NOT ALL ATTRIBUTES COMPOSE EQUALLY: FAMILY-AWARE STRUCTURED CALIBRATION FOR COMPOSITIONAL ZERO-SHOT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Compositional zero-shot learning (CZSL) aims to recognize unseen attribute-object compositions by recombining known primitives. Existing methods mainly improve representations before pair scores are produced, but the final ranking usually does not explicitly use relations among attributes. However, attributes do not compose equally: semantically related attributes share family-level visual evidence, while their appearance and compatibility remain object dependent. We study the complementary problem of post-hoc pair calibration and introduce FASC (Family-Aware Structured Calibration), which preserves the backbone candidate space while using semantic attribute families as intermediate structure. FASC derives an image-specific family posterior by marginalizing the softmax distribution over averaged candidate-pair logits, then combines the original logits with training-frequency priors, a ranked family bias, a centered family-prior correction, and a pseudo-unseen ranking objective. On C-GQA, MIT-States, and UT-Zappos, FASC improves closed-world AUC over the same-source uncalibrated scores and exhibits low variation across five calibrator seeds. Fixed-source ensemble, component, and taxonomy ablations isolate the contribution of semantic family structure, while evaluations on two additional CZSL backbones demonstrate transfer beyond the main backbone.

1 INTRODUCTION

Humans can recognize unfamiliar visual compositions by recombining known concepts, such as inferring a red tomato or a sliced apple from red apples and sliced tomatoes. Compositional zero-shot learning (CZSL) formalizes this ability by training on a subset of attribute-object pairs and testing on novel compositions of known primitives (Misra et al., 2017; Nagarajan & Grauman, 2018; Purushwalkam et al., 2019).

Large vision-language models (VLMs), especially CLIP (Radford et al., 2021), provide a strong foundation for CZSL because candidate labels can be expressed as natural language. Recent CZSL methods improve pair scoring with prompt learning and decomposition (Nayak et al., 2023; Lu et al., 2023), branch alignment (Huang et al., 2024), semantic relations constrained by logic (Wu et al., 2025b), or visual adaptation (Zhang et al., 2025). However, after pair logits are produced, the final ranking often does not explicitly use relations among attributes, leaving broader regularities underused.

Attribute evidence has coarse semantic regularities: color and material cues tend to be local surface evidence (Christoff & Tonchev, 2024; Balestra et al., 2025), whereas physical state, pose, interaction, and facial affect often require object geometry or context (Stavrothanasopoulos et al., 2025; Yongsiriwit et al., 2025). Families can therefore calibrate heterogeneous candidate pairs (Kim & Choi, 2025), but should not replace scoring at the pair level, since *open* can look different for a door, laptop, or mouth. We use families as structured calibration over pair scores, not as standalone visual templates.

Figure 1 illustrates this motivation. The upper part shows why attribute families are useful but should not be treated as rigid templates: the same family may manifest differently across object categories, while attributes from the same family can share more consistent evidence for a fixed object.

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.

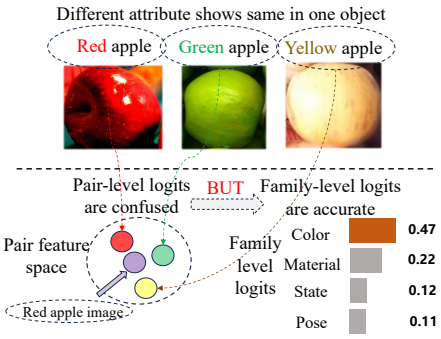


Figure 1: Why attribute families are useful as calibration structure. The same family can have object-dependent appearances (top), while ambiguous pair scores can still share coherent posterior mass at the family level (bottom). FASC uses this family posterior to rerank the original candidate pairs rather than predicting families as new labels.

The lower part shows the resulting scoring ambiguity. Flat pair scores can confuse candidates with similar logits, whereas marginalization at the family level exposes a stronger cue for the image and allows FASC to calibrate the ranking (Fan et al., 2022; Kim & Choi, 2025).

Following this view, we propose FASC for post-hoc calibration of CZSL candidate-pair logits while keeping the backbone fixed. A fixed taxonomy maps attributes to semantic families, and the softmax distribution over averaged candidate-pair logits is marginalized into an image-specific family posterior. The calibrator combines the original pair logits with lightweight frequency priors, a ranked family bias, a centered family-prior correction, and a pseudo-unseen objective (Fan et al., 2022; Liu et al., 2024). Because the family posterior is an intermediate statistic rather than the output space, object-specific differences within a family remain represented by the original pair logits.

Our contributions are:

- We identify a family-level regularity in CZSL: attributes do not compose equally; semantically related attributes share visual evidence, while their manifestations remain object dependent. This motivates semantic families as intermediate structure for pair ranking.
- We introduce FASC, a lightweight post-hoc method that calibrates existing candidate logits through an image-specific family posterior with negligible additional cost.
- Across three benchmarks, FASC consistently improves closed-world performance. Fixed-source ablations, random-taxonomy controls, and further analyses validate its effectiveness, stability, and generality.

2 Related Work

CZSL studies how to recognize unseen attribute-object pairs by composing known primitives. Early work composes classifiers with attribute transformations (Misra et al., 2017), models attributes as operators on object representations (Nagarajan & Grauman, 2018), assembles modular classifiers for new compositions (Purushwalkam et al., 2019), or regularizes attribute-object transformations with symmetry (Li et al., 2020). Subsequent methods introduce graph propagation among attributes, objects, and compositions (Naeem et al., 2021), open world scoring with feasibility estimates (Mancini et al., 2021), and primitive modeling guided by knowledge (Karthik et al., 2022). Recent work further shows that attributes should not always be treated as fixed visual tokens: CANet learns attributes conditioned on the image and object (Wang et al., 2023), CDS-CZSL studies attribute specificity across object contexts (Li et al., 2024), and CPF decomposes composition probability into object likelihood and conditional attribute likelihood (Wu et al., 2025a).

Vision-language pretraining has shifted CZSL toward composition scoring with text prompts. CLIP (Radford et al., 2021) enables candidate pairs to be represented as prompts written in natural language, while CSP (Nayak et al., 2023), DFSP (Lu et al., 2023), and ADE (Hao et al., 2023) adapt prompt or concept representations for compositional recognition. Troika jointly models state, object, and composition branches with visual adaptation (Huang et al., 2024); LOGICZSL injects semantic relations among primitives and compositions through objectives induced by logic (Wu et al., 2025b); and recent visual proxy methods further align textual prototypes with composition evidence on the image side (Zhang et al., 2025).

Prior CZSL work therefore already establishes that attributes are object-conditional rather than interchangeable primitives. Separately, long-tail recognition uses frequency priors and posterior cor-

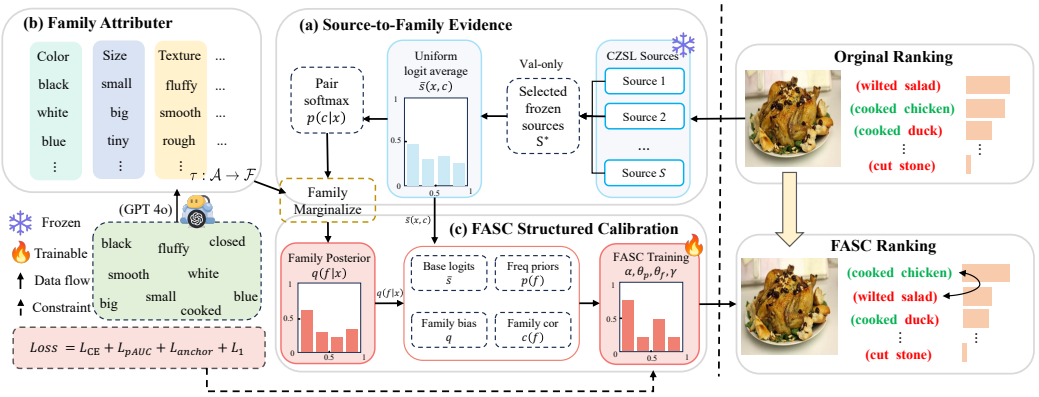


Figure 2: Overview of FASC. (a) Validation selects frozen sources $\mathcal{S}^\star$, whose uniformly averaged logits yield a pair softmax and, after marginalization, the family posterior $q(f \mid x)$. (b) The family attributer constructs the fixed map τ offline. (c) Base logits, frequency priors, q-derived family bias, and centered family correction are jointly calibrated to rerank the original candidates. The figure uses compact labels; their full definitions and loss weights are given in the text.

rection to repair biased scores (Fan et al., 2022; Liu et al., 2024). FASC connects these directions through a candidate-logit interface: semantic families become image-dependent intermediate statistics for post-hoc CZSL ranking, and their effect is tested separately from source averaging and non-family calibration.

