FACT: FACTORIZED ATTRIBUTE CONDITIONING FOR COMPOSITIONAL ZERO-SHOT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Compositional Zero-Shot Learning (CZSL) recognizes unseen attribute-object compositions, given training that covers all primitives but only some of their pairings. Its central difficulty is the context-dependence of attributes, which manifest differently across objects and make attribute recognition the bottleneck. In CLIP-based methods, causal attention leaves the attribute token object-invariant, so every composition receives the same context-free attribute signal, and prior methods compensate visually by compressing an attribute's diverse appearances into one composition-invariant feature. We instead adapt the text side with FACT (**F**actorized **A**ttribute **C**onditioning for prompt **T**uning), which casts the per-composition offsets as a correction field over the composition space and solves it by a low-rank Canonical Polyadic (CP) decomposition into per-primitive factors, giving every composition, seen or unseen, its own offset at negligible cost. As these factors are fitted only by the seen-composition loss, we add Centripetal Regularization, a global-interaction regularizer against overfitting those pairings, and Compositional Smoothing, which exempts the ground truth's primitive-sharing neighbors from one-hot suppression. As a text-side plug-in, FACT improves two bases and achieves the best AUC on three benchmarks under both closed- and open-world settings.

1 INTRODUCTION

Compositional Zero-Shot Learning (CZSL) Purushwalkam et al. (2019); Mancini et al. (2021); Munir et al. (2025) aims to recognize novel attribute-object compositions from images, given training coverage of all individual primitives but only a subset of their possible pairings. A fundamental difficulty arises from the context-dependent nature of attributes Wu et al. (2025a); Qu et al. (2025): the visual manifestation of an attribute varies substantially across objects. For example, "young" applied to a tree indicates fresh foliage, while for a horse it refers to juvenile proportions. This variability makes attribute recognition the central bottleneck: unlike object identity, attribute appearance must be interpreted relative to its accompanying object.

Recent CZSL methods Nayak et al. (2023); Huang et al. (2024) adopt a CLIP-based multi-branch framework that learns per-primitive prompt tokens to separately align attribute, object, and composition features with their visual counterparts, and generalizes to unseen compositions by composing the trained attribute and object tokens into novel pairings. However, CLIP treats the object as the primary entity and the attribute as a subordinate modifier, both in caption-level pre-training and at the structural level of the text encoder's causal attention. Since the attribute token precedes the object, its hidden states depend only on the prefix and itself, and thus remain identical across all compositions containing that attribute at every layer (Fig. 1(a)). The object token, in contrast, absorbs the attribute as context layer by layer. The EOS-pooled text feature therefore receives the attribute only through these object-invariant states and thus carries the same context-free attribute signal across every composition.

Existing methods address this challenge from the visual side Wu et al. (2025a); Qu et al. (2025); Yang et al. (2025), compressing the diverse visual appearances of an attribute across objects into one composition-invariant feature that aligns with the single text soft token (Fig. 1(b)). While this aids generalization by disentangling attributes from observed pairings, extracting such unified cues

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.

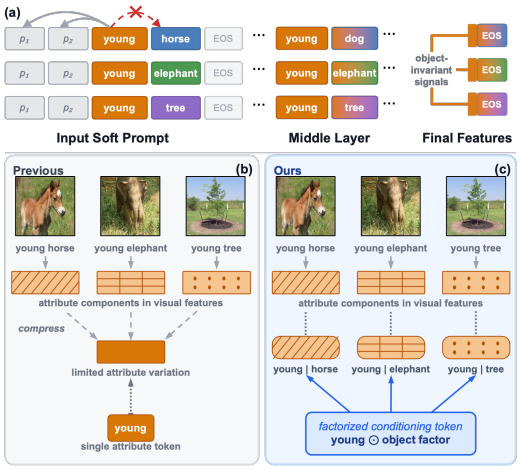

Figure 1: **Why FACT.** (a) Causal attention keeps the attribute token's hidden states identical across objects at every layer, so each composition receives the same object-free attribute signal. (b) Prior methods compress an attribute's diverse visual manifestations into one unified feature to match the single attribute token. (c) FACT instead expands the attribute into per-composition variants by solving a low-rank correction field, covering unseen pairs by factor recombination.

from inherently polysemous attributes remains challenging and struggles to capture every specific manifestation.

Therefore, taking a complementary perspective, we adapt the text side (Fig. 1(c)) by producing a distinct attribute text representation for each composition, so that the visual features can retain their per-image details and only need to align with a composition-aware text prototype. The key challenge is efficiently producing such a representation for every composition while covering unseen ones.

To address this, we first make the learning target precise. Writing $\boldsymbol{\delta}_{ij} \in \mathbb{R}^D$ for the offset that the composition of attribute a_i and object o_j adds to its attribute token, the offsets of all $M \times N$ compositions together form a *correction field* over the composition space, namely a third-order tensor $T \in \mathbb{R}^{M \times N \times D}$ whose (i, j)-th vector entry is $\boldsymbol{\delta}_{ij}$. Learning composition-specific conditioning is then the problem of solving this MND-entry field from its values on the seen compositions $\mathcal{C}^s$.

Determining a field from only a subset of its entries requires exploiting structure shared across coordinates. In our setting, such structure arises naturally from the compositional nature of attributes and objects: the correction induced by a composition can be decomposed into factors associated with the attribute and object, rather than learned independently for every attribute-object pair. This motivates modeling the correction field with a separable low-rank representation, in which each composition-specific offset is expressed through a small number of attribute and object factors. Specifically, a separable field of rank r is accordingly a superposition of r product modes, with each entry taking the form $\sum_{k=1}^{r} f_k(i)\, g_k(j)\, h_k(d)$, where mode k couples one profile per coordinate multiplicatively and the profiles along a coordinate are indexed by that coordinate alone. Applied to our correction field, this means that each offset is generated from profiles attached to its three coordinates rather than stored independently as a separate entry.

For the discretized three-coordinate field, this is exactly a Canonical Polyadic (CP) decomposition of rank at most r Kolda & Bader (2009), with rank-one component corresponding to one product mode Grasedyck et al. (2013). The learning problem thus becomes estimating these shared factors from the observed compositions, after which they can be recombined to construct offsets for unseen attribute-object pairs. Equivalently, the correction field amounts to estimating a rank-r factorization

$$T = \sum_{k=1}^{r} \mathbf{u}_{:,k} \circ \mathbf{v}_{:,k} \circ \mathbf{w}_{:,k}, \tag{1}$$

where $\circ$ denotes the outer product and the profiles of the three modes are collected into factor matrices $\mathbf{U} \in \mathbb{R}^{M \times r}$, $\mathbf{V} \in \mathbb{R}^{N \times r}$, and $\mathbf{W} \in \mathbb{R}^{D \times r}$. Such factorization reduces the number of free

parameters from $M \times N \times D$ entries to a quantity that scales with the numbers of attributes and objects, while enabling the factors learned from seen compositions to be recombined for unseen ones. Thus, the core problem is not to memorize the missing entries of T, but to recover its underlying factorized structure from the observed compositions. Appendix A makes precise how few parameters a rank-r field has and how its observed part constrains the rest.

This decomposition endows its factor matrices with a concrete meaning in the CZSL setting. The rows of $\mathbf{U}$ attach to the M attributes and the rows of $\mathbf{V}$ to the N objects, so a single factor $\mathbf{u}_i$ enters the correction of every composition that involves a_i, and the columns of $\mathbf{W}$ provide r shared directions in the token space through which every correction is written. The factorization therefore shares statistical strength across compositions through their constituent attributes and objects, rather than learning each composition independently. This structure aligns directly with the CZSL setting: attributes and objects observed during training are reused to form novel compositions at test time, allowing the learned primitive-level factors to be recombined for unseen attribute-object pairs.

