Beyond Global Alignment: Structured Compositional Reasoning for Vision-Language Models

Zhoujun Ye[1] **Yiwei Fu**[1,†] **Qiyun Huang**[2] **Jie Yang**[3] **Dongjie Wang**[4] **Xiao Luo**[5]
[1]East China Normal University [2]Peking University [3]University of Wollongong
[4]Northeast Normal University [5]University of Wisconsin–Madison
zhoujunye0712@gmail.com, ywfu@sfs.ecnu.edu.cn [†]**Corresponding author**

Abstract

Despite remarkable progress in image-text understanding, vision-language models (VLMs) still struggle with compositional reasoning. In particular, they often fail to resolve relation direction and bind attributes to the correct objects, leading to similar representations for semantically different image-text pairs. This problem mainly stems from the reliance on global image-text alignment, which captures coarse correspondence but overlooks fine-grained compositional structures. To address these limitations, we propose an evidence-aware framework, termed Directional Relation Bucketing with Binding Localization (DELTA), to improve compositional understanding in VLMs. The key idea of our DELTA is to improve compositionality of VLMs from two complementary perspectives: *direction modeling* and *binding localization*. More specifically, we first leverage a learnable gate to model the cumulative contextual distance between anchor terms, thereby assigning relation concepts to direction-aware discrete buckets. To enrich textual representations with fine-grained relational semantics, we calibrate global semantic representations by incorporating intermediate-layer hidden states. Furthermore, DELTA leverages textual cues to ground visual evidence for object attributes, pulling image patches closer to their matched attribute descriptions while pushing them away from mismatched attribute descriptions. Extensive experiments on four compositional datasets demonstrate the effectiveness of the proposed DELTA. Code is available at https://github.com/Zhoujun-Ye/DELTA.

1 Introduction

Vision-language models (VLMs) have achieved significant progress in image-text understanding [25, 72, 77, 36, 15, 17], establishing themselves as a foundational paradigm for a wide range of downstream multimodal tasks, including image classification [74, 1, 23, 14], image-text retrieval [5, 71, 45], and image captioning [6, 58, 67]. Among various VLMs, CLIP is regarded as one of the pioneering frameworks, aligning multimodal representations through pre-training on large-scale image-text pairs [52]. Despite their strong performance on a wide range of established vision-language benchmarks, VLMs such as CLIP still face challenges in accurately capturing and reasoning about fine-grained compositional semantics [27, 19].

More specifically, when two sentences describe the same objects but differ in attributes or relations, VLMs generally produce highly similar textual representations despite their compositional differences [70, 43, 29, 4, 21]. One possible explanation is their reliance on multimodal alignment of pooled representations, which may overlook fine-grained compositional semantics [11, 38, 32, 8, 64]. Several recent studies have sought to address this issue by improving the compositional reasoning capabilities of VLMs [24]. For instance, NegCLIP [68] introduces composition-aware hard negatives through textual perturbations to help distinguish subtle semantic differences. Nevertheless, its

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

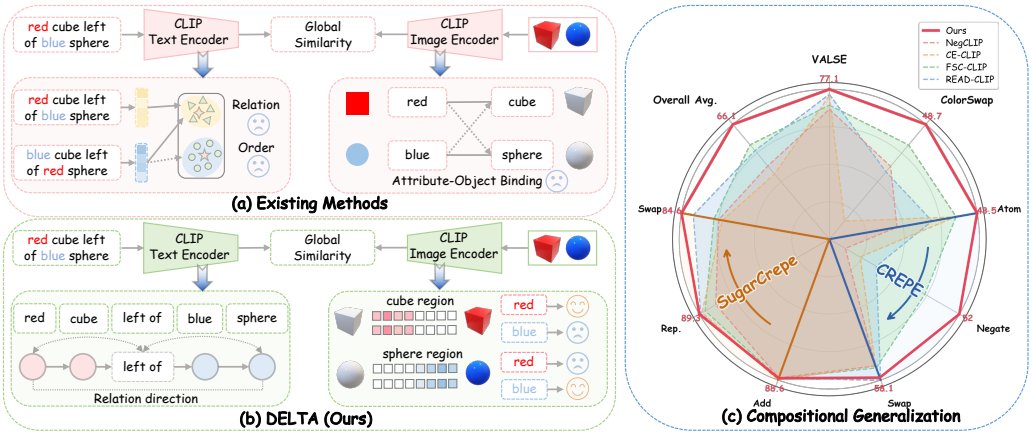


Figure 1: A comparison of previous approaches in (a) and our DELTA in (b) for compositional reasoning in VLMs. Previous approaches struggle to explicitly model relation directions and attribute-object binding, while our DELTA addresses these challenges through two components: *direction-aware bucketing* and *object-conditioned binding localization*. As shown in (c), DELTA achieves superior performance over competitive baselines across diverse benchmarks.

learning objective remains centered on global image-text similarity rather than explicit compositional modeling. Other approaches, such as CE-CLIP [73] and FSC-CLIP [47], employ refined contrastive objectives to learn representations that better preserve fine-grained semantics.

As shown in Figure 1, most existing approaches primarily focus on global alignment using contrastive or reconstruction-based objectives, leaving compositional relations without explicit and interpretable modeling. Generally, relation directions are implicitly encoded in pooled representations, making relation semantics difficult to calibrate. This raises a crucial question: ❶ ***Can we design an explicit module to better distinguish similar representations that encode different relations?*** Furthermore, we find that visual grounding evidence is often insufficient, causing attribute-object binding to still rely heavily on global representations. This leads to a second question: ❷ ***How can we localize binding evidence within images to improve attribute-object binding?***

To address these limitations, we propose a novel framework named Directional Relation Bucketing with Binding Localization (DELTA) to improve compositional understanding in vision-language models. To represent relations between objects, DELTA leverages a gated relation bucketing module to map continuous direction-aware relations into discrete learnable buckets. Based on the bucket vocabulary, DELTA interpolates relation-aware key and value biases, which inject relational information into the attention module of the text encoder. Furthermore, we calibrate the final representations using hidden states from intermediate layers to better preserve relation semantics. We optimize relation modeling by incorporating a relation-aware ranking loss. To model attribute-object binding, DELTA localizes object-conditioned visual evidence from image patches. Conditioned on the object prior, DELTA retrieves evidence-aware patches through attribute augmentations, pulling patches closer to matched attribute descriptions while pushing them away from mismatched attribute-augmented descriptions. We conduct comprehensive experiments on four benchmarks to demonstrate the effectiveness of DELTA, achieving an average improvement of 5.1 percentage points over the best baseline.

Overall, the main contributions of our DELTA are summarized as follows:

❶ ***New Perspectives.*** We investigate compositional reasoning in VLMs from two complementary perspectives, i.e., direction modeling and binding localization, enabling a comprehensive and effective framework for modeling fine-grained compositionality beyond global image-text alignment.

❷ ***Novel Framework with Theoretical Analysis.*** We propose DELTA, an evidence-aware framework that combines relation bucketing with object-conditioned binding localization to improve compositional reasoning in VLMs. We further derive a binding-margin guarantee, showing that object-conditioned localization can suppress off-object distractors while preserving discriminative attribute-object evidence under the stated assumptions.

❸ ***Empirical Validation.*** We conduct extensive experiments across four benchmarks, showing that DELTA is: ***(i) accuracy-improving***, outperforming the strongest baseline by 5.1 pp on average, and ***(ii) evidence-grounded***, offering explicit visual evidence for attribute-object binding.

2 Related Work

Vision-Language Models. Vision-language models (VLMs) have become central to multimodal representation learning, with many approaches mapping visual and textual features into a shared embedding space [72, 36, 66, 2, 79, 22]. As an early representative framework, CLIP [52] employs contrastive learning on large-scale image-text pairs to establish cross-modal alignment. Subsequent architectures, including the BLIP family [40, 39] and SigLIP 2 [61], have further advanced vision-language representation learning. However, many existing VLMs still rely on pooled representations for global image-text alignment [65, 76, 46, 69, 7]. Although these representations provide effective global semantic alignment, they remain less sensitive to the compositional structure of visual scenes. For instance, models may correctly recognize the objects present in an image but fail to distinguish their relationships or associate specific attributes with the corresponding objects [75, 53, 60, 28]. Such failures have motivated growing interest in compositional reasoning for VLMs [13, 62].

Compositional Reasoning in VLMs. Compositional reasoning [59, 9, 41, 31, 18] concerns the ability of VLMs to distinguish fine-grained semantic differences, particularly in relation direction and attribute-object binding. Existing methods [41, 70, 21, 26, 37] often rely on rule-based perturbations to construct hard negatives, as in NegCLIP [68] and TripletCLIP [49]. Beyond hard negative construction, READ-CLIP [35] refines the text encoder through caption reconstruction and auxiliary alignment losses, while LABCLIP [33] introduces a lightweight projector to mitigate BoW-like behavior in attribute-object binding. Despite these advances, existing approaches still rely heavily on global image-text alignment, leaving compositional structures largely implicit. This motivates our `DELTA` framework, which explicitly models relation direction and attribute-object correspondences through direction-aware relation bucketing and object-conditioned binding localization.

3 The Proposed `DELTA`

Problem Formulation. Given a training dataset of image-text pairs $\mathcal{D} = \{(I_i, T_i)\}_{i=1}^{N}$, the representative VLM backbone CLIP [52] employs a visual encoder ϕ_I and a text encoder ϕ_T to project images and texts into a shared representation space, where multimodal features are aligned through self-supervised contrastive learning on large-scale image-text corpora. The CLIP objective is formulated as a symmetric image-text alignment loss:

$$\mathcal{L}_{clip} = \frac{1}{2}(\mathcal{L}_{I2T} + \mathcal{L}_{T2I}), \tag{1}$$

where the image-to-text and text-to-image losses are defined as

$$\mathcal{L}_{I2T} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(s(I_i,T_i)/\tau)}{\sum_{j=1}^{N}\exp(s(I_i,T_j)/\tau)}, \quad \mathcal{L}_{T2I} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(s(I_i,T_i)/\tau)}{\sum_{j=1}^{N}\exp(s(I_j,T_i)/\tau)}, \tag{2}$$

in which $s(I_i, T_j) = \cos(\phi_I(I_i), \phi_T(T_j))$ denotes the cosine similarity between the i-th image and j-th text in the embedding space, and τ denotes the logit temperature inherited from the CLIP backbone. Despite its impressive performance, CLIP still struggles to capture fine-grained linguistic structures in texts, which often leads to failures in compositional reasoning. Let $\{T_i^-\}_{i=1}^{N} \subseteq \mathcal{N}$ denote the set of compositional negative captions. Our goal is to inject compositional knowledge into VLMs, enabling them to assign higher compatibility to matched image-text pairs than to their compositional negative counterparts. Specifically, this objective can be formulated as follows:

$$D(I_i, T_i) \leq D(I_i, T_i^-), \quad T_i^- \in \mathcal{N}, \quad i = 1, 2, \cdots, N, \tag{3}$$

where D denotes the distance and T_i^- denotes the compositional negative text.

