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

Zacharie Bugaud
Astera Institute
zacharie@astera.org

Abstract

Ensembling Vision-Language Models (VLMs) from different providers maximizes benchmark accuracy, yet models from the same architectural family *share correlated errors that standard voting ignores. We study this structure across 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 correlated majority errors destroy accuracy to 0% despite the best model being correct.*

We propose three family-aware methods. ***Hierarchical Family Voting*** *(HFV) aggregates within families before voting across them, recovering +18–26 pp on the Misleading tier.* ***QualRCCV****, a training-free method weighting models by calibration, family quality, and inverse family size, is the first to beat calibrated voting on all three benchmarks (*p*<*0.05*).* ***Learned Candidate Scoring*** *(LCS) trains a cross-validated classifier to re-rank candidate answers using support breadth, family diversity, and model quality, achieving the largest gains:* $+0.68\%$ *VQAv2,* $+0.61\%$ *TextVQA,* $+2.45\%$ *GQA—all significant—and is the only learned method that never degrades any benchmark. On VQAv2 test-standard (EvalAI), LCS reaches* ***87.83%*** *with 12 models, confirming generalization.*

1. Introduction

Combining predictions from multiple models is the default strategy for maximizing accuracy in visual question answering (VQA) competitions [11, 24] and across machine learning more broadly. Condorcet's jury theorem [5] provides the theoretical motivation: majority voting improves with more voters, provided each voter is better than random and errors are independent. Conventional wisdom further holds that *diverse* ensembles—combining different architectures—outperform homogeneous ones [6].

In practice, state-of-the-art VLM ensembles are constructed from models belonging to a small number of *architectural families*—e.g. Qwen2.5-VL, Qwen3-VL, InternVL, Molmo, Phi-4, LLaVA-OneVision, Pixtral, Idefics3—where models within a family share training data, architecture, and pre-training methodology. This creates a hidden structure that standard ensemble methods ignore: **within-family errors are strongly correlated**, violating the independence assumption that makes voting powerful.

We present the first multi-benchmark study of this family structure, spanning 17 VLMs from 8 families across VQAv2 (N=20,001), TextVQA (N=5,000), and GQA (N=12,578). Our contributions are:

1. **A multi-benchmark analysis** of family-correlated errors revealing that eigenvalue structure reduces 17 models to only 2–4 effective voters, and a difficulty taxonomy identifying a *Misleading tier* (1.5–6.5% of questions) where calibrated voting collapses to **0%** despite the best model being correct (Sec. 4).
2. **Hierarchical Family Voting (HFV)**, a training-free method that aggregates within families first and then across them, recovering the Misleading tier by **+18–26 pp**. **HFV-sharp**, with cross-validation for α, achieves **87.19%** on VQAv2 ($+0.49\%$, p<0.0001) and **64.27%** on GQA ($+0.25\%$, p=0.087), remaining entirely training-free (Sec. 5).
3. **Quality-weighted Redundancy-Corrected Calibrated Voting (QualRCCV)**, a training-free single-level vote that weights each model by $w_m \cdot q_f^{\gamma}/|F(m)|^{\rho}$ where q_f is the family's best-member accuracy. QualRCCV is the first method to beat calibrated voting on *all three benchmarks*: $+0.17\%$ VQAv2 (p=0.003), $+0.21\%$ TextVQA (p=0.034), $+0.31\%$ GQA (p=0.003), remaining training-free with two hyperparameters (Sec. 5.4).
4. **Learned Candidate Scoring (LCS)**, a cross-validated method that scores individual candidate answers based on per-answer features (support breadth, family diversity, supporter quality). LCS achieves the largest gains of any method: $+0.68\%$ on VQAv2 (p<0.0001) and $+2.45\%$ on GQA (p<0.0001), while remaining posi-

tive on TextVQA (+0.61%, p<0.0001). LCS is the only learned method that never degrades any benchmark (Sec. 6).