3 Method

FASC follows the three-stage flow in Fig. 2. The offline family attributer in panel (b) provides a fixed attribute-to-family map. In panel (a), validation selects frozen CZSL sources $\mathcal{S}^\star$; their pair logits are uniformly averaged into $\bar{s}(x, c)$, normalized into $p(c \mid x)$, and marginalized through the family map to obtain $q(f \mid x)$. Panel (c) combines the base logits, training-frequency priors, a q-derived family bias, and a centered family correction to rerank the same candidate set. Only the compact calibrator is trained; the source models, label space, and family map remain fixed.

3.1 Problem Formulation

Let $\mathcal{A}$ be the attribute set, $\mathcal{O}$ the object set, and $\mathcal{C} \subseteq \mathcal{A} \times \mathcal{O}$ the candidate pair set. In closed world evaluation, $\mathcal{C}$ is the benchmark candidate composition set; in open world evaluation, $\mathcal{C}$ can be the full Cartesian product, optionally constrained by the benchmark feasibility protocol. Each image is labeled with an attribute-object pair $c^\star = (a^\star, o^\star)$. The final prediction ranks all candidate pairs:

$$\hat{c} = \arg\max_{c \in \mathcal{C}} S(x, c). \tag{1}$$

FASC does not change this prediction space; attribute families serve only as intermediate calibration structure.

3.2 Source Logits and Attribute Families

Let $s_m(x, c)$ be the logit from source checkpoint m for image x and candidate pair $c = (a, o)$. When multiple checkpoints or seeds are available, we select a subset $\mathcal{S}$ by validation performance and average their logits uniformly:

$$\bar{s}(x, c) = \frac{1}{|\mathcal{S}|} \sum_{m \in \mathcal{S}} s_m(x, c). \tag{2}$$

Let $\mathcal{F} = \{1, \ldots, K\}$ be a fixed set of semantic attribute families, let $\mathcal{A}_F \subseteq \mathcal{A}$ denote the attributes covered by the family branch, and let

$$\tau : \mathcal{A}_F \to \mathcal{F} \tag{3}$$

map each attribute to its family. For readability, Fig. 2(b) writes $\mathcal{A}$ for this covered inventory; attributes in $\mathcal{A} \setminus \mathcal{A}_F$ have no family assignment. We construct this taxonomy offline with GPT-4o

(OpenAI, 2024): given the dataset attribute vocabulary, the model assigns each attribute to a concise semantic family, after which the names are canonicalized and the mapping is fixed. The taxonomy is not a prediction target; it defines structured statistics and family evidence for calibrating pair scores. Candidates whose attributes are outside $\mathcal{A}_F$ remain in the evaluator and receive no family-specific correction.

3.3 Family Evidence

The primary family posterior is defined directly from the selected averaged pair logits. We first normalize over the candidate pairs,

$$p(c \mid x) = \frac{\exp(\bar{s}(x, c))}{\sum_{c' \in \mathcal{C}} \exp(\bar{s}(x, c'))}, \tag{4}$$

and aggregate probability mass for all pairs whose attribute belongs to family f:

$$q(f \mid x) = \sum_{(a,o) \in \mathcal{C}: \tau(a) = f} p((a, o) \mid x). \tag{5}$$

This pair-derived quantity is the primary family posterior shown in Fig. 2(a). An independently computed compatible family cache, when available, supplies optional evidence $q^{\text{aux}}(f \mid x)$ to the ranked family term; this auxiliary path is omitted from the figure, and otherwise the method uses q alone. Marginalizing over all candidate objects exposes coherent family evidence even when the top pair is ambiguous.

3.4 Family-Aware Structured Calibration

Panel (c) summarizes four inputs to the calibrated score. "Base logits" denotes $\bar{s}(x, c)$; "Freq priors $p(f)$" compactly denotes the full pair/attribute/object/family frequency vector; "Family bias q" denotes the ranked transformation of $q(f \mid x)$; and "Family cor $c(f)$" denotes its centered correction against the training family prior. To avoid overloading p, q, and the composition symbol c, the equations below write these three calibration terms as g_{prior}, g_{family}, and h_{prior}. They only rescore existing pairs:

$$\begin{aligned} S(x, c) = \alpha\, \bar{s}(x, c) + g_{\text{prior}}(c; \theta_p) \\ + g_{\text{family}}(x, c; \theta_f) + \gamma\, h_{\text{prior}}(x, c). \end{aligned} \tag{6}$$

Here, α is a learnable base logit scale, g_{prior} is a training frequency structural prior, g_{family} is a ranked family bias, and h_{prior} is a family prior correction for each sample.

Training frequency prior. Let $\hat{P}_{\text{tr}}$ denote smoothed empirical frequencies from the training split. For $c = (a, o)$, define the log-frequency feature vector and its linear prior as

$$\begin{aligned} \boldsymbol{\ell}(c) = \big[&\log \hat{P}_{\text{tr}}(a, o), \log \hat{P}_{\text{tr}}(a), \log \hat{P}_{\text{tr}}(o), \\ &\log \hat{P}_{\text{tr}}(\tau(a)), \log \hat{P}_{\text{tr}}(\tau(a), o)\big]^\top, \\ g_{\text{prior}}(c; \theta_p) = \boldsymbol{\beta}^\top &\boldsymbol{\ell}(c). \end{aligned} \tag{7}$$

Thus, $\boldsymbol{\ell}(c)$ jointly represents pair, attribute, object, family, and family-object frequencies; Fig. 2 abbreviates this complete block as $p(f)$.

Ranked family bias. The pair-derived and optional auxiliary family posteriors are converted into a small ranked bias over candidate families. Let $T_r(x)$ and $T_r^{\text{aux}}(x)$ denote the family at rank r under $q(\cdot \mid x)$ and $q^{\text{aux}}(\cdot \mid x)$, respectively. Let $\kappa(x) = \max_f q(f \mid x)$ and $\rho(x) = \mathbb{I}[\kappa(x) \geq \theta]$. For $f_c = \tau(a)$, the bias is

$$\begin{aligned} g_{\text{family}}(x, c; \theta_f) = \rho(x) \sum_{r=1}^{R} \eta_r \mathbf{1}[f_c = T_r(x)] \\ + \sum_{r=1}^{R_{\text{aux}}} \eta_r^{\text{aux}} \mathbf{1}[f_c = T_r^{\text{aux}}(x)]. \end{aligned} \tag{8}$$

The auxiliary sum is omitted when q^{aux} is unavailable; pair-specific differences remain governed by $\bar{s}(x, c)$ and g_{prior}.

Family prior correction for each sample. We also compare the pair-derived family posterior with the training family prior. For $c = (a, o)$, define

$$z(x, c) = \log q(\tau(a) \mid x) - \log \hat{P}_{\mathrm{tr}}(\tau(a)). \tag{9}$$

We center this quantity over candidate pairs for the same image:

$$h_{\mathrm{prior}}(x, c) = z(x, c) - \frac{1}{|\mathcal{C}|} \sum_{c' \in \mathcal{C}} z(x, c'). \tag{10}$$

The scalar γ is learned with the rest of the calibrator.

