GROUND, RELATE, AND MATCH: STRUCTURED UNBALANCED OPTIMAL TRANSPORT FOR COMPOSITIONAL IMAGE-TEXT MATCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Dual-encoder vision-language models such as CLIP achieve strong zero-shot image-text matching, but global similarity can overlook fine-grained compositional structure, including entity binding, attributes, and directed semantic relations. Existing training-free methods introduce local evidence, yet often score components independently without explicitly modeling structured consistency across entities and relations. We introduce LEM (Local Evidence Matters), a training-free framework for compositional image-text matching through structured graph alignment. LEM parses paired captions into directed semantic graphs and constructs a visual region graph from localized image proposals. A regularized unbalanced optimal transport formulation aligns textual entities with visual regions while allowing noisy or irrelevant regions to receive reduced transport mass. Node costs combine entity identity with candidate-conditioned attribute evidence, while relation costs evaluate pairwise interaction context together with directed endpoint consistency. The resulting transport plan is used to compute weighted entity and relation similarities, which are combined into a structured local alignment score. Candidate-conditioned comparisons distinguish competing attributes, predicates, and argument directions, while the structured local score is fused with the original global image-text similarity for final matching. LEM improves Winoground Image and Group Score and SugarCrepe attribute-binding accuracy over both CLIP and NegCLIP backbones, though gains on object swaps and relation replacement remain mixed.

1 INTRODUCTION

Vision-language models learn shared representations from large-scale image-caption data, enabling effective image-text matching and zero-shot transfer (Radford et al., 2021; Jia et al., 2021). Subsequent work improves fine-grained image-text interaction (Yao et al., 2022) and multimodal pretraining (Li et al., 2022a), while open-vocabulary grounding enables language-guided localization of visual concepts (Liu et al., 2025). Yet the global similarity used by dual encoders does not explicitly check which image regions support individual entities, whether attributes belong to the correct entities, or whether their relations are visually supported. Recognizing the same concepts in an image and a caption is therefore not enough for compositional matching. Figure 1 illustrates this distinction: standard dual-encoders assign a higher global similarity score to the role-reversed caption (*A tv monitor looking at a cat on a desk*) than to the correct one, simply because they share identical lexical content.

While extracting local regional evidence might intuitively solve this ambiguity, evaluating these regions independently also falls short. As shown in the top half of Figure 1, an independent localization model can confidently map the text entities to their corresponding visual regions (r_1 and r_2). However, without a structural mechanism to verify the directed relation (*looking at*) between them, the model cannot computationally distinguish the true caption from its negative role-reversal. We introduce **LEM** (Local Evidence Matters), a training-free framework that complements global similarity with structured local evidence. LEM parses a caption into a directed graph whose nodes represent entities with their attributes and whose edges represent relations. It constructs a caption-guided image graph whose nodes are candidate regions obtained using entity phrases and whose edges rep-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

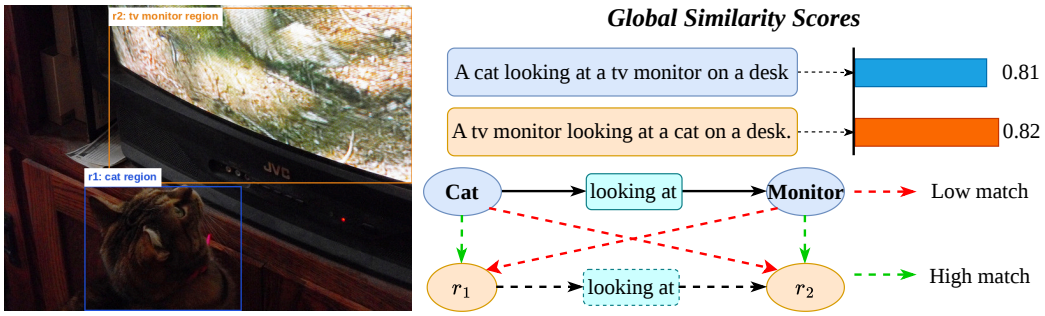

Figure 1: **Why structured local evidence is necessary for compositional matching.** Global dual encoders (bottom) capture holistic features but are blind to semantic roles, incorrectly assigning a higher score to the role-reversed negative caption. While independent local extraction (top) successfully grounds individual entities (e.g., matching the *cat* to r_1 and the *monitor* to r_2), evaluating these components without verifying the directed predicate (*looking at*) leaves the competing semantic role configurations indistinguishable. This necessitates a structured alignment approach.

resent candidate region pairs for relation evaluation. Drawing on entropic optimal transport (Cuturi, 2013; Peyré & Cuturi, 2019), LEM aligns these graphs through a structured unbalanced formulation. Its relaxed marginal constraints downweight ungrounded transport mass, making alignment robust to visual clutter, duplicate boxes, and false-positive region proposals.

The alignment objective combines a factorized node cost with a quadratic relation cost. The node cost assesses entity-region compatibility and contrasts the stated attributes with alternatives from the caption. The relation cost combines forward–reverse prompt comparisons on crops containing both candidate regions with endpoint-role consistency on the individual regions. Together, these terms assess whether the visual evidence supports both the stated relation and the assignment of its subject and object. The relation term couples pairs of transport assignments, so relational evidence guides the alignment itself. LEM then combines the resulting local score with global image-text similarity. All pretrained components remain frozen. Our evaluation focuses on Winoground (Thrush et al., 2022) and SugarCrepe (Hsieh et al., 2023), examining matching under changes in attribute binding and semantic role assignment. Our contributions are:

1. We formulate compositional image-text matching as structured unbalanced optimal transport between caption and region graphs, jointly accounting for entity correspondence and relation consistency to complement global dual-encoder similarity.
2. We design compositional transport costs for attribute binding and directed relations, combining contrastive attribute evidence with pair-context and endpoint-role evidence within a single alignment objective.
3. We introduce LEM as a training-free, plug-and-play scoring framework for compatible pretrained dual encoders, integrating structured local evidence at inference time without updating model parameters or modifying backbone architectures.

2 Related Work

Compositional Vision-Language Evaluation. Contrastive dual encoders achieve strong zero-shot image-text retrieval (Radford et al., 2021; Jia et al., 2021), yet diagnostic studies show that their global similarity is often insensitive to how concepts are composed (Thrush et al., 2022; Yuksekgonul et al., 2022). Winoground (Thrush et al., 2022) isolates this failure with items that pair two images with two captions containing the same words in a different order. It reports Text, Image, and Group Scores, whose chance levels are 25%, 25%, and 16.7%, and most models evaluated in the original study remain below chance on Group Score. Diwan et al. (2022) show that many Winoground items additionally require commonsense reasoning or resolution of ambiguous captions, meaning aggregate scores can conflate compositional matching with other skills. ARO (Yuksekgonul et al., 2022) extends the diagnosis to attribution, relation, and word order, finding that

models often behave as bag-of-words embedders. Complementary benchmarks probe this weakness from different angles: VALSE targets specific linguistic phenomena (Parcalabescu et al., 2022), VL-CheckList separates objects, attributes, and relations (Zhao et al., 2022), and SugarCrepe (Hsieh et al., 2023) uses language-model-generated hard negatives with human validation to test attribute binding and object swaps without linguistic artifacts.

Learning Compositional Representations. To mitigate these bag-of-words failures, NegCLIP (Yuksekgonul et al., 2022) fine-tunes CLIP with word-swapped negative captions and nearest-neighbor negative images, explicitly introducing compositional mismatches during training. However, Kamath et al. (2024) show that improvements on hard negatives can coincide with reduced consistency on meaning-preserving hard positives, a limitation they address by jointly training on both to force the model to distinguish compositional changes from meaning-preserving variations. Other approaches learn structural alignment components on top of frozen backbones. BindCLIP (Song et al., 2026) uses a balanced token-patch-depth transport coupling to score semantic and spatial evidence, optimizing learned projections and scoring parameters. Similarly, Structure-CLIP (Huang et al., 2024) uses scene graphs to guide cross-modal contrastive learning.

Region-Level Grounding and Scene Graphs. Global dual-encoder similarity fails to verify whether specific image regions support individual entities. Open-vocabulary detection and region-level vision-language models, such as RegionCLIP (Zhong et al., 2022), GLIP (Li et al., 2022b), and Grounding DINO (Liu et al., 2023), address this by aligning textual phrases with localized visual proposals. While these models excel at grounding isolated entities, they do not inherently verify directed semantic relations between the grounded regions. Traditionally, Scene Graph Generation (SGG) models have been used to explicitly map objects and their relations (Xu et al., 2017), but they rely on closed-set relation vocabularies and heavy supervision.

Training-Free Alignment and Optimal Transport. A growing line of work seeks to improve compositional matching at inference time without additional training. ComCLIP (Jiang et al., 2024) associates textual components with corresponding visual subimages and combines subject and object regions to construct predicate subimages. It computes component-level similarities and uses them to weight the corresponding subimage embeddings when augmenting the global image representation. However, individual components are aggregated through similarity-based weighting rather than a joint structural assignment. Optimal transport (OT) provides a principled mathematical framework for soft feature alignment, with entropic regularization enabling efficient Sinkhorn-based computation (Cuturi, 2013). Gromov-Wasserstein compares pairwise structure within two domains (Mémoli, 2011; Peyré et al., 2016), while fused Gromov–Wasserstein combines feature and structural costs (Vayer et al., 2019). While Graph Optimal Transport (Chen et al., 2020) uses a shared transport plan as a training regularizer, recent works like ConceptOT (Kumar, 2026) and Geometry-Aware CLIP Retrieval (Prakash et al., 2026) apply OT for inference-time reranking. LEM differs from these approaches by using a shared *unbalanced* optimal transport (Chizat et al., 2017) to align caption entities with visual regions while coupling directed pairwise relation evidence directly into the alignment objective.

3 METHODOLOGY

3.1 OVERVIEW AND GRAPH CONSTRUCTION

Given an image I and a caption c, LEM computes a compositional matching score by combining global similarity with local graph alignment. The detector, parser, and vision–language encoders remain frozen throughout inference. Let f_{v} and f_{t} denote normalized image and text encoders, respectively, and write $s(v,t) = f_{\mathrm{v}}(v)^{\top} f_{\mathrm{t}}(t)$. The global score is $S_{\mathrm{global}} = s(I,c)$, as mentioned in CLIP (Radford et al., 2021).

Caption graph. We parse c into a directed graph $\mathcal{G}_{\mathrm{txt}} = (\mathcal{V}_{\mathrm{txt}}, \mathcal{E}_{\mathrm{txt}})$ with M entity nodes. Each node u_a stores an entity name n_a and its modifying attributes $\mathcal{A}_a$, including adjectival and numerical modifiers. An edge $e = (a, r, b)$ assigns entities a and b the subject and object roles of relation r. Compound nouns remain single entities, and passive constructions are normalized to a consistent subject-object convention. Parser settings and prompt construction are specified in Appendix C.1.