2. Related Work

Ensemble theory and diversity. Condorcet's jury theorem [5] shows majority voting improves with more independent, better-than-random voters. Extensions to correlated voters [1, 2, 14] predict ensemble degradation when errors are positively correlated. The bias–variance–covariance decomposition [3, 19] formalizes how ensemble error depends on both individual accuracy and pairwise diversity. Kuncheva & Whitaker [13] survey diversity measures and show that no single measure reliably predicts ensemble accuracy. Our work contributes an *empirical* diversity analysis for VLM ensembles, revealing that architectural family membership is the dominant source of correlation structure.

LLM and VLM ensembles. LLM-Blender [10] trains a ranking model to select the best response from multiple LLMs. Mixture-of-Agents [21] iteratively refines outputs by passing responses through multiple LLMs. RouteLLM [16] and FrugalGPT [4] train routers or cascades to optimize cost–quality trade-offs. More-Agents-Is-All-You-Need [15] shows scaling the number of LLM agents improves performance on reasoning tasks through majority voting. Wang et al. [22] introduce self-consistency decoding, sampling multiple reasoning paths from a single model and voting. In contrast to methods requiring training data or iterative generation, our HFV method is training-free and operates on answer-level outputs from heterogeneous models.

Structured and hierarchical aggregation. Hierarchical voting appears in social choice theory (e.g., electoral colleges [5]) and in ensemble learning via stacking [23] and mixture of experts [9, 17]. Nested cross-validation and meta-learning approaches [20] aggregate base learners in stages. To our knowledge, we are the first to apply hierarchical *architecture-family-level* aggregation to VLM ensembles and to analyze when it helps versus hurts.

VQA benchmarks and evaluation. VQAv2 [7] introduced balanced image pairs to reduce language bias; TextVQA [18] requires OCR reasoning; GQA [8] tests compositional reasoning via scene graphs. Prior ensemble work on VQA focuses on homogeneous ensembles of task-specific models [11, 24]; we study heterogeneous ensembles of general-purpose VLMs.

3. Experimental Setup

Models. We assemble 17 VLMs from 8 architectural families (Tab. 1): 5 Qwen2.5-VL variants (7B fine-tuned on VQAv2, two 7B LoRA variants, 32B and 72B zero-shot), 2 Qwen3-VL variants (8B and 32B zero-shot), 2 InternVL variants (InternVL2-8B and InternVL3-8B, both zero-shot), 2 Molmo2-8B variants (one with prompt engineering, one raw), Phi-4-multimodal (14B zero-shot), 2 LLaVA variants (OneVision-7B and LLaVA-NeXT-Mistral-7B zero-shot), Pixtral-12B zero-shot, and 2 Idefics variants (Idefics3-8B and SmolVLM-2B zero-shot). All inference uses vLLM v0.11 or HuggingFace Transformers.

Benchmarks. We evaluate on three VQA benchmarks:

- **VQAv2** [7]: minival split, N=20,001 questions evenly split across yes/no, number, and other types (33.3% each). Soft accuracy with 10 annotators.
- **TextVQA** [18]: val split, N=5,000 questions requiring OCR and text reasoning. Soft accuracy with 10 annotators.
- **GQA** [8]: testdev split, N=12,578 questions from scene-graph-based compositional reasoning. Exact-match accuracy.

Aggregation baselines. We compare: *majority voting* (unweighted), *calibrated voting* (per-model log-odds weights based on overall accuracy), *deduplication* (best model per family, then calibrated vote), *correlation-aware weighting* (inverse-agreement weights), and the *per-question oracle* (selecting the best answer for each question).

Statistical testing. All confidence intervals are 95% bootstrap CIs (2,000 resamples). Significance of HFV vs. calibrated voting is assessed via a paired bootstrap test: we resample questions with replacement and compute the fraction of resamples where calibrated voting outperforms HFV. We report this as a one-sided p-value.

4. Analysis: Family Structure in VLM Ensembles

We first characterize the family structure of model errors and identify when standard ensembling fails catastrophically. All analysis in this section uses VQAv2 as the primary benchmark; Sec. 6 extends key findings across all three benchmarks.