3.5 Learning Objective

For each source subset, we fit the parameters $\Theta = \{\alpha, \theta_p, \theta_f, \gamma\}$ from labeled training pairs. For each training image, we take the top-K candidates under $\bar{s}(x, c)$ and keep the sample when the true pair is included in this candidate set. The primary objective is cross entropy over the calibrated top-K scores:

$$\mathcal{L}_{\mathrm{CE}} = -\log \frac{\exp(S(x, c^\star))}{\sum_{c \in \mathcal{T}_K(x)} \exp(S(x, c))}. \tag{11}$$

To align learning with generalized CZSL evaluation, we deterministically permute the distinct training compositions and designate 25% as pseudo unseen for an auxiliary ranking loss. This partition affects only the auxiliary objective: the top-K cross-entropy still uses all eligible training images. The differentiable surrogate combines soft seen/unseen accuracies across a bias grid, their trapezoidal AUC, a smooth best harmonic mean, and within-domain pairwise margins. The final objective is

$$\begin{aligned} \mathcal{L} = \mathcal{L}_{\mathrm{CE}} &+ \lambda_{\mathrm{auc}} \mathcal{L}_{\mathrm{pAUC}} \\ &+ \lambda_{\mathrm{anchor}} \mathcal{L}_{\mathrm{anchor}} + \lambda_1 \|\Theta\|_1. \end{aligned} \tag{12}$$

Figure 2 suppresses the λ weights and displays the four loss components only. The anchor term limits deviation from initialization, and the L_1 term shrinks weak calibration coefficients. The complete definition of $\mathcal{L}_{\mathrm{pAUC}}$, the seeded pseudo-unseen construction, smoothing temperatures, bias grid, and all optimization hyperparameters are given in the supplementary section "Pseudo-Unseen AUC Surrogate." After fitting each source subset, its validation AUC is used to choose the final source subset.

3.6 Training and Inference Procedure

Algorithm 1 summarizes the full procedure. During training, for each source subset $\mathcal{S}$, FASC averages its pair logits, computes the pair-derived family posterior (plus an optional compatible auxiliary cache), and learns the calibrator parameters on the training images. Each subset therefore has its own calibrated scorer. After training, the validation AUC selects the final subset and its parameters.

At inference time, no parameters are updated. FASC averages the selected source logits, computes the same family evidence, adds the learned calibration terms to every candidate pair, and ranks candidates by Eq. 6. Closed world prediction and the generalized CZSL seen/unseen bias sweep both use this same calibrated score.

4 Experiments

4.1 Experimental Setting

Datasets. Following standard CZSL evaluation protocols, we evaluate FASC on three widely used benchmarks: MIT-States (Isola et al., 2015), UT-Zappos (Yu & Grauman, 2014), and C-GQA (Naeem et al., 2021). Detailed dataset statistics are reported in the supplementary material. For C-GQA, the family branch uses a fixed taxonomy over 413 attributes. The other 40 attributes do not admit reliable assignments to the available semantic families and are therefore excluded only from family-based correction; evaluation still retains the full benchmark inventory of 453 attributes and 870 objects.

Algorithm 1 FASC Calibration and Inference

Require: Source logits $\{s_m(x, c)\}$, optional auxiliary family cache, training set $\mathcal{D}_{tr}$, validation set $\mathcal{D}_{val}$, candidate set $\mathcal{C}$, family map τ, maximum subset size B.
Ensure: Calibrated ranking score $S(x, c)$.
Compute smoothed training-frequency statistics $\ell^P, \ell^A, \ell^O, \ell^F, \ell^{FO}$ from $\mathcal{D}_{tr}$.
Initialize $v^\star \leftarrow -\infty$ and $\mathcal{S}^\star \leftarrow \emptyset$.
for all source subsets $\mathcal{S}$ with $1 \leq |\mathcal{S}| \leq B$ **do**
  Average selected source logits: $\bar{s}(x, c) \leftarrow |\mathcal{S}|^{-1} \sum_{m \in \mathcal{S}} s_m(x, c)$.
  Compute $p(c \mid x)$ and $q(f \mid x)$ by pair marginalization; average a compatible auxiliary cache if provided.
  Initialize $\Theta = \{\alpha, \theta_p, \theta_f, \gamma\}$.
  for epoch $= 1, \ldots, E$ **do**
    Build $\mathcal{T}_K(x)$ under $\bar{s}(x, c)$ for each $x \in \mathcal{D}_{tr}$.
    Compute $S(x, c)$ by Eq. 6 on $\mathcal{T}_K(x)$.
    Update Θ using $\mathcal{L}_{CE} + \lambda_{auc}\mathcal{L}_{pAUC} + \lambda_{anchor}\mathcal{L}_{anchor} + \lambda_1\|\Theta\|_1$.
  end for
  Evaluate validation AUC $v_{\mathcal{S}}$.
  if $v_{\mathcal{S}} > v^\star$ **then**
    $v^\star \leftarrow v_{\mathcal{S}}$; $\mathcal{S}^\star \leftarrow \mathcal{S}$; store $\Theta^\star \leftarrow \Theta$.
  end if
end for
Inference: average logits from $\mathcal{S}^\star$, compute q, score every $c \in \mathcal{C}$ using Eq. 6, and rank by $S(x, c)$.

Metrics. Following standard closed- and open-world protocols (Huang et al., 2024; Zhang et al., 2025), we report seen (S) and unseen (U) composition accuracy, their best harmonic mean (HM) over the seen/unseen bias sweep, and the area under that curve (AUC).

Implementation details. We use VP-CMJL (Zhang et al., 2025) as the CZSL backbone and apply FASC to its candidate pair logits. The backbone uses CLIP ViT-L/14 (Radford et al., 2021) as the image/text encoder. For each dataset, we construct a fixed pool of independent backbone checkpoints and select a source subset by validation AUC after uniformly averaging its candidate logits and fitting the same calibrator. Source selection and uniform averaging are distinct from the proposed family calibration, so Sec. 4.3 reports Base-logit ensemble baselines and fixed-source ablations. Unless stated otherwise, results use the validation-selected source subset and one calibrator seed. For statistical reliability, we additionally freeze each selected subset and refit only the calibrator with five seeds. The exact source pools, subset-search protocol, optimization hyperparameters, and hardware are reported in the supplementary material.

4.2 Main Results

Table 1 reports the results under both closed and open world settings, comparing FASC with recent CZSL methods based on CLIP. The VP-CMJL row is a single released backbone, whereas FASC can select and average several source checkpoints before calibration. Consequently, that comparison measures the complete pipeline rather than the family effect alone; the controlled analyses in Sec. 4.3 separate source averaging, generic calibration, and family structure. In the closed world setting, FASC achieves the best HM and AUC on all three datasets and obtains the best result in 11 out of 12 reported metrics. Compared with VP-CMJL, FASC improves AUC by +0.8, +0.6, and +1.0 on C-GQA, MIT-States, and UT-Zappos, respectively, while also improving HM by +0.1, +0.5, and +1.1. These improvements are concentrated on HM and AUC, the metrics that summarize the seen/unseen tradeoff.

In the open world setting, FASC achieves the best AUC on C-GQA and MIT-States and improves MIT-States HM/AUC from 22.0/8.3 to 22.5/8.8. On C-GQA, AUC increases slightly while HM decreases from 15.5 to 15.3. On UT-Zappos, FASC matches VP-CMJL in U (66.6), but its lower S (69.8 vs. 71.9) yields lower HM/AUC (53.9/39.9 vs. 54.5/41.4). Its attributes are dominated by material cues, and the open world expansion introduces many plausible but incorrect pairs such as *leather boot*, *suede shoe*, and *rubber sandal*. These results show that family calibration is most ef-

Table 1: Experimental results for both closed and open world settings. The best and second best performance are highlighted in bold and underlined, respectively.