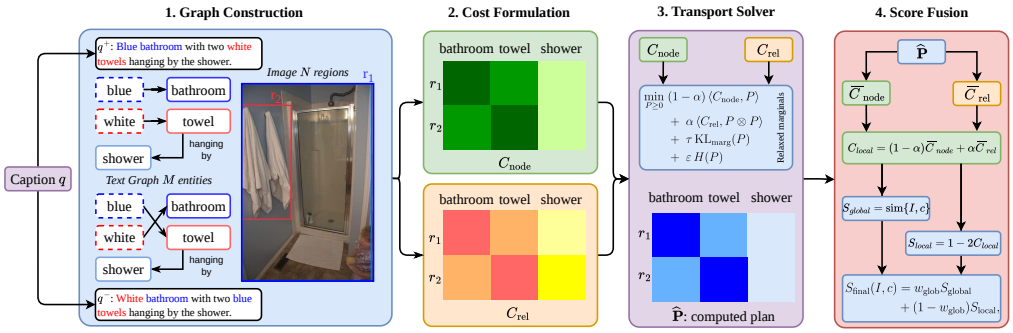

Figure 2: Overview of LEM. Caption-derived entities guide region extraction. Factorized node and directed relation costs define a structured unbalanced transport problem, whose alignment yields mass-normalized local evidence fused with global similarity. All components remain frozen. Notably, *shower* is missed by Grounding DINO here, a failure case (§5) that the unbalanced formulation still handles by withholding transport mass from the ungrounded entity.

Region graph. Grounding DINO (Liu et al., 2025) uses caption-derived entity phrases to extract N candidate boxes filtered by confidence thresholding and NMS. Their context-padded crops v_i form the nodes of $\mathcal{G}_{\text{img}}$. Its edges are ordered candidate pairs $(i, j) \in \mathcal{P}$ $(i \neq j)$. For each pair, v_{ij} is a crop of the rectangle covering both boxes, cached over unordered pairs since $v_{ij} \equiv v_{ji}$. The node counts N and M need not coincide. Cropping details are given in Appendix C.2.

3.2 Factorized Node Cost

The node cost separates entity identity from attribute binding. Entity compatibility is measured using the bare entity name:

$$C_{\text{ent}}(i, a) = \frac{1 - s(v_i, n_a)}{2}. \tag{1}$$

For an attribute-bound phrase q_a^+ and a contrasting phrase q_a^- obtained by replacing its attribute with an alternative from the caption, we compute

$$\begin{aligned} \Delta_{\text{attr}}(i, a) &= s(v_i, q_a^+) - s(v_i, q_a^-), \\ C_{\text{contrast}}(i, a) &= \sigma(-\beta_{\text{attr}} \Delta_{\text{attr}}(i, a)), \\ C_{\text{bound}}(i, a) &= \frac{1 - s(v_i, q_a^+)}{2}, \\ C_{\text{attr}}(i, a) &= \gamma_{\text{attr}} C_{\text{bound}}(i, a) + (1 - \gamma_{\text{attr}}) C_{\text{contrast}}(i, a), \end{aligned} \tag{2}$$

where σ is the logistic sigmoid, $\beta_{\text{attr}} > 0$ controls sensitivity to the contrast, and $\gamma_{\text{attr}} \in [0, 1]$ balances direct bound-phrase compatibility and contrastive evidence. For example, *blue bathroom* is compared with *white bathroom* when white is an alternative attribute in the caption. The combined node cost is

$$C_{\text{node}}(i, a) = (1 - \lambda_a) C_{\text{ent}}(i, a) + \lambda_a C_{\text{attr}}(i, a), \tag{3}$$

where $\lambda_a = \begin{cases} \lambda_{\text{attr}}, & \mathcal{A}_a \neq \emptyset, \\ 0, & \mathcal{A}_a = \emptyset, \end{cases}$ governed by global weight $\lambda_{\text{attr}} \in [0, 1]$. When an attributed entity has no valid contrasting attribute elsewhere in the caption, we use $C_{\text{attr}}(i, a) = C_{\text{bound}}(i, a)$ for that entity. Attribute construction and aggregation details are specified in Appendix C.3, with numerical settings given in Appendix C.8.

3.3 Directed Relation Cost

For each text edge $e = (a, r, b)$, let $q_e^{\rightarrow}$ express the stated relation in a standardized template (e.g., *"a photo of* n_a r n_b*"*) and $q_e^{\leftarrow}$ exchange its subject and object phrases (n_b and n_a). The pair crop provides a relative measure of support for the stated relation:

$$s_{\text{ctx}}(i, j, e) = \sigma(\beta_{\text{ctx}} [s(v_{ij}, q_e^{\rightarrow}) - s(v_{ij}, q_e^{\leftarrow})]). \tag{4}$$

Since $v_{ij} = v_{ji}$, this comparison alone cannot distinguish the two assignments of regions to a fixed text edge. We therefore add endpoint-role evidence. Let $t_a = n_a$ and $t_b = n_b$ denote the bare entity phrases used to score the individual regions:

$$\Delta_{\mathrm{dir}}(i,j,e) = \tfrac{1}{2}\left[s(v_i,t_a) - s(v_i,t_b) + s(v_j,t_b) - s(v_j,t_a)\right],$$
$$s_{\mathrm{dir}}(i,j,e) = \begin{cases} 1, & \Delta_{\mathrm{dir}}(i,j,e) \geq 0, \\ 2\sigma(\beta_{\mathrm{dir}}\Delta_{\mathrm{dir}}(i,j,e)), & \text{otherwise.} \end{cases} \tag{5}$$

This gate reduces support when the individual regions favor the reversed endpoint assignment. The relation cost is

$$C_{\mathrm{rel}}(i,j,e) = 1 - s_{\mathrm{ctx}}(i,j,e)s_{\mathrm{dir}}(i,j,e). \tag{6}$$

Both node and relation costs lie in $[0,1]$. The endpoint gate makes the relation cost sensitive to region order when the endpoint evidence distinguishes the assignments. Exact prompt construction is provided in Appendix C.1, with directional-cost details in Appendix C.4.

3.4 Structured Unbalanced Optimal Transport

Let $\mathbf{P} \in \mathbb{R}_+^{N\times M}$ be a soft alignment, where P_{ia} assigns transport mass from region i to entity a. We use uniform reference marginals $p_i = 1/N$ and $q_a = 1/M$. Unlike fixed-marginal transport (Peyré & Cuturi, 2019), unbalanced transport penalizes deviations from these references rather than enforcing exact equality (Chizat et al., 2017).

Define $\mathcal{S} = \{(i,j,e) : (i,j) \in \mathcal{P},\ e = (a,r,b) \in \mathcal{E}_{\mathrm{txt}}\}$. Combining node and pairwise terms follows the structured transport perspective of Vayer et al. (2019); here, the pairwise cost directly measures cross-modal relation compatibility. With entropic regularization (Cuturi, 2013), the objective is

$$\begin{aligned} \min_{\mathbf{P}\geq 0} \quad & (1-\alpha)\sum_{i,a} C_{\mathrm{node}}(i,a)P_{ia} + \alpha \sum_{(i,j,e)\in\mathcal{S}} C_{\mathrm{rel}}(i,j,e)P_{ia}P_{jb} \\ & + \tau\left[\mathrm{KL}(\mathbf{P}\mathbf{1}_M\|\mathbf{p}) + \mathrm{KL}(\mathbf{P}^\top\mathbf{1}_N\|\mathbf{q})\right] \\ & + \varepsilon\sum_{i,a} P_{ia}(\log P_{ia} - 1), \end{aligned} \tag{7}$$

where $\alpha \in [0,1]$, $\tau, \varepsilon > 0$, and $\mathrm{KL}(\mathbf{x}\|\mathbf{y}) = \sum_k [x_k \log(x_k/y_k) - x_k + y_k]$ is the generalized KL divergence, with $0\log 0 = 0$.

The relaxed marginals allow reduced mass on poorly supported correspondences, while the quadratic term couples pairs of assignments. The quadratic assignment term makes the objective generally non-convex, so the numerical procedure is not guaranteed to recover a global minimizer. In practice, LEM computes a numerical alignment plan $\widehat{\mathbf{P}}$ using an alternating procedure that combines linearized exponentiated mirror-descent updates with generalized Sinkhorn scaling. Optimization details, initialization, and stopping criteria are provided in Appendix C.5.

3.5 Local Scoring and Global Fusion

We score the resulting alignment using transport-weighted mean costs. Each term is normalized by its own contributing mass, removing the direct scale dependence on total transported mass

$$\overline{C}_{\mathrm{node}} = \frac{\sum_{i,a} C_{\mathrm{node}}(i,a)\widehat{P}_{ia}}{\sum_{i,a}\widehat{P}_{ia}} \tag{8}$$

and

$$\overline{C}_{\mathrm{rel}} = \frac{\sum_{(i,j,e)\in\mathcal{S}} C_{\mathrm{rel}}(i,j,e)\widehat{P}_{ia}\widehat{P}_{jb}}{\sum_{(i,j,e)\in\mathcal{S}}\widehat{P}_{ia}\widehat{P}_{jb}} \tag{9}$$

These normalizations are applied after optimization; the KL and entropy terms are excluded from the matching score. We combine the costs and map the result to $[-1,1]$, where

$$C_{\mathrm{local}} = (1-\alpha)\overline{C}_{\mathrm{node}} + \alpha\overline{C}_{\mathrm{rel}} \tag{10}$$

and

$$S_{\text{local}} = 1 - 2C_{\text{local}}. \tag{11}$$

When relational evidence is unavailable or its transport mass is negligible, we use $C_{\text{local}} = \overline{C}_{\text{node}}$. In the end,

$$S_{\text{final}}(I, c) = w_{\text{glob}} S_{\text{global}} + (1 - w_{\text{glob}}) S_{\text{local}}, \tag{12}$$

where $w_{\text{glob}} \in [0, 1]$ is fixed for each evaluation configuration. If no entities or regions are available, or the node transport mass is negligible, we use S_{global}. Numerical thresholds and fallback handling are specified in Appendix C.6, while evaluation-specific weight selection is described in Appendix C.7.

4 Experiments

4.1 Experimental Setup

Benchmarks. We evaluate LEM on two compositional benchmarks covering four diagnostic tasks: **(i) Winoground** (Thrush et al., 2022) (400 pairs / 800 instances) testing visio-linguistic counterfactual reasoning on paired $(I_0, I_1) \times (c_0, c_1)$ instances sharing identical lexical vocabulary; and **(ii) SugarCrepe** (Hsieh et al., 2023) assessing binary caption retrieval against artifact-free hard negatives across *swap attribute* (666 samples), *swap object* (246 samples), and *replace relation* (1,406 samples).

Baselines. We benchmark against: **(i) Global VLMs**, including zero-shot **CLIP** (ViT-B/32) (Radford et al., 2021) and hard-negative fine-tuned **NegCLIP** (Yuksekgonul et al., 2022); and **(ii) Region-based compositional models**, specifically **ComCLIP** (Jiang et al., 2024). LEM evaluates the fused score S_{final} combining global representations with structured local evidence across both backbones.