We adopt this factorized structure as our modeling hypothesis for compositional conditioning. Its validity is supported by Proposition B.1 in Appendix B, which shows that the proposed factorization is consistent with the modifier-as-operator view of conceptual combination Baroni & Zamparelli (2010); Coecke et al. (2010). Specifically, when the attribute anchors are linearly independent, each field represented by our factorization induces an object-indexed linear operator of rank at most r on the attribute anchors. This provides a principled interpretation of the shared factors while keeping the representation explicitly compositional. We instantiate Eq. (1) as our Factorized Conditioning module, which holds the primitive factor tables $\mathbf{U}$ and $\mathbf{V}$ and the shared directions $\mathbf{W}$ as learnable parameters. For a composition (a_i, o_j), the offset $\boldsymbol{\delta}_{ij}$ follows the product rule of Eq. (1) from $\mathbf{u}_i$ and $\mathbf{v}_j$ and is added to its attribute token before the CLIP text encoder. Since the factors are shared across compositions, unseen pairs are conditioned by recombining factors learned from their constituent primitives, covering $M \times N$ compositions at a complexity that scales as $M+N+D$.

Factorized Conditioning provides the framework for modeling the correction field, but the factors are learned solely through the cross-entropy on the seen compositions. Consequently, the gradient received by primitive factors is determined by the primitives with which it is paired in training. This raises two problems. First, although every factor receives gradients, the interactions among factors are supervised only through the seen compositions, which can lead to overfits to the observed pairings and limit generalization. Our proposed Centripetal Regularization therefore trains offsets in the two primitive branches from the uniform average of the factors on the opposing side, which supplies a global-interaction regularizer for the optimization of Factorized Conditioning. Second, the cross-entropy also suppresses the activations of all non-target compositions equally in the recognition of each composition, including the neighbors that deserve to be exempt, namely the seen compositions that share an attribute or an object with the ground truth. Our proposed Compositional Smoothing therefore redistributes a fixed fraction of the target mass to those neighbors, divided evenly between the two kinds and leaving the remainder of the label space at zero. This structured label smoothing relieves the damage that the cross-entropy otherwise does to the training of Factorized Conditioning, and the two yield a joint gain (§4).

Factorized Conditioning, Centripetal Regularization, and Compositional Smoothing together constitute FACT (**F**actorized **A**ttribute **C**onditioning for prompt **T**uning), which operates entirely on the text prompt side and leaves the visual pipeline untouched.

Our contributions can be summarized as:

(1) We formulate composition-specific attribute conditioning in CZSL as the recovery of a separable correction field from seen compositions, and address it with a low-rank CP decomposition;

(2) We introduce Factorized Conditioning, which assigns learnable factors to individual attributes and objects and constructs each composition-specific offset by combining its primitive factors, enabling direct recombination to unseen compositions.

(3) To address the inherent training challenges under sparse composition coverage, we propose Centripetal Regularization and Compositional Smoothing, which jointly ensure reliable factor training and improve the generalization of Factorized Conditioning;

(4) As a text-side plug-in, FACT yields consistent gains on multiple base models and surpasses the state of the art on MIT-States*, C-GQA, and C-Fashion.

2 Related Work

Compositional zero-shot learning. Early methods compose primitive word embeddings or attentions Purushwalkam et al. (2019); Mancini et al. (2021), and CLIP-based approaches tune per-primitive prompt tokens Zhou et al. (2022); Nayak et al. (2023); Lu et al. (2023) or decompose recognition into attribute, object, and composition branches Huang et al. (2024); Qu et al. (2025); Yan & Feng (2025), composing the trained primitives at inference to recognize unseen pairs. A recent line addresses the context dependence of attributes directly, since an attribute manifests differently under different objects. CPF Wu et al. (2025a) models the conditional probability of an attribute given the object, ClusPRO Qu et al. (2025) mines multi-prototype clusters to capture within-primitive diversity, CDS-CZSL Li et al. (2024) exploits context-based diversity, LOGICZSL Wu et al. (2025b) injects LLM-generated logic rules into the training loss, and WARM-CAT Yan et al. (2026) adapts at test time with unlabeled data. These methods act on the visual features or on the supervision, and on the text side they all represent each primitive with a single context-free embedding, leaving the attribute-object dependency unmodeled there. FACT instead conditions the attribute prompt token itself through factorized interactions, addressing the attribute-object dependency exactly on the text side where these methods leave it unmodeled.

Factorized interaction models. Estimating unobserved combinations from per-entity factors is a classical route to interaction modeling, from factorization machines Rendle (2010) to low-rank tensor approximations such as the CP decomposition Kolda & Bader (2009); Grasedyck et al. (2013), and the same factorized form underlies low-rank adaptation of frozen networks Hu et al. (2021) and the treatment of a modifier as an operator on a head noun in compositional distributional semantics Baroni & Zamparelli (2010); Coecke et al. (2010). FACT formulates composition-specific attribute conditioning in CZSL as the recovery of a separable correction field from seen compositions and addresses it with a low-rank CP decomposition. It further designs Centripetal Regularization and Compositional Smoothing to help fit the factors from the seen compositions alone.

3 Method

3.1 Preliminaries

Given an attribute set $\mathcal{A}$ ($|\mathcal{A}|{=}M$) and an object set $\mathcal{O}$ ($|\mathcal{O}|{=}N$), the compositional label space is $\mathcal{C} = \mathcal{A} \times \mathcal{O}$, partitioned into seen $\mathcal{C}^s$ and unseen $\mathcal{C}^u$ with $\mathcal{C}^s \cap \mathcal{C}^u = \emptyset$. The model is trained on $\mathcal{T} = \{(x, c) \mid x \in \mathcal{X}, c \in \mathcal{C}^s\}$ and evaluated on $\mathcal{C}^{test} \supseteq \mathcal{C}^u$.

Recent CLIP-based CZSL methods Huang et al. (2024); Qu et al. (2025); Wu et al. (2025a) commonly adopt a three-path framework (Fig. 2(a)) that jointly models attribute, object, and composition recognition. On the visual side, a frozen CLIP image encoder produces $\mathbf{f} = \phi^{vis}(x) \in \mathbb{R}^D$, with lightweight adapters yielding branch-specific features $\mathbf{f}^a$, $\mathbf{f}^o$, and $\mathbf{f}^c{=}\mathbf{f}$. On the text side, three learnable prompts are constructed: $\mathbf{S}_i^a = [\mathbf{p}_{1:m}^a, \mathbf{w}_i^a]$, $\mathbf{S}_j^o = [\mathbf{p}_{1:m}^o, \mathbf{w}_j^o]$, and $\mathbf{S}_{ij}^c = [\mathbf{p}_{1:m}^c, \mathbf{w}_i^a, \mathbf{w}_j^o]$, where $\mathbf{p}_{1:m}$ are learnable prefix tokens and $\mathbf{w}_i^a, \mathbf{w}_j^o \in \mathbb{R}^D$ are per-primitive vocabulary tokens. A frozen CLIP text encoder ϕ^{txt} with causal attention maps each prompt to a text feature: $\mathbf{t}_i^a = \phi^{txt}(\mathbf{S}_i^a)$, $\mathbf{t}_j^o = \phi^{txt}(\mathbf{S}_j^o)$, $\mathbf{t}_{ij}^c = \phi^{txt}(\mathbf{S}_{ij}^c)$. Each branch computes cross-modal similarity and is supervised by cross-entropy: $\mathcal{L}^{base} = \mathcal{L}_{CE}^a + \mathcal{L}_{CE}^o + \mathcal{L}_{CE}^c$.

FACT operates entirely on the text prompt side of this framework and is agnostic to the visual pipeline, making it compatible with different models such as Troika Huang et al. (2024) and CLEAR Li et al. (2026).