Method	Venue	C-GQA				MIT-States				UT-Zappos			
		S	U	HM	AUC	S	U	HM	AUC	S	U	HM	AUC
Closed World Results													
CLIP (2021)	ICML'21	7.5	25.0	8.6	1.4	30.2	46.0	26.1	11.0	15.8	49.1	15.6	5.0
CoOp (2022)	IJCV'22	20.5	26.8	17.1	4.4	34.4	47.6	29.8	13.5	52.1	49.3	34.6	18.8
CSP (2023)	ICLR'23	28.8	26.8	20.5	6.2	46.6	49.9	36.3	19.4	64.2	66.2	46.6	33.0
DFSP(i2t) (2023)	CVPR'23	35.6	29.3	24.3	8.7	47.4	52.4	37.2	20.7	64.2	66.4	45.1	32.1
DFSP(BiF) (2023)	CVPR'23	36.5	32.0	26.2	9.9	47.1	52.8	37.7	20.8	63.3	69.2	47.1	33.5
DFSP(t2i) (2023)	CVPR'23	38.2	32.0	27.1	10.5	46.9	52.0	37.3	20.6	66.7	71.7	47.2	36.0
DLM (2024)	AAAI'24	32.4	28.5	21.9	7.3	46.3	49.8	37.4	20.0	67.1	72.5	52.0	39.6
ProLT (2024)	AAAI'24	39.5	32.9	27.7	11.0	49.1	51.0	38.2	21.1	66.0	70.1	49.4	36.1
PLID (2024)	ECCV'24	38.8	33.0	27.9	11.0	49.7	52.4	39.0	22.1	67.3	68.8	52.4	38.7
CDS-CZSL (2024)	CVPR'24	38.3	34.2	28.1	11.1	50.3	52.9	39.2	22.4	63.9	74.8	52.7	39.5
Troika (2024)	CVPR'24	41.0	35.7	29.4	12.4	49.0	53.0	39.3	22.1	66.8	73.8	54.6	41.7
IMAX (2025)	TPAMI'25	39.7	35.8	29.8	12.8	48.7	53.8	39.1	21.9	69.3	70.7	54.2	40.6
VP-CMJL (2025)	ICCV'25	46.0	40.2	34.9	16.3	**51.8**	52.6	40.4	23.3	71.9	76.3	58.5	47.9
FASC (Ours)	–	**46.9**	**41.0**	**35.0**	**17.1**	51.5	**53.9**	**40.9**	**23.9**	**72.1**	**77.7**	**59.6**	**48.9**
Open World Results													
CLIP (2021)	ICML'21	7.5	4.6	4.0	0.3	30.1	14.3	12.8	3.0	15.7	20.6	11.2	2.2
CoOp (2022)	IJCV'22	21.0	4.6	5.5	0.7	34.6	9.3	12.3	2.8	52.1	31.5	28.9	13.2
CSP (2023)	ICLR'23	28.7	5.2	6.9	1.2	46.3	15.7	17.4	5.7	64.1	44.1	38.9	22.7
DFSP(i2t) (2023)	CVPR'23	35.6	5.6	9.0	1.9	47.2	18.2	19.1	6.7	64.3	53.8	41.2	26.4
DFSP(BiF) (2023)	CVPR'23	36.5	7.6	10.6	2.4	47.1	18.1	19.2	6.7	63.5	57.2	42.7	27.6
DFSP(t2i) (2023)	CVPR'23	38.2	7.2	10.4	2.4	47.5	18.5	19.3	6.8	66.8	60.0	44.0	30.3
PLID (2024)	ECCV'24	39.1	7.5	10.6	2.5	49.1	18.7	20.4	7.3	67.6	55.5	46.6	30.8
CDS-CZSL (2024)	CVPR'24	37.6	8.2	11.6	2.7	49.4	**21.8**	22.1	8.5	64.7	61.3	48.2	32.3
Troika (2024)	CVPR'24	40.8	7.9	10.9	2.7	48.8	18.4	20.1	7.2	66.4	61.2	47.8	33.0
IMAX (2025)	TPAMI'25	38.7	7.9	11.2	2.5	50.2	18.6	21.4	7.6	68.4	57.3	47.5	32.3
VP-CMJL (2025)	ICCV'25	46.0	**11.5**	**15.5**	4.6	**51.8**	19.9	22.0	8.3	**71.9**	**66.6**	**54.5**	**41.4**
FASC (Ours)	–	**47.0**	11.2	15.3	**4.7**	51.6	20.8	**22.5**	**8.8**	69.8	**66.6**	53.9	39.9

Table 2: Closed-world AUC attribution. Ensemble-only variants exclude the taxonomy, structural priors, ranked family correction, learnable base-logit scale, pseudo-unseen objective, and calibrator fitting.

Variant	C-GQA	MIT-States	UT-Zappos
VP-CMJL (single)	16.30	23.30	47.90
Val.-selected ensemble	16.90	23.82	47.48
Same-source Base	16.88	23.53	48.60
Full FASC	**17.12**	**23.94**	**48.91**

Table 3: Fixed-source grouped ablation (closed-world AUC). Generic is a matched non-family calibrator; every column uses identical source logits.

Dataset	Base	Generic	No scale	No pAUC	Full
C-GQA	16.88	17.11	16.87	16.93	**17.15**
MIT-States	23.53	23.71	23.80	23.92	**23.94**
UT-Zappos	48.60	48.68	48.42	48.45	**48.91**

fective when the candidate scorer provides reliable object evidence, while highly ambiguous object–material candidates remain a key source of open-world error.

4.3 Controlled Analysis

Each controlled table holds the stated source logits and calibrator seed fixed so that its rows are directly paired. The main run and the separate control suites use different fixed calibrator seeds; small cross-table differences between entries labeled "Full" or "Semantic" therefore reflect run choice, and the intended comparisons are within each table.

Separating source averaging from calibration. Table 2 distinguishes four stages of the closed-world pipeline. The Base-validation ensemble selects the source subset by uncalibrated validation AUC and only averages logits. The same-source Base variant disables all calibration while using the subset selected for full FASC. Thus, the difference between the last two rows measures calibration on identical source logits, whereas the first two rows expose the contribution of checkpoint averaging and selection.

Base-logit ensembling is dataset dependent: relative to one VP-CMJL checkpoint, it changes AUC by $+0.60$, $+0.52$, and -0.42. On the fixed FASC source subsets, calibration adds $+0.24$, $+0.41$,

Table 4: Closed-world AUC for the Base, semantic, randomly permuted, and no-family FASC variants.

Dataset	Base	Semantic	Random	No family
C-GQA	16.30	**17.15**	16.99	16.81
MIT-States	23.30	**23.86**	23.26	23.35
UT-Zappos	47.90	**48.87**	48.47	48.60

Table 5: Closed-world AUC over five calibrator seeds; Base values are reproduced from the main table.

Dataset	Base	FASC mean $\pm$ std	95% CI
C-GQA	16.300	17.150 ± 0.037	[17.104,17.196]
MIT-States	23.300	23.842 ± 0.038	[23.794,23.890]
UT-Zappos	47.900	48.813 ± 0.093	[48.698,48.928]

and $+0.31$ AUC. The main-table gains reflect both source-logit averaging and calibration; the following fixed-source analyses isolate the contribution of family structure.

Fixed-source component attribution. We next freeze the selected source logits and remove groups of calibration components. The "Generic" calibration baseline removes q, q^{aux}, the family and family-object priors, ranked family correction, and sample-level family-prior correction, while retaining the learnable base scale, pair/attribute/object frequency terms, top-K cross-entropy, and pseudo-unseen objective. Thus it uses the identical source logits, training budget, and validation protocol without semantic family information.

