majority voting. Self-consistency [15] samples multiple reasoning paths and votes. These methods assume approximate independence; we show that family structure violates this assumption and propose family-aware corrections. Hierarchical aggregation appears in stacking [16] and mixture of experts [6]; we are the first to apply architecture-family-level hierarchical voting to VLM ensembles.

3. Setup

We assemble 17 VLMs from 8 families: 5 Qwen2.5-VL (7B–72B, including fine-tuned and LoRA variants), 2 Qwen3-VL (8B, 32B), 2 InternVL (v2, v3), 2 Molmo2-8B, Phi-4-multimodal, 2 LLaVA (OneVision, NeXT), Pixtral-12B, and 2 Idefics (Idefics3, SmolVLM). Individual accuracy ranges from 60.4% (Phi-4 on VQAv2) to 86.3% (Molmo on VQAv2), with family dominance varying by benchmark: Molmo leads on VQAv2, Qwen2.5-VL LoRA variants lead on TextVQA, and LLaVA-NeXT leads on GQA.

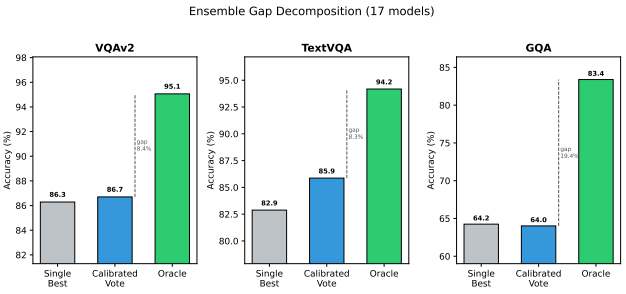

Figure 1. Gap decomposition across benchmarks. Calibrated voting captures only a small fraction of the gap between single-best and oracle accuracy, especially on VQAv2 and GQA.

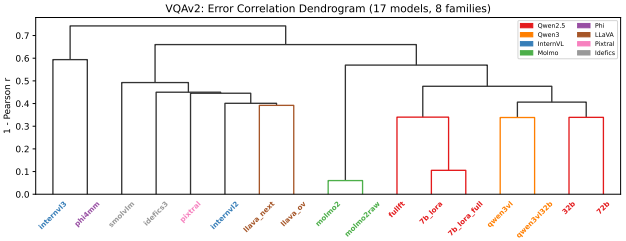

Figure 2. Hierarchical clustering on error correlation distance. Family-colored leaves show architecture families cluster together.

We evaluate on VQAv2 minival [4] (N=20,001; soft accuracy, 10 annotators), TextVQA val [13] (N=5,000; OCR reasoning), and GQA testdev [5] (N=12,578; compositional reasoning, exact match). Baselines include majority voting, calibrated voting (per-model log-odds weights $w_m = \log(p_m/(1-p_m))$), and the per-question oracle. All significance tests use paired bootstrap (2,000 resamples, 95% CIs).

4. Family Structure in VLM Ensembles

The ensemble ceiling gap. On VQAv2, calibrated voting reaches 86.70%—only 0.41% above the best model (86.29%)—yet the oracle achieves 95.06%, an **8.8% gap** (Fig. 1). Only 4.7% of this gap is captured by voting; the remainder requires per-question model selection. This ceiling motivates investigating *why* voting fails.

Error correlations cluster by family. Pearson correlation of per-question accuracy vectors across all 136 model pairs shows within-family $r = 0.67 \pm 0.12$ vs. cross-family $r = 0.53 \pm 0.07$ (Mann-Whitney $p < 0.001$). Hierarchical clustering on the correlation matrix recovers family-aligned groups (Fig. 2), and spectral clustering at k=8 achieves ARI = 0.42, NMI = 0.82 without any label information. Eigenvalue analysis reveals effective dimensionality of only **2.86** on VQAv2 (3.59 TextVQA, 2.49 GQA): 17 models carry the statistical power of ∼3 independent voters [9].

Table 1. Difficulty taxonomy on VQAv2 (17 models). T2 (Misleading) shows catastrophic failure: calibrated voting achieves 0%, HFV recovers +26 pp.