4.1. The Ensemble Ceiling

On VQAv2, calibrated voting reaches 86.70%—just 0.41% above the best model (86.29%). Yet the oracle achieves 95.06%, an **8.8% gap**. Only 4.7% of this gap is captured by voting; 31% by routing (choosing between model and ensemble per question); 64% requires per-question model selection (Fig. 1).

Table 1. Model inventory and individual accuracy across three benchmarks. Family dominance varies: Molmo leads on VQAv2, while Qwen2.5-VL LoRA variants lead on TextVQA; LLaVA-NeXT leads on GQA. Models fine-tuned on VQAv2 (fullft) transfer poorly to TextVQA (67%). InternVL3 and Phi-4 collapse below 50% on TextVQA. All 17 models evaluated on all benchmarks.

Model	Family	Size	VQAv2	TextVQA	GQA
molmo2raw	Molmo	8B	**86.3**	77.9	59.0
molmo2	Molmo	8B	85.0	77.9	59.0
fullft	Qwen2.5	7B	84.8	67.2	60.6
7b_lora	Qwen2.5	7B	83.6	**82.9**	61.2
7b_lora_full	Qwen2.5	7B	83.6	82.4	61.3
72b_zs	Qwen2.5	72B	82.8	81.2	59.7
qwen3vl32b	Qwen3	32B	82.6	79.9	60.4
qwen3vl	Qwen3	8B	82.0	80.0	60.9
llava_ov	LLaVA	7B	80.5	73.0	60.6
32b_zs	Qwen2.5	32B	79.7	77.3	60.1
llava_next	LLaVA	7B	78.9	64.7	**64.3**
idefics3	Idefics	8B	77.8	72.5	52.6
internvl2	InternVL	8B	77.5	74.5	61.3
pixtral	Pixtral	12B	77.3	74.6	57.5
smolvlm	Idefics	2B	74.4	70.2	49.1
internvl3	InternVL	8B	60.6	49.3	50.3
phi4mm	Phi	14B	60.4	46.4	41.4

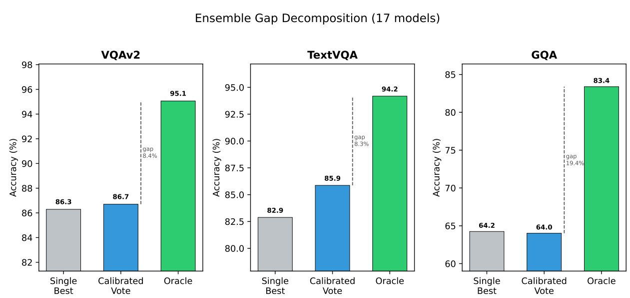

Figure 1. Gap decomposition across benchmarks. Calibrated voting captures only a small fraction of the gap between single-best and oracle accuracy, especially on VQAv2 and GQA.

4.2. Difficulty Taxonomy

We classify questions into five tiers (Tab. 2): **Trivial** (all correct), **Easy** (best model and majority correct), **Misleading** (best model correct, majority wrong), **Hard** (best model wrong, some correct), and **Impossible** (none correct).

The most striking finding is tier T2: the best model's correct answer is *outvoted* by correlated errors from same-family models (5/17 from Qwen2.5-VL).

4.3. Error Correlation Has Family Structure

Pearson correlation of per-question accuracy vectors across all 136 model pairs (Fig. 2) shows: within-family $r = 0.67 \pm 0.12$, cross-family $r = 0.53 \pm 0.07$.

This gap is significant (Mann-Whitney $p < 0.001$) and is

Table 2. Difficulty taxonomy on VQAv2 (17 models). The Misleading tier (T2, 2.5%) shows catastrophic failure: the best model achieves 79% but calibrated voting collapses to **0%**. HFV recovers +26.0 pp of this tier.