Protocol and Significance Testing. LEM is strictly **training-free**. Region candidates are generated via Grounding DINO Tiny (Liu et al., 2025) with box threshold 0.35 and padding $\rho = 0.10$. Dependency graphs $\mathcal{G}_{\text{txt}}$ are parsed via Stanza v1.14.0 under Universal Dependencies. Unbalanced transport plans are optimized using the alternating procedure described in Appendix C.5, combining linearized exponentiated mirror-descent updates with generalized Sinkhorn scaling ($\tau = 1.0$, $\epsilon = 0.005$, $T_{\max} = 200$). To eliminate proposal bias on SugarCrepe, proposals are extracted once per image from the joint prompt $c^+ \cup c^-$. We report **Group**, **Image**, and **Text Score** on Winoground, and pairwise **Accuracy** on SugarCrepe. Statistical validation against global baselines is measured via two-sided paired **McNemar's exact tests** and non-parametric percentile **Bootstrap 95% Confidence Intervals** ($B = 10{,}000$ resamples) on all net margins.

4.2 Main Results

As summarized in Table 1, LEM improves several entity- and attribute-sensitive compositional settings, with its clearest gains on Winoground image-level matching and SugarCrepe attribute binding, while performance remains mixed on object-swap and relation-replacement cases.

Fine-Grained Visual Grounding on Winoground. On Winoground, LEM delivers strong improvements on image-level visual grounding. Coupled with CLIP, LEM raises the Image Score from 9.25% to **14.00%** ($\Delta = +4.75\%$, net $+19$ examples rescued), achieving rigorous statistical significance ($p = 0.016 < 0.05$) with a strictly positive 95% confidence interval ($[+1.00\%, +8.50\%]$). This enhancement also transfers to the overall Group Score, reaching **10.50%** ($\Delta = +3.00\%$, $p = 0.088$), outperforming both NegCLIP and ComCLIP (8.00%).

Unlike similarity-based aggregation, LEM uses unbalanced optimal transport to permit reduced transport mass on weakly supported region–entity correspondences while preserving structured entity alignment.

Overcoming Attribute Binding Failures on SugarCrepe. On SugarCrepe *swap attribute*, where models must correctly associate modifiers with entities (e.g., *"the red car and blue bike"*), LEM achieves the highest accuracy across both backbones. With NegCLIP, LEM reaches **77.33%**

Table 1: **Main Results and Statistical Significance on Winoground and SugarCrepe.** All models evaluate frozen ViT-B/32 encoders. Best accuracy per block is **bolded**. Shaded rows highlight our proposed LEM framework along with paired Bootstrap 95% Confidence Intervals (CI) and two-sided McNemar p-values of paired differences against global baselines. † denotes statistically significant difference ($p < 0.05$); * denotes marginal significance ($p < 0.10$).

	Winoground (%)			SugarCrepe (%)		
Method	**Group**	**Image**	**Text**	**Swap-Att**	**Swap-Obj**	**Repl-Rel**
Backbone: OpenAI CLIP ViT-B/32						
Global Baseline (Radford et al., 2021)	7.50	9.25	**30.25**	66.52	63.01	71.34
ComCLIP (Jiang et al., 2024)	8.00	12.00	29.25	65.77	**63.41**	**73.19**
LEM (Ours)	**10.50***	**14.00**†	26.25	**68.32**	61.38	67.21
95% CI	$[-0.2, +6.3]$	$[+\mathbf{1.0}, +\mathbf{8.5}]$	$[-8.0, -0.3]$	$[-2.0, +5.6]$	$[-9.4, +6.1]$	$[-6.5, -1.8]$
McNemar p	0.088^{*}	$\mathbf{0.016}^{\dagger}$	0.060^{*}	0.388	0.747	$< 0.001^{\dagger}$
Backbone: NegCLIP ViT-B/32						
Global Baseline (Yuksekgonul et al., 2022)	8.00	11.25	**30.50**	75.23	**75.20**	76.46
ComCLIP (Jiang et al., 2024)	8.00	11.00	25.50	75.08	71.14	**77.10**
LEM (Ours)	**8.75**	**13.00**	28.50	**77.33***	72.76	74.25
95% CI	$[-1.5, +3.0]$	$[-1.0, +4.5]$	$[-5.3, +1.3]$	$[-0.2, +4.4]$	$[-6.9, +2.0]$	$[-3.8, -0.5]$
McNemar p	0.664	0.265	0.291	0.081^{*}	0.392	$0.013^{\dagger}$

($\Delta = +2.10\%$, $p = 0.081$, net $+14$ samples), outperforming ComCLIP+NegCLIP (75.08%). Under the CLIP backbone, ComCLIP degrades compared to global CLIP ($66.52\% \rightarrow 65.77\%$) when the counterfactual captions share identical words. In contrast, LEM explicitly binds attributes to candidate entity proposals via factorized node matching, lifting performance to **68.32%** ($+2.55\%$ over ComCLIP).

Understanding Relational and Object Reasoning Limits. On SugarCrepe *swap object*, LEM reaches 72.76% with NegCLIP, compared with 71.14% for ComCLIP+NegCLIP, while remaining below the 75.20% global NegCLIP baseline. On *replace relation*, global backbones remain stronger than their LEM counterparts. One plausible explanation is that nominal entities are directly localized by region proposals, whereas abstract predicates such as *"riding"* or *"walking near"* must be inferred from interaction context. Consequently, similar entity proposals may provide limited discrimination between competing relation captions. These results suggest that structured local evidence is more effective for some entity- and attribute-sensitive decisions than for relation replacement, where preserving global scene context remains important.

4.3 Ablation Studies

Entity–attribute balance in the node cost. We sweep $\rho \in \{0.0, 0.3, 0.5, 0.7, 1.0\}$ in $C^{\text{abl}}_{\text{node}}(i, a) = \rho C_{\text{ent}}(i, a) + (1 - \rho) C_{\text{attr}}(i, a)$ (Eq. equation 3), where $\rho = 0$ and $\rho = 1$ are the attribute-aware and entity-only endpoints, respectively. Relative to the entity-only endpoint, $\rho = 0.7$ is the only setting that improves all four tasks: Winoground group, swap-object, swap-attribute, and replace-relation increase by 1.00, 0.41, 5.41, and 0.50 points, respectively. Winoground alone peaks at $\rho = 0.3$ ($+1.50$ points), but this setting reduces swap-object and replace-relation by 2.03 and 0.21 points. Thus, the entity-dominant mixture at $\rho = 0.7$ provides the most consistent cross-task trade-off, showing that neither endpoint is sufficient.

Effect of relation weight. We sweep $\alpha \in \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$, which weights node and relation terms in both the structured UOT objective and the local matching cost, where $\alpha = 0$ uses node costs only and $\alpha = 1$ uses relation costs only. Figure 3(b) reports the change in final accuracy relative to the node-only endpoint. On Winoground, $\alpha = 0.40$ raises group accuracy from 8.00% to 10.50% ($+2.50$ points), while $\alpha = 0.20$ improves swap-object by 4.47 points. Swap-attribute and replace-relation generally deteriorate as relation evidence receives more weight, with replace-relation decreasing from 70.70% at $\alpha = 0$ to 64.30% at $\alpha = 1$. Thus, moderate relation weighting benefits selected compositional decisions, but relying too strongly on relation costs is brittle and motivates task-aware calibration.

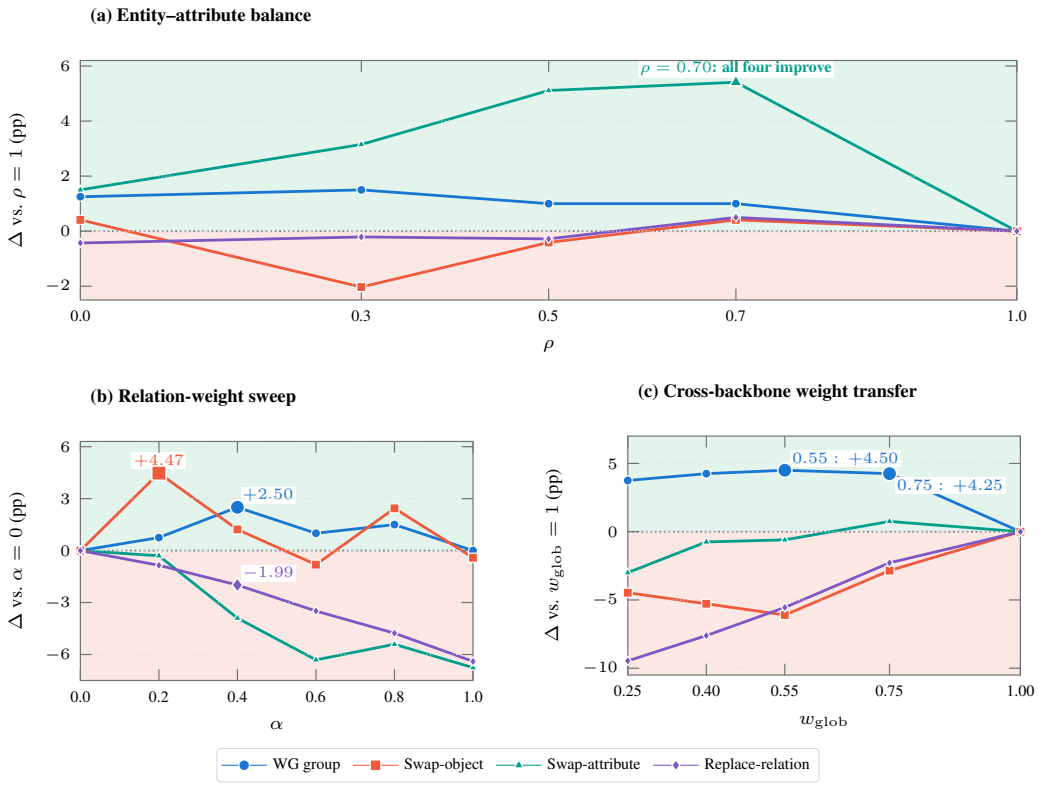

Figure 3: Three complementary ablations using final fused predictions. (a) Entity–attribute node-cost balance relative to the entity-only endpoint ($\rho = 1$), where $C^{\text{abl}}_{\text{node}}(i, a) = \rho C_{\text{ent}}(i, a) + (1 - \rho)C_{\text{attr}}(i, a)$. (b) Relation-weight sweep relative to node-only OT ($\alpha = 0$). (c) Cross-backbone evaluation of the two candidate weights ($w_{\text{glob}} = 0.55$ and 0.75) identified on CLIP ViT-B/32 and NegCLIP; intermediate weights show sensitivity. The pure-local endpoint ($w_{\text{glob}} = 0$) is omitted for readability (-28.09 points on replace-relation). Results are relative to global-only inference ($w_{\text{glob}} = 1$), and decreasing w_{glob} assigns more weight to the local score. Across all three panels, values above zero denote improvements over the corresponding reference, whereas values below zero denote performance degradation.