Relative to full FASC, removing the family block reduces AUC by 0.04, 0.23, and 0.23; removing the base scale reduces it by 0.28, 0.14, and 0.49; and removing the pseudo-unseen objective reduces it by 0.22, 0.02, and 0.46. Family structure contributes on every benchmark, demonstrating a consistent benefit beyond source averaging and generic calibration. The supplementary section "Fixed-Source Family-Signal Isolation" defines explicit modes for Generic, q only, q^{aux} only, and their combination. In those modes, q is always recomputed from the averaged pair logits and current taxonomy, while q^{aux} must come from a separate aligned cache; disabled branches are zeroed and frozen rather than replaced by another signal.

Taxonomy controls. Table 4 compares the semantic taxonomy with a family-size-preserving random taxonomy and a no-family control using the same selected source subset and calibrator seed. The Base column reproduces the corresponding single-backbone closed-world AUC from Table 1.

The semantic map exceeds the random taxonomy by 0.16, 0.60, and 0.40 AUC and the no-family control by 0.34, 0.51, and 0.27. The consistent advantage further supports the effectiveness of semantically structured attribute families.

Calibrator-seed reliability. We freeze the final selected source combination and semantic taxonomy, then repeat only the train-time calibrator fitting with five seeds. Table 5 reports the resulting variation.

All five-run confidence intervals remain above the Base references in the main table, and the small standard deviations show that the calibration gains are stable across fitting seeds.

Transfer across candidate-logit backbones. To test whether the interface is specific to VP-CMJL, we apply the same calibration procedure to two additional C-GQA scorers using one checkpoint and no ensembling: a region-map FamilyCZSL scorer and the original Troika scorer (Huang et al., 2024).

Table 6: C-GQA closed-world AUC on two non-VP-CMJL sources.

Candidate-logit source	Base	+FASC	Δ
FamilyCZSL region map	6.48	7.01	+0.53
Troika original	12.41	13.10	+0.70

FASC improves both sources, supporting compatibility with different candidate-logit backbones. The larger gains arise when the underlying scorer assigns sufficient support to the correct pair, highlighting the complementary roles of candidate scoring and family calibration.

The supplementary material documents its construction, inventory, dataset-specific settings, sensitivity analyses, and the complete efficiency protocol. After backbone scoring, FASC adds only 0.0705 ms per image over base argmax ranking on C-GQA.



Figure 3: Qualitative comparisons on C-GQA and MIT-States. "w/o family" removes family calibration. Text labels ensure grayscale readability and mark cross-family, within-family, and object-score cases.

4.4 Quality Analysis

Case analysis. Figure 3 compares FASC with its no-family variant. When object evidence is reliable, flat pair scores may still follow a salient but inappropriate cue, producing *white tie* instead of *plaid tie* or *wet dog* instead of *brown dog*. Family calibration aggregates evidence shared by related attributes, improving cue-family consistency while retaining pair-level distinctions. MIT-States shows the same pattern, suppressing the cross-family distractor *wilted salad* for *cooked chicken* and separating close alternatives like *crumpled plastic* and *crinkled paper*.

The failures reveal two main boundaries: *object-recognition failures* arise when the backbone assigns high mass to pairs containing the wrong object; because FASC preserves the base pair score and chiefly adjusts attribute evidence, it rarely overturns a strong object error. *Family-calibration failures* arise when $q(f \mid x)$ favors the wrong family and promotes attributes with an inappropriate semantic cue; even with the correct family, multiple within-family object-material pairs can remain plausible, especially in open-world UT-Zappos. The failures *snowy snowboarder→white jacket* and *dry forest→huge plant* illustrate object confusion, whereas *peeled eggs→cooked eggs* is a within-object attribute ambiguity. Family calibration thus complements rather than replaces the candidate scorer, and works best for cross-family ambiguity with reliable object evidence.

Score space visualization. Fig. 4 further visualizes family-level rank shifts in score space.

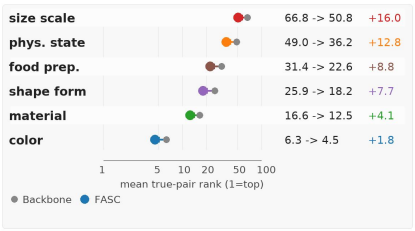


Figure 4: Mean true pair rank by true attribute family on C-GQA before and after FASC; leftward arrows mark rank gains.

We rank all candidate pairs for each selected C-GQA test image, group images by true attribute family, and compare the mean true composition rank before and after FASC; smaller ranks indicate better ordering. Gray and colored markers denote backbone and calibrated ranks; leftward arrows indicate rank gains. For example, "66.8 → 50.8" reports a mean rank gain of 16.0 positions within the size family. All six families improve consistently (e.g., physical state 49.0 → 36.2). Family-level evidence thus yields better candidate ordering, complementing Fig. 3.

5 Conclusion

We presented FASC, a lightweight structured calibrator using a pair-derived family posterior as intermediate calibration structure over the original CZSL candidate pair space. Controlled experiments isolate this contribution from source averaging and generic calibration: semantic assignments beat random and no-family controls, gains persist across five seeds, and the interface improves two additional backbones. Family evidence is thus an effective complementary signal for CZSL calibration, and our analyses show how it repairs ambiguous candidate rankings.

AI USE STATEMENT

In this work, we used generative AI tools for constructing the offline attribute-to-family taxonomy from the dataset attribute vocabularies: GPT-4o (OpenAI, 2024) assigns each attribute to a semantic family, after which the outputs are canonicalized, audited, and fixed before any experiment, as described in the Method section and the appendix. We also used generative AI tools for language polishing of this paper. We have not used generative AI tools for research ideation, experimental design, or result analysis, and these tasks are not applicable to this work. The taxonomy contribution is validated by random-taxonomy and no-family controls; all taxonomy assignments and all AI-polished text were reviewed and verified by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work complies with the ICLR Code of Ethics. It does not involve human subjects or sensitive personal data: all datasets used in this work (C-GQA, MIT-States, and UT-Zappos) are publicly available benchmarks released for research use, and the backbone checkpoints and language resources we build upon are publicly released research artifacts used in accordance with their original licenses and terms. The proposed method is a post-hoc calibration procedure intended for research purposes, and we are not aware of any foreseeable negative societal impact or potential misuse beyond those already associated with general image-recognition systems.

REPRODUCIBILITY STATEMENT

We are committed to ensuring the reproducibility of our results. The model architecture, calibration objective, and inference procedure are fully specified in the Method section, and the supplementary material provides the complete experimental protocol, including dataset statistics, source-pool construction, validation-based source selection, calibrator hyperparameters, parameter sensitivity analyses, and the efficiency protocol. All datasets used are publicly available. Our code is submitted as part of the supplementary material, and we will release it publicly upon publication.

REFERENCES

Mattia Balestra, Marina Paolanti, and Roberto Pierdicca. Whu-rs19 abzsl: An attribute-based dataset for remote sensing image understanding. *Remote Sensing*, 2025. URL https://api.semanticscholar.org/CorpusID:280089184.

Wentao Bao, Lichang Chen, Heng Huang, and Yu Kong. Prompting language-informed distribution for compositional zero-shot learning. In *European Conference on Computer Vision*, 2024.

Nicole Christoff and Krasimir Tonchev. Material attribute estimation as part of telecommunication augmented reality, virtual reality, and mixed reality system: Systematic review. *Electronics*, 2024. URL https://api.semanticscholar.org/CorpusID:270749284.

Xinyue Fan, Teng Liu, Hong Bao, Weiguo Pan, Tianjiao Liang, and Han Li. Long-tail instance segmentation based on memory bank and confidence calibration. *Applied Sciences*, 2022. URL https://api.semanticscholar.org/CorpusID:252419668.

Shaozhe Hao, Kai Han, and Kwan-Yee Kenneth Wong. Learning attention as disentangler for compositional zero-shot learning. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15315–15324, 2023. URL https://api.semanticscholar.org/CorpusID:257766704.

Xiaoming Hu and Zilei Wang. A dynamic learning method towards realistic compositional zero-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 2265–2273, 2024. doi: 10.1609/aaai.v38i3.28000.