3.1 Framework Overview

In this work, we address compositional understanding from the dual perspectives of textual and visual modalities, thereby mitigating misalignment in relation direction and attribute-object binding. Specifically, `DELTA` first estimates the gated contextual distance between token pairs, and then leverages learnable bucket embeddings to calibrate the attention weights, thereby enabling direction-aware relation modeling. Subsequently, we refine the final textual representations by incorporating hidden states from intermediate layers. On the visual side, `DELTA` localizes attribute-binding evidence within image patches, pulling patches closer to matching attribute descriptions while pushing them away from incorrect attribute-based augmentations. An illustration of `DELTA` is shown in Figure 2.

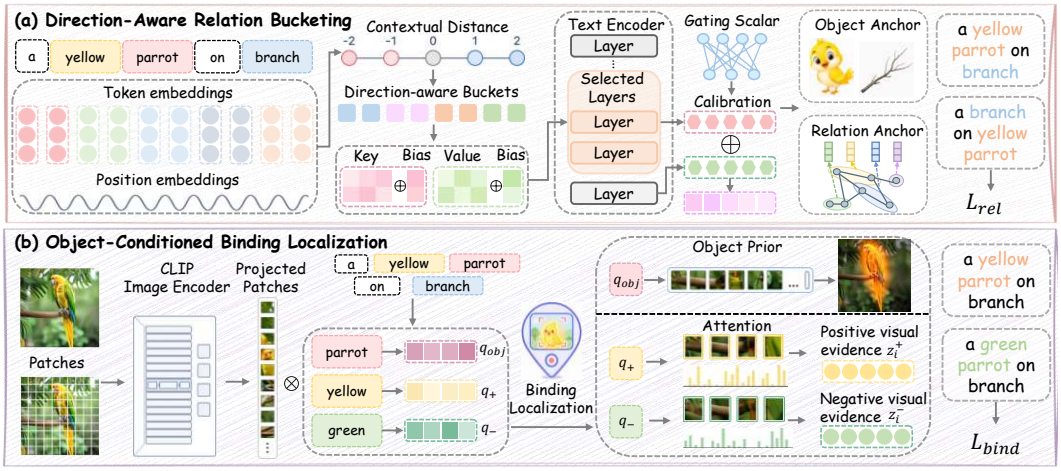


Figure 2: Overview of DELTA. DELTA addresses compositional misalignment from both textual and visual perspectives: (a) the text branch estimates gated contextual distances between token pairs and combines bucket embeddings with intermediate-layer hidden states to inject direction-aware relational cues; (b) the vision branch uses object and attribute cues to retrieve patch-level evidence, grounding matched attributes while suppressing attribute-based negatives.

3.2 Direction-Aware Relation Bucketing

Previous studies [68, 33] have shown that CLIP tends to behave like a bag-of-words model, exhibiting limited capability in capturing relational structures among objects. For instance, two sentences with the same object tokens but reversed relational directions, such as "a dog chasing a cat" and "a cat chasing a dog", may still be mapped to similar textual embeddings. Building on contextual position encoding [16, 55], we introduce a gated contextual distance with direction information to model relative relationships between anchor tokens. Furthermore, as relational semantics are often better preserved in intermediate layers [63, 44], we refine the final textual representation by incorporating relation-calibrated hidden states from these layers. Overall, this design enhances relational reasoning on the textual side by explicitly encoding relation direction and contextual distance.

Gated Relation Bucketing. To model pairwise token relations, we construct a gated contextual distance with directional information inspired by relative position encoding. Specifically, for the l-th layer, let the hidden states be denoted as $H^l = (h_1^l, h_2^l, \cdots, h_T^l)$ of length T. For the r-th attention head, the standard query, key, and value vectors are computed as $q_t^{l,r} = W_q^{l,r} h_t^l$, $k_u^{l,r} = W_k^{l,r} h_u^l$, and $v_u^{l,r} = W_v^{l,r} h_u^l$. The gated weight between h_t^l and h_u^l is defined as

$$g_{t,u}^{l,r} = \sigma\left(\frac{(W_{gq}^{l,r} h_t^l)^\top (W_{gk}^{l,r} h_u^l)}{\sqrt{d_h}}\right), \tag{4}$$

where $W_{gq}^{l,r}$ and $W_{gk}^{l,r}$ denote the gating query and key projection matrices for the r-th head in the l-th layer, d_h is the head dimension, and $\sigma(\cdot)$ represents the sigmoid activation function. Subsequently, we employ these gated weights to measure the contextual distance between the t-th and u-th tokens. Therefore, the gated contextual distance is formulated as follows:

$$d_{t,u}^{l,r} = \left| \sum_{m \le t} g_{t,m}^{l,r} - \sum_{m \le u} g_{t,m}^{l,r} \right| = \sum_{u < m \le t} g_{t,m}^{l,r}, \quad u \le t. \tag{5}$$

In this way, we measure pairwise token relations via a relative distance. Under causal attention, the gated contextual distance captures relative token distances, while $s_{t,u} = \mathbb{I}(t > u)$ distinguishes preceding tokens from the current token. However, it is difficult to directly learn embeddings for such continuous values. To address this issue, we introduce a distance vocabulary with B buckets and discretize the continuous distance into bucketized embeddings. Specifically, we map $d_{t,u}^{l,r}$ to a continuous coordinate within the finite bucket range:

$$\rho_{t,u}^{l,r} = \frac{B-1}{D_{\max}} \min(d_{t,u}^{l,r}, D_{\max}), \tag{6}$$

where B is the number of distance buckets and $D_{\max}$ denotes the maximum contextual distance. For this finite bucket vocabulary, we learn the direction-aware key-side and value-side relation embeddings $e^{k,l,r}_{s,c}, e^{v,l,r}_{s,c} \in \mathbb{R}^{d_h}$ to calibrate the attention weights, where $s \in \{0,1\}$ denotes the direction and $c \in \{0,\dots,B-1\}$ corresponds to the bucket index. Based on these embeddings, we compute an interpolated key-side bias for each token pair (t,u):

$$\widetilde{e}^{k,l,r}_{t,u} = \left(1-\lambda^{l,r}_{t,u}\right)(q^{l,r}_t)^\top e^{k,l,r}_{s_{t,u},c^{-,l,r}_{t,u}} + \lambda^{l,r}_{t,u}(q^{l,r}_t)^\top e^{k,l,r}_{s_{t,u},c^{+,l,r}_{t,u}}, \tag{7}$$

where $c^{-,l,r}_{t,u} = \left\lfloor \rho^{l,r}_{t,u} \right\rfloor$ and $c^{+,l,r}_{t,u} = \min\left(c^{-,l,r}_{t,u}+1, B-1\right)$ denote the lower and upper bucket indices, respectively. $\lambda^{l,r}_{t,u} = \rho^{l,r}_{t,u} - c^{-,l,r}_{t,u}$. The key-side embeddings serve as additive biases to the attention logits, which are defined as

$$\alpha^{l,r}_{t,u} = \mathrm{softmax}_{u\le t}\left(\frac{(q^{l,r}_t)^\top k^{l,r}_u}{\sqrt{d_h}} + \widetilde{e}^{k,l,r}_{t,u}\right), \tag{8}$$

where $q^{l,r}_t$ and $k^{l,r}_u$ denote the query and key vectors for tokens t and u, respectively. Similarly, the interpolated value-side bias is

$$\widetilde{e}^{v,l,r}_{t,u} = \left(1-\lambda^{l,r}_{t,u}\right) e^{v,l,r}_{s_{t,u},c^{-,l,r}_{t,u}} + \lambda^{l,r}_{t,u} e^{v,l,r}_{s_{t,u},c^{+,l,r}_{t,u}}. \tag{9}$$

By incorporating the gated relation-aware attention, the final output is computed as

$$o^{l,r}_t = \sum_{u=1}^{t} \alpha^{l,r}_{t,u}\left(v^{l,r}_u + \widetilde{e}^{v,l,r}_{t,u}\right). \tag{10}$$

This design yields a gated contextual bucketing mechanism for encoding directional relations.

Contextual Semantics Calibration. Since relational semantics are often better preserved in intermediate layers than in the final representation, we extract the [EOS] token features $z^{i,l}_{eos}$ and z^{i}_{eos} for the i-th sample from the l-th intermediate layer and the final layer, respectively. The calibrated hidden states are computed as

$$\widehat{H}^l_i = (1-g^l_i)\cdot H^{\mathrm{top}}_i + g^l_i \cdot H^l_i, \tag{11}$$

where H^l_i and H^{top}_i denote the hidden states from the l-th intermediate layer and the final layer, respectively, and $g^l_i = \sigma\left(\mathrm{MLP}([z^{i,l}_{eos} \oplus z^i_{eos}])\right)$ is an adaptive gating scalar that controls the contribution of the intermediate-layer representation. Subsequently, a cross-attention readout summarizes $\widehat{H}^l_i$ as $\bar{z}^i_{eos} = \mathrm{Readout}(z^i_{eos}, \widehat{H}^l_i)$. To better distinguish captions with different relational structures, we introduce a relation-aware ranking loss. For the i-th sample, let T_i denote the positive relational text and T^{r-}_i denote its relational negative. After the calibrated readout, we obtain their final text embeddings as $t^+_i = \mathrm{Proj}_T(\bar{z}^{i,+}_{eos})$ and $t^{r-}_i = \mathrm{Proj}_T(\bar{z}^{i,r-}_{eos})$, where $\mathrm{Proj}_T(\cdot)$ denotes the text projection. The relation-aware ranking loss is defined as

$$\mathcal{L}_{rel} = -\frac{1}{\sum_i m_i}\sum_i m_i \log\sigma\left((s^+_i - s^{r-}_i)/\tau_{rel}\right), \tag{12}$$

where $s^+_i = \cos(\phi_I(I_i), t^+_i)$, $s^{r-}_i = \cos(\phi_I(I_i), t^{r-}_i)$, and $m_i \in \{0,1\}$ indicates whether T^{r-}_i is a valid relational negative. Here, $\tau_{rel} > 0$, and we set $\mathcal{L}_{rel} = 0$ when $\sum_i m_i = 0$.