3.2 Factorized Conditioning

Factorized Conditioning solves the correction field of §1 by holding one learnable factor per primitive and computing each composition's offset through the product rule of Eq. (1).

Therefore, following the modeling hypothesis of §1, we assign each primitive an interaction factor, $\mathbf{u}_i \in \mathbb{R}^r$ for attribute $a_i \in \mathcal{A}$ and $\mathbf{v}_j \in \mathbb{R}^r$ for object $o_j \in \mathcal{O}$, where $r \ll D$. The offset for

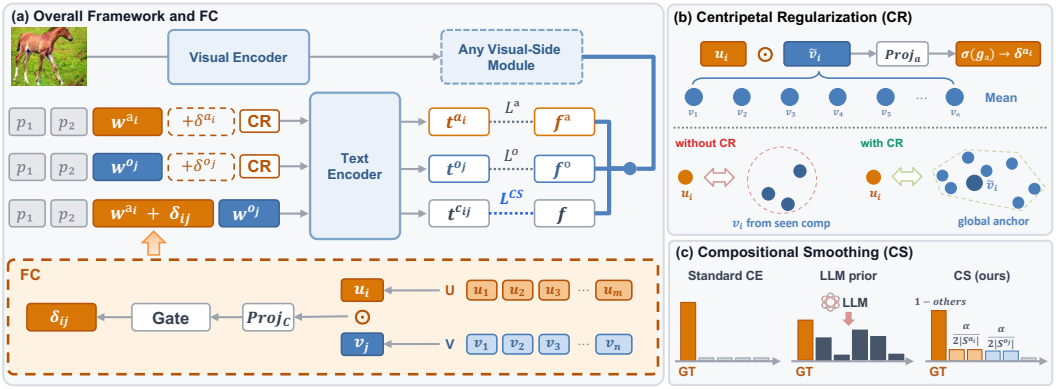

Figure 2: **The FACT framework.** (a) Three-branch CLIP-based CZSL with Factorized Conditioning (FC): each primitive gets a low-rank factor, and the offset $\boldsymbol{\delta}_{ij} = \sigma(g_c)\mathbf{W}_c(\mathbf{u}_i \odot \mathbf{v}_j)$ is injected into the attribute token of the composition prompt *before* the frozen text encoder. (b) Centripetal Regularization (CR): $\mathbf{u}_i$ interacts with a uniformly averaged summary $\tilde{\mathbf{v}}$ of *all* object factors, supplying a global-interaction regularizer against overfitting to the seen pairings. (c) Compositional Smoothing (CS): a fixed fraction of the one-hot target mass is redistributed to the attr-/obj-sharing seen neighbors, derived purely from the split structure, unlike LLM-prior soft labels. Examples are real seen compositions of MIT-States⋆.

composition (a_i, o_j) is the field of Eq. (1) evaluated at the cell (i, j),

$$\boldsymbol{\delta}_{ij} = \sigma(g_c)\,\mathbf{W}_c(\mathbf{u}_i \odot \mathbf{v}_j) = \sigma(g_c)\sum_{k=1}^{r} u_{ik}\, v_{jk}\, \mathbf{w}_{:,k}, \tag{2}$$

where $\mathbf{W}_c \in \mathbb{R}^{D\times r}$ is a learned projection whose columns are the shared directions $\mathbf{w}_{:,k}$, g_c a scalar gate, σ the sigmoid, and $\odot$ the Hadamard product. The forward pass is therefore the evaluation of a rank-r CP field. The composition prompt becomes

$$\tilde{\mathbf{S}}^c_{ij} = [\mathbf{p}^c_{1:m},\ \mathbf{w}^a_i + \boldsymbol{\delta}_{ij},\ \mathbf{w}^o_j]. \tag{3}$$

As shown in Fig. 2(a), $\boldsymbol{\delta}_{ij}$ is added directly to the attribute token, so that even within the frozen causal CLIP encoder the text feature carries a composition-specific attribute signal.

This structure satisfies our modeling assumptions from three perspectives. **(i)** *Composition-specificity*: the Hadamard product lets $\boldsymbol{\delta}_{ij}$ vary with $\mathbf{v}_j$, allowing the attribute signal to adapt to each object instead of remaining object-invariant. **(ii)** *Unseen coverage*: since every primitive participates in at least one seen composition, all factors $\mathbf{u}_i, \mathbf{v}_j$ are trained, and the offset $\boldsymbol{\delta}_{ij^*}$ of any unseen pair (i, j^*) is obtained by evaluating the same product rule on its two trained factors, with nothing stored per pair. **(iii)** *Parameter efficiency*: the factorized interaction uses $(M+N)\times r + r\times D$ parameters, scaling linearly with the number of primitives rather than with their Cartesian product. The low rank r further acts as implicit regularization, constraining the factorized interaction space and preventing overfitting to individual compositions.

3.3 Centripetal Regularization

Factorized Conditioning provides the framework in which the correction field is solved, but the factors are fitted only through the loss on the seen compositions. The composition branch covers only $\mathcal{C}^s$, so each factor $\mathbf{u}_i$ receives gradient from at most $|S_i|$ partner objects, where $S_i = \{j : (i, j) \in \mathcal{C}^s\}$. As an attribute co-occurs with only a small fraction of the objects, the gradient

$$\frac{\partial \mathcal{L}^c}{\partial \mathbf{u}_i} = \sigma(g_c)\sum_{j\in S_i} \mathrm{diag}(\mathbf{v}_j)\,\mathbf{W}_c^\top\,\mathbf{g}_{ij} \tag{4}$$

only involves $\{\mathbf{v}_j : j \in S_i\}$, so the interaction $\mathbf{u}_i \odot \mathbf{v}_j$ is exercised on those pairings alone. Every factor does receive a gradient, but an interaction fitted only where it is observed can overfit to those pairings and limit generalization to the unseen pairs that reuse the same factors (Fig. 2(b)).

Therefore, we leverage the primitive branches (Fig. 2(b)), whose classification objectives naturally span *all* N objects (resp. M attributes). In the attribute branch, we have $\mathbf{u}_i$ interact with the aggregate of all object-side factors:

$$\boldsymbol{\delta}_i^a = \sigma(g_a) \cdot \mathbf{W}_a \left(\mathbf{u}_i \odot \tilde{\mathbf{v}}\right), \tag{5}$$

where $\tilde{\mathbf{v}} = \frac{1}{N}\sum_{j=1}^{N} \mathbf{v}_j$ is the uniform average over all object factors. The attribute prompt becomes $\tilde{\mathbf{S}}_i^a = [\mathbf{p}_{1:m}^a, \mathbf{w}_i^a + \boldsymbol{\delta}_i^a]$. The object branch operates symmetrically with $\tilde{\mathbf{u}} = \frac{1}{M}\sum_i \mathbf{u}_i$: $\boldsymbol{\delta}_j^o = \sigma(g_o) \cdot \mathbf{W}_o(\mathbf{v}_j \odot \tilde{\mathbf{u}})$.

This introduces an additional gradient on $\mathbf{u}_i$, $\partial\mathcal{L}^a/\partial\mathbf{u}_i = \sigma(g_a)\,\mathrm{diag}(\tilde{\mathbf{v}})\,\mathbf{W}_a^\top\,\mathbf{g}_i^a$, in which $\mathbf{u}_i$ is coupled to every object factor at once. Since every weight is $1/N$, the aggregate does not depend on which pairs are observed, so this term supplies interaction signal beyond the seen pairings and acts as a global-interaction regularizer on the optimization of Factorized Conditioning, keeping $\mathbf{u}_i \odot \mathbf{v}_{j^*}$ meaningful for any unseen j^*.

Sec. 4 compares this aggregation with averages over only the seen or only the unseen opposing factors and with a learnable weighting; the uniform average is best in this comparison, consistent with its independence from the observation pattern.