Cross-backbone evaluation of fusion weights. We test whether the two fusion weights previously identified on CLIP ViT-B/32 and NegCLIP, $w_{\text{glob}} = 0.55$ and 0.75, transfer to OpenCLIP ViT-L/14 without backbone-specific retuning. The former favors the strongest Winoground improvement, whereas the latter is a more conservative setting intended to balance Winoground and SugarCrepe. Figure 3(c) additionally shows intermediate weights to characterize sensitivity. We use $S_{\text{final}} = w_{\text{glob}} S_{\text{global}} + (1 - w_{\text{glob}}) S_{\text{local}}$, with $w_{\text{glob}} = 1$ as the global-only reference. Both transferred settings improve Winoground group accuracy, although neither transfers uniformly across the SugarCrepe tasks. At $w_{\text{glob}} = 0.55$, Winoground group accuracy improves by 4.50 points (paired bootstrap 95% CI $[1.50, 7.75]$; exact McNemar $p = 0.008$). At $w_{\text{glob}} = 0.75$, the gain remains comparable at 4.25 points, while the SugarCrepe trade-off becomes less severe: swap-attribute improves by 0.75 points, and the swap-object and replace-relation losses decrease to 2.85 and 2.28 points, respectively. Thus, the local signal transfers to Winoground, but not uniformly across tasks.

5 Grounding DINO Localization Error Analysis on SugarCrepe

§2 (Region-Level Grounding and Scene Graphs) argues that open-vocabulary detectors such as Grounding DINO (Liu et al., 2023) reliably localize nominal entities but do not themselves verify directed relations. We quantify the *entity-localization* side of this claim directly: we measure how often Grounding DINO Tiny, under the exact detector configuration used throughout our experiments ($\theta_{\text{box}} = 0.35$, $\theta_{\text{text}} = 0.25$; Appendix C.2), recovers the correct region for entities mentioned

in SugarCrepe (Hsieh et al., 2023) captions, using COCO-2017 validation annotations (Lin et al., 2014) as ground truth (sugacrepe is derived from the COCO-2017 validation set). Error analysis protocol in Appendix D.

Table 2: Grounding DINO localization performance on SugarCrepe entities against COCO-2017 val ground truth (IoU ≥ 0.5, greedy per-category matching). Evaluated on 6,108 samples with at least one COCO-mapped target (81.3% of the 7,511-sample manifest).

Metric	Value
True Positives (TP)	12,600
False Positives (FP)	2,641
False Negatives (FN)	10,967
Precision $= \mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$	82.67%
Recall $= \mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$	53.46%
F1	64.94%
Miss Rate $= \mathrm{FN}/(\mathrm{TP}+\mathrm{FN})$	46.54%
False Discovery Rate $= \mathrm{FP}/(\mathrm{TP}+\mathrm{FP})$	17.33%
Overall Error Rate $= (\mathrm{FP}+\mathrm{FN})/(\mathrm{TP}+\mathrm{FP}+\mathrm{FN})$	51.92%
Mean IoU on matched pairs	91.35%

Two results directly support the region-grounding assumptions underlying LEM. First, the mean IoU on matched pairs is 91.35%, indicating that when Grounding DINO does localize the entity referred to in a caption, the resulting box is tightly aligned with the true object extent – consistent with entity-level grounding being a comparatively reliable signal (§ 4.2). Second, the false discovery rate is a non-trivial 17.33%: roughly one in six predicted boxes does not correspond to any ground-truth instance of the stated category. This empirically motivates the relaxed-marginal formulation of § 3.4 (Eq. 7), which allows unbalanced transport mass to be withheld from such spurious, ungrounded proposals rather than forcing them into the alignment.

The recall figure (53.46%) should be read with one important caveat. Our matching protocol treats *every* ground-truth instance of a mentioned category within an image as a valid target, whereas a SugarCrepe caption typically refers to a single specific instance (e.g., one person in a crowd scene). Consequently, correctly localizing the referenced instance while missing other, unmentioned instances of the same category still counts as multiple false negatives here. This is a structural property of category-level (rather than instance-level) matching and is expected to underestimate Grounding DINO's true referential recall; we report it primarily for its bearing on the aggregate false-negative volume rather than as a precise estimate of per-caption grounding failure. Coverage is similarly restricted to the 81.3% of samples whose caption entities map onto the fixed 80-category COCO vocabulary, so these statistics characterize localization behavior on commonly-named, COCO-nameable objects rather than the full lexical diversity of SugarCrepe.

6 Conclusion

We presented LEM, a training-free framework that complements global image–text similarity with structured local evidence for compositional matching. LEM parses captions into directed semantic graphs, grounds them against caption-guided region proposals, and casts entity correspondence and relation consistency as a single structured unbalanced optimal transport problem, combining a factorized node cost for entity identity and attribute binding with a directed relation cost for pairwise interaction, while relaxed marginal constraints downweight noisy or ungrounded proposals. Across Winoground and SugarCrepe, this structured local evidence, fused with global similarity, improves image-level visual grounding and attribute-binding accuracy over both CLIP and NegCLIP backbones, and our ablations show that a moderate, entity-leaning node cost and directed relation costs drive most of these gains, with fusion weights transferring reasonably well across backbones. At the same time, gains on object swaps and relation replacement are inconsistent, and our localization error analysis on SugarCrepe reveals non-trivial detector misses and false positives, indicating that relation-level composition remains more dependent on proposal quality and holistic context.

AI USE STATEMENT

Generative AI tools were used to assist with drafting and editing the manuscript. All AI-assisted suggestions and outputs were reviewed by the authors for correctness, relevance, and originality. The authors take full responsibility for the final methodology, experiments, claims, and presentation.

REFERENCES

Liqun Chen, Zhe Gan, Yu Cheng, Linjie Li, Lawrence Carin, and Jingjing Liu. Graph optimal transport for cross-domain alignment. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1542–1553. PMLR, 13–18 Jul 2020.

Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced transport problems, 2017.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

Anuj Diwan, Layne Berry, Eunsol Choi, David Harwath, and Kyle Mahowald. Why is winoground hard? investigating failures in visuolinguistic compositionality, 2022.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 31096–31116. Curran Associates, Inc., 2023. doi: 10.52202/075280-1355.

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, pp. 2417–2425, 2024.

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 Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4904–4916. PMLR, 18–24 Jul 2021.

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)*, pp. 6639–6659, 2024.

Amita Kamath, Cheng-Yu Hsieh, Kai-Wei Chang, and Ranjay Krishna. The hard positive truth about vision-language compositionality. In *European Conference on Computer Vision*, pp. 37–54. Springer, 2024.

Pawan Kumar. ConceptOT: Fine-grained vision-language alignment via low-rank unbalanced optimal transport. In *Third Workshop on Visual Concepts*, 2026.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12888–12900. PMLR, 17–23 Jul 2022a.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10965–10975, 2022b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll'ar, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 38–55, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-72970-6.

Facundo Mémoli. Gromov–wasserstein distances and the metric approach to object matching. *Foundations of Computational Mathematics*, 11:417–487, 2011.

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)*, pp. 8253–8280, 2022.

Gabriel Peyré and Marco Cuturi. Computational optimal transport with applications to data sciences. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 02 2019. ISSN 1935-8237. doi: 10.1561/2200000073.

Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-wasserstein averaging of kernel and distance matrices. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2664–2672, New York, New York, USA, 20–22 Jun 2016. PMLR.

Nirmalendu Prakash, Narmeen Fatimah Oozeer, Xin Su, Phillip Howard, Shaan Shah, Zoe Wanying He, Shuang Wu, Shivam Raval, Roy Ka-Wei Lee, Meenakshi Khosla, and Amir Abdullah. Geometry-aware clip retrieval via local cross-modal alignment and steering, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021.

Liuyang Song, Yi Zhang, Zhongyi Deng, Daqian Yang, and Hongbo Zhang. Bindclip: One balanced coupling for compositional vision language scoring, 2026.

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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5228–5238, Los Alamitos, CA, USA, June 2022. IEEE Computer Society. doi: 10.1109/CVPR52688.2022.00517.

Titouan Vayer, Laetitia Chapel, Rémi Flamary, Romain Tavenard, and Nicolas Courty. Optimal transport for structured data with application on graphs, 2019.

Danfei Xu, Yuke Zhu, Christopher B. Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5410–5419, 2017.

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. In *International Conference on Learning Representations*, 2022.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Y. Zou. When and why vision-language models behave like bags-of-words, and what to do about it? *ArXiv*, abs/2210.01936, 2022.

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.

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16793–16803, 2022.

A FORMAL DEFINITIONS

Throughout, Σ^* denotes the set of finite natural-language strings (phrases), and $\sigma(z) = 1/(1+e^{-z})$ denotes the standard logistic sigmoid function.

Definition 1 (Similarity function)**.** *Let $f_{\mathrm{v}} : \mathcal{I} \to \mathbb{S}^{D-1}$ and $f_{\mathrm{t}} : \Sigma^* \to \mathbb{S}^{D-1}$ denote the frozen, ℓ_2-normalized image and text encoders, respectively, where $\mathcal{I}$ is the space of visual image crops and $\mathbb{S}^{D-1} \subset \mathbb{R}^D$ is the unit hypersphere. The multimodal similarity function $s : \mathcal{I} \times \Sigma^* \to [-1, 1]$ is defined by*

$$s(v, t) \;=\; f_{\mathrm{v}}(v)^\top f_{\mathrm{t}}(t). \tag{13}$$

All local matching costs in § 3 are obtained by composing s with affine or sigmoid transformations mapping into $[0, 1]$.

Definition 2 (Caption graph)**.** *The caption graph is a directed, vertex- and edge-labeled graph $\mathcal{G}_{\mathrm{txt}} = (\mathcal{V}_{\mathrm{txt}}, \mathcal{E}_{\mathrm{txt}})$ with entity node set $\mathcal{V}_{\mathrm{txt}} = \{u_1, \ldots, u_M\}$. Each node is a tuple $u_a = (n_a, \mathcal{A}_a)$, where $n_a \in \Sigma^*$ is the bare entity head noun and $\mathcal{A}_a$ is a (possibly empty) finite set of entity-bound attribute tokens (adjectival or numerical modifiers) extracted from caption c. A directed relation edge $e = (a, r, b) \in \mathcal{E}_{\mathrm{txt}} \subseteq \{1, \ldots, M\} \times \mathcal{R} \times \{1, \ldots, M\}$, with $a \neq b$ and relation predicate $r \in \mathcal{R}$, assigns entity u_a the subject role and entity u_b the object role. Compound nouns are contracted into single entities prior to graph construction, and passive constructions are canonicalized to this active subject–object convention.*

Definition 3 (Region graph and pair crops)**.** *Given image I and caption-derived entity queries, Grounding DINO returns N candidate bounding boxes $B_1, \ldots, B_N \subset \mathbb{R}^4$ filtered by confidence thresholding and non-maximum suppression (NMS). The region graph is defined as $\mathcal{G}_{\mathrm{img}} = (\mathcal{V}_{\mathrm{img}}, \mathcal{P})$, where $\mathcal{V}_{\mathrm{img}} = \{v_1, \ldots, v_N\}$ contains the context-padded visual crops corresponding to $\{B_i\}_{i=1}^N$ with padding ratio $\rho = 0.10$. The visual edge set $\mathcal{P}$ comprises all ordered pairs of distinct visual regions:*

$$\mathcal{P} \;=\; \big\{(i, j) \in \{1, \ldots, N\}^2 \;:\; i \neq j\big\}, \tag{14}$$

yielding $|\mathcal{P}| = N(N-1)$ directed candidate edges. For each pair $(i, j) \in \mathcal{P}$, the pair crop $v_{ij} \in \mathcal{I}$ is the visual crop of the minimal unpadded axis-aligned bounding box enclosing $B_i \cup B_j$. By construction of the bounding union, $v_{ij} = v_{ji}$, allowing visual embeddings to be cached across the $\binom{N}{2}$ unordered pairs.