3.3 Object-Conditioned Binding Localization

To enrich attribute-level semantics of VLMs, we ground attribute-object binding in images under textual guidance [3]. `DELTA` extracts three types of spans from the i-th caption: the object span, a positive attribute span, and a negative attribute span. The representations o_i, a^+_i, and a^-_i are obtained via masked pooling over token-level features, followed by a binding-specific text projector. Here, a^+_i corresponds to the ground-truth attribute for the image, while a^-_i does not match the target object. Moreover, the patch-level embeddings of the i-th image are represented as $\{p_{i,1}, p_{i,2}, \dots, p_{i,L}\}$, which are then projected into the same binding space as the text spans via a binding-specific visual projector. Subsequently, we model attribute-object binding as a process of retrieving attribute-specific

visual evidence conditioned on the object. Prior to attribute retrieval, we first localize object-related regions by matching the object representation with image patches in the binding space:

$$\pi_{i\ell} = \mathrm{softmax}_\ell \left(\frac{(W_q^{obj} o_i)^\top W_k p_{i\ell}}{\tau_{\mathrm{loc}}} \right), \tag{13}$$

where τ_{loc} denotes the local retrieval temperature, and we fix it to 0.07 following the CLIP-style temperature scale [52]. This yields an object prior that localizes regions where the object is likely to appear. Given the object prior, we retrieve attribute-specific visual evidence as

$$\alpha_{i\ell}^{\pm} = \mathrm{softmax}_\ell \left(\frac{(W_q^{att} a_i^{\pm})^\top W_k p_{i\ell}}{\tau_{\mathrm{loc}}} + \gamma \log \pi_{i\ell} \right), \tag{14}$$

where a_i^+ and a_i^- denote positive and negative attribute representations, respectively, and γ is a scaling hyperparameter. Intuitively, the object-prior term $\log \pi_{i\ell}$ injects object-conditioned spatial information into attribute retrieval, encouraging attention weights to favor attribute-relevant, object-consistent patches. Furthermore, the attribute-augmented binding supervision is formulated as

$$\mathcal{L}_{bind} = -\frac{1}{\sum_i n_i} \sum_i n_i \log \sigma \left(((a_i^+)^\top W_b z_i^+ - (a_i^-)^\top W_b z_i^-)/\tau_{bind} \right), \tag{15}$$

where W_b is a learnable binding matrix, and n_i indicates whether a_i^- is a valid binding negative for the i-th sample. We set $\mathcal{L}_{bind} = 0$ when $\sum_i n_i = 0$. The resulting visual evidence z_i^+ and z_i^- are computed as

$$z_i^+ = \sum_{\ell=1}^{L} \alpha_{i\ell}^+ p_{i\ell} \quad \text{and} \quad z_i^- = \sum_{\ell=1}^{L} \alpha_{i\ell}^- p_{i\ell}. \tag{16}$$

This design enables the grounding of object-conditioned attributes under textual guidance, facilitating the learning of visual binding relationships.

3.4 Model Optimization

Hard Negative Construction. Prior studies [68, 57, 35] have demonstrated that incorporating hard negative captions can improve CLIP's compositional reasoning ability. These captions preserve much of the lexical content of their positive counterparts while introducing subtle semantic changes, making them particularly challenging to distinguish. Following this paradigm, we construct four negative captions online for each positive caption through rule-based perturbations, including attribute replacement, object replacement, relation replacement, and attribute-binding swaps.

Contrastive Optimization. Motivated by this observation, we adopt a variant of the standard contrastive loss in Eq. 1 with multiple types of compositional negatives. For each positive text T_j, let $\{\widetilde{T}_j^{(m)}\}_{m=1}^{M_j}$ denote the set of constructed negative captions. By incorporating these negatives into the denominator, the image-to-text loss in Eq. 2 is modified as

$$\widetilde{\mathcal{L}}_{I2T} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(s(I_i, T_i)/\tau)}{\sum_{j=1}^{N} \left[\exp(s(I_i, T_j)/\tau) + \sum_{m=1}^{M_j} \exp(s(I_i, \widetilde{T}_j^{(m)})/\tau) \right]}. \tag{17}$$

Accordingly, the contrastive loss used in DELTA is $\mathcal{L}_{contrastive} = \frac{1}{2}(\widetilde{\mathcal{L}}_{I2T} + \mathcal{L}_{T2I})$. Overall, the fine-tuning objective of DELTA is defined as

$$\mathcal{L} = \mathcal{L}_{contrastive} + \mathcal{L}_{rel} + \mathcal{L}_{bind}. \tag{18}$$

The overall optimization procedure is summarized in Algorithm 1.

3.5 Theoretical Analysis

We provide a theoretical analysis for object-conditioned binding localization. The result below establishes a conditional guarantee under two assumptions: the object prior in Eq. 13 assigns consistently higher weights to target-object patches than to off-object patches, and matched attributes

Algorithm 1 DELTA: Directional Relation Bucketing with Binding Localization

Require: Training dataset $\mathcal{D} = \{(I_i, T_i)\}_{i=1}^{N}$, pretrained CLIP encoders ϕ_I, ϕ_T.

for each mini-batch $\{(I_i, T_i)\}_{i=1}^{B}$ **do**
  Encode the texts to obtain hidden states H^l in the l-th layer;
  Calculate the contextual distance $d_{t,u}^{l,r}$ using Eq. 5; // Obtain relative distance
  Interpolate relation biases $\widetilde{e}_{t,u}^{k,l,r}, \widetilde{e}_{t,u}^{v,l,r}$ using Eq. 7 and Eq. 9, respectively;
  Calibrate attention logits with $\widetilde{e}_{t,u}^{k,l,r}$ using Eq. 8 and compute outputs by Eq. 10;
  Obtain the calibrated hidden states $\widehat{H}_i^l$ using Eq. 11; // Calibrate textual embeddings
  Compute attribute attention weights $\alpha_{i\ell}^{+}$ and $\alpha_{i\ell}^{-}$ by Eq. 14;
  Obtain visual evidence z_i^{+} and z_i^{-} using Eq. 16; // Object-conditioned visual evidence
  Optimize all parameters using Eq. 18;
end for
return the fine-tuned DELTA model; // Optimized model for compositional alignment

are more distinguishable from mismatched ones on the target object. Under these conditions, increasing γ tightens an upper bound on off-object attention and yields the binding-margin lower bound below. Bounding the positive and negative retrieval scores separately is essential because their attention distributions need not coincide. The resulting margin bound quantifies how uniformly separable target-object evidence can be weakened by off-object attention, rather than asserting that localization creates separability.

Theorem 3.1 (Object-conditioned localization preserves a binding margin)**.** *Fix the model parameters and consider one training sample with a valid binding negative. Let $\mathcal{O}_i \subseteq \{1, \dots, L\}$ be the target-object patch set with $1 \le |\mathcal{O}_i| = m \le L$. Assume $\tau_{\text{loc}} > 0$, $R \ge 0$, and $M > 0$, with bounded attention logits and patch-level binding scores: $\forall \ell$, $\left|(W_q^{\text{att}} a_i^{\pm})^\top W_k p_{i\ell}\right| \le R$, $\left|(a_i^{\pm})^\top W_b p_{i\ell}\right| \le M$. Further assume that the object prior separates object patches from off-object patches, i.e.,*

$$\pi_{i\ell} \ge c_{\text{in}} \quad (\ell \in \mathcal{O}_i), \qquad \pi_{i\ell} \le c_{\text{out}} \quad (\ell \notin \mathcal{O}_i),$$

with $0 < c_{\text{out}} < c_{\text{in}}$ and $\frac{c_{\text{out}}}{c_{\text{in}}} = \rho \in (0, 1)$, and assume that the object patches are attribute-separable:

$$(a_i^{+})^\top W_b p_{i\ell} \ge \mu_+, \qquad (a_i^{-})^\top W_b p_{i\ell} \le \mu_-, \qquad \forall \ell \in \mathcal{O}_i,$$

where $\Delta = \mu_+ - \mu_- > 0$. Define the constant $C_0 = 4M \frac{L-m}{m} \exp\left(\frac{2R}{\tau_{\text{loc}}}\right)$. Then for $\gamma \ge 0$, the binding margin in Eq. 15 satisfies

$$(a_i^{+})^\top W_b z_i^{+} - (a_i^{-})^\top W_b z_i^{-} \ge \Delta - C_0 \rho^{\gamma}.$$

Remark 3.1. *The target-object patch set is used only for analysis; training does not require ground-truth object annotations. The assumptions require uniform prior separation between target and off-object patches, along with uniform cross-patch attribute separability on the target object, rather than a binary mask. With model parameters fixed, $C_0 \rho^{\gamma}$ decreases as γ grows, yielding a tighter lower bound but not necessarily a monotonic improvement in the actual margin. The guarantee weakens as ρ approaches one or Δ decreases. A larger off-object-to-object patch ratio or logit range relative to τ_{loc} increases C_0, making the bound less informative for a given γ. When $m = L$, the distractor term vanishes.*

Corollary 3.2 (Positive margin under sufficient object conditioning)**.** *Under the assumptions of Theorem 3.1 and $\tau_{\text{bind}} > 0$, there exists a finite $\gamma_0 \ge 0$ (proof deferred to Appendix C) such that, for any $\gamma \ge \gamma_0$, the single-sample binding loss in Eq. 15 is upper bounded by*

$$\ell_{\text{bind}} \le \log\left(1 + \exp\left(-\frac{\Delta}{2\tau_{\text{bind}}}\right)\right).$$

Remark 3.2. *The corollary bounds the auxiliary binding loss conditionally; it does not guarantee that training learns the assumed priors or separability, or that a low loss certifies better grounding or test accuracy. For example, scaling W_b can reduce the loss for a positive margin without changing retrieval attention; loss reduction alone is therefore not a localization metric. The bound's informativeness at the operating point requires empirical validation.*

Table 1: Main results on multiple benchmarks. Best results are in bold, second-best results are underlined, and arrows indicate the change of each method relative to CLIP in percentage points (pp).

Models	VALSE	ColorSwap	CREPE				SugarCrepe				Avg.
	Avg.	ITT	Atom	Negate	Swap	Avg.	Add	Rep.	Swap	Avg.	
CLIP (ICML'21)	67.4	35.7	18.9	35.3	17.3	23.8	73.3	80.1	62.7	72.0	49.7
NegCLIP (ICLR'23)	73.7$^{\uparrow 6.3}$	35.7$^{0.0}$	32.1$^{\uparrow 13.2}$	16.6$^{\downarrow 18.7}$	42.8$^{\uparrow 25.5}$	30.5$^{\uparrow 6.7}$	87.7$^{\uparrow 14.4}$	85.6$^{\uparrow 5.5}$	76.7$^{\uparrow 14.0}$	83.3$^{\uparrow 11.3}$	55.8$^{\uparrow 6.1}$
GNM-CLIP (WACV'24)	70.7$^{\uparrow 3.3}$	13.0$^{\downarrow 22.7}$	21.7$^{\uparrow 2.8}$	13.3$^{\downarrow 22.0}$	17.3$^{0.0}$	17.4$^{\downarrow 6.4}$	84.2$^{\uparrow 10.9}$	82.0$^{\uparrow 1.9}$	65.6$^{\uparrow 2.9}$	77.2$^{\uparrow 5.2}$	44.6$^{\downarrow 5.1}$
CE-CLIP (CVPR'24)	76.0$^{\uparrow 8.6}$	18.7$^{\downarrow 17.0}$	40.4$^{\uparrow 21.5}$	21.2$^{\downarrow 14.1}$	42.7$^{\uparrow 25.4}$	34.8$^{\uparrow 11.0}$	**92.1**$^{\uparrow 18.8}$	88.7$^{\uparrow 8.6}$	75.9$^{\uparrow 13.2}$	85.6$^{\uparrow 13.6}$	53.8$^{\uparrow 4.1}$
FSC-CLIP (EMNLP'24)	74.4$^{\uparrow 7.0}$	<u>42.3</u>$^{\uparrow 6.6}$	<u>40.5</u>$^{\uparrow 21.6}$	<u>41.3</u>$^{\uparrow 6.0}$	45.5$^{\uparrow 28.2}$	<u>42.4</u>$^{\uparrow 18.6}$	88.5$^{\uparrow 15.2}$	88.0$^{\uparrow 7.9}$	77.7$^{\uparrow 15.0}$	84.7$^{\uparrow 12.7}$	<u>61.0</u>$^{\uparrow 11.3}$
CLIC (NeurIPS'25)	71.1$^{\uparrow 3.7}$	39.7$^{\uparrow 4.0}$	28.0$^{\uparrow 9.1}$	26.8$^{\downarrow 8.5}$	29.9$^{\uparrow 12.6}$	28.2$^{\uparrow 4.4}$	84.3$^{\uparrow 11.0}$	83.4$^{\uparrow 3.3}$	69.2$^{\uparrow 6.5}$	79.0$^{\uparrow 7.0}$	54.5$^{\uparrow 4.8}$
READ-CLIP (NeurIPS'25)	<u>76.2</u>$^{\uparrow 8.8}$	34.0$^{\downarrow 1.7}$	37.1$^{\uparrow 18.2}$	26.2$^{\downarrow 9.1}$	**61.2**$^{\uparrow 43.9}$	41.5$^{\uparrow 17.7}$	<u>89.0</u>$^{\uparrow 15.7}$	<u>88.8</u>$^{\uparrow 8.7}$	<u>82.2</u>$^{\uparrow 19.5}$	<u>86.7</u>$^{\uparrow 14.7}$	59.6$^{\uparrow 9.9}$
LABCLIP (ICLR'26)	71.3$^{\uparrow 3.9}$	28.0$^{\downarrow 7.7}$	30.5$^{\uparrow 11.6}$	22.6$^{\downarrow 12.7}$	33.9$^{\uparrow 16.6}$	29.0$^{\uparrow 5.2}$	81.0$^{\uparrow 7.7}$	83.3$^{\uparrow 3.2}$	73.8$^{\uparrow 11.1}$	79.3$^{\uparrow 7.3}$	51.9$^{\uparrow 2.2}$
DELTA	**77.1**$^{\uparrow 9.7}$	**48.7**$^{\uparrow 13.0}$	**43.5**$^{\uparrow 24.6}$	**52.0**$^{\uparrow 16.7}$	<u>58.1</u>$^{\uparrow 40.8}$	**51.2**$^{\uparrow 27.4}$	88.6$^{\uparrow 15.3}$	**89.3**$^{\uparrow 9.2}$	**84.6**$^{\uparrow 21.9}$	**87.5**$^{\uparrow 15.5}$	**66.1**$^{\uparrow 16.4}$