3.4 Compositional Smoothing

The factors are fitted by the loss target, which for the composition branch is one-hot and assigns zero mass to every non-target composition. Consequently, the CE loss suppresses all of them equally, including the neighbors (compositions sharing an attribute or an object with the ground truth) that should be exempt (Fig. 2(c)). Under Eq. (2) those neighbors reuse one of the two factors of the ground truth, so suppressing them drives that factor away from the partner it is paired with in them, and the same factor generates the offsets of every other composition of that primitive.

We therefore redistribute a fraction of probability mass to primitive-sharing neighbors:

$$y[k] = \begin{cases} 1-\alpha & k = (i,j) \\ \frac{\alpha}{2|S_i^a|} & k \in S_i^a \\ \frac{\alpha}{2|S_j^o|} & k \in S_j^o \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

where S_i^a (resp. S_j^o) is the set of seen compositions that share attribute a_i (resp. object o_j) with the ground truth, excluding the GT pair itself, and α controls the smoothing strength. Unlike previous methods Wu et al. (2025b), which rely on external LLM-generated knowledge (Fig. 2(c)), we build our soft targets purely from the intrinsic structure of observed compositions, and this structured label smoothing relieves the damage that the suppression does to the training of Factorized Conditioning.

3.5 Training and Inference

The overall training objective of FACT is:

$$\mathcal{L} = \mathcal{L}_{CS}^c + \mathcal{L}_{CE}^a + \mathcal{L}_{CE}^o + \lambda\left(\frac{\|\mathbf{U}\|_F^2}{M} + \frac{\|\mathbf{V}\|_F^2}{N}\right), \tag{7}$$

where $\mathcal{L}_{CS}^c$ is the Compositional Smoothing target (Eq. 6) replacing the composition cross-entropy, and $\mathcal{L}_{CE}^a, \mathcal{L}_{CE}^o$ are the standard cross-entropies of the primitive branches whose prompts carry the CR offsets.

At inference, $\boldsymbol{\delta}_{ij}$ for any test composition is computed directly from the learned factors, adding no overhead. The prediction score for each candidate composition aggregates all three branches:

$$\hat{c} = \arg\max_{c_{ij} \in \mathcal{C}^{test}} \left[s(\mathbf{f}^c, \mathbf{t}_{ij}^c) + s(\mathbf{f}^a, \mathbf{t}_i^a) + s(\mathbf{f}^o, \mathbf{t}_j^o)\right], \tag{8}$$

where $s(\cdot,\cdot)$ denotes cosine similarity scaled by the CLIP temperature. FACT introduces only a small number of parameters, and although its factor tables grow with the number of primitives, even on C-GQA, the largest CZSL benchmark, the whole of FACT adds 0.26M parameters, about 0.04% of the 606M-parameter base model.

Table 1: Closed-world (CW) and open-world (OW) results on MIT-States⋆ and C-Fashion (CLIP ViT-L/14). Best AUC per setting in **bold**, second best underlined.

		MIT-States⋆								C-Fashion							
		CW				OW				CW				OW			
Method	Venue	S	U	HM	AUC	S	U	HM	AUC	S	U	HM	AUC	S	U	HM	AUC
DFSP Lu et al. (2023)	CVPR'23	46.7	46.2	46.2	31.4	46.4	26.0	25.5	9.8	69.0	70.2	58.2	44.0	69.2	50.1	47.1	30.7
PLID Bao et al. (2024)	ECCV'24	45.6	76.1	47.8	31.5	45.1	15.3	17.6	5.9	76.3	64.3	54.8	43.0	76.7	41.3	45.6	28.4
RAPR Jing et al. (2024)	AAAI'24	47.9	78.8	49.2	34.2	47.8	18.4	21.7	7.7	64.8	66.8	53.1	39.7	64.9	47.3	43.8	27.3
IMAX Jiang et al. (2025)	TPAMI'25	49.3	82.1	52.0	36.7	49.7	34.2	30.8	13.9	64.5	67.7	55.7	40.5	65.3	48.1	45.6	28.6
ClusPRO Qu et al. (2025)	ICLR'25	50.2	80.1	52.0	36.8	50.2	34.0	31.2	14.2	73.5	74.6	59.5	47.9	69.3	57.0	50.2	34.7
TOMCAT Yan & Feng (2025)	NeurIPS'25	55.0	79.7	52.7	39.2	52.6	35.2	30.7	15.2	74.3	72.9	59.7	48.0	74.2	58.4	54.8	39.5
Troika Huang et al. (2024)	CVPR'24	49.7	80.2	50.6	35.7	49.6	28.4	27.8	11.6	74.3	72.8	59.3	47.5	68.4	56.5	48.7	33.7
+ FACT (ours)	–	52.9	83.2	54.5	40.4	34.8	24.3	28.6	13.0	72.1	74.5	62.4	49.6	58.0	48.9	53.1	37.6
CLEAR Li et al. (2026)	ICME'26	52.3	81.2	53.6	38.6	52.3	25.0	26.5	10.7	72.4	75.4	62.0	49.8	71.7	58.1	53.4	38.0
+ FACT (ours)	–	54.0	82.8	55.8	**41.5**	54.0	38.6	32.6	**16.3**	73.3	76.6	63.4	**51.2**	73.3	59.4	55.1	**39.5**

Table 2: Closed-world (CW) and open-world (OW) results on C-GQA and UT-Zappos (CLIP ViT-L/14). Best AUC per setting in **bold**, second best underlined.

		C-GQA								UT-Zappos							
		CW				OW				CW				OW			
Method	Venue	S	U	HM	AUC	S	U	HM	AUC	S	U	HM	AUC	S	U	HM	AUC
DFSP Lu et al. (2023)	CVPR'23	27.1	38.2	27.1	10.5	38.3	7.2	10.4	2.4	66.7	71.7	47.2	36.0	66.8	60.0	44.0	30.3
PLID Bao et al. (2024)	ECCV'24	38.8	33.0	27.9	11.0	39.1	7.5	10.6	2.5	67.3	68.8	52.4	38.7	67.6	55.5	46.6	30.8
RAPR Jing et al. (2024)	AAAI'24	45.6	36.0	32.0	14.4	45.5	11.2	14.6	4.4	69.4	72.8	56.5	44.5	69.4	59.4	47.9	33.3
IMAX Jiang et al. (2025)	TPAMI'25	39.7	35.8	29.8	12.8	38.7	7.9	11.2	2.6	69.3	70.7	54.2	40.6	68.4	57.3	47.5	32.3
ClusPRO Qu et al. (2025)	ICLR'25	44.3	37.8	32.8	14.9	41.6	8.3	11.6	3.0	70.7	76.0	58.5	46.6	71.0	66.2	54.1	39.5
TOMCAT Yan & Feng (2025)	NeurIPS'25	45.3	40.1	34.0	16.0	45.1	10.6	14.2	4.2	74.5	72.8	60.2	**48.3**	74.1	65.8	57.9	**43.7**
LOGICZSL Wu et al. (2025b)	CVPR'25	44.4	39.4	33.3	15.3	43.7	9.3	12.6	3.4	69.6	74.9	57.8	45.8	69.6	63.7	50.8	36.2
FlowComposer He et al. (2026)	CVPR'26	44.8	40.7	34.0	15.9	43.5	10.2	12.6	3.5	71.1	74.9	58.6	46.8	70.1	61.2	51.0	35.5
Troika Huang et al. (2024)	CVPR'24	41.0	35.7	29.4	12.4	40.8	7.9	10.9	2.7	66.8	73.8	54.6	41.7	66.4	61.2	47.8	33.0
+ FACT (ours)	–	46.4	41.0	34.8	16.8	46.4	11.3	15.3	**4.7**	68.4	75.6	55.1	43.2	68.5	61.7	49.4	34.3
CLEAR Li et al. (2026)	ICME'26	45.5	41.9	34.9	16.6	45.8	11.5	11.3	4.6	69.3	76.4	58.2	46.3	68.7	65.8	52.6	38.1
+ FACT (ours)	–	46.6	41.8	35.3	**17.1**	46.6	11.2	15.0	4.6	69.1	77.4	58.4	46.7	68.6	68.4	51.9	39.3

4 EXPERIMENTS

4.1 SETUP

Datasets. We evaluate FACT on four benchmarks. **MIT-States⋆** Yan et al. (2026) refines MIT-States Isola et al. (2015) with corrected labels (141 attributes, 266 objects, 1,074 seen compositions); **C-GQA** Naeem et al. (2021) (413 attributes, 674 objects) is the largest; **UT-Zappos** Yu & Grauman (2014) (16 attributes, 12 objects) is the smallest; **C-Fashion** Yan et al. (2026) (76 attributes, 28 objects) covers fine-grained clothing.