Tier	% Q's	Single	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

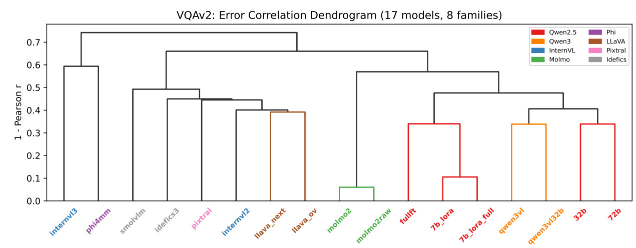

Figure 2. Hierarchical clustering (Ward linkage) on error correlation distance. Family-colored leaves reveal that architecture families cluster together, confirming correlated within-family errors.

the root cause of the Misleading tier: same-family models share systematic biases that amplify incorrect consensus.

Effective number of voters. Eigenvalue analysis of the error correlation matrix reveals 58.0% of variance in a single component and 75.2% in the top 5. The effective dimensionality (participation ratio) is only **2.86**, meaning 17 models have the statistical power of fewer than $\sim$3 independent voters [12].

Data-driven family discovery. If family structure is a real property of the error landscape, unsupervised clustering should recover architecture-aligned groups without any label information. We apply spectral clustering on the error correlation affinity matrix (Fig. 4). At k=8 (matching the true number of families), spectral clustering recovers architecture-aligned groups (ARI = 0.42, NMI = 0.82), with the modest ARI reflecting that some families (e.g. InternVL with heterogeneous members) are harder to separate from cross-family neighbours. At k=9, further splitting yields ARI = 0.43, NMI = 0.82, and the score continues to improve at higher k (ARI = 0.54 at k=12), suggesting that sub-family structure (e.g. Qwen2.5 scale groups) is also discoverable. This confirms that architecture families are *discoverable* from the data, not assumed (Fig. 4; see supplementary material for full spectral clustering results).

5. Hierarchical Family Voting (HFV)

The analysis above reveals that standard calibrated voting treats all models as independent voters, ignoring the family structure that causes correlated errors. We propose **Hierarchical Family Voting** (HFV), a training-free aggregation method that explicitly accounts for this structure.

5.1. Standard Calibrated Voting

In calibrated voting, each model is weighted by its log-odds accuracy $w_m = \log(p_m/(1-p_m))$ and the ensemble selects $\hat{a} = \arg\max_a \sum_m w_m \cdot \mathbf{1}[a_m = a]$. This ignores *correlation*: five Qwen2.5-VL models with similar errors collectively dominate the vote.

5.2. HFV: Two-Level Aggregation

HFV aggregates in two stages:

Stage 1: Within-family aggregation. For each family f, compute a family-level answer using calibrated voting *within* the family:

$$\hat{a}_f = \arg\max_a \sum_{m \in f} w_m \cdot \mathbf{1}[a_m = a] \tag{1}$$

Stage 2: Cross-family voting. Aggregate family-level answers using family-level weights. Each family's weight is the log-odds of its Stage 1 accuracy:

$$W_f = \log\left(\frac{P_f}{1 - P_f}\right), \quad \hat{a} = \arg\max_a \sum_{f=1}^{F} W_f \cdot \mathbf{1}[\hat{a}_f = a] \tag{2}$$

where P_f is the calibrated-vote accuracy of family f's internal ensemble.

Why HFV works. By collapsing each family to a single vote, HFV *decorrelates* the voting pool. Standard voting gives the Qwen2.5 family 5 of 17 votes—a 5:2:2:2:2:2:1:1 ratio—but when within-family errors are highly correlated, these 5 votes carry little more information than 1. HFV reduces to F=8 effectively independent voters weighted by family quality, properly reflecting the true degrees of freedom. On the Misleading tier, where Qwen2.5's five models all agree on the wrong answer, HFV correctly resolves by giving other families' votes equal standing.