Definition 4 (Phrase templates)**.** *Let $\pi : \bigcup_k \mathcal{X}_k \to \Sigma^*$ denote deterministic templating functions rendering structured graph elements into natural-language phrases for encoding by f_{t}:*

- $n_a = \pi_{\mathrm{ent}}(u_a) \in \Sigma^*$*: the bare entity name (e.g.,* "dog"*).*
- $q_a^+ = \pi_{\mathrm{attr}}(u_a, \mathrm{attr}) \in \Sigma^*$*, for* $\mathrm{attr} \in \mathcal{A}_a$*: an attribute-bound phrase asserted in caption c (e.g.,* "black dog"*).*
- $q_a^- = \pi_{\mathrm{attr}}(u_a, \mathrm{attr}') \in \Sigma^*$*, for a competing attribute* $\mathrm{attr}' \in \mathcal{A}_a^{\mathrm{alt}}$*, where $\mathcal{A}_a^{\mathrm{alt}} = \bigcup_{b \neq a} \mathcal{A}_b \setminus \mathcal{A}_a$ is the set of alternative attributes attached to other entities in c (e.g.,* "white dog"*). When an attributed entity has no valid contrasting attribute elsewhere in the caption ($\mathcal{A}_a^{\mathrm{alt}} = \emptyset$), we set $C_{\mathrm{attr}}(i, a) = C_{\mathrm{bound}}(i, a)$ for that entity. The attribute weight is $\lambda_a = \lambda_{\mathrm{attr}}$ if $\mathcal{A}_a \neq \emptyset$ and 0 otherwise, governed by global hyperparameter $\lambda_{\mathrm{attr}} \in [0, 1]$.*
- $q_e^{\rightarrow} = \pi_{\mathrm{rel}}(n_a, r, n_b) \in \Sigma^*$ *and* $q_e^{\leftarrow} = \pi_{\mathrm{rel}}(n_b, r, n_a) \in \Sigma^*$*: natural-language relational prompts expressing edge $e = (a, r, b)$ in forward and reversed subject–object order, respectively.*

Remark 1 (Endpoint representations t_a, t_b)**.** *In Eq. equation 5, visual endpoints v_i and v_j are evaluated against endpoint representations t_a, t_b. We adopt bare entity names $t_a := n_a$ and $t_b := n_b$. This formulation isolates semantic identity and spatial roles from adjectival modifiers, ensuring that the directional gate s_{dir} remains orthogonal to attribute binding while avoiding artificial coupling between relation scoring and the attribute weight λ_a.*

Definition 5 (Structured alignment index set)**.** *The structured alignment index set couples directed visual region pairs with directed textual relations:*

$$\mathcal{S} \;=\; \big\{\,(i, j, e) \;:\; (i, j) \in \mathcal{P},\; e = (a, r, b) \in \mathcal{E}_{\mathrm{txt}}\,\big\}. \tag{15}$$

Each element $(i, j, e) \in \mathcal{S}$ names an alignment hypothesis wherein region i is assigned to subject a, region j is assigned to object b, and pair crop v_{ij} evaluates predicate r. The total number of relational candidates evaluated is $|\mathcal{S}| = |\mathcal{P}| \cdot |\mathcal{E}_{\text{txt}}| = N(N-1)|\mathcal{E}_{\text{txt}}|$.

Definition 6 (Transport plans and reference marginals)**.** *For N visual regions and M text entities, the space of non-negative transport plans is $\mathbb{R}_+^{N\times M}$, where $P_{ia} \geq 0$ denotes the transport mass assigned between region i and entity a. The reference marginals are uniform vectors $\mathbf{p} = (1/N, \ldots, 1/N)^\top \in \Delta_N$ and $\mathbf{q} = (1/M, \ldots, 1/M)^\top \in \Delta_M$, where Δ_K denotes the probability simplex in $\mathbb{R}^K$. For non-negative vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^K$, the generalized Kullback–Leibler divergence is defined as*

$$\mathrm{KL}(\mathbf{x}\|\mathbf{y}) \;=\; \sum_{k=1}^{K}\left[x_k \log\left(\frac{x_k}{y_k}\right) - x_k + y_k\right], \tag{16}$$

with the standard continuous convention $0 \log 0 = 0$ and $\mathrm{KL}(\mathbf{x}\|\mathbf{y}) = +\infty$ if $x_k > 0 = y_k$. Rather than constraining the marginals of $\mathbf{P}$ to strictly satisfy $\mathbf{P}\mathbf{1}_M = \mathbf{p}$ and $\mathbf{P}^\top \mathbf{1}_N = \mathbf{q}$, unbalanced optimal transport penalizes marginal discrepancies via $\tau\,\mathrm{KL}(\cdot\|\cdot)$, naturally permitting mass deflation on poorly grounded visual proposals or textual entities.

Definition 7 (Cost functions and range invariance)**.** *The node cost matrix $C_{\text{node}} \in [0,1]^{N\times M}$ and sparse relation cost tensor $C_{\text{rel}} \in [0,1]^{|\mathcal{S}|}$ are given in Eq. equation 3 and Eq. equation 6. Both costs are strictly bounded in $[0, 1]$:*

- *(i) $C_{\text{ent}}(i, a) = \frac{1-s(v_i,n_a)}{2} \in [0, 1]$ since $s \in [-1, 1]$;*
- *(ii) $C_{\text{bound}}(i, a) \in [0, 1]$ is an affine rescaling of s, and $C_{\text{contrast}}(i, a) = \sigma(-\beta_{\text{attr}}\Delta_{\text{attr}}) \in (0, 1)$. When $\mathcal{A}_a^{\text{alt}} \neq \emptyset$, their convex combination satisfies $C_{\text{attr}}(i, a) = \gamma_{\text{attr}} C_{\text{bound}} + (1 - \gamma_{\text{attr}})C_{\text{contrast}} \in [0, 1]$; otherwise, $C_{\text{attr}}(i, a) = C_{\text{bound}}(i, a) \in [0, 1]$;*
- *(iii) $C_{\text{node}}(i, a) = (1 - \lambda_a)C_{\text{ent}}(i, a) + \lambda_a C_{\text{attr}}(i, a)$ is a convex combination of two $[0, 1]$-valued costs, hence $C_{\text{node}}(i, a) \in [0, 1]$;*
- *(iv) $s_{\text{ctx}}(i, j, e) = \sigma(\cdot) \in (0, 1)$, and $s_{\text{dir}}(i, j, e) \in (0, 1]$ by piecewise definition (1.0 when $\Delta_{\text{dir}} \geq 0$, and $2\sigma(\beta_{\text{dir}}\Delta_{\text{dir}}) \in (0, 1)$ when $\Delta_{\text{dir}} < 0$). Therefore, the product $s_{\text{ctx}}s_{\text{dir}} \in (0, 1)$, ensuring $C_{\text{rel}}(i, j, e) = 1 - s_{\text{ctx}}s_{\text{dir}} \in (0, 1) \subset [0, 1]$.*

Definition 8 (Structured unbalanced transport problem)**.** *Given costs C_{node}, C_{rel}, index set $\mathcal{S}$, reference marginals $\mathbf{p}, \mathbf{q}$, and hyperparameters $\alpha \in [0, 1]$, $\tau, \varepsilon > 0$, the theoretical structured UOT problem is:*

$$\mathbf{P}^\star \;\in\; \operatorname*{arg\,min}_{\mathbf{P}\in\mathbb{R}_+^{N\times M}} \mathcal{F}(\mathbf{P}), \tag{17}$$

where $\mathcal{F}$ is the regularized objective in Eq. equation 7. Since the node and relation costs are non-negative and the KL and entropic terms grow superlinearly as the transport mass diverges, $\mathcal{F}$ is coercive and lower semicontinuous on $\mathbb{R}_+^{N\times M}$; hence a global minimizer $\mathbf{P}^\star$ exists. Because the quadratic relation term makes $\mathcal{F}$ generally non-convex, the numerical solver is not guaranteed to recover $\mathbf{P}^\star$; in practice, LEM computes a numerical alignment plan $\widehat{\mathbf{P}}$ via the iterative procedure in Appendix C.5.

Definition 9 (Mass-normalized local evidence costs and thresholding)**.** *Let $\xi_{\text{mass}} = 10^{-6}$ denote the threshold for negligible transport mass (Table 3). For the computed plan $\widehat{\mathbf{P}}$, the total transported node mass and relation mass are:*

$$m_{\text{node}} \;=\; \sum_{i=1}^{N}\sum_{a=1}^{M} \widehat{P}_{ia}, \qquad m_{\text{rel}} \;=\; \sum_{(i,j,e)\in\mathcal{S}} \widehat{P}_{ia}\widehat{P}_{jb}, \quad \text{with } e = (a, r, b). \tag{18}$$

When $m_{\text{node}} > \xi_{\text{mass}}$ and $m_{\text{rel}} > \xi_{\text{mass}}$, the normalized costs $\overline{C}_{\text{node}}$ and $\overline{C}_{\text{rel}}$ are given by Eq. equation 8. The local evidence cost is defined piecewise:

$$C_{\text{local}} \;=\; \begin{cases} (1-\alpha)\,\overline{C}_{\text{node}} + \alpha\,\overline{C}_{\text{rel}}, & \text{if } \mathcal{S} \neq \emptyset \text{ and } m_{\text{rel}} > \xi_{\text{mass}}, \\ \overline{C}_{\text{node}}, & \text{otherwise}, \end{cases} \tag{19}$$

and the local similarity score is $S_{\text{local}} = 1 - 2C_{\text{local}} \in [-1, 1]$. For numerical safety in finite-precision arithmetic, the implementation evaluates $S_{\text{local}} = \mathrm{clip}(1 - 2C_{\text{local}}, -1.0, 1.0)$; analytically, the unclipped value already lies in this interval.