Metrics. Following standard protocol Mancini et al. (2021); Huang et al. (2024), we report the top-1 accuracy on seen (S) and unseen (U) compositions, their harmonic mean (HM), and the area under the seen–unseen curve (AUC) traced by sweeping the seen-class bias, under both the closed-world (CW) and open-world (OW) settings; AUC is the primary metric.

Implementation. We use CLIP ViT-L/14 Radford et al. (2021) as the frozen backbone, factor rank r=32, Compositional Smoothing α=0.2, and Text LoRA Hu et al. (2021) rank 4 on all 12 layers, trained with Adam (lr 2e-4, step decay) on a single NVIDIA A6000 (48GB).

Baselines. We compare against recent CZSL methods spanning embedding composition Purushwalkam et al. (2019), prompt tuning Lu et al. (2023); Nayak et al. (2023), multi-branch recognition Huang et al. (2024); Qu et al. (2025), language-prior supervision Wu et al. (2025b); Bao et al. (2024), and test-time adaptation Yan & Feng (2025); Yan et al. (2026), and attach FACT to two architecturally different bases, Troika Huang et al. (2024) and CLEAR Li et al. (2026), to verify that it behaves as a text-side plug-in rather than a base-specific design.

Table 3: Component ablation on MIT-States⋆ (CW, AUC×100); Δ = gain over the previous row.

FC	CS	CR	AUC	Δ
			38.6	–
✓			39.2	+0.6
✓	✓		40.4	+1.2
✓	✓	✓	**41.5**	+1.1

Table 4: Interaction form comparison on MIT-States⋆ (CW, AUC×100). All variants include CR and CS.

Interaction form	AUC
Hadamard (ours)	**41.1**
Bilinear $(\mathbf{u}^\top \mathbf{G}) \odot \mathbf{v}$	40.6
MLP(concat$(\mathbf{u}, \mathbf{v})$)	40.3
Random factors (untrained)	38.8
Per-primitive bias (no interaction)	38.9

Table 5: Factor rank r on MIT-States⋆ (CW); r=32 (bold) is the value used throughout.

r	S	U	HM	AUC
8	52.6	82.7	55.0	40.2
16	52.7	83.5	55.2	40.8
32	54.0	82.8	55.8	**41.5**
64	53.8	82.5	55.4	41.3
128	54.1	82.2	55.7	41.4

Table 6: Aggregation of the opposing-side factors on MIT-States⋆ (CW).

Aggregation	S	U	HM	AUC
Seen only	56.8	74.6	53.8	39.5
Unseen only	52.7	78.3	53.1	39.3
All (ours)	54.0	82.8	55.8	**41.5**
Learnable	53.0	82.8	54.7	40.4

4.2 Comparison with State-of-the-Art

Tables 1 and 2 report closed- and open-world results grouped by dataset; as a text-side plug-in, FACT lifts both bases by +2.2 CW AUC on average and raises OW AUC +2.1 on average. Across MIT-States⋆, C-Fashion, and C-GQA, FACT achieves the best closed-world AUC. In the open world, FACT achieves the best AUC on MIT-States⋆ and C-GQA and ties the best on C-Fashion. On UT-Zappos FACT still improves both bases, but TOMCAT Yan & Feng (2025) stays ahead. This can be attributed to the shoe-specific material attributes (e.g., suede, canvas) in UT-Zappos, which admit only a few combinable objects and present consistent, readily inducible visual manifestations. The gains transfer across two architecturally different bases (Troika, CLEAR), confirming that FACT behaves as a general text-side plug-in rather than a base-specific design.

4.3 Ablation Study

Main Components. Table 3 ablates the three components of FACT on MIT-States⋆, where Δ is the gain over the previous row. Notably, the gains accumulate across the components in the order predicted by the method: CS relieves the one-hot suppression of primitive-sharing neighbors, and CR further regularizes factor interactions beyond the seen pairings. The complete combination is the best of the four and is the full model used throughout the paper.

Interaction structure of FC. As shown in Table 4, under the same setup (all with CS and CR), the Hadamard product outperforms the bilinear and MLP forms. The bilinear form mixes the attribute factor by a core matrix before the product, which does not enlarge the class of correction fields, since the mixed factors range over the same set as the free factor table, yet it adds r^2 parameters that are fitted on the seen pairs alone. The MLP form leaves the product structure of Eq. (1) altogether, so its offsets for unseen pairs are no longer generated by the rule under which the factors were trained. As two controls, disabling the interaction entirely (per-primitive bias) or replacing the learned factors with random ones leaves only the modest gains of the two regularizers, confirming that the gain stems mainly from the learned factorized interaction working with the regularizers. In particular, the random-factors row shows that the multiplicative form alone is not sufficient: with $\mathbf{u}, \mathbf{v}$ fixed at random the score falls below every trained interaction, so the improvement requires the factors to be trained rather than merely composed.

Factor rank. The rank r sets the capacity of the separable field: it fixes how many product modes the correction field of Eq. (1) may superpose, from a single mode at r=1 towards an unrestricted field as r grows, and hence how expressive the attribute–object interaction can be. Table 5 varies



Figure 3: **Qualitative cases on MIT-States⋆.** Predictions are marked green when correct and red when wrong. Left eight (columns 1–4): attribute-confusion corrections, where the CLEAR baseline binds a wrong attribute to the correctly recognized object and FACT recovers the ground truth. Right four (columns 5–6): remaining failures, where the baseline is correct but FACT still confuses a fine-grained attribute, marking the limit of the correction.

r from 8 to 128 on MIT-States⋆. Where the rank is too small to carry the interaction, the AUC falls below the r=32 value, and where it is larger the extra modes bring no further gain, so the curve plateaus once the rank suffices and the residual spread is comparable to single-seed noise. The effective interaction complexity of the task is therefore modest, and we use r=32 for all four benchmarks rather than tuning it per dataset.

Aggregation Strategies in CR. Table 6 compares four ways of forming the aggregate of opposing-side factors in the primitive branches. Averaging all opposing primitives, the setting used throughout, is the best of the four, since it is the one aggregation whose value does not depend on which pairs happen to be observed and so cannot drift towards the seen partners. Restricting the average to the seen partners, or to the unseen ones, is worse in both cases: the former is exactly the quantity the composition branch already fits and so adds no new interaction signal, while the latter discards the evidence the seen pairs provide about the opposing coordinate. Making the weights learnable is likewise worse than the uniform mean: because a factor is observed only together with its seen partners, the recognition loss can reward the weights only for concentrating on those partners, so the learned aggregate collapses onto the seen subset and overfits.

Qualitative results. Fig. 3 shows representative attribute-confusion cases corrected by FACT together with its remaining failures.