4 Experiments

4.1 Experimental Setup

Datasets and Evaluation. Following prior compositional reasoning works [68, 57, 47, 73, 35], we fine-tune our model on a 100K MS-COCO subset [42] under the Karpathy split [30]. We conduct zero-shot evaluation on VALSE [48], ColorSwap [10], CREPE [46], and SugarCrepe [20], where each task requires selecting the positive caption for a given image. Details are provided in Appendix A.1.

Compared Baselines. To comprehensively evaluate our DELTA, we compare against the original CLIP model and recent state-of-the-art fine-tuning baselines for compositional reasoning in VLMs, including NegCLIP [68], GNM-CLIP [54], CE-CLIP [73], FSC-CLIP [47], CLIC [50], READ-CLIP [35], and LABCLIP [33]. More details are provided in Appendix A.2.

Implementation Details. Following prior works [68, 35], we initialize DELTA from the OpenAI CLIP ViT-B/32 backbone and fine-tune all parameters on the COCO Karpathy subset for 5 epochs. We use AdamW with a batch size of 256 and a learning rate of 1×10^{-5}. For our auxiliary objectives, we set the relation-aware ranking temperature to $\tau_{\text{rel}} = 1/3$ and the binding-loss temperature to $\tau_{\text{bind}} = 1/5$. All experiments are conducted using a single NVIDIA A100 80GB GPU.

4.2 Main Results

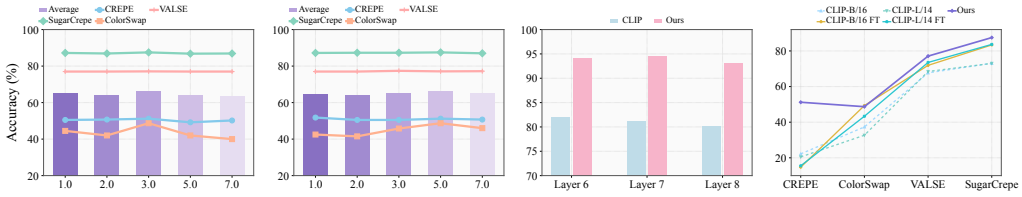
We report the main results on four compositional benchmarks in Table 1. According to the results, we make three observations (Obs.) as follows: **Obs. ❶** DELTA *consistently improves the original CLIP across diverse compositional reasoning scenarios.* Specifically, DELTA improves the average score over the CLIP baseline by 16.4 pp, with consistent gains on VALSE (+9.7 pp), ColorSwap (+13.0 pp), CREPE (+27.4 pp), and SugarCrepe (+15.5 pp), demonstrating the effectiveness of object-conditioned binding localization and relation-aware textual calibration for compositional image-text matching. **Obs. ❷** DELTA *achieves state-of-the-art performance among recent fine-tuning methods.* Compared with the competing baselines, DELTA obtains the highest overall average score of 66.1%, surpassing FSC-CLIP and READ-CLIP by 5.1 and 6.5 pp, respectively. Moreover, DELTA achieves the best benchmark-level average on all four datasets, reflecting a more balanced compositional reasoning ability. **Obs. ❸** DELTA *provides more effective binding-aware reasoning.* Compared with LABCLIP, which focuses on attribute-object binding failures in CLIP, DELTA improves the overall average by 14.2 pp, with notable gains on ColorSwap (+20.7 pp) and CREPE (+22.2 pp). These results highlight the benefit of localizing object-conditioned visual evidence for binding-aware reasoning.

4.3 Ablation Study

We evaluate the contributions of components in DELTA using five variants, denoted as V_1–V_5. V_1 removes all proposed components, V_2 adds object-conditioned binding localization, and V_3 further introduces the relation-aware ranking loss. To examine the effectiveness of discrete relation bucketing, V_4 removes the relation bias, while V_5 replaces the discrete buckets in the full DELTA with a learnable continuous signed-distance bias, keeping other

Table 2: Ablation study of component variants and relation bias designs.

Variants	VALSE	Color Swap	CREPE	Sugar Crepe	Avg.
V_1	76.6$^{\downarrow 0.5}$	39.6$^{\downarrow 9.1}$	42.9$^{\downarrow 8.3}$	86.5$^{\downarrow 1.0}$	61.4$^{\downarrow 4.7}$
V_2	76.4$^{\downarrow 0.7}$	43.3$^{\downarrow 5.4}$	46.2$^{\downarrow 5.0}$	87.5$^{0.0}$	63.4$^{\downarrow 2.7}$
V_3	77.0$^{\downarrow 0.1}$	44.0$^{\downarrow 4.7}$	49.4$^{\downarrow 1.8}$	86.6$^{\downarrow 0.9}$	64.3$^{\downarrow 1.8}$
V_4	76.4$^{\downarrow 0.7}$	44.3$^{\downarrow 4.4}$	49.5$^{\downarrow 1.7}$	86.1$^{\downarrow 1.4}$	64.1$^{\downarrow 2.0}$
V_5	76.1$^{\downarrow 1.0}$	45.7$^{\downarrow 3.0}$	48.8$^{\downarrow 2.4}$	**87.7**$^{\uparrow 0.2}$	64.6$^{\downarrow 1.5}$
DELTA	**77.1**	**48.7**	**51.2**	87.5	**66.1**



(a) Sensitivity to $1/\tau_{\text{rel}}$ **(b) Sensitivity to $1/\tau_{\text{bind}}$** **(c) Text-layer Probe** **(d) Backbone Capacity**

Figure 3: Further analysis of compositional reasoning. (a)–(b) Sensitivity to inverse temperatures $1/\tau_{\text{rel}}$ and $1/\tau_{\text{bind}}$; the plotted axes use the inverse-temperature convention. (c) Text-layer probe of relation-aware textual representations. (d) Comparison of DELTA with CLIP backbone scaling and standard fine-tuning on compositional performance.

settings unchanged. The results are shown in Table 2, from which we derive the following findings: **Obs. ❶** *Object-conditioned binding localization improves attribute-object binding.* Introducing this module helps localize object-related regions and retrieve attribute-specific visual evidence, enabling more accurate attribute-object correspondence. **Obs. ❷** *Relation-aware ranking loss enhances discrimination of structured compositional negatives.* This supervision encourages the model to rank matched captions above relationally corrupted ones, making image-text similarity more sensitive to compositional perturbations. **Obs. ❸** *Discrete relation bucketing enhances relational modeling.* Compared with both the variant without relation bias and the continuous bias variant, the full DELTA achieves stronger overall performance. We conjecture that, under limited fine-tuning data, discrete bucketing provides a more effective inductive bias, enabling more stable learning of direction-aware relational structures than continuous bias modeling. Overall, these findings demonstrate the complementary contributions of the components in enhancing fine-grained compositional reasoning.

4.4 Further Analysis

Sensitivity Analysis. To evaluate the robustness of DELTA, we analyze its sensitivity to the inverse relation-aware ranking temperature $1/\tau_{\text{rel}}$ and inverse binding-loss temperature $1/\tau_{\text{bind}}$. As shown in Figure 3(a)–(b), we find that **Obs. ❶** DELTA *maintains stable performance under different hyperparameter configurations.* Specifically, these inverse temperatures regulate the scaling of their corresponding auxiliary objectives, affecting the discrimination of relational semantics and attribute-object bindings. Across the tested ranges of $1/\tau_{\text{rel}}$ and $1/\tau_{\text{bind}}$, DELTA shows only slight variations in overall performance. These results show that relation-aware supervision and binding localization remain effective under different temperature settings, suggesting that DELTA is robust to temperature variations and does not require extensive hyperparameter tuning.

Text-layer Probing Analysis. To examine relation-aware text representations, we freeze the text encoder and train a lightweight linear probe on EOS representations from relation-calibrated layers 6–8 to distinguish original captions from subject–object-swapped counterfactuals. As these counterfactuals retain the same lexical content but alter semantic roles, the probe measures sensitivity to relational structures beyond word-level information. As shown in Figure 3(c), **Obs. ❷** DELTA *captures more discriminative relational semantics.* Compared with vanilla CLIP, DELTA consistently achieves higher probing accuracy across the examined layers, suggesting that direction-aware relation bucketing and contextual calibration improve the encoding of relation direction and semantic roles. The improvements observed across intermediate layers further suggest that textual calibration helps preserve relational information in the learned representations.

Capacity-Controlled Compositional Analysis. To investigate whether increasing CLIP backbone capacity alone can improve compositional reasoning, we compare DELTA with larger CLIP variants, including ViT-B/16 and ViT-L/14, along with their counterparts fine-tuned using the standard image-text contrastive objective. All models are evaluated under the same settings to assess how backbone capacity and conventional fine-tuning affect compositional understanding. The results are presented in Figure 3(d), which suggests that **Obs. ❸** DELTA *captures compositional cues that backbone scaling alone fails to recover.* While larger backbones and standard fine-tuning yield moderate improvements, their performance remains below that of DELTA, particularly in fine-grained relational understanding and attribute-object binding. This suggests that explicit compositional modeling provides benefits beyond simply increasing model capacity or applying conventional fine-tuning.