Definition 10 (Final matching score and fallback guarantee)**.** *For global fusion weight* $w_{\text{glob}} \in [0,1]$*, the unified compositional matching score is:*

$$S_{\text{final}}(I,c) \;=\; \begin{cases} S_{\text{global}}, & \text{if } N=0 \text{ or } M=0 \text{ or } m_{\text{node}} \le \xi_{\text{mass}}, \\ w_{\text{glob}}\, S_{\text{global}} + (1-w_{\text{glob}})\, S_{\text{local}}, & \text{otherwise}. \end{cases} \tag{20}$$

Because $S_{\text{global}}, S_{\text{local}} \in [-1,1]$ *and the non-fallback branch is a convex combination of both,* $S_{\text{final}}(I,c)$ *is strictly bounded in* $[-1,1]$ *across both operational branches, ensuring range compatibility between the two operational branches.*

Definition 11 (Hyperparameter configuration)**.** *Given frozen feature encoders* $(f_{\text{v}}, f_{\text{t}})$*, a zero-shot detector, and a parser, the inference pipeline is fully characterized by the configuration tuple:*

$$\Theta \;=\; \big(\alpha,\ \tau,\ \varepsilon,\ \beta_{\text{attr}},\ \beta_{\text{ctx}},\ \beta_{\text{dir}},\ \gamma_{\text{attr}},\ \lambda_{\text{attr}},\ w_{\text{glob}},\ \xi_{\text{mass}}\big), \tag{21}$$

together with the visual context padding $\rho = 0.10$*, detector thresholds* $(\theta_{\text{box}}, \theta_{\text{text}}) = (0.35, 0.25)$*, and NMS threshold* $\theta_{\text{NMS}} = 0.50$ *specified in Appendix C.2. The structural parameters* $(\alpha, \lambda_{\text{attr}}, w_{\text{glob}})$ *are evaluation-specific (Appendix C.7); all remaining parameters in* Θ *are fixed across all benchmarks as reported in Table 3.*

B Preliminaries and Theoretical Background

This section provides the mathematical foundation and conceptual background underlying LEM, spanning discrete optimal transport, unbalanced entropic regularizations, structured quadratic transport, and the compositional limits of contrastive vision–language models.

B.1 Discrete and Entropic Optimal Transport

Kantorovich formulation. Let $\mathbf{p} \in \Delta_N$ and $\mathbf{q} \in \Delta_M$ denote discrete probability distributions over visual proposals and textual entities, where $\Delta_K = \{\mathbf{x} \in \mathbb{R}_+^K : \sum_{k=1}^K x_k = 1\}$ represents the standard probability simplex. Given a ground cost matrix $\mathbf{C} \in \mathbb{R}_+^{N\times M}$, the classical Kantorovich optimal transport problem (Peyré & Cuturi, 2019) seeks a coupling plan $\mathbf{P} \in \mathbb{R}_+^{N\times M}$ that minimizes total transport cost:

$$\min_{\mathbf{P}\in\Pi(\mathbf{p},\mathbf{q})} \langle \mathbf{P}, \mathbf{C}\rangle \;=\; \sum_{i=1}^{N}\sum_{a=1}^{M} C_{ia}P_{ia}, \tag{22}$$

where $\Pi(\mathbf{p},\mathbf{q}) = \{\mathbf{P} \in \mathbb{R}_+^{N\times M} : \mathbf{P}\mathbf{1}_M = \mathbf{p},\ \mathbf{P}^\top \mathbf{1}_N = \mathbf{q}\}$ defines the bounded Kantorovich transportation polytope enforcing strict marginal conservation.

Entropic regularization and Sinkhorn scaling. To avoid repeatedly solving the full linear program and enable efficient iterative optimization, entropic regularization yields the Sinkhorn scaling formulation (Cuturi, 2013):

$$\min_{\mathbf{P}\in\Pi(\mathbf{p},\mathbf{q})} \langle \mathbf{P}, \mathbf{C}\rangle + \varepsilon \sum_{i,a} P_{ia}(\log P_{ia} - 1), \tag{23}$$

where $\varepsilon > 0$ controls entropic smoothing. By Fenchel duality, the unique optimal solution takes the factorized Gibbs form:

$$P^*_{ia} \;=\; u_i K_{ia} v_a, \qquad \mathbf{K} = \exp(-\mathbf{C}/\varepsilon), \tag{24}$$

where $\mathbf{u} \in \mathbb{R}_+^N$ and $\mathbf{v} \in \mathbb{R}_+^M$ are non-negative scaling dual vectors. They are computed efficiently via Sinkhorn–Knopp alternating fixed-point iterations:

$$\mathbf{u} \leftarrow \frac{\mathbf{p}}{\mathbf{K}\mathbf{v}}, \qquad \mathbf{v} \leftarrow \frac{\mathbf{q}}{\mathbf{K}^\top \mathbf{u}}. \tag{25}$$

B.2 Unbalanced Optimal Transport via Marginal Relaxation

Motivation for unbalanced alignment. In open-world vision–language reasoning, bounding-box detectors inherently produce variable proposal counts N, frequently capturing background clutter,

spurious objects, or redundant overlaps. Conversely, text captions contain abstract or ungrounded concepts (e.g., *"mood"*, *"atmosphere"*) that have no physical counterpart in the image. Enforcing hard marginal constraints $\mathbf{P}\mathbf{1}_M = \mathbf{p}$ and $\mathbf{P}^\top \mathbf{1}_N = \mathbf{q}$ forces transport mass onto spurious correspondences, distorting alignment fidelity.

Csiszár marginal penalties. Unbalanced Optimal Transport (UOT) resolves this pathology by relaxing hard equality constraints into convex marginal divergence penalties (Chizat et al., 2017; Peyré & Cuturi, 2019). Using generalized Kullback–Leibler (KL) divergence penalties, the entropic UOT problem is:

$$\min_{\mathbf{P}\geq 0} \langle \mathbf{P}, \mathbf{C}\rangle + \tau_1 \,\mathrm{KL}(\mathbf{P}\mathbf{1}_M \| \mathbf{p}) + \tau_2 \,\mathrm{KL}(\mathbf{P}^\top \mathbf{1}_N \| \mathbf{q}) + \varepsilon \sum_{i,a} P_{ia}(\log P_{ia} - 1), \tag{26}$$

where $\tau_1, \tau_2 > 0$ control marginal softness. For uniform references $\mathbf{p} = \frac{1}{N}\mathbf{1}_N$ and $\mathbf{q} = \frac{1}{M}\mathbf{1}_M$, Chizat et al. (2017) demonstrated that the dual scaling updates preserve the matrix-scaling structure of Sinkhorn's algorithm, with scaled exponentiation:

$$u_i \leftarrow \left(\frac{p_i}{[\mathbf{K}\mathbf{v}]_i}\right)^{\frac{\tau_1}{\tau_1+\varepsilon}}, \qquad v_a \leftarrow \left(\frac{q_a}{[\mathbf{K}^\top \mathbf{u}]_a}\right)^{\frac{\tau_2}{\tau_2+\varepsilon}}. \tag{27}$$

As $\tau \to \infty$, the scaling exponent $\kappa = \frac{\tau}{\tau+\varepsilon} \to 1$, smoothly recovering standard balanced Sinkhorn transport. For finite τ, $\kappa < 1$ relaxes exact marginal matching, allowing the transported mass to adapt to the underlying costs.

B.3 Structured Quadratic Transport and Graph Alignment

While linear optimal transport captures first-order feature similarities $\mathbf{C}_{\mathrm{node}}$, it is invariant to relational structure and geometric transformations. The Gromov–Wasserstein (GW) discrepancy (Mémoli, 2011; Peyré et al., 2016) addresses this by seeking an alignment that matches intra-graph pairwise affinities:

$$\min_{\mathbf{P}\in\Pi} \sum_{i,j,a,b} \mathcal{L}(D^{\mathrm{img}}_{ij}, D^{\mathrm{txt}}_{ab}) P_{ia} P_{jb}, \tag{28}$$

where D^{img} and D^{txt} measure intra-domain geodesic or structural distances. Fused Gromov–Wasserstein (FGW) (Vayer et al., 2019) combines linear node costs and quadratic intra-domain structure:

$$\min_{\mathbf{P}\in\Pi} (1-\alpha)\langle \mathbf{P}, \mathbf{C}_{\mathrm{node}}\rangle + \alpha \sum_{i,j,a,b} |D^{\mathrm{img}}_{ij} - D^{\mathrm{txt}}_{ab}|^2 P_{ia} P_{jb}. \tag{29}$$

LEM's Cross-Modal Pairwise Formulation. Unlike standard FGW, which compares ungrounded intra-modal distances (D^{img}_{ij} vs D^{txt}_{ab}), LEM directly formulates a *cross-modal relational compatibility* $C_{\mathrm{rel}}(i, j, e)$ conditioned on a contextual pair crop v_{ij} and directional relation templates $(q^{\rightarrow}_e, q^{\leftarrow}_e)$. This couples pairs of node assignments through explicit multimodal relation semantics rather than abstract geometric metric comparisons.

B.4 Compositional Limitations of Pretrained Vision–Language Models

Contrastive vision–language backbones such as CLIP (Radford et al., 2021) embed entire images I and captions c into a shared unit sphere $\mathbb{S}^{D-1}$ via encoders f_{v} and f_{t}, trained with InfoNCE loss over web-scraped image–text pairs:

$$\mathcal{L}_{\mathrm{InfoNCE}} \;=\; -\log \frac{\exp(f_{\mathrm{v}}(I)^\top f_{\mathrm{t}}(c)/\tau_{\mathrm{clip}})}{\sum_{c'} \exp(f_{\mathrm{v}}(I)^\top f_{\mathrm{t}}(c')/\tau_{\mathrm{clip}})}. \tag{30}$$

Despite strong zero-shot retrieval accuracy, diagnostic benchmarks (Winoground (Thrush et al., 2022), ARO (Yuksekgonul et al., 2022), SugarCrepe (Hsieh et al., 2023)) show that global image–text similarities can be insensitive to fine-grained compositional binding:

- **Attribute binding deficit:** Swapping attributes among objects (e.g., *"the red car and the blue bike"* vs *"the blue car and the red bike"*) yields near-identical global similarities.

- **Directed relation deficit:** Reversing directed action predicates (e.g., *"the horse kicks the man"* vs *"the man kicks the horse"*) cannot be resolved reliably because representations are dominated by entity presence rather than directed syntactic structure.

These diagnostic findings motivate LEM's structured decomposition: extracting localized evidence via factorized node costs and directed relation gates, grounded through unbalanced graph alignment.