5 Conclusion

We present FACT, a text-side plug-in for CLIP-based CZSL that produces composition-specific attribute representations by solving a low-rank correction field through per-primitive factors, and that fits these factors with two corrections to the loss on the seen compositions, Centripetal Regularization and Compositional Smoothing. Experiments on four benchmarks show that FACT consistently improves two architecturally different bases under both the closed- and open-world settings, and surpasses the state of the art on three of them. Because FACT operates only on the prompt side and leaves the visual pipeline untouched, it is compatible with CLIP-based CZSL models that follow the common three-branch framework.

AI USE STATEMENT

In this work, generative AI tools were used only as assistive tooling in the four capacities below; the research questions, method, experiments, proofs, and all final text originate with the authors. **(i) Writing aid and polishing.** AI assisted with grammar, phrasing, and terminology consistency of paragraphs that the authors had drafted; every such passage was revised and approved by the authors before use. **(ii) Retrieval and discovery.** AI helped surface candidate references and summarize publicly available abstracts during the related-work survey; every cited work was selected, read, and checked by the authors, and all compared numbers were taken from the original publications or from our own re-runs. **(iii) Drafting selected passages.** AI produced first drafts of a few expository passages (related-work summaries and figure/table captions) from author-supplied bullet points, equations, and measured numbers; the authors rewrote, fact-checked, and finalized every sentence, and no section of the paper was generated end-to-end by AI. **(iv) Checking mathematical derivations.** AI performed symbolic and numeric cross-checks of the identities in Appendices A and B on random tensors (e.g., the operator construction of Proposition B.1 and the degrees-of-freedom count of Proposition A.2); one such check motivated a boundary-condition clarification in Proposition A.2, which the authors re-derived and verified independently. All definitions, propositions, proof strategies, and final proofs are the authors' work. We have not used generative AI tools for research ideation or for generating synthetic datasets, and no other disclosure-required use applies. We have reviewed all AI-assisted work and take full responsibility for the final content of this paper, including any text, claims, or artifacts produced with the aid of AI.

REPRODUCIBILITY STATEMENT

Sections 3 and 3.5 specify the FACT architecture, training objective, inference rule, and all model-specific settings, including the factor rank, smoothing strength, and Text LoRA configuration. Section 4 describes the four public benchmarks, evaluation protocol, frozen backbone, optimizer, learning-rate schedule, and hardware. Tables 3–6 report component, interaction, rank, and aggregation ablations, while the appendix provides the supporting structural analysis, together providing the information needed to reimplement and evaluate FACT under the stated protocol.

REFERENCES

Wentao Bao, Lichang Chen, Heng Huang, and Yu Kong. Prompting language-informed distribution for compositional zero-shot learning. In *Computer Vision – ECCV 2024*, pp. 107–123, 2024. ISBN 9783031726309. doi: 10.1007/978-3-031-72630-9_7. URL http://dx.doi.org/10.1007/978-3-031-72630-9_7.

Marco Baroni and Roberto Zamparelli. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 1183–1193, 2010. URL https://aclanthology.org/D10-1115/.

Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. Mathematical foundations for a compositional distributional model of meaning. *Linguistic Analysis*, 36(1-4):345–384, 2010. ISSN 0098-9053.

Lars Grasedyck, Daniel Kressner, and Christine Tobler. A literature survey of low-rank tensor approximation techniques. *GAMM-Mitteilungen*, 36(1):53–78, 2013. ISSN 1521-4001. doi: 10.1002/gamm.201310004. URL http://dx.doi.org/10.1002/gamm.201310004.

Zhenqi He, Lin Li, and Long Chen. Flowcomposer: Composable flows for compositional zero-shot learning. *arXiv preprint arXiv:2603.16641*, 2026. URL https://arxiv.org/abs/2603.16641.

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

Siteng Huang, Biao Gong, Yutong Feng, Min Zhang, Yiliang Lv, and Donglin Wang. Troika: Multi-path cross-modal traction for compositional zero-shot learning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24005–24014, 2024. doi: 10.1109/cvpr52733.2024.02266. URL http://dx.doi.org/10.1109/cvpr52733.2024.02266.

Phillip Isola, Joseph J. Lim, and Edward H. Adelson. Discovering states and transformations in image collections. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1383–1391, 2015. doi: 10.1109/cvpr.2015.7298744. URL http://dx.doi.org/10.1109/cvpr.2015.7298744.

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. ISSN 1939-3539. doi: 10.1109/tpami.2024.3487631. URL http://dx.doi.org/10.1109/tpami.2024.3487631.

Chenchen Jing, Yukun Li, Hao Chen, and Chunhua Shen. Retrieval-augmented primitive representations for compositional zero-shot learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(3):2652–2660, 2024. ISSN 2159-5399. doi: 10.1609/aaai.v38i3.28043. URL http://dx.doi.org/10.1609/aaai.v38i3.28043.

Tamara G. Kolda and Brett W. Bader. Tensor decompositions and applications. *SIAM Review*, 51 (3):455–500, 2009. ISSN 0036-1445. doi: 10.1137/07070111x. URL http://dx.doi.org/10.1137/07070111x.

Joseph B. Kruskal. Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra and its Applications*, 18 (2):95–138, 1977. ISSN 0024-3795. doi: 10.1016/0024-3795(77)90069-6. URL http://dx.doi.org/10.1016/0024-3795(77)90069-6.

Weize Li, Zhicheng Zhao, and Fei Su. From model patterns to abstract semantics in compositional zero-shot learning. *arXiv preprint arXiv:2609.15649*, 2026. URL https://arxiv.org/abs/2609.15649.

Yun Li, Zhe Liu, Hang Chen, and Lina Yao. Context-based and diversity-driven specificity in compositional zero-shot learning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17037–17046, 2024. doi: 10.1109/cvpr52733.2024.01612. URL http://dx.doi.org/10.1109/cvpr52733.2024.01612.

Xiaocheng Lu, Song Guo, Ziming Liu, and Jingcai Guo. Decomposed soft prompt guided fusion enhancing for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23560–23569, 2023.

Massimiliano Mancini, Muhammad Ferjad Naeem, Yongqin Xian, and Zeynep Akata. Open world compositional zero-shot learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5218–5226, 2021. doi: 10.1109/cvpr46437.2021.00518. URL http://dx.doi.org/10.1109/cvpr46437.2021.00518.

Ans Munir, Faisal Z. Qureshi, Mohsen Ali, and Muhammad Haris Khan. Compositional zero-shot learning: A survey. *arXiv preprint arXiv:2510.11106*, 2025. URL https://arxiv.org/abs/2510.11106.

Muhammad Ferjad Naeem, Yongqin Xian, Federico Tombari, and Zeynep Akata. Learning graph embeddings for compositional zero-shot learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 953–962, 2021. doi: 10.1109/cvpr46437.2021.00101. URL http://dx.doi.org/10.1109/cvpr46437.2021.00101.

Nihal V. Nayak, Peilin Yu, and Stephen H. Bach. Learning to compose soft prompts for compositional zero-shot learning. *arXiv preprint arXiv:2204.03574*, 2023. URL https://arxiv.org/abs/2204.03574.

Senthil Purushwalkam, Maximillian Nickel, Abhinav Gupta, and Marc'aurelio Ranzato. Task-driven modular networks for zero-shot compositional learning. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3592–3601, 2019. doi: 10.1109/iccv.2019.00369. URL http://dx.doi.org/10.1109/iccv.2019.00369.

Hongyu Qu, Jianan Wei, Xiangbo Shu, and Wenguan Wang. Learning clustering-based prototypes for compositional zero-shot learning. In *International Conference on Learning Representations*, volume 2025, pp. 5554–5575, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/10b99813d5992673da90a499b6bcd5f0-Paper-Conference.pdf.