Siteng Huang, Biao Gong, Yutong Feng, Min Zhang, Yiliang Lv, and Donglin Wang. Troika: Multi-path cross-modal traction for compositional zero-shot learning. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24005–24014, 2024. URL https://api.semanticscholar.org/CorpusID:257766582.

Phillip Isola, Joseph J. Lim, and Edward H. Adelson. Discovering states and transformations in image collections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1383–1391, 2015.

Chenyi Jiang and Haofeng Zhang. Revealing the proximate long-tail distribution in compositional zero-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 2498–2506, 2024. doi: 10.1609/aaai.v38i3.28026.

Chenyi Jiang, Shidong Wang, Yang Long, Zechao Li, Haofeng Zhang, and Ling Shao. Imaginary-connected embedding in complex space for unseen attribute-object discrimination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3):1395–1413, 2025. doi: 10.1109/TPAMI.2024.3487631.

Shyamgopal Karthik, Massimiliano Mancini, and Zeynep Akata. Kg-sp: Knowledge guided simple primitives for open world compositional zero-shot learning. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9326–9335, 2022. URL https://api.semanticscholar.org/CorpusID:248798544.

Soohyeong Kim and Yong Suk Choi. Contextual feature expansion with superordinate concept for compositional zero-shot learning. *Applied Sciences*, 2025. URL https://api.semanticscholar.org/CorpusID:281285789.

Yong-Lu Li, Yue Xu, Xiaohan Mao, and Cewu Lu. Symmetry and group in attribute-object compositions. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11313–11322, 2020. URL https://api.semanticscholar.org/CorpusID:214743077.

Yun Li, Zhe Liu, Hang Chen, and Lina Yao. Context-based and diversity-driven specificity in compositional zero-shot learning. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17037–17046, 2024. URL https://api.semanticscholar.org/CorpusID:268032034.

Xiaodong Liu, Weixing Luo, Jiale Du, Xinshuo Wang, Yuhao Dang, and Yang Liu. A robust generalized zero-shot learning method with attribute prototype and discriminative attention mechanism. *Electronics*, 2024. URL https://api.semanticscholar.org/CorpusID:272866893.

Xiaocheng Lu, Song Guo, Ziming Liu, and Jingcai Guo. Decomposed soft prompt guided fusion enhancing for compositional zero-shot learning. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23560–23569, 2023. URL https://api.semanticscholar.org/CorpusID:253734570.

Massimiliano Mancini, Muhammad Ferjad Naeem, Yongqin Xian, and Zeynep Akata. Open world compositional zero-shot learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5218–5226, 2021. URL https://api.semanticscholar.org/CorpusID:231728349.

Ishan Misra, Abhinav Kumar Gupta, and Martial Hebert. From red wine to red tomato: Composition with context. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1160–1169, 2017. URL https://api.semanticscholar.org/CorpusID:19886856.

Muhammad Ferjad Naeem, Yongqin Xian, Federico Tombari, and Zeynep Akata. Learning graph embeddings for compositional zero-shot learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 953–962, 2021. URL https://api.semanticscholar.org/CorpusID:231786709.

Tushar Nagarajan and Kristen Grauman. Attributes as operators: Factorizing unseen attribute-object compositions. *Proceedings of the European Conference on Computer Vision (ECCV)*, abs/1803.09851, 2018. URL https://api.semanticscholar.org/CorpusID:52000169.

Nihal V. Nayak, Peilin Yu, and Stephen H. Bach. Learning to compose soft prompts for compositional zero-shot learning. *International Conference on Learning Representations*, abs/2204.03574, 2023. URL https://api.semanticscholar.org/CorpusID:248006494.

OpenAI. Gpt-4o system card, 2024. URL https://cdn.openai.com/gpt-4o-system-card.pdf. Accessed: 2026-05-26.

Senthil Purushwalkam, Maximilian Nickel, Abhinav Kumar Gupta, and Marc'Aurelio Ranzato. Task-driven modular networks for zero-shot compositional learning. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3592–3601, 2019. URL https://api.semanticscholar.org/CorpusID:155089387.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *Proceedings of the 38th International Conference on Machine Learning*, abs/2103.00020:8748–8763, 2021. URL https://api.semanticscholar.org/CorpusID:231591445.

Klearchos Stavrothanasopoulos, Konstantinos Gkountakos, Konstantinos Ioannidis, Theodora Tsikrika, Stefanos Vrochidis, and Ioannis Kompatsiaris. Cylinders: A benchmark visual dataset for robust gas cylinder detection and attribute classification in real-world scenes. *Sensors (Basel, Switzerland)*, 25, 2025. URL https://api.semanticscholar.org/CorpusID:276270129.

Qingsheng Wang, Lingqiao Liu, Chenchen Jing, Hao Chen, Guoqiang Liang, Peng Wang, and Chunhua Shen. Learning conditional attributes for compositional zero-shot learning. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11197–11206, 2023. URL https://api.semanticscholar.org/CorpusID:258960120.

Peng Wu, Qiuxia Lai, Hao Fang, Guo-Sen Xie, Yilong Yin, Xiankai Lu, and Wenguan Wang. A conditional probability framework for compositional zero-shot learning. *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3673–3683, 2025a. URL https://api.semanticscholar.org/CorpusID:280227308.

Peng Wu, Xiankai Lu, Hao Hu, Yongqin Xian, Jianbing Shen, and Wenguan Wang. Logiczsl: Exploring logic-induced representation for compositional zero-shot learning. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30301–30311, 2025b. URL https://api.semanticscholar.org/CorpusID:280157745.

Karn Yongsiriwit, Parkpoom Chaisiriprasert, Thannob Aribarg, and Sokliv Kork. Context-aware visual emotion recognition through hierarchical fusion of facial micro-features and scene semantics. *Applied Sciences*, 2025. URL https://api.semanticscholar.org/CorpusID:284012800.

Aron Yu and Kristen Grauman. Fine-grained visual comparisons with local learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 192–199, 2014.

Shiyu Zhang, Cheng Yan, Yang Liu, Chenchen Jing, Lei Zhou, and Wenjun Wang. Learning visual proxy for compositional zero-shot learning. *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2793–2802, 2025. URL https://api.semanticscholar.org/CorpusID:275820107.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. doi: 10.1007/s11263-022-01653-1.

A LIMITATIONS

FASC is a calibration method built on top of candidate pair logits, and its performance therefore depends on the candidate coverage and ranking quality of the underlying scorer. This dependence is most visible in the UT-Zappos open world setting in Table 1 of the main paper. Its material- and state-centric vocabulary creates many plausible but incorrect material–footwear pairs. Family calibration strengthens consistency, while reliable object evidence remains important when several candidates share similar family evidence.

The method also relies on a fixed attribute family taxonomy constructed offline with GPT-4o. Canonicalization and auditing reduce naming errors, and the random-taxonomy and no-family controls show that semantically coherent assignments improve calibration. UT-Zappos uses a legacy cache with fine-grained material channels, while C-GQA and MIT-States use richer coarse semantic families. Future work should learn uncertain or hierarchical assignments and evaluate freshly generated, taxonomy-aligned caches. Finally, the main results use VP-CMJL, and two additional C-GQA sources demonstrate that the interface transfers to different candidate-logit backbones.

B DATASET DETAILS

Table 7 summarizes the dataset statistics used by prior CZSL work (Lu et al., 2023; Huang et al., 2024; Zhang et al., 2025). Here, s and o denote the number of attribute/state and object concepts, c_s and c_u denote seen and unseen composition pairs for each split, and i denotes images.

Table 7: Summary statistics of the three CZSL datasets.

Dataset	s	o	Split	c_s	c_u	i
MIT-States	115	245	Train	1262	–	30338
			Val.	300	300	10420
			Test	400	400	12995
UT-Zappos	16	12	Train	83	–	22998
			Val.	15	15	3214
			Test	18	18	2914
C-GQA	453	870	Train	6963	–	26920
			Val.	1173	1368	7280
			Test	1022	1047	5098

The C-GQA evaluator retains the full benchmark inventory of 453 attributes and 870 objects. The family-evidence branch uses a fixed taxonomy over 413 attributes. The remaining 40 attributes do not have reliable assignments to the available semantic families and are therefore omitted only from family-based correction, without reducing the evaluator label space.