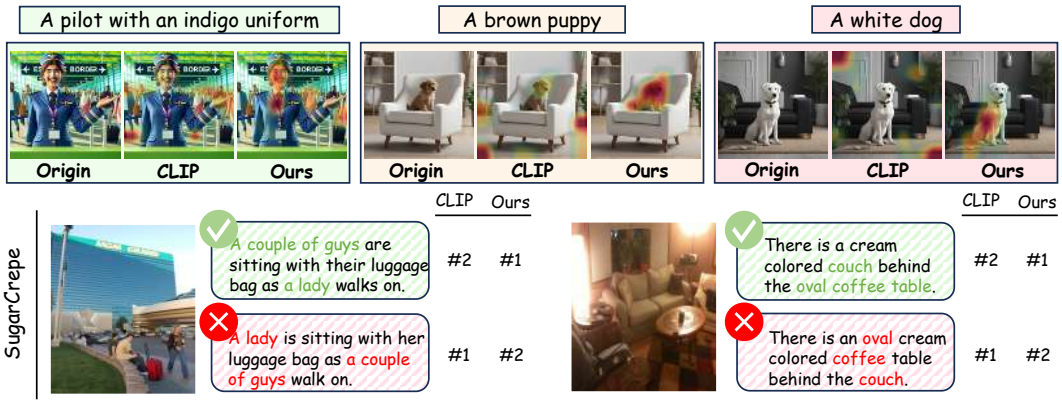


Figure 4: Qualitative case studies comparing CLIP and our proposed DELTA. Top: attribute-object binding maps. Bottom: image-caption matching examples with similarity-based ranks.

4.5 Qualitative Case Study

To examine compositional reasoning qualitatively, we compare CLIP and DELTA from visual and textual perspectives in Figure 4. These examples provide intuitive evidence for how DELTA improves compositional understanding beyond numerical performance. We make two observations as follows:

Object-Conditioned Visual Localization. As shown in the top panels, **Obs. ❶** DELTA *provides more focused visual evidence for attribute-object binding.* CLIP often produces diffuse responses or highlights off-target regions, whereas DELTA yields more localized responses around the queried objects. To achieve this, DELTA uses object-conditioned soft attention over image patches to identify attribute-relevant evidence without requiring explicit object annotations. When a caption contains multiple objects, separate object queries can produce distinct patch distributions, allowing the model to localize evidence for each object rather than relying on a single global representation.

Compositional Caption Matching. As shown in the bottom panels, **Obs. ❷** DELTA *better distinguishes compositional differences in image-text matching.* Unlike the object-conditioned visualizations, these examples examine full-caption matching, where multiple objects and their relations must be considered jointly. CLIP often assigns higher scores to hard negatives that share similar wording with the correct captions but differ in meaning, consistent with its bag-of-words behavior [68, 33]. In comparison, DELTA assigns higher scores to matched captions, reflecting its improved ability to capture relational semantics and attribute-object correspondences. Taken together, these results provide further evidence of the complementary roles played by object-conditioned visual evidence and relation-aware textual representations in compositional reasoning.

5 Conclusion

This paper investigates an important question: how can VLMs move beyond global image-text alignment to capture compositional structures? To address this, we propose DELTA, an evidence-aware framework integrating direction-aware relation bucketing and object-conditioned binding localization. By jointly modeling relational semantics and localized visual evidence, DELTA better captures fine-grained compositional structures often overlooked by global representations, improving relation discrimination and attribute-object binding. Extensive experiments on four challenging compositional benchmarks demonstrate the effectiveness of our method, highlighting the importance of explicit compositional modeling for robust vision-language understanding.

Limitations. Although DELTA achieves strong results on compositional reasoning benchmarks, whether these improvements extend to more complex and interactive vision-language tasks remains unclear. Object-conditioned binding localization also depends on the quality of learned object priors, which may become less reliable when objects overlap or attributes are ambiguous. We further observe a decline in general zero-shot classification performance, suggesting a possible trade-off between improved compositional reasoning and general visual recognition. Future work will examine more robust visual grounding strategies and investigate how to improve fine-grained compositional understanding while preserving general recognition performance.

Acknowledgments

We thank the anonymous reviewers for their valuable comments and constructive suggestions. This work received no specific external funding. The authors declare no competing interests.

References

[1] Rabab Abdelfattah, Qing Guo, Xiaoguang Li, Xiaofeng Wang, and Song Wang. Cdul: Clip-driven unsupervised learning for multi-label image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1348–1357, 2023.

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[3] Rim Assouel, Pietro Astolfi, Florian Bordes, Michal Drozdzal, and Adriana Romero-Soriano. Object-centric binding in contrastive language-image pretraining. *Advances in Neural Information Processing Systems*, 38:131418–131458, 2026.

[4] Rabiul Awal, Saba Ahmadi, Le Zhang, and Aishwarya Agrawal. Vismin: Visual minimal-change understanding. *Advances in Neural Information Processing Systems*, 37:107795–107829, 2024.

[5] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Effective conditioned and composed image retrieval combining clip-based features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21466–21474, 2022.

[6] Manuele Barraco, Marcella Cornia, Silvia Cascianelli, Lorenzo Baraldi, and Rita Cucchiara. The unreasonable effectiveness of clip features for image captioning: an experimental analysis. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4662–4670, 2022.

[7] Davide Berasi, Matteo Farina, Massimiliano Mancini, Elisa Ricci, and Nicola Strisciuglio. Not only text: Exploring compositionality of visual representations in vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24917–24927, 2025.

[8] Ioana Bica, Anastasija Ilic, Matthias Bauer, Goker Erdogan, Matko Bošnjak, Christos Kaplanis, Alexey A Gritsenko, Matthias Minderer, Charles Blundell, Razvan Pascanu, et al. Improving fine-grained understanding in image-text pre-training. In *International Conference on Machine Learning*, pages 3974–3995. PMLR, 2024.

[9] Emanuele Bugliarello, Laurent Sartran, Aishwarya Agrawal, Lisa Anne Hendricks, and Aida Nematzadeh. Measuring progress in fine-grained vision-and-language understanding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1559–1582, 2023.

[10] Jirayu Burapacheep, Ishan Gaur, Agam Bhatia, and Tristan Thrush. Colorswap: A color and word order dataset for multimodal evaluation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1716–1726, 2024.

[11] Ian Connick Covert, Tony Sun, James Zou, and Tatsunori Hashimoto. Locality alignment improves vision-language models. In *The Thirteenth International Conference on Learning Representations*.

[12] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335, 2017.

[13] Sri H Dumpala, Aman Jaiswal, Chandramouli Sastry, Evangelos Milios, Sageev Oore, and Hassan Sajjad. Sugarcrepe++ dataset: Vision-language model sensitivity to semantic and lexical alterations. *Advances in Neural Information Processing Systems*, 37:17972–18018, 2024.

[14] Yiwei Fu, Hui Wan, Xiao Luo, and Minghua Deng. Star: Test-time adaptation can enhance universal prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 31482–31492, 2026.

[15] Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Vinija Jain, and Aman Chadha. Exploring the frontier of vision-language models: A survey of current methodologies and future directions. *arXiv preprint arXiv:2404.07214*, 2024.

[16] Olga Golovneva, Tianlu Wang, Jason Weston, and Sainbayar Sukhbaatar. Contextual position encoding: Learning to count what's important. *arXiv preprint arXiv:2405.18719*, 2024.

[17] Yiyang Gu, Jiayue Fan, Kaili Liu, Bohan Wu, Binqi Chen, Zequn Liu, Zhiping Xiao, Rong-Cheng Tu, Xiao Luo, and Ming Zhang. Evirag: Evidence-guided retrieval-augmented generation for medical vision-language models. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3753–3758, 2026.

[18] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14953–14962, 2023.

[19] Lisa Anne Hendricks and Aida Nematzadeh. Probing image-language transformers for verb understanding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3635–3644, 2021.

[20] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *Advances in neural information processing systems*, 36:31096–31116, 2023.

[21] Irene Huang, Wei Lin, Muhammad Jehanzeb Mirza, Jacob A Hansen, Sivan Doveh, Victor I Butoi, Roei Herzig, Assaf Arbelle, Hilde Kuehne, Trevor Darrell, et al. Conme: Rethinking evaluation of compositional reasoning for modern vlms. *Advances in Neural Information Processing Systems*, 37:22927–22946, 2024.

[22] Jiazhen Huang, Zhiming Liu, Changhu Wang, Wei Ju, Ziyue Qiao, and Xiao Luo. Von mises-Fisher mixture model with dynamic shrinkage for realistic test-time transduction. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 47178–47211. PMLR, 06–11 Jul 2026.

[23] Tianyu Huang, Bowen Dong, Yunhan Yang, Xiaoshui Huang, Rynson WH Lau, Wanli Ouyang, and Wangmeng Zuo. Clip2point: Transfer clip to point cloud classification with image-depth pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22157–22167, 2023.

[24] Yufeng Huang, Jiji Tang, Zhuo Chen, Rongsheng Zhang, Xinfeng Zhang, Weijie Chen, Zeng Zhao, Zhou Zhao, Tangjie Lv, Zhipeng Hu, et al. Structure-clip: Towards scene graph knowledge to enhance multi-modal structured representations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 2417–2425, 2024.

[25] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[26] Kenan Jiang, Xuehai He, Ruize Xu, and Xin Wang. Comclip: Training-free compositional image and text matching. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6639–6659, 2024.

[27] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017.

[28] Amita Kamath, Jack Hessel, and Kai-Wei Chang. Text encoders bottleneck compositionality in contrastive vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4933–4944, 2023.

[29] Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9161–9175, 2023.

[30] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

[31] Fucai Ke, Joy Hsu, Zhixi Cai, Zixian Ma, Xin Zheng, Xindi Wu, Sukai Huang, Weiqing Wang, Pari Delir Haghighi, Gholamreza Haffari, et al. Explain before you answer: A survey on compositional visual reasoning. *arXiv preprint arXiv:2508.17298*, 2025.

[32] Zaid Khan, BG Vijay Kumar, Xiang Yu, Samuel Schulter, Manmohan Chandraker, and Yun Fu. Single-stream multi-level alignment for vision-language pretraining. In *European Conference on Computer Vision*, pages 735–751. Springer, 2022.

[33] Darina Koishigarina, Arnas Uselis, and Seong Joon Oh. Clip behaves like a bag-of-words model cross-modally but not uni-modally. *arXiv preprint arXiv:2502.03566*, 2025.

[34] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

[35] Jihoon Kwon, Kyle Min, and Jy-yong Sohn. Enhancing compositional reasoning in clip via reconstruction and alignment of text descriptions. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

[36] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *Advances in Neural Information Processing Systems*, 37:87874–87907, 2024.

[37] Haoxin Li and Boyang Li. Enhancing vision-language compositional understanding with multimodal synthetic data. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24849–24861, 2025.

[38] Juncheng Li, Xin He, Longhui Wei, Long Qian, Linchao Zhu, Lingxi Xie, Yueting Zhuang, Qi Tian, and Siliang Tang. Fine-grained semantically aligned vision-language pre-training. *Advances in neural information processing systems*, 35:7290–7303, 2022.