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. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Steffen Rendle. Factorization machines. In *2010 IEEE International Conference on Data Mining*, pp. 995–1000, 2010. doi: 10.1109/icdm.2010.127. URL http://dx.doi.org/10.1109/icdm.2010.127.

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. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3673–3683, 2025a. doi: 10.1109/iccv51701.2025.00350. URL http://dx.doi.org/10.1109/iccv51701.2025.00350.

Peng Wu, Xiankai Lu, Hao Hu, Yongqin Xian, Jianbing Shen, and Wenguan Wang. Logiczsl: Exploring logic-induced representation for compositional zero-shot learning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30301–30311, 2025b. doi: 10.1109/cvpr52734.2025.02821. URL http://dx.doi.org/10.1109/cvpr52734.2025.02821.

Xudong Yan and Songhe Feng. Tomcat: Test-time comprehensive knowledge accumulation for compositional zero-shot learning. *arXiv preprint arXiv:2510.20162*, 2025. URL https://arxiv.org/abs/2510.20162.

Xudong Yan, Songhe Feng, Jiaxin Wang, Xin Su, and Yi Jin. Warm-cat: Warm-started test-time comprehensive knowledge accumulation for compositional zero-shot learning. *arXiv preprint arXiv:2602.23114*, 2026. URL https://arxiv.org/abs/2602.23114.

Pan Yang, Cheng Deng, Jing Yang, Han Zhao, Yun Liu, Yuling Chen, Xiaoli Ruan, and Yanping Chen. Cams: Towards compositional zero-shot learning via gated cross-attention and multi-space disentanglement. *arXiv preprint arXiv:2511.16378*, 2025. URL https://arxiv.org/abs/2511.16378.

Aron Yu and Kristen Grauman. Fine-grained visual comparisons with local learning. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 192–199, 2014. doi: 10.1109/cvpr.2014.32. URL http://dx.doi.org/10.1109/cvpr.2014.32.

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. ISSN 1573-1405. doi: 10.1007/s11263-022-01653-1.

A PROPERTIES OF SEPARABLE FIELDS

Throughout this section, a *separable field of rank* r on the three coordinates (M, N, D) is a tensor $T \in \mathbb{R}^{M\times N\times D}$ of the form

$$T[i, j, d] = \sum_{k=1}^{r} f_k(i)\, g_k(j)\, h_k(d), \tag{9}$$

where the profiles of the three modes are collected into $\mathbf{U} \in \mathbb{R}^{M\times r}$, $\mathbf{V} \in \mathbb{R}^{N\times r}$, and $\mathbf{W} \in \mathbb{R}^{D\times r}$ through $\mathbf{U}_{ik} = f_k(i)$, $\mathbf{V}_{jk} = g_k(j)$, and $\mathbf{W}_{dk} = h_k(d)$. Equivalently, $T = [\![\mathbf{U}, \mathbf{V}, \mathbf{W}]\!] = \sum_{k=1}^{r} \mathbf{u}_{:,k} \circ \mathbf{v}_{:,k} \circ \mathbf{w}_{:,k}$, which is a tensor of CP rank at most r Kolda & Bader (2009), and a single rank-one term is the discretization of one product mode Grasedyck et al. (2013). We record three properties of this class, and the rank r controls its capacity, which expands from a single product mode at $r{=}1$ towards unrestricted fields as r grows.

Proposition A.1 (Value subspace)**.** *Every vector entry of T lies in the column space of $\mathbf{W}$. Explicitly, $T[i, j, :] = \mathbf{W}\,\mathbf{c}_{ij}$ with $\mathbf{c}_{ij}[k] = f_k(i)\, g_k(j)$, so the set of values $\{T[i, j, :] : 1 \le i \le M, 1 \le j \le N\}$ spans a subspace of $\mathbb{R}^D$ of dimension at most $\min\{r, \operatorname{rank}(\mathbf{W})\}$.*

Proof. Substituting Eq. (9), $T[i, j, :] = \sum_{k=1}^{r} f_k(i) g_k(j)\, h_k = \mathbf{W}\,\mathbf{c}_{ij}$, which is a linear combination of the columns of $\mathbf{W}$. □

Proposition A.2 (Degrees of freedom)**.** *The field T in Eq. (9) is determined by the $(M{+}N{+}D)r$ entries of $(\mathbf{U}, \mathbf{V}, \mathbf{W})$, so the separable fields of rank at most r form a subset of $\mathbb{R}^{M\times N\times D}$ of dimension at most $(M{+}N{+}D)r$, strictly smaller than MND whenever $(M{+}N{+}D)r < MND$. The parameterization is redundant, and the effective number of degrees of freedom is $(M{+}N{+}D)r{-}2r$, capped at MND once the secant variety fills the ambient space; every shape and rank used in this paper lies in the uncapped regime.*

Proof. The first statement is immediate from Eq. (9). For the second, let $\alpha_k, \beta_k, \gamma_k$ be nonzero scalars with $\alpha_k\beta_k\gamma_k = 1$ for every k, and let $\mathbf{P}$ be an $r \times r$ permutation matrix. Then

$$\sum_{k=1}^{r} (\mathbf{U}\operatorname{diag}(\boldsymbol{\alpha})\,\mathbf{P})_{:,k} \circ (\mathbf{V}\operatorname{diag}(\boldsymbol{\beta})\,\mathbf{P})_{:,k} \circ (\mathbf{W}\operatorname{diag}(\boldsymbol{\gamma})\,\mathbf{P})_{:,k} = \sum_{k=1}^{r} \alpha_k\beta_k\gamma_k\ \mathbf{u}_{:,k} \circ \mathbf{v}_{:,k} \circ \mathbf{w}_{:,k} = T, \tag{10}$$

since the permutation only relabels the modes of the sum. The continuous part of this symmetry group has dimension $2r$, two free scalars per mode, so the fiber over a generic field is $2r$-dimensional and the image has dimension $(M{+}N{+}D)r - 2r$, or MND in the capped regime where the secant variety fills the ambient space. This agrees with the standard count for the affine cone over the r-th secant variety of the Segre variety, whose expected dimension is $r(M{+}N{+}D{-}3) + (r{-}1)$ in projective space. □

Proposition A.3 (Slice coupling)**.** *Fixing the first coordinate at i gives the slice $T[i, :, :] = \sum_{k=1}^{r} f_k(i)\,(\mathbf{g}_k \circ \mathbf{h}_k) \in \mathbb{R}^{N\times D}$, which lies in $\operatorname{span}\{\mathbf{g}_k \circ \mathbf{h}_k\}_{k=1}^{r}$, a subspace of dimension at most r that does not depend on i. The analogous statement holds for the slices of the other two coordinates.*

Proof. Immediate from Eq. (9), since fixing i turns the coefficients $f_k(i)$ into scalars while the profiles $\mathbf{g}_k$ and $\mathbf{h}_k$ are unchanged. □

Remark A.1. *Proposition A.3 says that the slices of one coordinate share a common mode space generated by the profiles of the other two, so observations on some slices carry information about the space in which all remaining slices are evaluated. We make no claim that a field is uniquely determined by its observed part. Such a claim would require assumptions on the observation pattern and on the field itself that our setting does not supply, since the correction field is never observed directly and its seen part is fixed by the benchmark split rather than sampled at random.*

B Operator form of the correction field

Proposition B.1 (Modifier as a low-rank operator)**.** *Let the anchors $\{\mathbf{w}_i^a\}_{i=1}^M \subset \mathbb{R}^D$ be linearly independent, which implies $M \le D$, and let $\mathbf{B} = [\mathbf{w}_1^a \mid \cdots \mid \mathbf{w}_M^a] \in \mathbb{R}^{D\times M}$. Then for every field $T = [\![\mathbf{U}, \mathbf{V}, \mathbf{W}]\!]$ of CP rank at most r there exist matrices $\mathbf{L}_j \in \mathbb{R}^{D\times D}$, $j = 1, \ldots, N$, with $\operatorname{rank}(\mathbf{L}_j) \le r$, such that*

$$\mathbf{w}_i^a + T[i, j, :] = (\mathbf{I} + \mathbf{L}_j)\, \mathbf{w}_i^a \qquad \text{for all } i \text{ and } j. \tag{11}$$

One may take $\mathbf{L}_j = \mathbf{W} \operatorname{diag}(\mathbf{v}_j)\, \mathbf{\Theta}$ with $\mathbf{\Theta} = \mathbf{U}^\top \mathbf{B}^+$ and $\mathbf{B}^+ = (\mathbf{B}^\top \mathbf{B})^{-1} \mathbf{B}^\top$, and this $\mathbf{\Theta}$ satisfies $\mathbf{\Theta}\, \mathbf{w}_i^a = \mathbf{u}_i$ for every i.