Tier	% Q's	Best	Cal	HFV	Δ
T0: Trivial	41.6	97.8	97.9	98.0	+0.1
T1: Easy	47.7	91.5	91.7	90.2	−1.5
T2: Misleading	2.5	78.9	**0.0**	**26.0**	**+26.0**
T3: Hard	7.1	0.0	30.6	29.5	−1.1
T4: Impossible	1.1	0.0	0.0	0.0	0

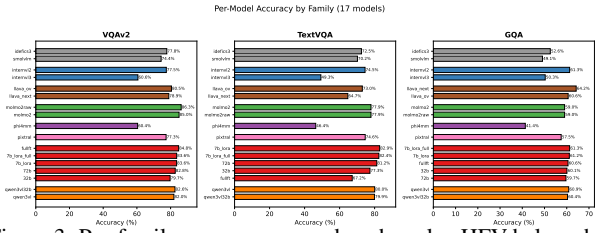

Figure 3. Per-family accuracy across benchmarks. HFV helps when family quality is balanced (VQAv2) but hurts when one family is dramatically weaker (InternVL3 at 49% on TextVQA).

The Misleading tier. We classify questions into five difficulty tiers (Tab. 1). The key finding is tier T2 (*Misleading*): the best model is correct but its answer is outvoted by correlated errors from same-family models. On VQAv2, this affects 2.5% of questions where calibrated voting drops to **0%** despite the best model achieving 79%. This pathology appears on all three benchmarks (1.5–6.5%).

5. Family-Aware Aggregation Methods

Hierarchical Family Voting (HFV). HFV aggregates in two stages. *Stage 1*: within each family f, compute a family answer via calibrated voting: $\hat{a}_f = \arg\max_a \sum_{m \in f} w_m \cdot \mathbf{1}[a_m{=}a]$, where $w_m = \log(p_m/(1-p_m))$. *Stage 2*: aggregate family answers using family-level weights $W_f = \log(P_f/(1-P_f))$, where P_f is family f's internal ensemble accuracy: $\hat{a} = \arg\max_a \sum_f W_f \cdot \mathbf{1}[\hat{a}_f{=}a]$. By collapsing each family to a single vote, HFV decorrelates the voting pool: the Qwen2.5 family's 5 correlated votes become 1 effective vote, properly reflecting the true degrees of freedom. **HFV-sharp** raises W_f to a power $\alpha{>}1$ to down-weight weak families, selected via cross-validation.

When HFV helps and fails. HFV's aggregate effect depends on family quality balance (Fig. 3). On VQAv2, HFV-sharp achieves +0.49% ($p{<}0.0001$) and on GQA +0.25% ($p{=}0.087$). On TextVQA (−0.60%), InternVL3 collapses to 49% and Phi-4 to 46%—near random for OCR tasks—so equalizing families gives poor predictions undue influence. QualRCCV's soft correction avoids this pathology.

QualRCCV. Rather than the hard two-level split of HFV, QualRCCV applies a soft redundancy correction by dividing each model's calibrated weight by its family size and scaling by family quality:

$$\hat{a} = \arg\max_{a} \sum_{m} \frac{w_m \cdot q_{F(m)}^{\gamma}}{|F(m)|^{\rho}} \cdot \mathbf{1}[\hat{a}_m{=}a] \qquad (1)$$

where $q_f = \max_{m \in f} \mathrm{acc}(m)$, ρ=0.4, γ=1.0. This preserves high-quality family contributions while correcting for numerical dominance.

Learned Candidate Scoring (LCS). LCS scores *individual candidate answers* rather than choosing between aggregation methods. For each question, it generates top-K candidates ranked by QualRCCV weight, extracts per-candidate features (support breadth, family diversity, supporter quality, margin), and trains a gradient-boosted classifier (LightGBM) via 5-fold cross-validation to predict $P(\text{correct} \mid \text{features})$. The dominant feature is the QualRCCV margin (importance >0.77).

6. Experiments

Tab. 3 presents results across all three benchmarks. **QualRCCV** is the first training-free method to beat calibrated voting on all three benchmarks simultaneously: +0.17% VQAv2 (p=0.003), +0.21% TextVQA (p=0.034), +0.31% GQA (p=0.003). **LCS** achieves the largest gains: +0.68% VQAv2, +0.61% TextVQA, +2.45% GQA—all statistically significant (p<0.0001). LCS is the only learned method that never degrades any benchmark (FAAR-learn degrades TextVQA by −0.87%).

The GQA puzzle. GQA is the benchmark where standard calibrated voting *underperforms* the single best model (64.02% vs. 64.25%): correlated family errors overwhelm the weaker models' contributions. LCS recovers to **66.47%**—more than 2.2 pp above the best individual model—demonstrating that family correlation causes real performance degradation that answer-level scoring can reverse.