Proposition 1 (When HFV outperforms flat voting)**.** *Consider a binary question with F families of sizes $n_1, \ldots, n_F$. Let ρ_w be the average within-family error correlation and ρ_b the average between-family correlation, and let P_f be the accuracy of each family's internal ensemble. HFV outperforms flat voting when all of the following hold: (i) the correlation gap $\rho_w - \rho_b > 0$ (family structure exists); (ii) $\min_f P_f > 0.5$ (all families are better than random); (iii) the family size distribution is imbalanced (so flat voting overweights the largest family).*

Algorithm 1 Hierarchical Family Voting (HFV)

Require: Models $m_1, \ldots, m_M$ partitioned into families $\mathcal{F}_1, \ldots, \mathcal{F}_F$; accuracy-based weights w_m (log-odds of model accuracy)
Ensure: Ensemble answer $\hat{a}$ for each question
for each family $f = 1, \ldots, F$ **do**
  $\hat{a}_f \leftarrow \arg\max_a \sum_{m \in \mathcal{F}_f} w_m \cdot \mathbf{1}[a_m = a]$ {Within-family vote}
  $W_f \leftarrow \log(P_f/(1 - P_f))$ {Family-level weight}
end for
$\hat{a} \leftarrow \arg\max_a \sum_{f=1}^{F} W_f \cdot \mathbf{1}[\hat{a}_f = a]$ {Cross-family vote}

Intuition. Under flat voting, a family of size n_f casts n_f highly correlated votes, effectively inflating its influence to $\sim n_f/(1 + (n_f - 1)\rho_w)$ independent votes via the Kish effective sample size. When ρ_w is high, these n_f votes behave like a single vote but are *counted* n_f times, distorting the majority. HFV collapses each family to one vote, removing this distortion. Condition (ii) ensures that no family vote is adversarial; when it fails (e.g., InternVL3 at 49.3% on TextVQA), giving that family equal standing introduces quality dilution that exceeds the correlation benefit. Condition (iii) is the "trigger": with equal-size families, flat voting already approximates HFV.

5.3. HFV-sharp: Sharpened Cross-Family Weights

Standard HFV gives each family a weight proportional to its log-odds accuracy W_f. When the model pool includes weak families (e.g., InternVL3 at 60.6% or Phi-4 at 60.4% on VQAv2), equalizing influence can hurt aggregate accuracy. HFV-sharp addresses this by raising cross-family weights to a power $\alpha > 1$:

$$\hat{a} = \arg\max_a \sum_{f=1}^{F} W_f^{\alpha} \cdot \mathbf{1}[\hat{a}_f = a] \tag{3}$$

When α=1 this recovers standard HFV; as α grows, stronger families increasingly dominate the cross-family vote.

5.4. Extensions

Redundancy-Corrected Calibrated Voting (RCCV). HFV addresses family correlation through hard two-level aggregation, but this equalisation can amplify weak families. We propose a softer alternative: **RCCV** divides each model's calibrated weight by its family size raised to a power ρ, producing a single-level weighted vote with built-in redundancy correction:

$$\hat{a} = \arg\max_{a} \sum_{m=1}^{M} \frac{w_m(t_q)}{|F(m)|^{\rho}} \cdot \mathbf{1}[\hat{a}_m = a] \tag{4}$$

where $F(m)$ denotes the family of model m and $|F(m)|$ its size.

Quality-weighted RCCV (QualRCCV). RCCV corrects for redundancy but treats all families equally regardless of quality. We extend it by additionally scaling each model's weight by its family's quality—measured as the maximum accuracy among family members:

$$\hat{a} = \arg\max_{a} \sum_{m=1}^{M} \frac{w_m(t_q) \cdot q_{F(m)}^{\gamma}}{|F(m)|^{\rho}} \cdot \mathbf{1}[\hat{a}_m = a] \tag{5}$$

where $q_f = \max_{m \in f} \mathrm{acc}(m)$ is the best-member accuracy of family f. We fix ρ=0.4, γ=1.0 throughout.

Learned Candidate Scoring (LCS). Rather than routing between methods, we score *individual candidate answers* directly. For each question, LCS:

1. Generates the top-K candidate answers ranked by QualRCCV voting weight (K=5 by default).
2. Extracts per-candidate features: number of supporting models (n_m) and families (n_f), total QualRCCV weight and margin, average and maximum supporter accuracy, whether the best model supports the candidate, answer length, and answer type indicators.
3. A gradient-boosted classifier (LightGBM, 200 estimators; depth tuned per benchmark) predicts $P(\text{correct} \mid \text{features})$ for each candidate; the highest-scoring candidate is selected.