[39] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[40] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[41] Zhaoyi Li, Gangwei Jiang, Hong Xie, Linqi Song, Defu Lian, and Ying Wei. Understanding and patching compositional reasoning in llms. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9668–9688, 2024.

[42] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[43] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023.

[44] Chen-Yi Lu, Yueh-Shao Chen, and Somali Chaterji. What clip knows but cannot say: Recovering negation from frozen intermediate features. In *European Conference on Computer Vision*, pages 19–36. Springer, 2026.

[45] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM international conference on multimedia*, pages 638–647, 2022.

[46] Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. Crepe: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10910–10921, 2023.

[47] Youngtaek Oh, Jae Won Cho, Dong-Jin Kim, In So Kweon, and Junmo Kim. Preserving multi-modal capabilities of pre-trained vlms for improving vision-linguistic compositionality. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19060–19076, 2024.

[48] Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8253–8280, 2022.

[49] Maitreya Patel, Abhiram Kusumba, Sheng Cheng, Changhoon Kim, Tejas Gokhale, Chitta Baral, and Yezhou Yang. Tripletclip: Improving compositional reasoning of clip via synthetic vision-language negatives. *Advances in neural information processing systems*, 37:32731–32760, 2024.

[50] Amit Peleg, Naman Deep Singh, and Matthias Hein. Advancing compositional awareness in clip with efficient fine-tuning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

[51] Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. Grounded situation recognition. In *European Conference on Computer Vision*, pages 314–332. Springer, 2020.

[52] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[53] Arijit Ray, Filip Radenovic, Abhimanyu Dubey, Bryan Plummer, Ranjay Krishna, and Kate Saenko. Cola: A benchmark for compositional text-to-image retrieval. *Advances in Neural Information Processing Systems*, 36:46433–46445, 2023.

[54] Ugur Sahin, Hang Li, Qadeer Khan, Daniel Cremers, and Volker Tresp. Enhancing multimodal compositional reasoning of visual language models with generative negative mining. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5563–5573, 2024.

[55] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468, 2018.

[56] Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. Foil it! find one mismatch between image and language caption. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 255–265, 2017.

[57] Harman Singh, Pengchuan Zhang, Qifan Wang, Mengjiao Wang, Wenhan Xiong, Jingfei Du, and Yu Chen. Coarse-to-fine contrastive learning in image-text-graph space for improved vision-language compositionality. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 869–893, 2023.

[58] Mingkang Tang, Zhanyu Wang, Zhenhua Liu, Fengyun Rao, Dian Li, and Xiu Li. Clip4caption: Clip for video caption. In *Proceedings of the 29th ACM international conference on multimedia*, pages 4858–4862, 2021.

[59] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022.

[60] Matthew Trager, Pramuditha Perera, Luca Zancato, Alessandro Achille, Parminder Bhatia, and Stefano Soatto. Linear spaces of meanings: compositional structures in vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15395–15404, 2023.

[61] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

[62] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems*, 37:121475–121499, 2024.

[63] Monika Wysoczańska, Oriane Siméoni, Michaël Ramamonjisoa, Andrei Bursuc, Tomasz Trzciński, and Patrick Pérez. Clip-dinoiser: Teaching clip a few dino tricks for open-vocabulary semantic segmentation. In *European Conference on Computer Vision*, pages 320–337. Springer, 2024.

[64] Rui Xiao, Sanghwan Kim, Mariana-Iuliana Georgescu, Zeynep Akata, and Stephan Alaniz. Flair: Vlm with fine-grained language-informed image representations. In *Proceedings of the computer vision and pattern recognition conference*, pages 24884–24894, 2025.

[65] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021.

[66] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

[67] Jiarui Yu, Haoran Li, Yanbin Hao, Bin Zhu, Tong Xu, and Xiangnan He. Cgt-gan: Clip-guided text gan for image captioning. In *Proceedings of the 31st ACM international conference on multimedia*, pages 2252–2263, 2023.

[68] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*, 2023.

[69] Yunan Zeng, Yan Huang, Jinjin Zhang, Zequn Jie, Zhenhua Chai, and Liang Wang. Investigating compositional challenges in vision-language models for visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14141–14151, 2024.

[70] Aimen Zerroug, Mohit Vaishnav, Julien Colin, Sebastian Musslick, and Thomas Serre. A benchmark for compositional visual reasoning. *Advances in neural information processing systems*, 35:29776–29788, 2022.

[71] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. In *European conference on computer vision*, pages 310–325. Springer, 2024.

[72] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5625–5644, 2024.

[73] Le Zhang, Rabiul Awal, and Aishwarya Agrawal. Contrasting intra-modal and ranking cross-modal hard negatives to enhance visio-linguistic compositional understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13774–13784, 2024.

[74] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European conference on computer vision*, pages 493–510. Springer, 2022.

[75] Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations. *arXiv preprint arXiv:2207.00221*, 2022.

[76] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16793–16803, 2022.

[77] 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.

[78] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004, 2016.

[79] Shunge Zou, Changhu Wang, Wei Ju, Ziyue Qiao, and Xiao Luo. Dance: Diversity-attended dynamic caching with asymmetric quantization for test-time adaptation of vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 37329–37343, 2026.

A Details of the Experimental Setup

A.1 More Details of Evaluation

We evaluate DELTA on four compositional reasoning benchmarks: VALSE [48], ColorSwap [10], CREPE [46], and SugarCrepe [20]. Together, they cover different aspects of compositional understanding, including relational reasoning, attribute-object binding, and sensitivity to semantic perturbations. We follow the standard image-to-text matching protocol for all benchmarks, reporting accuracy based on whether the positive caption receives a higher similarity score than the corresponding negative caption. Table 3 summarizes the benchmark statistics, evaluation focuses, licenses, and image sources. All four benchmarks are publicly available, and we use them under their respective release terms without redistributing the original images or annotations.

Table 3: Details of compositional reasoning benchmarks, including evaluation focuses, licenses, and image sources.

Benchmark	#Examples	Compositionality	License	Image Source
VALSE [48]	8,309	Existence, Plurality, Counting, Spatial Relations, Actions, Entity Coreference	MIT	Visual7W [78], COCO [42], SWiG [51], VisDial v1.0 [12], FOIL-it [56]
ColorSwap [10]	1,000	Color–Object Binding, Word Order	MIT	Synthetic images generated with human-in-the-loop
CREPE [46]	17K	Atomic Replacement, Negation, Semantic Swap	Unspecified	Visual Genome [34]
SugarCrepe [20]	7,511	Addition, Replacement, Swapping	MIT	COCO [42]

A.2 More Details of Baselines

We compare DELTA with the original CLIP model [52] and seven representative fine-tuning methods for compositional reasoning. Among methods that rely on challenging negative samples, NegCLIP [68] constructs hard negative captions through rule-based perturbations, while GNM-CLIP [54] explores generative negative mining to provide more informative supervision. CE-CLIP [73] combines intra-modal contrastive learning with cross-modal ranking objectives to improve sensitivity to compositional differences. Other approaches focus on different aspects of CLIP adaptation. FSC-CLIP [47] aims to preserve general multimodal capabilities during compositional fine-tuning, whereas CLIC [50] emphasizes efficient adaptation with limited model modifications. READ-CLIP [35] improves textual representations through caption reconstruction and auxiliary alignment objectives, while LABCLIP [33] introduces a lightweight projector to address attribute-object binding. These baselines cover different strategies for improving compositional understanding, providing a basis for evaluating DELTA against existing CLIP adaptation methods.

B Supplementary Experimental Results

B.1 Detailed Benchmark Results

Beyond the overall scores in Table 1, we report per-category results for VALSE and SugarCrepe in Tables 4 and 5. These results provide a finer-grained view of compositional reasoning, with SugarCrepe averaged over three groups in the Table 1 and seven categories in the appendix.

B.2 Zero-Shot Image Classification

To assess whether compositional fine-tuning preserves general visual recognition ability, we evaluate zero-shot classification performance across ten datasets. The results are presented in Table 6. Although DELTA achieves the best performance on KITTI, its overall classification accuracy remains below that of the original CLIP, suggesting a trade-off between compositional reasoning and general visual recognition. Nevertheless, DELTA remains competitive with several fine-tuned baselines.

Table 4: Detailed results on VALSE. Best results are in bold and second-best results are underlined.

Model	Actions	Coreference	Counting	Existence	Noun Phrases	Plurals	Relations	Avg.
CLIP (ICML'21)	74.3	54.4	61.7	69.3	90.4	57.9	66.0	67.4
NegCLIP (ICLR'23)	84.1	60.2	65.7	75.3	93.4	70.3	69.6	73.7
GNM-CLIP (WACV'24)	72.1	61.1	66.5	76.0	90.8	68.7	62.6	70.7
CE-CLIP (CVPR'24)	85.0	59.9	67.6	78.2	94.4	**78.8**	74.0	76.0
FSC-CLIP (EMNLP'24)	82.9	59.4	66.3	77.6	93.5	72.7	75.3	74.4
CLIC (NeurIPS'25)	76.5	**61.5**	63.6	74.9	92.0	68.9	66.5	71.1
READ-CLIP (NeurIPS'25)	86.3	55.7	**69.0**	80.8	**95.8**	73.8	75.0	76.2
LABCLIP (ICLR'26)	78.7	54.6	64.6	79.6	92.8	67.0	66.7	71.3
DELTA	**87.8**	60.3	68.8	**81.4**	94.7	75.4	**76.4**	**77.1**

Table 5: Detailed results on SugarCrepe. Att., Obj., and Rel. denote Attribute, Object, and Relation, respectively. Best results are in bold and second-best results are underlined.

Model	Add Att.	Add Obj.	Replace Att.	Replace Obj.	Replace Rel.	Swap Att.	Swap Obj.	Avg.
CLIP (ICML'21)	69.5	77.0	80.3	90.7	69.4	64.1	61.2	73.2
NegCLIP (ICLR'23)	85.3	90.0	88.2	94.0	74.6	77.9	75.5	83.6
GNM-CLIP (WACV'24)	79.9	88.4	84.9	93.2	67.8	70.0	61.2	77.9
CE-CLIP (CVPR'24)	**91.9**	**92.3**	90.2	94.4	81.4	76.7	75.1	86.0
FSC-CLIP (EMNLP'24)	86.7	90.2	89.2	94.3	80.4	77.8	77.6	85.2
CLIC (NeurIPS'25)	81.6	87.0	85.3	93.0	72.0	71.5	66.9	79.6
READ-CLIP (NeurIPS'25)	87.7	90.3	**91.0**	94.9	80.6	82.7	81.6	87.0
LABCLIP (ICLR'26)	77.9	84.0	84.6	92.4	72.8	74.9	72.7	79.9
DELTA	87.4	89.7	90.9	**95.1**	**81.9**	**85.1**	**84.1**	**87.7**

B.3 Generalization Beyond COCO

To assess whether the improvements of DELTA depend on the COCO training distribution, we further conduct experiments using Flickr30K as the training corpus. For a fair comparison, all methods adopt the same backbone and evaluation protocol. As shown in Table 7, DELTA consistently outperforms all baselines across four compositional benchmarks, achieving an average accuracy of 61.8%, surpassing the strongest baseline by 4.2 pp. These results suggest that the effectiveness of DELTA extends beyond COCO-specific training patterns, demonstrating its generalizability across different training distributions.