C Additional Methodological Details

C.1 Caption Parsing and Prompt Construction

Caption parsing and dependency rules. We construct the caption graph using Stanza v1.14.0 with its English dependency-parsing pipeline (`tokenize`, `pos`, `lemma`, `depparse`). Graph construction follows deterministic linguistic rules over the Universal Dependencies (UD) tree:

- **Entities:** Extracted from head words with POS tag `NOUN` or `PROPN`. Multi-word nominal expressions connected via `compound` or `flat` dependencies are merged into a single entity string (e.g., *"water bottle"*). Groundable indefinite human pronouns (*"someone"*, *"somebody"*, *"anyone"*) are canonicalized to the entity name *"person"*. Repeated mentions receive distinct instance indices ($u_0, u_1, \dots$) to support duplicate entities.
- **Attributes:** Extracted from adjectival (`amod`) and numerical (`nummod`) dependents. Comparative and superlative adjectives (`Degree=Cmp`, `Degree=Sup`) retain surface inflections, whereas descriptive adjectives take lowercase lemmas.
- **Verb relations:** Verbs with POS `VERB` (or participial clauses `acl`, `advcl`) governing direct objects (`obj`) and nominal subjects (`nsubj`) form active relation edges $(u_{\text{subj}}, r, u_{\text{obj}})$. Participial modifiers without an explicit subject attach to their governing noun head.
- **Passive voice normalization:** When a verb governs a passive subject (`nsubj:pass`) and an agent nominal introduced via an oblique prepositional phrase (`obl` with `case = "by"`), the relation is canonicalized to active voice: $(u_{\text{agent}}, r, u_{\text{subj:pass}})$.
- **Prepositional and spatial pivot relations:** Prepositional dependents marked by `case` are attached to objects. Locative pivot nouns (*"front"*, *"back"*, *"top"*, *"bottom"*, *"side"*, *"middle"*, *"center"*) are absorbed into spatial predicate phrases (e.g., *"in front of"*, *"on top of"*), and the relation subject is resolved by ascending the dependency tree to the nearest entity head.

If parsing yields no groundable entity, an empty graph is returned, and the scoring pipeline falls back to global similarity S_{global}.

Detection phrases. Region proposal is conditioned on caption-derived entity phrases. When an entity has associated attributes, an attribute-bound noun phrase is used; otherwise, the bare entity name is retained. The resulting phrases are concatenated with period delimiters (e.g., *"black dog . white cat ."*) and supplied to Grounding DINO.

Attribute contrast construction. For entity u_a with bare name n_a, let $\mathcal{Q}_a^+ = \{q_{a,k}^+\}_{k=1}^{K_a}$ denote its valid attribute-bound phrases. Competing attributes are taken from attributes assigned to other entities in the same caption and rebound to the current entity name, yielding $\mathcal{Q}_a^- = \{q_{a,m}^-\}_{m=1}^{L_a}$. Thus, if the caption contains a *black dog* and a *white cat*, the positive phrase *black dog* can be contrasted with *white dog*. If no valid competing attribute is available, attribute contrast is disabled for that entity.

Relation prompts. For directed edge $e = (a, r, b)$, let $\chi(n)$ denote the rendered entity phrase with an indefinite article when appropriate. We construct

$$\begin{aligned} q_e^{\rightarrow} &= \text{“a photo of ”} \oplus \chi(n_a) \oplus r \oplus \chi(n_b), \\ q_e^{\leftarrow} &= \text{“a photo of ”} \oplus \chi(n_b) \oplus r \oplus \chi(n_a), \end{aligned} \tag{31}$$

where $\oplus$ denotes whitespace-separated phrase concatenation. If the full rendering is unavailable, the corresponding bare subject–predicate–object phrases are used.

The endpoint phrases used in Eq. equation 5 are the bare entity names:

$$t_a = n_a, \qquad t_b = n_b. \tag{32}$$

Attributes therefore contribute through the factorized node cost rather than the endpoint-role gate.

C.2 Visual Proposal and Cropping Details

Region proposals. We use Grounding DINO Tiny (Liu et al., 2025) for caption-guided region proposal. Candidate boxes are extracted with confidence thresholds

$$\theta_{\text{box}} = 0.35, \qquad \theta_{\text{text}} = 0.25, \tag{33}$$

and overlapping redundant proposals are suppressed via non-maximum suppression (NMS) with IoU threshold $\theta_{\text{NMS}} = 0.50$. Predicted boxes are converted to absolute image coordinates and clipped to the image boundary. Boxes with non-positive width or height are discarded, and the remaining proposals are ordered by detector confidence.

Visual edge set. For N valid region proposals, the image graph contains all ordered pairs of distinct regions:

$$\mathcal{P} = \left\{(i,j) \in \{1,\dots,N\}^2 : i \neq j\right\}. \tag{34}$$

Thus, both (i,j) and (j,i) are retained, whereas self-pairs are excluded. The resulting image graph contains $N(N-1)$ candidate directed edges.

Single-region cropping. Let $B_i = (x_{i,1}, y_{i,1}, x_{i,2}, y_{i,2})$ denote a valid proposal with width $w_i = x_{i,2} - x_{i,1}$ and height $h_i = y_{i,2} - y_{i,1}$. We expand each proposal by a fixed context ratio $\rho = 0.10$. The padded coordinates are

$$\begin{aligned} x_{\min} &= \max(0, \lfloor x_{i,1} - \rho w_i \rfloor), & x_{\max} &= \min(W_{\text{img}}, \lceil x_{i,2} + \rho w_i \rceil), \\ y_{\min} &= \max(0, \lfloor y_{i,1} - \rho h_i \rfloor), & y_{\max} &= \min(H_{\text{img}}, \lceil y_{i,2} + \rho h_i \rceil). \end{aligned} \tag{35}$$

The resulting context-preserving crop is denoted by v_i.

Pair-view cropping and symmetry caching. For $(i,j) \in \mathcal{P}$, the pair view is extracted from the smallest axis-aligned rectangle covering the two original proposal boxes:

$$B_{ij} = \big[\min(x_{i,1}, x_{j,1}), \min(y_{i,1}, y_{j,1}), \max(x_{i,2}, x_{j,2}), \max(y_{i,2}, y_{j,2})\big]. \tag{36}$$

No additional context padding is applied to the pair-view rectangle. The resulting crop is denoted by v_{ij}.

Since the covering operation is symmetric,

$$v_{ij} = v_{ji}. \tag{37}$$

We exploit this symmetry by caching visual features across the $\binom{N}{2}$ unordered pairs. Consequently, the pair crop itself does not encode endpoint order; directional information is introduced through the endpoint-role evidence in Eq. equation 5.

Vision–language encoding. A single frozen vision–language backbone is shared across global image encoding, individual region encoding, pair-view encoding, and text encoding. We evaluate CLIP ViT-B/32 (Radford et al., 2021) and NegCLIP configurations. Backbone-specific preprocessing is applied consistently to all visual inputs. All visual and textual representations are L2-normalized before cosine similarity is computed:

$$\mathbf{z} = \frac{\mathbf{f}}{\|\mathbf{f}\|_2}. \tag{38}$$

C.3 Factorized Node-Cost Details

The entity-identity component follows Eq. equation 1. The attribute component combines direct support for the correctly bound phrase with contrastive evidence against attributes attached to other entities in the same caption.

When multiple positive or competing phrases are available, the pairwise definition in Eq. equation 2 is extended by taking the strongest positive and competing similarities:

$$\Delta_{\mathrm{attr}}(i,a) = \max_{1\le k\le K_a} s(v_i, q^+_{a,k}) - \max_{1\le m\le L_a} s(v_i, q^-_{a,m}). \tag{39}$$

For a single positive and a single competing phrase, this reduces directly to the contrast defined in Eq. equation 2.

The contrastive attribute cost is

$$C_{\mathrm{contrast}}(i,a) = \sigma(-\beta_{\mathrm{attr}}\Delta_{\mathrm{attr}}(i,a)), \qquad \beta_{\mathrm{attr}} = 15.0. \tag{40}$$

Direct compatibility with the correctly bound attribute phrases is computed as

$$C_{\mathrm{bound}}(i,a) = \frac{1}{K_a}\sum_{k=1}^{K_a} \frac{1 - s(v_i, q^+_{a,k})}{2}. \tag{41}$$

The implemented attribute cost is

$$C_{\mathrm{attr}}(i,a) = \gamma_{\mathrm{attr}} C_{\mathrm{bound}}(i,a) + (1-\gamma_{\mathrm{attr}}) C_{\mathrm{contrast}}(i,a), \qquad \gamma_{\mathrm{attr}} = 0.30. \tag{42}$$

Thus, direct bound-phrase compatibility and contrastive attribute evidence receive weights 0.30 and 0.70, respectively.

The resulting node cost follows Eq. equation 3. If entity u_a possesses no attributes ($\mathcal{A}_a = \emptyset$), we set $\lambda_a = 0$, reducing the node cost to $C_{\mathrm{ent}}(i,a)$. When an attributed entity has no valid contrasting attribute elsewhere in the caption ($\mathcal{A}^{\mathrm{alt}}_a = \emptyset$), we set $C_{\mathrm{attr}}(i,a) = C_{\mathrm{bound}}(i,a)$ for that entity.

C.4 Directed Relation-Cost Details

The contextual relation comparison in Eq. equation 4 uses $\beta_{\mathrm{ctx}} = 10.0$. The endpoint-role gate in Eq. equation 5 uses the bare entity phrases defined in Eq. equation 32, with $\beta_{\mathrm{dir}} = 10.0$.

Because the pair-view crop is symmetric,

$$s_{\mathrm{ctx}}(i,j,e) = s_{\mathrm{ctx}}(j,i,e). \tag{43}$$

By contrast, the endpoint margin satisfies

$$\Delta_{\mathrm{dir}}(j,i,e) = -\Delta_{\mathrm{dir}}(i,j,e). \tag{44}$$

Hence exchanging the two visual endpoints changes the directional gate whenever the individual region evidence distinguishes the two entity identities. The contextual and endpoint terms are combined through Eq. equation 6.

C.5 Structured UOT Optimization Procedure

The objective in Eq. equation 7 couples linear node matching with a non-convex quadratic relation assignment term $\mathcal{R}(\mathbf{P}) = \sum_{(i,j,e)\in\mathcal{S}} C_{\mathrm{rel}}(i,j,e) P_{ia} P_{jb}$. Rather than an exact convex program solver, LEM optimizes this objective via an alternating numerical procedure that combines linearized entropic mirror-descent steps with generalized Sinkhorn scaling.