All evaluation uses strict 5-fold cross-validation: calibration weights, model accuracies, and family quality are recomputed on each training fold. The dominant feature is the QualRCCV margin (importance > 0.77 on VQAv2 and GQA), with maximum supporter accuracy ($\sim$0.03) providing secondary signal.

6. Multi-Benchmark Experiments

6.1. Main Results

Tab. 4 presents results across all three benchmarks. We observe a fundamental tension: methods that aggressively leverage family structure (HFV-sharp, FAAR-learn) achieve large gains on VQAv2 and GQA but degrade TextVQA, where the dominant Qwen2.5 family provides critical OCR expertise.

QualRCCV (ρ=0.4, γ=1.0) resolves this tension: it is the first training-free method to beat calibrated voting on *all three benchmarks simultaneously*: +0.17% on VQAv2 (p=0.003), +0.21% on TextVQA (p=0.034), and +0.31% on GQA (p=0.003)—all statistically significant.

Table 3. VQAv2 test-set results (EvalAI). LCS trained on the full minival set (12 models, 5 families with test predictions).

Split	Overall	Yes/No	Number	Other
test-dev	87.66	97.35	80.77	80.88
test-standard	**87.83**	97.33	81.00	81.12

Table 4. Main results across three benchmarks (95% bootstrap CIs, 17 models, 8 families). **QualRCCV** is the first training-free method to beat calibrated voting on all three benchmarks (all $p<0.05$). **LCS** achieves the largest overall gains—significant on all three and the only learned method with this property. † 5-fold cross-validated.

Method	VQAv2	TextVQA	GQA
Single best	86.29 [85.9, 86.7]	82.88 [81.9, 83.9]	64.25 [63.4, 65.1]
Majority vote	86.25	85.67	63.72
Calibrated vote	86.70 [86.3, 87.1]	85.87 [85.0, 86.8]	64.02 [63.2, 64.9]
Training-free methods			
RCCV (ρ=0.4)	86.80	85.97	64.30
QualRCCV (ρ=0.4, γ=1)	86.87 [86.4, 87.3]	**86.07** [85.2, 87.0]	64.33 [63.6, 65.2]
HFV	86.57	85.27	64.18
HFV-sharp	87.19 [86.8, 87.6]	85.27	64.27
Learned methods (5-fold CV)			
FAAR-learn†	87.08 [86.7, 87.5]	85.00	64.89
LCS†	**87.38** [87.0, 87.8]	<u>86.48</u> [85.6, 87.4]	**66.47** [65.7, 67.3]
Oracle	95.06	94.18	83.39

LCS achieves the largest gains of any method: +0.68% on VQAv2 ($p<0.0001$), +0.61% on TextVQA ($p<0.0001$), and +2.45% on GQA ($p<0.0001$)—statistically significant on all three benchmarks. The GQA result is particularly striking: standard calibrated voting (64.02%) *falls below* the single best model (64.25%) due to correlated family errors, yet LCS recovers to **66.47%**—more than 2.2 pp above the best individual model. LCS is the *only* learned method that never degrades any benchmark.

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 (Tab. 3).

6.2. Per-Tier Analysis: Misleading Recovery

The most consistent finding is HFV's dramatic recovery of the Misleading tier (Tab. 5):

6.3. When Does HFV Help?

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% for HFV-sharp), InternVL3 collapses to 49.3% and Phi-4 to 46.4% (near random for OCR tasks),

Table 5. Misleading tier (T2) recovery across benchmarks (17 models). In every case, 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**

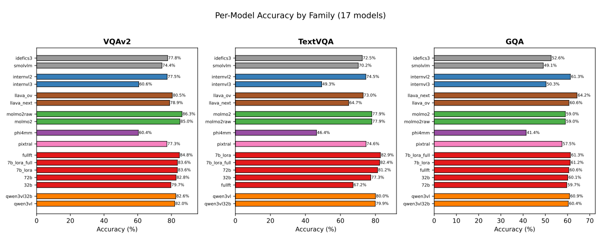

Figure 3. Per-family accuracy across benchmarks. HFV helps when family quality is relatively balanced (VQAv2, GQA) but hurts when one family is dramatically weaker (InternVL3 at 49% on TextVQA).