Proof. Because $\mathbf{B}$ has full column rank, $\mathbf{B}^+\mathbf{B} = \mathbf{I}_M$ and therefore $\mathbf{B}^+ \mathbf{w}_i^a = \mathbf{e}_i$, the i-th coordinate vector of $\mathbb{R}^M$. Define $\mathbf{\Theta} = \mathbf{U}^\top \mathbf{B}^+ \in \mathbb{R}^{r\times D}$, which is well defined for any $\mathbf{U}$ under the independence assumption alone, and note that $\mathbf{\Theta}\, \mathbf{w}_i^a = \mathbf{U}^\top \mathbf{e}_i = \mathbf{u}_i$. Set $\mathbf{L}_j = \mathbf{W} \operatorname{diag}(\mathbf{v}_j)\, \mathbf{\Theta}$, whose dimensions are $(D \times r)(r \times r)(r \times M)(M \times D)$. Writing $\boldsymbol{\theta}_k^\top$ for the k-th row of $\mathbf{\Theta}$, we have $\mathbf{L}_j = \sum_{k=1}^r v_{jk}\, \mathbf{w}_{:,k}\, \boldsymbol{\theta}_k^\top$, a sum of r rank-one matrices, so $\operatorname{rank}(\mathbf{L}_j) \le r$. Finally, $\mathbf{L}_j\, \mathbf{w}_i^a = \mathbf{W} \operatorname{diag}(\mathbf{v}_j)\, \mathbf{\Theta}\, \mathbf{w}_i^a = \mathbf{W} \operatorname{diag}(\mathbf{v}_j)\, \mathbf{u}_i = \sum_{k=1}^r u_{ik} v_{jk}\, \mathbf{w}_{:,k} = T[i, j, :]$, which establishes Eq. (11). □

Remark B.1 (The roles are independent of the factorization)**.** *A given field admits many factor triples, yet the operator form does not depend on which one is chosen. Under the rescaling $\mathbf{u}_{:,k} \mapsto \alpha_k \mathbf{u}_{:,k}$, $\mathbf{v}_{:,k} \mapsto \beta_k \mathbf{v}_{:,k}$, $\mathbf{w}_{:,k} \mapsto \gamma_k \mathbf{w}_{:,k}$ with $\alpha_k \beta_k \gamma_k = 1$, the factors transform as $\mathbf{U} \mapsto \mathbf{U} \operatorname{diag}(\boldsymbol{\alpha})$, $\mathbf{V} \mapsto \mathbf{V} \operatorname{diag}(\boldsymbol{\beta})$, $\mathbf{W} \mapsto \mathbf{W} \operatorname{diag}(\boldsymbol{\gamma})$, so $\mathbf{\Theta} \mapsto \operatorname{diag}(\boldsymbol{\alpha})\, \mathbf{\Theta}$ and $\operatorname{diag}(\mathbf{v}_j) \mapsto \operatorname{diag}(\boldsymbol{\beta}) \operatorname{diag}(\mathbf{v}_j)$, whence $\mathbf{L}_j \mapsto \mathbf{W} \operatorname{diag}(\boldsymbol{\gamma}\boldsymbol{\beta}\boldsymbol{\alpha}) \operatorname{diag}(\mathbf{v}_j)\, \mathbf{\Theta} = \mathbf{L}_j$. Under a permutation $\mathbf{P}$ of the r modes, $\mathbf{\Theta} \mapsto \mathbf{P}^\top \mathbf{\Theta}$ and $\mathbf{v}_j \mapsto \mathbf{P}^\top \mathbf{v}_j$, and $\mathbf{L}_j \mapsto \mathbf{W}\mathbf{P} \operatorname{diag}(\mathbf{P}^\top \mathbf{v}_j)\, \mathbf{P}^\top \mathbf{\Theta} = \mathbf{W} \operatorname{diag}(\mathbf{v}_j)\, \mathbf{\Theta} = \mathbf{L}_j$ because $\mathbf{P} \operatorname{diag}(\mathbf{P}^\top \mathbf{v}_j)\, \mathbf{P}^\top = \operatorname{diag}(\mathbf{v}_j)$. The three statements that $\mathbf{\Theta}$ reads each anchor into its attribute factor, that $\mathbf{v}_j$ supplies the r modulation coefficients, and that the columns of $\mathbf{W}$ are shared output directions are therefore properties of the field rather than of a particular parameterization. By Kruskal's criterion, the factorization is unique up to exactly this symmetry whenever the sum of the Kruskal ranks of $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{W}$ is at least $2r+2$ Kruskal (1977); Kolda & Bader (2009). We accordingly attach no meaning to individual factor coordinates and interpret only the roles of the three factor families.*

Remark B.2 (Converse direction)**.** *Not every family of object-indexed operators of rank at most r arises from a separable field, since Eq.* (11) *additionally requires the $\mathbf{L}_j$ to share one set of r output directions and one readout $\mathbf{\Theta}$, with only the coefficients v_{jk} varying across objects. The inclusion established in Proposition B.1 is one-directional, which is what §1 uses.*

C Limitations

The separable structure is a modeling hypothesis, tested rather than assumed. We posit the rank-r separable form of the correction field and subject it to the tests that a hypothesis affords: the rank sweep of Table 5, the interaction-form comparison of Table 4, and the no-interaction and random-factor controls, all of which support the product structure at a small rank. What the plateau shows is that the effective interaction complexity of the task is modest enough for a small rank to suffice, which is the only property our use of the hypothesis requires; whether attribute–object interactions are low-rank in any intrinsic sense lies outside the reach of any empirical CZSL study, since the field itself is latent.

The gain is bounded by how compositional a benchmark is. On UT-Zappos, FACT improves both base models, by $+1.5$ and $+0.4$ CW AUC over Troika and CLEAR, while the gap to TOMCAT persists. The shoe attributes of this benchmark admit only a few combinable objects and present visually consistent, readily inducible manifestations, so a context-free attribute token already carries most of the information and the composition-specific correction has little left to add; the marginal value of modeling composition-specificity shrinks as the attribute–object dependency weakens, and the polysemous benchmarks are where FACT sets the state of the art.

The correction field is latent, so its mechanisms are validated through consequences. As with other learned representations, the correction field is a latent construct: no ground-truth field exists

in CZSL, and its seen part is fixed by the benchmark split rather than sampled, so we make no unique-recovery or identifiability claim (Remark A.1). Its validity is established by the route standard for latent structures, namely observable consequences: the component, interaction, rank, and aggregation ablations of Tables 3–5 and the structural properties of Appendices A–B.

Fine-grained attributes are the residual hard cases. Where two attributes differ only by a subtle visual cue, the corrected attribute token can still be confused or over-corrected; on MIT-States* these residual cases are a small minority, e.g. FACT errs on 118 of the $2{,}143$ test compositions on which the baseline is correct, and they concentrate in fine-grained attributes.