C SOURCE POOL DETAILS

Table 8 reports the fixed checkpoint pool with four sources used to generate candidate pair logits for each benchmark. Each pool is a deterministic checkpoint grid over backbone learning rates and random seeds, and each source uses the retained validation checkpoint snapshot from the corresponding backbone run. After the pool is fixed, every benchmark follows the same procedure: enumerate all nonempty source subsets with size at most four, uniformly average their logits, fit the same calibrator, and select the final subset by validation AUC. For $\mathcal{M} = \{S1, S2, S3, S4\}$, the search space is $\{\mathcal{S} : \emptyset \neq \mathcal{S} \subseteq \mathcal{M}\}$ and contains $2^4 - 1 = 15$ subsets. Source labels, rather than learning rate or seed alone, identify cached checkpoints. Each subset receives an independently fitted calibrator on the training split. The subset and calibrator are selected only by validation AUC, after which the corresponding scorer is evaluated on the test split. Closed- and open-world settings perform this selection separately because their candidate spaces differ. Test labels are not used for source selection.

All sources use the same VP-CMJL backbone architecture. The optimizer is Adam; batch size and training epochs follow the corresponding backbone recipe. C-GQA and MIT-States use the same seed/lr grid to keep source pool construction aligned across the two benchmarks.

Table 8: Backbone source pool used for logit averaging. Repeated UT-Zappos learning rates and seeds refer to distinct snapshot/cache identifiers, not duplicate files. The release manifest records the corresponding cache paths.

Dataset	Src.	LR	Seed	Snapshot/cache ID
C-GQA	S1	1.0×10^{-4}	51	retained val. snapshot
	S2	1.5×10^{-4}	52	retained val. snapshot
	S3	2.0×10^{-4}	53	retained val. snapshot
	S4	3.0×10^{-4}	54	retained val. snapshot
MIT-St.	S1	1.0×10^{-4}	51	retained val. snapshot
	S2	1.5×10^{-4}	52	retained val. snapshot
	S3	2.0×10^{-4}	53	retained val. snapshot
	S4	3.0×10^{-4}	54	retained val. snapshot
UT-Zap.	S1	5.0×10^{-4}	0	`official_e0` (ep. 0)
	S2	5.0×10^{-4}	0	`seed0_e1` (ep. 1)
	S3	5.0×10^{-4}	0	distinct seed-0 snapshot
	S4	5.0×10^{-4}	6	`seed6_e2` (ep. 2)

Table 9: Validation-selected source subsets used in the main experiments.

Dataset	Closed world	Open world
C-GQA	S1, S2, S3, S4	S1, S3, S4
MIT-States	S1, S4	S1, S4
UT-Zappos	S1, S2, S4	S1, S2, S3

Table 9 lists the subsets selected by validation AUC under the closed and open world settings. Closed and open world candidate spaces are evaluated separately, so the validation selected subset can also differ between the two settings. In all cases, the selection rule and the calibrator are unchanged; a smaller subset is selected when adding more sources does not improve validation AUC.

D Implementation and Optimization Details

We use VP-CMJL (Zhang et al., 2025) as the candidate-logit backbone, with CLIP ViT-L/14 (Radford et al., 2021) as its image and text encoder. The backbone checkpoints are fixed before calibrator training. For every candidate source subset described in Sec. C, FASC uniformly averages the cached pair logits and fits a separate calibrator.

The calibrator is optimized for four epochs with AdamW, learning rate 5×10^{-3}, zero optimizer weight decay, and batch size 512. Training uses the top-K candidates under the averaged source logits with $K = 50$; an image enters the cross-entropy term only when its true composition is retained in this set. The source weights remain uniform, the learnable base-logit scale is initialized to 1, and the final state is selected by train loss rather than validation labels. Validation is used only to compare the 15 fitted source subsets and select the final subset.

The pair, object-tail, attribute-tail, family, and family-object coefficients are initialized to 0.10, 0.10, 0.10, 0.05, and 0.05, respectively; the sign constraints described in the main paper are enforced throughout optimization. The three main ranked-family coefficients are initialized to 0.10, the two auxiliary coefficients to 0.02, the sample family-prior coefficient to 0, and the learnable main confidence threshold to 0.50. At training time, the hard gate is relaxed as $\sigma((\max_f q(f \mid x) - \theta)/0.05)$;

Table 10: Calibrator training and pseudo-AUC hyperparameters used for the reported experiments.

Item	Paper setting
Optimizer / epochs	AdamW / 4
Learning rate / weight decay	5×10^{-3} / 0
Batch size / top-K	512 / 50
Pseudo-unseen fraction	0.25 of distinct training pairs
$\lambda_{\text{auc}}, \lambda_{\text{anchor}}, \lambda_1$	0.02, 0.001, 0.05
τ_d, τ_p, τ_H	0.20, 0.10, 0.02
$w_{\text{AUC}}, w_{\text{HM}}, w_{\text{margin}}$	1, 1, 0.05
Bias grid	31 points, quantiles 0.02–0.98
Family ranks / gate relaxation	$R = 3$, $R_{\text{aux}} = 2$ / temperature 0.05
Source averaging / selection	uniform / exhaustive validation selection

inference uses $\mathbf{1}[\max_f q(f \mid x) \geq \theta]$. Unless otherwise stated, the main results use one fixed calibrator seed. For the five-run reliability analysis, we freeze the selected source subset and all hyperparameters and change only the calibrator seed. All backbone inference and calibrator fitting are conducted on NVIDIA A100 GPUs.

E Pseudo-Unseen AUC Surrogate

The pseudo-unseen split is constructed entirely within the training split and is used only by the auxiliary ranking objective. Let $\mathcal{P}_{\mathrm{tr}}$ be the distinct training compositions present in the cache. For each source subset, a seeded permutation of $\mathcal{P}_{\mathrm{tr}}$ is generated once. With $\rho = 0.25$, the first

$$h = \min\{\max(1, \mathrm{round}(\rho|\mathcal{P}_{\mathrm{tr}}|)), |\mathcal{P}_{\mathrm{tr}}| - 1\} \tag{13}$$

compositions form $\mathcal{P}_u$, and the remainder form $\mathcal{P}_s$. The partition, source logits, and bias grid are held fixed throughout a run. All eligible training images still contribute to top-K cross-entropy; "pseudo unseen" describes only their role in the auxiliary surrogate.

For training image i, let $r_i = S(x_i, c_i^\star)$ be the true-pair score. Let M_i^s and M_i^u be the maximum scores over pseudo-seen and pseudo-unseen candidates, and let m_i^s and m_i^u be the corresponding best within-domain competitors after excluding the true pair when it belongs to that domain. For a seen/unseen bias b, we use

$$\pi_i^s = \sigma((r_i - m_i^s)/\tau_p), \qquad d_i^s(b) = \sigma((r_i - M_i^u - b)/\tau_d), \tag{14}$$

$$\pi_i^u = \sigma((r_i - m_i^u)/\tau_p), \qquad d_i^u(b) = \sigma((r_i + b - M_i^s)/\tau_d). \tag{15}$$

The soft accuracies are

$$\widetilde{A}_s(b) = \frac{1}{|\mathcal{I}_s|} \sum_{i \in \mathcal{I}_s} \pi_i^s d_i^s(b), \tag{16}$$

$$\widetilde{A}_u(b) = \frac{1}{|\mathcal{I}_u|} \sum_{i \in \mathcal{I}_u} \pi_i^u d_i^u(b). \tag{17}$$