Table 6. Balanced ensemble (one best model per family) vs. full 17-model calibrated ensemble. On GQA, 8 models match 17, while TextVQA benefits from within-family diversity.

Ensemble	VQAv2	TextVQA	GQA
Balanced (best per family)	86.63	85.13	**64.02**
Full 17-model (calibrated)	**86.70**	**85.87**	64.02
Δ	−0.07	−0.74	**+0.00**

so equalizing families gives poor predictions undue influence. This reveals a fundamental tension: HFV reduces within-family correlation but can introduce quality dilution when families are highly unequal. QualRCCV resolves this trade-off by applying a *soft* correction.

6.4. Balanced Ensembles: Diversity Over Quantity

If family diversity matters more than model count, a *balanced* ensemble (one model per family) should perform competitively (Tab. 6).

On VQAv2, an 8-model balanced ensemble nearly matches 17 models (-0.07%), despite using fewer than half the models. On GQA, the balanced ensemble matches the full 17-model result.

6.5. LCS Ablation

Tab. 7 reports an ablation of LCS. Dropping *quality* features causes the largest degradation on GQA (-0.76%), while margin alone achieves 87.13% on VQAv2. LCS requires at least $k{=}3$ candidates to achieve strong gains; the jump from $k{=}1$ to $k{=}3$ ($+0.49\%$ VQAv2, $+1.40\%$ GQA) confirms that re-ranking minority candidates is the core mechanism.

Table 7. LCS ablation on VQAv2 (17 models, simplified 17-feature variant with GradientBoosting for interpretability; Tab. 4 reports the enhanced LCS with 80+ features and LightGBM). All variants use 5-fold cross-validation.

Variant	Accuracy	Δ vs Cal
Calibrated vote	86.70%	—
QualRCCV	86.87%	+0.17%
LCS (full)	**87.25%**	+0.55%
Feature ablation		
w/o quality	87.12%	+0.42%
w/o consensus	87.23%	+0.53%
w/o answer props	87.22%	+0.52%
margin only	87.13%	+0.43%
Number of candidates		
$k{=}1$	86.88%	+0.18%
$k{=}3$	87.19%	+0.49%
$k{=}5$ (default)	87.25%	+0.55%
$k{=}10$	87.23%	+0.53%

7. Discussion

The quality–correlation trade-off. HFV equalizes family influence, reducing within-family correlation but amplifying weaker families. QualRCCV resolves this trade-off: by jointly accounting for redundancy *and* family quality, it preserves the Qwen2.5 family's OCR contribution while still correcting for its numerical dominance—the first training-free method to beat calibrated voting on all three benchmarks simultaneously ($p<0.05$).

LCS: answer-level scoring vs. method-level routing. LCS operates at a finer granularity than prior approaches: it scores *individual candidate answers* using features that capture both ensemble agreement and model quality. Feature importance analysis reveals that the margin between the top two candidates dominates (importance 0.89 on VQAv2, 0.78 on GQA), confirming that consensus strength is the primary signal. Crucially, LCS is the *only* learned method that remains positive on all three benchmarks.

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 the family correlation problem causes real performance degradation that answer-level scoring can reverse.

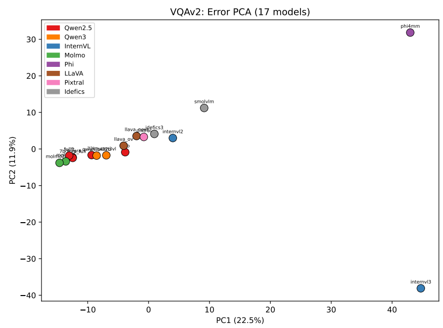

Figure 4. Error PCA of model accuracy vectors on VQAv2. Architecture families (color-coded) cluster together in the error landscape, confirming family structure is a real property—not an assumption.