Table 7: Compositional reasoning results with Flickr30K training.

Model	VALSE	Color Swap	CREPE	Sugar Crepe	Avg.
CLIC	69.9	40.7	24.6	78.9	53.5
READ-CLIP	74.0	36.3	38.8	81.3	57.6
LABCLIP	67.0	13.0	40.4	70.4	47.7
DELTA	**74.4**	**42.0**	**49.1**	**81.5**	**61.8**

B.4 Scalability to Larger Backbones

We also investigate whether DELTA maintains its effectiveness with larger model architectures by extending the evaluation to the CLIP ViT-L/14 backbone. All methods are trained on the same COCO corpus and evaluated under an identical protocol to ensure a fair comparison. As reported in Table 8, DELTA outperforms all competing methods across the four compositional benchmarks, reaching an average accuracy of 67.3%, which exceeds the strongest baseline by 5.1 pp. On CREPE, DELTA achieves 54.7%, surpassing READ-CLIP by 10.9 pp and showing stronger discrimination of fine-grained compositional differences. For ColorSwap,

Table 8: Compositional reasoning results with the CLIP ViT-L/14 backbone.

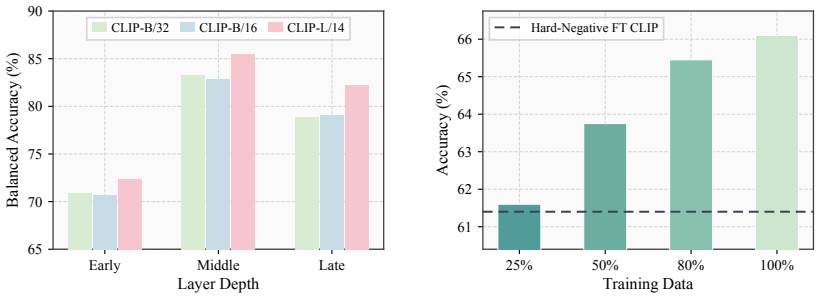
Model	VALSE	Color Swap	CREPE	Sugar Crepe	Avg.
CLIC	73.6	41.3	31.8	81.1	57.0
READ-CLIP	78.8	39.0	43.8	87.3	62.2
LABCLIP	73.2	30.3	34.0	82.1	54.9
DELTA	**79.8**	**45.6**	**54.7**	**88.9**	**67.3**

Table 6: Zero-shot image classification results across ten datasets. Best results are in bold and second-best results are underlined.

Model	Caltech	C10	C100	Aircraft	EuroSAT	KITTI	MNIST	PCAM	Cars	ImageNet	Avg.
CLIP (ICML'21)	81.4	90.0	**65.1**	**19.6**	50.4	27.1	44.2	**62.1**	**59.3**	**63.4**	**56.3**
NegCLIP (ICLR'23)	**82.6**	88.9	63.0	16.8	49.7	27.7	49.7	54.9	53.9	60.9	54.8
GNM-CLIP (WACV'24)	81.5	88.5	65.0	17.2	50.7	25.3	**54.3**	55.6	53.1	61.4	55.3
CE-CLIP (CVPR'24)	78.3	85.9	60.1	10.0	42.8	34.6	40.6	50.0	35.3	49.9	48.8
FSC-CLIP (EMNLP'24)	81.8	89.1	64.9	15.8	51.6	29.4	50.4	51.0	51.8	59.2	54.5
DAC-LLM (NeurIPS'23)	77.7	**90.4**	63.9	11.3	**52.3**	16.6	42.2	50.0	39.4	51.0	49.5
DAC-SAM (NeurIPS'23)	75.7	89.9	63.7	9.0	51.2	18.7	45.3	50.0	39.9	52.4	49.6
Triplet-CLIP (NeurIPS'24)	80.6	89.1	61.5	12.7	35.2	15.3	34.3	49.6	23.9	54.8	45.7
READ-CLIP (NeurIPS'25)	77.2	89.5	57.0	15.4	41.6	26.9	30.7	50.0	37.2	51.5	47.7
LABCLIP (ICLR'26)	78.0	86.0	56.7	15.1	36.5	29.4	11.6	50.0	42.0	49.4	45.5
DELTA	79.2	89.3	59.2	15.6	47.5	**35.4**	40.6	52.4	40.6	52.4	51.2

DELTA obtains the highest accuracy of 45.6%, consistent with improved attribute-object binding. It also achieves competitive results on VALSE and SugarCrepe, with accuracies of 79.8% and 88.9%, respectively. These results suggest that the improvements observed with ViT-B/32 are retained when DELTA is applied to the larger ViT-L/14 backbone.

B.5 Layer-wise Probing



(a) Layer-wise Probing **(b) Data Scaling**

Figure 5: (a) Balanced probing accuracy across early, middle, and late layers of three CLIP backbones. (b) Compositional reasoning accuracy of DELTA with varying proportions of COCO training data. The dashed line indicates hard-negative fine-tuned CLIP trained on the full dataset.

We further examine how compositional information varies across the text layers of CLIP backbones, including ViT-B/32, ViT-B/16, and ViT-L/14. We train lightweight linear probes on representations extracted from early, middle, and late layers to assess their ability to distinguish fine-grained compositional differences. As shown in Figure 5(a), middle-layer representations achieve the highest balanced probing accuracy for all three backbones, while early and late layers perform relatively less well. This suggests that intermediate layers contain more readily accessible compositional information, which may be less evident in the final representations. The similar pattern observed across different backbones also supports our use of intermediate-layer features to preserve relational semantics and improve compositional modeling.

B.6 Data Scaling

We further investigate how DELTA performs with different amounts of training data. Following the main experimental setup, we use the same fine-tuning configuration and train the model on 25%, 50%, 80%, and 100% of the COCO training split. We compare against hard-negative fine-tuned CLIP trained on the full training subset, shown as the dashed line in Figure 5(b). DELTA already outperforms this baseline with 25% of the training data, and its accuracy increases as the training proportion grows to 100%. These results indicate that DELTA can exceed the baseline without using the full training subset, while benefiting further from additional training data.

B.7 Robustness to Noisy Binding Cues

We use spaCy dependency parsing to extract attribute-object pairs from training captions. Specifically, we identify adjectives linked to nouns or proper nouns through the amod relation and align the extracted spans with CLIP tokens. Samples without valid pairs or token alignments are excluded from the auxiliary binding loss. We further examine how errors in these cues affect performance by randomly replacing 50% of valid object or positive-attribute cues with alternatives from the same caption. As shown in Table 9, the average accuracy drops from 66.1% to 65.2% with corrupted object cues and to 64.2% with corrupted attribute cues. The respective decreases of 0.9 pp and 1.9 pp indicate that DELTA remains relatively robust to noisy cues. Since these cues are only used for auxiliary supervision during training, inference relies on full captions without requiring dependency parsing.

Table 9: Robustness to noisy object and attribute cues.

Variants	VALSE	Color Swap	CREPE	Sugar Crepe	Avg.
50% Object-cue corruption	**77.3**	46.7	50.3	86.5	65.2
50% Attribute-cue corruption	76.8	45.3	49.0	85.7	64.2
DELTA	77.1	**48.7**	**51.2**	**87.5**	**66.1**

B.8 Extended Case Study

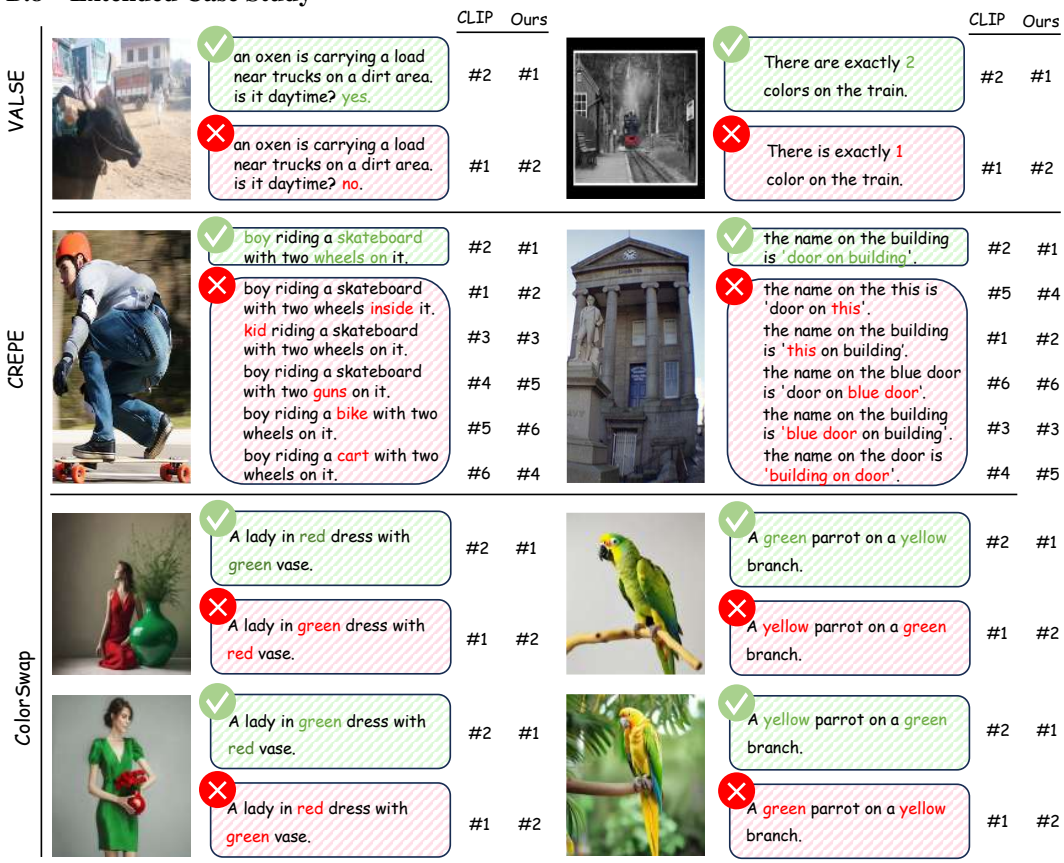


Figure 6: Extended qualitative case studies comparing CLIP with our method on VALSE, CREPE, and ColorSwap. Positive ✓ and negative ✗ captions are shown with their rankings based on image-caption cosine similarity. Across diverse compositional scenarios, CLIP often assigns higher similarity to hard negatives that preserve most lexical content but alter fine-grained semantics, whereas DELTA more consistently ranks the matched captions above the negatives. These examples further illustrate the effectiveness of DELTA in modeling relation-aware semantics and attribute-object binding.

We further provide extended qualitative case studies in Figure 6 on VALSE, CREPE, and ColorSwap. These examples complement the qualitative results in the main paper by illustrating how DELTA better distinguishes fine-grained compositional differences across diverse benchmarks. Compared with CLIP, DELTA more consistently ranks the matched captions above hard compositional negatives, demonstrating its stronger attribute-object binding and relation-aware compositional reasoning ability.