We evaluate these quantities on 31 ordered bias values. The grid uses the 0.02–0.98 quantiles of the pseudo-unseen gap $M_i^s - r_i$; when fewer than four pseudo-unseen examples are available, it falls back to a uniform grid on $[-3, 3]$. The soft AUC and smooth best harmonic mean are

$$\widetilde{\mathrm{AUC}} = \mathrm{Trapz}_b\Big(\widetilde{A}_s(b), \widetilde{A}_u(b)\Big), \tag{18}$$

$$\widetilde{H} = \tau_H \log \sum_b \exp\left(\frac{2\widetilde{A}_s(b)\widetilde{A}_u(b)}{\tau_H(\widetilde{A}_s(b) + \widetilde{A}_u(b) + 10^{-6})}\right). \tag{19}$$

The notation Trapz_b follows the implementation: it integrates soft seen accuracy with respect to soft unseen accuracy along the ordered bias sweep. The within-domain margin penalty is

$$\mathcal{M} = \frac{1}{|\mathcal{I}_s|} \sum_{i \in \mathcal{I}_s} \mathrm{softplus}\left(\frac{m_i^s - r_i}{\tau_p}\right) \tag{20}$$

$$+ \frac{1}{|\mathcal{I}_u|} \sum_{i \in \mathcal{I}_u} \mathrm{softplus}\left(\frac{m_i^u - r_i}{\tau_p}\right). \tag{21}$$

Finally,

$$\mathcal{L}_{\mathrm{pAUC}} = -w_{\mathrm{AUC}}\widetilde{\mathrm{AUC}} - w_{\mathrm{HM}}\widetilde{H} + w_{\mathrm{margin}}\mathcal{M}, \tag{22}$$

with the temperatures and weights in Table 10. The anchor regularizer is

$$\mathcal{L}_{\mathrm{anchor}} = \sum_{\theta \in \Theta} \mathrm{mean}\Big[(\theta - \theta^{(0)})^2\Big], \tag{23}$$

where $\theta^{(0)}$ is the initialization used for that source subset. No validation or test labels enter the cross-entropy, pseudo-unseen split, or surrogate.

F Fixed-Source Family-Signal Isolation

The main posterior q is deterministically recomputed from the uniformly averaged pair logits and the current taxonomy. The optional q^{aux} must be supplied by a separate sample-aligned, family-

Table 12: UT-Zappos parameter sensitivity. Bold settings are used in the main experiments.

Factor	Setting	Val	HM	AUC
Top-K	20	54.00	59.64	48.93
	50	53.72	59.60	48.91
	100	53.70	59.07	47.37
λ_{auc}	0.00	53.78	58.89	47.47
	0.02	53.72	59.60	48.91
	0.05	53.80	59.60	48.92
λ_1	0.00	53.89	59.50	48.70
	0.05	53.72	59.60	48.91
	0.10	54.02	59.23	47.62

channel-aligned cache; an absent auxiliary cache is never replaced by a copy of q. At evaluation time neither signal uses validation or test labels. Table 11 defines matched modes that hold source logits, generic priors, top-K, pseudo split, optimizer, and seed fixed.

Table 11: Definition of the fixed-source family-signal ablations. Disabled coefficients are set to zero and frozen.

Mode	Ranked q	Sample q	Ranked q^{aux}	Generic
Generic	No	No	No	Yes
q only	Yes	Yes	No	Yes
q^{aux} only	No	No	Yes	Yes
$q + q^{\text{aux}}$	Yes	Yes	Yes	Yes

The Generic endpoint corresponds to the "Generic" column of the fixed-source table in the main paper. It retains pair, attribute, and object frequency calibration, the base-logit scale, cross-entropy, and pseudo-AUC, but excludes all taxonomy-dependent statistics and corrections. This prevents improvements from generic calibration capacity from being attributed to semantic family evidence.

G Parameter Sensitivity

Table 12 varies one factor at a time on UT-Zappos while holding the remaining settings fixed. The sweep covers candidate-set size during training, the pseudo-AUC objective, and calibrator shrinkage. $K = 20$ and $K = 50$ perform similarly, whereas $K = 100$ admits weaker candidates and reduces AUC. A nonzero pseudo-unseen loss is important, but increasing its weight from 0.02 to 0.05 provides no meaningful gain. Moderate L_1 shrinkage performs best among the tested capacities. Overall, the selected configuration provides a stable tradeoff across the reported metrics and avoids the degradation observed with excessive candidate coverage or shrinkage.

H Efficiency Analysis

We measure the scoring cost after the backbone for FASC on C-GQA, which has the largest candidate space among our benchmarks. Once candidate pair logits are available, FASC only applies vectorized score scaling, family marginalization, offsets conditioned on family, and final ranking. On C-GQA with 5,098 test images and 9,378 cached candidate logits per image, a single-thread Apple M1 Pro CPU benchmark adds 0.0705 ms per image over base argmax ranking. This number is the median across three independent process runs, each of which reports the median of 21 timed repetitions after five warm-up repetitions. This indicates that inference cost is dominated by the backbone scorer, while the proposed family-structured calibration introduces negligible additional overhead.

I Taxonomy Audit

The released taxonomy files contain the fixed attribute-to-family assignments used by the family-evidence pipeline. The audited inventories are: 413 C-GQA family attributes mapped into 20 popu-

lated families (13 core, five auxiliary, and two noise families); all 115 MIT-States attributes mapped into 10 populated families within a 20-channel vocabulary; and all 16 UT-Zappos attributes mapped into six core families in the semantic release. For C-GQA, 40 of the 453 evaluator attributes do not admit reliable assignments to the available semantic families and are therefore not used by the family branch. Their candidates remain in the complete evaluation space and retain the backbone score and non-family calibration terms, while family-dependent corrections are applied to the 413 reliably categorized attributes.

Representative assignments include `red`→`color`, `plaid`→`pattern`, `angry`→`facial_expression_affect`, `cooked`→`food_preparation_edibility`, and `closed`/`wilted`→`physical_state`. For UT-Zappos, the semantic release groups materials into six families such as `textile_fabric`, `leather_finish`, and `suede_nubuck`.

The legacy UT-Zappos family cache used by the reported main experiment exposes 16 fine-grained channels, one for each material attribute. It is paired with the matching 16-channel mapping; the six-family semantic release is used for newly generated taxonomy-aligned caches. The UT-Zappos control confirms the importance of correct cache-to-family alignment, while the C-GQA and MIT-States random-taxonomy and no-family controls demonstrate the benefit of semantically coherent family grouping.

J Attribute Family Construction Prompt

We construct the attribute family taxonomy offline from the dataset attribute vocabulary. For each dataset, the placeholders below are replaced with the dataset attribute list, a small set of seed family names, and family names already produced in previous batches. The resulting model outputs are canonicalized before being fixed as the mapping τ used by FASC.

Prompt for Attribute Family Taxonomy Construction.

Role. You are a careful visual attribute taxonomy annotator for compositional zero-shot learning.

Task. Given a list of English attribute words, assign each attribute to three fields: `family`, a short and reusable semantic family name in snake case; `strength`, one of `core`, `aux`, or `noise`; and `mask_mode`, one of `local`, `global`, `hybrid`, or `none`.

Definitions. `local` means that the main visual evidence is usually on the foreground object; `global` means that it is mainly in the scene or background; `hybrid` means that object and context are both important; and `none` is used for noisy or unstable attributes. `core` families include color, material, pattern, shape, size, age, surface texture, physical state, pose, action, expression, cleanliness, and illumination. `aux` families include weather, environment, style, configuration, and food preparation. `noise` is used for object like modifiers, suspicious labels, or attributes without stable shared visual evidence.

Constraints. The family set is not closed: create a new family when needed, but prefer reusing semantically equivalent names from `[SEED_FAMILIES]` and `[EXISTING_FAMILIES]`. Use concise, stable, and general family names; do not use uninformative names such as `other`, `misc`, or `unknown`. Cover every attribute in `[ATTRIBUTE_LIST]` exactly once. Return JSON only, with no Markdown or extra explanation.