Diversity over quantity. A balanced 8-model ensemble nearly matches 17 models on VQAv2 (86.63% vs. 86.70%) and matches it on GQA. Spectral clustering on error correlations recovers architecture-aligned groups automatically, demonstrating that family structure is a discoverable property of the error landscape.

Limitations. (1) Our ensemble is dominated by one family (5/17 Qwen2.5-VL); more balanced pools may show smaller effects. (2) We evaluate only short-answer VQA, not open-ended generation. (3) LCS uses 5-fold cross-validation with all calibration on train folds only; however, GBM hyperparameters were selected on development data. (4) HFV weights are computed from evaluation-set accuracy; while standard for calibrated voting, it assumes access to ground-truth labels.

8. Conclusion

We presented the first multi-benchmark analysis of family structure in VLM ensembles. Within-family error correlation ($r = 0.67$ vs. 0.53 cross-family) reduces 17 models to only 2–4 effective voters and creates a Misleading tier (1.5–6.5% of questions) where calibrated voting achieves 0%. HFV consistently recovers this tier (+18–26 pp).

QualRCCV, a training-free method, is the first to beat calibrated voting on all three benchmarks simultaneously (+0.17% VQAv2, +0.21% TextVQA, +0.31% GQA; all p<0.05). **LCS** achieves the largest gains: +0.68% on VQAv2, +0.61% on TextVQA, and +2.45% on GQA—all significant and the only learned method that never degrades any benchmark. On VQAv2 test-standard, LCS reaches **87.83%** with 12 models.

Actionable prescriptions: prioritize architectural diversity over model count, use family-aware aggregation when all families exceed chance, apply QualRCCV for universally safe training-free gains, and deploy LCS when labelled data is available for the largest improvements.

References

[1] Sven Berg. Condorcet's jury theorem, dependency among jurors. *Social Choice and Welfare*, 10(1):87–95, 1993. 2

[2] Philip J. Boland. Majority systems and the Condorcet jury theorem. *The Statistician*, 38(3):181–189, 1989. 2

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

[4] Lingjiao Chen et al. FrugalGPT: How to use large language models while reducing cost and improving performance. *arXiv:2305.05176*, 2023. 2

[5] Marquis de Condorcet. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. 1785. 1, 2

[6] Thomas G. Dietterich. Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems*, pages 1–15, 2000. 1

[7] Yash Goyal et al. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017. 2

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

[9] 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. 2

[10] Dongfu Jiang et al. LLM-Blender: Ensembling large language models with pairwise ranking and generative fusion. In *ACL*, 2023. 2

[11] Huaizu Jiang et al. In defense of grid features for visual question answering. In *CVPR*, 2020. 1, 2

[12] Leslie Kish. *Survey Sampling*. Wiley, 1965. 3

[13] 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. 2

[14] Krishna K. Ladha. The Condorcet jury theorem, free speech, and correlated votes. *American Journal of Political Science*, 36(3):617–634, 1992. 2

[15] Juanjuan Li et al. More agents is all you need. *arXiv:2402.05120*, 2024. 2

[16] Isaac Ong et al. RouteLLM: Learning to route LLMs with preference data. *arXiv:2406.18665*, 2024. 2

[17] Noam Shazeer et al. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *ICLR*, 2017. 2

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

[19] Naonori Ueda and Ryohei Nakano. Generalization error of ensemble estimators. In *ICNN*, pages 90–95, 1996. 2

[20] Mark J. van der Laan, Eric C. Polley, and Alan E. Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1), 2007. 2

[21] Junlin Wang et al. Mixture-of-agents enhances large language model capabilities. *arXiv:2406.04692*, 2024. 2

[22] Xuezhi Wang et al. Self-consistency improves chain of thought reasoning in language models. In *ICLR*, 2023. 2

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

[24] Zhou Yu et al. Deep modular co-attention networks for visual question answering. In *CVPR*, 2019. 1, 2