C Proofs

Proof. We prove Theorem 3.1 and Corollary 3.2. For compactness, define the patch-level attention logits and binding scores as

$$u_{i\ell}^{\pm} = (W_q^{\mathrm{att}} a_i^{\pm})^\top W_k p_{i\ell}, \qquad s_{i\ell}^{\pm} = (a_i^{\pm})^\top W_b p_{i\ell}. \tag{19}$$

By Eq. 14, the object-conditioned attention weights can be written as

$$\alpha_{i\ell}^{\pm} = \frac{\pi_{i\ell}^{\gamma} \exp(u_{i\ell}^{\pm}/\tau_{\mathrm{loc}})}{\sum_{j=1}^{L} \pi_{ij}^{\gamma} \exp(u_{ij}^{\pm}/\tau_{\mathrm{loc}})}. \tag{20}$$

By the definition of the object prior in Eq. 13, we have $\pi_{i\ell} > 0$ for all ℓ.

We first bound the total attention mass assigned to off-object patches. Let

$$\delta_i^{\pm} = \sum_{\ell \notin \mathcal{O}_i} \alpha_{i\ell}^{\pm}. \tag{21}$$

For either sign $\pm$, define

$$Z_{\mathrm{out}}^{\pm} = \sum_{\ell \notin \mathcal{O}_i} \pi_{i\ell}^{\gamma} \exp(u_{i\ell}^{\pm}/\tau_{\mathrm{loc}}), \qquad Z_{\mathrm{in}}^{\pm} = \sum_{\ell \in \mathcal{O}_i} \pi_{i\ell}^{\gamma} \exp(u_{i\ell}^{\pm}/\tau_{\mathrm{loc}}). \tag{22}$$

Then

$$\delta_i^{\pm} = \frac{Z_{\mathrm{out}}^{\pm}}{Z_{\mathrm{in}}^{\pm} + Z_{\mathrm{out}}^{\pm}} \le \frac{Z_{\mathrm{out}}^{\pm}}{Z_{\mathrm{in}}^{\pm}}. \tag{23}$$

By the bounded-logit assumption, for all ℓ,

$$\exp(u_{i\ell}^{\pm}/\tau_{\mathrm{loc}}) \le \exp(R/\tau_{\mathrm{loc}}), \qquad \exp(u_{i\ell}^{\pm}/\tau_{\mathrm{loc}}) \ge \exp(-R/\tau_{\mathrm{loc}}). \tag{24}$$

By the object-prior separation assumption, we have

$$\pi_{i\ell}^{\gamma} \le c_{\mathrm{out}}^{\gamma} \quad (\ell \notin \mathcal{O}_i), \qquad \pi_{i\ell}^{\gamma} \ge c_{\mathrm{in}}^{\gamma} \quad (\ell \in \mathcal{O}_i). \tag{25}$$

Therefore,

$$Z_{\mathrm{out}}^{\pm} \le (L-m) c_{\mathrm{out}}^{\gamma} \exp(R/\tau_{\mathrm{loc}}), \qquad Z_{\mathrm{in}}^{\pm} \ge m c_{\mathrm{in}}^{\gamma} \exp(-R/\tau_{\mathrm{loc}}). \tag{26}$$

It follows that

$$\delta_i^{\pm} \le \frac{L-m}{m} \left(\frac{c_{\mathrm{out}}}{c_{\mathrm{in}}}\right)^{\gamma} \exp\left(\frac{2R}{\tau_{\mathrm{loc}}}\right) = \frac{C_0}{4M} \rho^{\gamma}. \tag{27}$$

We now bound the binding margin. From the definition of z_i^+,

$$(a_i^+)^\top W_b z_i^+ = \sum_{\ell=1}^{L} \alpha_{i\ell}^+ s_{i\ell}^+. \tag{28}$$

Using the object-local separability assumption on $\mathcal{O}_i$ and the bounded-score assumption outside $\mathcal{O}_i$, we obtain

$$\begin{aligned}
(a_i^+)^\top W_b z_i^+ &= \sum_{\ell \in \mathcal{O}_i} \alpha_{i\ell}^+ s_{i\ell}^+ + \sum_{\ell \notin \mathcal{O}_i} \alpha_{i\ell}^+ s_{i\ell}^+ \\
&\ge \mu_+ \sum_{\ell \in \mathcal{O}_i} \alpha_{i\ell}^+ - M \sum_{\ell \notin \mathcal{O}_i} \alpha_{i\ell}^+ \\
&= \mu_+ (1 - \delta_i^+) - M \delta_i^+ \\
&= \mu_+ - (\mu_+ + M) \delta_i^+.
\end{aligned} \tag{29}$$

Because $\mathcal{O}_i$ is nonempty and $|s_{i\ell}^+| \le M$ while $s_{i\ell}^+ \ge \mu_+$ on $\mathcal{O}_i$, the assumptions imply $\mu_+ \le M$. Hence,

$$(a_i^+)^\top W_b z_i^+ \ge \mu_+ - 2M\delta_i^+ \ge \mu_+ - \frac{C_0}{2} \rho^{\gamma}. \tag{30}$$

Similarly, from the definition of z_i^-,

$$(a_i^-)^\top W_b z_i^- = \sum_{\ell=1}^{L} \alpha_{i\ell}^- s_{i\ell}^-. \tag{31}$$

Using object-local separability and boundedness again,

$$\begin{aligned}(a_i^-)^\top W_b z_i^- &= \sum_{\ell \in \mathcal{O}_i} \alpha_{i\ell}^- s_{i\ell}^- + \sum_{\ell \notin \mathcal{O}_i} \alpha_{i\ell}^- s_{i\ell}^- \\ &\le \mu_- \sum_{\ell \in \mathcal{O}_i} \alpha_{i\ell}^- + M \sum_{\ell \notin \mathcal{O}_i} \alpha_{i\ell}^- \\ &= \mu_-(1-\delta_i^-) + M\delta_i^- \\ &= \mu_- + (M-\mu_-)\delta_i^-. \end{aligned} \tag{32}$$

Because $\mathcal{O}_i$ is nonempty and $|s_{i\ell}^-| \le M$ while $s_{i\ell}^- \le \mu_-$ on $\mathcal{O}_i$, the assumptions imply $\mu_- \ge -M$. Therefore,

$$(a_i^-)^\top W_b z_i^- \le \mu_- + 2M\delta_i^- \le \mu_- + \frac{C_0}{2}\rho^\gamma. \tag{33}$$

Combining the bounds above, we obtain

$$(a_i^+)^\top W_b z_i^+ - (a_i^-)^\top W_b z_i^- \ge \left(\mu_+ - \frac{C_0}{2}\rho^\gamma\right) - \left(\mu_- + \frac{C_0}{2}\rho^\gamma\right) \tag{34}$$

$$= \Delta - C_0\rho^\gamma. \tag{35}$$

This proves Theorem 3.1.

It remains to prove Corollary 3.2. Since $\rho < 1$, we have $\log(1/\rho) > 0$. Let

$$\gamma_0 = \begin{cases} 0, & C_0 = 0, \\ \max\left\{\dfrac{\log(2C_0/\Delta)}{\log(1/\rho)}, 0\right\}, & C_0 > 0. \end{cases} \tag{36}$$

For any $\gamma \ge \gamma_0$, we have

$$C_0\rho^\gamma \le \frac{\Delta}{2}. \tag{37}$$

Thus, Theorem 3.1 gives

$$(a_i^+)^\top W_b z_i^+ - (a_i^-)^\top W_b z_i^- \ge \Delta - C_0\rho^\gamma \ge \frac{\Delta}{2} > 0. \tag{38}$$

Finally, for the single-sample binding loss in Eq. 15, let

$$\mathcal{M}_i = (a_i^+)^\top W_b z_i^+ - (a_i^-)^\top W_b z_i^-. \tag{39}$$

Then

$$\ell_{\text{bind}} = -\log \sigma(\mathcal{M}_i/\tau_{\text{bind}}) = \log\left(1 + \exp\left(-\frac{\mathcal{M}_i}{\tau_{\text{bind}}}\right)\right). \tag{40}$$

Since the function

$$x \mapsto \log\left(1 + \exp\left(-\frac{x}{\tau_{\text{bind}}}\right)\right) \tag{41}$$

is nonincreasing for $\tau_{\text{bind}} > 0$ and $\mathcal{M}_i \ge \Delta/2$, we obtain

$$\ell_{\text{bind}} \le \log\left(1 + \exp\left(-\frac{\Delta}{2\tau_{\text{bind}}}\right)\right), \tag{42}$$

which proves Corollary 3.2. □

D Broader Impacts

Improved attribute–object binding and relation-sensitive image–text matching may benefit applications such as image retrieval and accessibility. However, gains on compositional benchmarks alone do not establish reliability in real-world or safety-critical settings. Errors in attribute binding or scene interpretation may lead to misleading retrieval results or descriptions, potentially affecting downstream decisions. DELTA may also inherit biases from pretrained CLIP and its training data, while its visual matching capabilities could be misused for privacy-invasive surveillance. Before deployment, the method should be evaluated under application-specific conditions, with particular attention to bias, privacy risks, and appropriate human oversight. These precautions are recommended but have not been validated in this work.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Both in the abstract and the introduction, we accurately describe the contributions and scope of the work.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of our method are discussed in Section 5 (Conclusion).

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We include all proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The detailed experimental settings are included.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We release the DELTA implementation, training and evaluation scripts, configurations, and setup instructions at https://github.com/Zhoujun-Ye/DELTA. However, the released VALSE VisualDialog image-preparation workflow is incomplete, and ColorSwap requires authorized Hugging Face access. The current release therefore does not yet provide a complete end-to-end reproduction path for all main results.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: Section 4.1 reports the training data split, backbone, optimizer, batch size, learning rate, epochs, and auxiliary-loss settings; Appendix A.1 describes the evaluation benchmarks and protocol.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We report single-run benchmark results. We do not estimate variability across random seeds or provide error bars, confidence intervals, or statistical significance tests; the reported differences should be interpreted with this limitation.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Section 4.1 reports a single NVIDIA A100 GPU with 80GB memory. Per-run execution times and the total compute used for the reported experiments and preliminary runs are not provided, so the resource disclosure is incomplete.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research strictly follows the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Appendix D discusses potential benefits for image retrieval and accessibility, as well as risks from incorrect binding and scene interpretation, inherited biases, and privacy-invasive misuse.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This work studies discriminative image–text matching using existing datasets and does not introduce a newly scraped dataset or a generative model requiring high-risk release safeguards. This does not imply that downstream uses are free of risk; see Appendix D.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No]

    Justification: The paper credits the original datasets and pretrained CLIP. Table 3 summarizes benchmark licenses and image sources, but lists the CREPE license as unspecified. The paper does not yet fully document the licenses and terms for all assets, including COCO and pretrained CLIP; citation alone does not establish complete license disclosure.

    Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No]

    Justification: The released DELTA implementation is a new code asset. Its repository provides installation, data-access, training, and evaluation instructions and configurations, but documentation is incomplete: the VALSE VisualDialog preparation workflow is unfinished, and an explicit code license is not provided in the current release. We therefore do not claim complete asset documentation.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing experiments or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: This paper does not include study participants.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs were used solely for writing assistance (e.g., wording, formatting), and did not contribute to the design, implementation, or evaluation of the proposed methods.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.