Test-set evaluation. We train LCS on the full VQAv2 minival set and submit predictions to the EvalAI leaderboard. Using 12 models from 5 families (5 models lack test-set predictions), LCS achieves **87.83%** on test-standard, confirming generalization beyond cross-validation.

HFV consistently recovers the Misleading tier across all benchmarks (Tab. 2): +26.0 pp (VQAv2), +18.3 pp (TextVQA), +23.7 pp (GQA).

A balanced 8-model ensemble (one per family) nearly matches 17 models on VQAv2 (−0.07%) and matches on GQA, confirming that **family diversity matters more than model count**.

Table 2. Misleading tier (T2) recovery across benchmarks. Calibrated voting achieves 0% on T2 questions where the best model is correct; HFV consistently recovers a large fraction.

Benchmark	T2 %	Cal	HFV	Δ
VQAv2	2.5%	0%	26.0%	**+26.0**
TextVQA	1.5%	0%	18.3%	**+18.3**
GQA	6.5%	0%	23.7%	**+23.7**

Table 3. Main results (95% bootstrap CIs, 17 models, 8 families). QualRCCV: first training-free method to beat calibrated voting on all three benchmarks. LCS: largest gains, only learned method that never degrades. †5-fold CV.

Method	VQAv2	TextVQA	GQA
Single best	86.29 [85.9, 86.7]	82.88 [81.9, 83.9]	64.25 [63.4, 65.1]
Calibrated vote	86.70 [86.3, 87.1]	85.87 [85.0, 86.8]	64.02 [63.2, 64.9]
Training-free			
QualRCCV	86.87 [86.4, 87.3]	**86.07** [85.2, 87.0]	64.33 [63.6, 65.2]
HFV-sharp	87.19 [86.8, 87.6]	85.27	64.27
Learned (5-fold CV)			
LCS†	**87.38** [87.0, 87.8]	86.48 [85.6, 87.4]	**66.47** [65.7, 67.3]
Oracle	95.06	94.18	83.39

7. Conclusion

Within-family error correlation (r=0.67 vs. 0.53 cross-family) reduces 17 VLMs to ∼3 effective voters and creates a Misleading tier (1.5–6.5% of questions) where calibrated voting achieves 0%. HFV recovers this tier (+18–26 pp). QualRCCV is the first training-free method to beat calibrated voting on all three benchmarks (p<0.05). LCS achieves the largest gains (+2.45% GQA) and never degrades any benchmark. Practitioners should prioritize architectural diversity over model count and use family-aware aggregation.

Limitations. Our ensemble is dominated by one family (5/17 Qwen2.5-VL); more balanced pools may show smaller effects. We evaluate only short-answer VQA, not open-ended generation. LCS uses 5-fold cross-validation with all calibration on train folds only; however, GBM hyperparameters were selected on development data.

References

[1] Sven Berg. Condorcet's jury theorem, dependency among jurors. *Social Choice and Welfare*, 10(1):87–95, 1993. 1

[2] Gavin Brown, Jeremy Wyatt, Rachel Harris, and Xin Yao. Diversity creation methods: A survey and categorisation. *Information Fusion*, 6(1):5–20, 2005. 1

[3] Marquis de Condorcet. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. 1785. 1

[4] Yash Goyal et al. Making the V in VQA matter: Elevating

the role of image understanding in visual question answering. In *CVPR*, 2017. 2

[5] Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019. 2

[6] Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991. 1

[7] Dongfu Jiang et al. LLM-Blender: Ensembling large language models with pairwise ranking and generative fusion. In *ACL*, 2023. 1

[8] Huaizu Jiang et al. In defense of grid features for visual question answering. In *CVPR*, 2020. 1

[9] Leslie Kish. *Survey Sampling*. Wiley, 1965. 2

[10] Ludmila I. Kuncheva and Christopher J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2):181–207, 2003. 1

[11] Krishna K. Ladha. The Condorcet jury theorem, free speech, and correlated votes. *American Journal of Political Science*, 36(3):617–634, 1992. 1

[12] Juanjuan Li et al. More agents is all you need. *arXiv:2402.05120*, 2024. 1

[13] Amanpreet Singh et al. Towards VQA models that can read. In *CVPR*, 2019. 2

[14] Naonori Ueda and Ryohei Nakano. Generalization error of ensemble estimators. In *ICNN*, pages 90–95, 1996. 1

[15] Xuezhi Wang et al. Self-consistency improves chain of thought reasoning in language models. In *ICLR*, 2023. 1

[16] David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992. 1

[17] Zhou Yu et al. Deep modular co-attention networks for visual question answering. In *CVPR*, 2019. 1