Linearized surrogate cost. At outer iteration t, the quadratic relation cost is linearized around the current plan $\mathbf{P}^{(t)}$ via its first-order gradient $G^{\mathrm{rel}}_{ia}(\mathbf{P}^{(t)}) = \frac{\partial \mathcal{R}}{\partial P_{ia}}$. This derivative captures both subject and object endpoint assignments:

$$G^{\mathrm{rel}}_{ia}(\mathbf{P}^{(t)}) = \sum_{\substack{(i,j,e)\in\mathcal{S}\\ e=(a,r,b)}} C_{\mathrm{rel}}(i,j,e) P^{(t)}_{jb} + \sum_{\substack{(j,i,e)\in\mathcal{S}\\ e=(b,r,a)}} C_{\mathrm{rel}}(j,i,e) P^{(t)}_{jb}. \tag{45}$$

The composite first-order cost at step t is:

$$G^{(t)}_{ia} = (1-\alpha) C_{\mathrm{node}}(i,a) + \alpha G^{\mathrm{rel}}_{ia}(\mathbf{P}^{(t)}). \tag{46}$$

Linearized proximal update. To update the transport mass, the solver performs an entropic mirror-descent step on the linearized objective. Specifically, with step size $\eta = 1.0$ and temperature parameter $\varepsilon_{\text{step}} = \max(\varepsilon, 0.01) = 0.01$ serving as step-size damping against gradient magnitudes, the unconstrained proximal subproblem under the generalized Kullback–Leibler divergence $D_{\text{KL}}(\mathbf{P}\|\mathbf{P}^{(t)}) = \sum_{i,a}[P_{ia}\log(P_{ia}/P_{ia}^{(t)}) - P_{ia} + P_{ia}^{(t)}]$ is:

$$\widetilde{\mathbf{P}}^{(t+1)} \;=\; \arg\min_{\mathbf{P}\geq 0} \sum_{i,a} G_{ia}^{(t)} P_{ia} + \frac{\varepsilon_{\text{step}}}{\eta} D_{\text{KL}}(\mathbf{P}\|\mathbf{P}^{(t)}). \tag{47}$$

Setting the first-order optimality condition of this subproblem to zero gives the unconstrained exponentiated mirror-descent update $\widetilde{P}_{ia}^{(t+1)} = P_{ia}^{(t)} \exp(-\eta G_{ia}^{(t)}/\varepsilon_{\text{step}})$. For numerical stability in finite-precision implementation, we clip the exponent argument to $[-50, 50]$:

$$\widetilde{P}_{ia}^{(t+1)} = P_{ia}^{(t)} \exp\left[-\eta \,\text{clip}\left(\frac{G_{ia}^{(t)}}{\varepsilon_{\text{step}}}, -50, 50\right)\right]. \tag{48}$$

Generalized Sinkhorn scaling for UOT. To account for the marginal relaxation penalties $\tau\,\text{KL}(\cdot\|\cdot)$ in Eq. equation 7, the intermediate plan $\widetilde{\mathbf{P}}^{(t+1)}$ is subsequently scaled via generalized Sinkhorn iterations (Chizat et al., 2017) with contraction exponent $\kappa = \frac{\tau}{\tau+\varepsilon}$. At each outer step, the scaling dual vector is initialized as $v = \mathbf{1}_M$, and ten alternating scaling updates are applied:

$$u_i \leftarrow \left(\frac{p_i}{\sum_a \widetilde{P}_{ia}^{(t+1)} v_a + 10^{-300}}\right)^{\kappa}, \qquad v_a \leftarrow \left(\frac{q_a}{\sum_i \widetilde{P}_{ia}^{(t+1)} u_i + 10^{-300}}\right)^{\kappa}, \tag{49}$$

producing the next iterate $P_{ia}^{(t+1)} = u_i \widetilde{P}_{ia}^{(t+1)} v_a$.

Numerical stopping criterion. The plan is initialized uniformly as $P_{ia}^{(0)} = \frac{1}{NM}$. The solver runs for at most $T_{\max} = 200$ outer iterations and terminates when

$$\max_{i,a} \left|P_{ia}^{(t+1)} - P_{ia}^{(t)}\right| < 10^{-6}. \tag{50}$$

We emphasize that due to the non-convex quadratic relation term and distinct regularizers, this alternating numerical procedure is an iterative optimization heuristic rather than an exact convex solver; it is not guaranteed to recover a global minimizer $\mathbf{P}^\star$. The resulting final numerical plan is denoted by $\widehat{\mathbf{P}}$.

C.6 Alignment Cost Calibration

The KL and entropy terms in Eq. equation 7 influence the transport plan but are excluded from the final local matching score. Instead, node and relation costs are normalized by their respective contributing transport masses, removing the direct scale dependence on total transported mass:

$$m_{\text{node}} = \sum_{i,a} \widehat{P}_{ia}, \qquad m_{\text{rel}} = \sum_{(i,j,e)\in\mathcal{S}} \widehat{P}_{ia}\widehat{P}_{jb}, \quad \text{with } e = (a, r, b). \tag{51}$$

We use $\xi_{\text{mass}} = 10^{-6}$ as the threshold for negligible transport mass. For non-negligible mass:

$$\overline{C}_{\text{node}} = \frac{\sum_{i,a} C_{\text{node}}(i,a)\widehat{P}_{ia}}{m_{\text{node}}}, \qquad \overline{C}_{\text{rel}} = \frac{\sum_{(i,j,e)\in\mathcal{S}} C_{\text{rel}}(i,j,e)\widehat{P}_{ia}\widehat{P}_{jb}}{m_{\text{rel}}}. \tag{52}$$

The local cost is

$$C_{\text{local}} = \begin{cases} (1-\alpha)\overline{C}_{\text{node}} + \alpha\overline{C}_{\text{rel}}, & \mathcal{S} \neq \emptyset \text{ and } m_{\text{rel}} > \xi_{\text{mass}}, \\ \overline{C}_{\text{node}}, & \text{otherwise.} \end{cases} \tag{53}$$

The local similarity is $S_{\text{local}} = 1 - 2C_{\text{local}} \in [-1, 1]$, evaluated with numerical safeguard $\text{clip}(1 - 2C_{\text{local}}, -1.0, 1.0)$. If no valid entity–region alignment is available, the local branch is discarded and the final score falls back to S_{global}. Otherwise, the final score follows Eq. equation 12.

C.7 Evaluation-Specific Configuration

The structural weight α, attribute weight λ_{attr}, and global fusion weight w_{glob} are the only hyperparameters we allow to vary across evaluation settings; every other component of the configuration tuple Θ (Definition 11) is held fixed at the values in Table 3 for all experiments in the paper. We restrict tunability to these three weights because they directly control the balance between complementary sources of evidence – entity identity versus attribute binding within the node cost, node versus relation evidence within the local score, and local versus global similarity in the final fused score – whereas the remaining parameters (detector thresholds, entropic and marginal-relaxation coefficients, solver step sizes, and convergence tolerances) primarily affect numerical behavior rather than what evidence is being weighted, and we did not observe qualitative sensitivity to them during preliminary development.

For each benchmark and backbone, α, λ_{attr}, and w_{glob} are selected once from the sweep values reported in §4.3 and then held fixed when producing the main results in Table 1; no additional per-instance or per-caption tuning is performed at test time. This mirrors standard practice for training-free methods, where the absence of a training signal precludes learning these weights directly and a small, interpretable hyperparameter surface is instead validated by ablation. We report results under a single configuration per backbone rather than the best configuration found post hoc across all sweep values, so that the main table reflects a setting that could plausibly be fixed in advance rather than one selected in hindsight for maximum benchmark performance.

C.8 Implementation Hyperparameters

Table 3 summarizes the implementation parameters that are shared, unchanged, across every evaluation configuration reported in this paper. These parameters fall into four groups: (i) detector thresholds (θ_{box}, θ_{text}, θ_{NMS}) and the context padding ratio ρ, which govern region proposal and are inherited directly from the default Grounding DINO Tiny configuration without benchmark-specific retuning; (ii) cost-shaping scales (β_{attr}, γ_{attr}, β_{ctx}, β_{dir}), which control the sharpness of the sigmoid gates used in the factorized node and directed relation costs; (iii) unbalanced-transport regularization parameters (τ, ε, $\varepsilon_{\text{step}}$, η), which govern the trade-off between marginal fidelity and entropic smoothing in the optimization objective (Eq. 7); and (iv) solver control parameters (inner scaling iterations, maximum outer iterations, and the convergence tolerance), which determine when the alternating mirror-descent and Sinkhorn-scaling procedure of Appendix C.5 terminates.

Because LEM is training-free, none of these parameters are learned; they are instead set once from preliminary experimentation on held-out examples disjoint from the evaluation splits reported in §4.2, and then frozen for all subsequent experiments, backbones, and benchmarks. We found this configuration to be reasonably robust in the sense that small perturbations to the cost-shaping and regularization parameters did not qualitatively change the ranking between LEM and the baselines, though we did not perform an exhaustive sensitivity sweep over this full set, as our ablations in §4.3 instead focus on the three evaluation-specific weights identified in Appendix C.7, which we found to have a substantially larger effect on downstream accuracy. Fixing this larger parameter set also keeps the overall hyperparameter surface exposed to the user small and interpretable, consistent with LEM's design goal of being a plug-and-play scoring layer on top of frozen pretrained components rather than a system requiring extensive per-deployment calibration.

D Error Analysis Evaluation Protocol

SugarCrepe × COCO-2017 val join. Each SugarCrepe instance references an image by filename (e.g. `COCO_val2017_000000397133.jpg`); we recover the COCO `image_id` by stripping the zero-padded prefix and casting to `int`, then query the COCO-2017 val annotation file (`instances_val2017.json`) via `pycocotools` to fetch every ground-truth box (`bbox_xywh`→`bbox_xyxy`) and category label attached to that `image_id`. Instances whose image falls outside the val2017 split are dropped. This produces a manifest of **7,511** (image, caption, GT-box) triples spanning all seven SugarCrepe hard-negative types (Figure 4).

Entity parsing and COCO-label matching. For each caption we extract noun-chunk head phrases with a dependency parser and lemmatize the head token. Each lemma is mapped onto

Table 3: Fixed implementation parameters used by LEM.

Parameter	Symbol	Value
Detector box threshold	θ_{box}	0.35
Detector text threshold	θ_{text}	0.25
Detector NMS IoU threshold	θ_{NMS}	0.50
Region context padding	ρ	0.10
Attribute contrast scale	β_{attr}	15.0
Direct attribute-cost weight	γ_{attr}	0.30
Relation-context scale	β_{ctx}	10.0
Endpoint-direction scale	β_{dir}	10.0
Marginal relaxation	τ	1.0
Entropy regularization	ε	0.005
Proximal step temperature	$\varepsilon_{\text{step}}$	0.01
Outer update step size	η	1.0
Inner scaling iterations	–	10
Maximum outer iterations	–	200
Transport convergence tolerance	–	10^{-6}
Mass threshold	ξ_{mass}	10^{-6}



Figure 4: Successful cases from joining SugarCrepe with COCO-2017 val to obtain ground-truth bounding boxes while retaining the original SugarCrepe caption (Appendix D). Each panel shows the SugarCrepe manifest tag and caption alongside the ground-truth box(es) matched to the caption's entity phrase(s) (e.g., *sheep*, *pizza*, *cat*/*bench*), illustrating correctly resolved image–caption–box triples used in our localization error analysis.

one of the 80 COCO category labels via a curated synonym table (e.g. *skier*→*person*, *pup*→*dog*); lemmas of length ≥ 5 left unmatched fall back to a length- and first-letter-constrained fuzzy match against the same table, which avoids short-token collisions (e.g. *coat* vs. *cat*). A phrase is kept as an evaluation *target* only if its mapped category also has ≥ 1 ground-truth instance in the paired image, giving **6,108**/7,511 (**81.3%**) evaluable samples.

Detection and matching. Grounding DINO Tiny ($\theta_{\text{box}} = 0.35$, $\theta_{\text{text}} = 0.25$) is prompted with the kept phrases per image. Predicted boxes are greedily matched to ground-truth boxes of the *same* COCO category (IoU ≥ 0.5), independently per category and per image; matches, unmatched predictions, and unmatched ground-truth boxes accumulate as TP/FP/FN over all 6,108 evaluated samples.