PROBABILISTIC SET REPRESENTATIONS WITH PARTIAL OPTIMAL TRANSPORT FOR IMAGE–TEXT RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

An image caption often describes only part of a scene, posing a partial-correspondence problem for image–text retrieval. We propose a framework that independently represents images and captions as weighted sets of von Mises–Fisher components and aligns them through one-sided partial optimal transport. Analytic distributional overlap defines component compatibility, while learned component masses constrain allocation. A fixed matched-mass ratio allows some image mass to remain unmatched while preserving the relative masses of caption components. The framework learns from image–text pairs and supports optional region–phrase supervision, with no entity inputs required at inference. Experiments examine retrieval, structural choices, caption coverage, and component behavior. Real-match allocation favors annotated target regions over unmatched allocation. Entity supervision improves local support but can reduce the corresponding components' matching participation, while coverage analysis does not establish a larger partial-matching benefit for low-coverage captions. These findings characterize how probabilistic compatibility and learned mass constraints jointly shape partial alignment, and distinguish component support from contribution to retrieval scores.

1 INTRODUCTION

Image–text retrieval learns cross-modal relevance from paired visual and linguistic data (Radford et al., 2021). Yet an image and its caption often differ in scope: an image contains multiple objects and activities, whereas a caption may describe only a subset. Different captions can therefore be valid descriptions of the same scene. Figure 1 (left) illustrates this with captions mentioning both dogs or only the black dog. Retrieval must identify the described content while accounting for unmentioned content, whose omission does not itself imply a contradiction.

This partial correspondence raises questions about both representation and scoring. Probabilistic embeddings model ambiguous associations, while local alignment and multiple embeddings retain distinct matching cues (Chun et al., 2021; Chun, 2024; Lee et al., 2018; Song & Soleymani, 2019). These representations provide evidence for matching, but a scoring rule must determine how that evidence contributes to overall relevance. A locally compatible correspondence alone does not specify its share of the aggregate score; that share also depends on other correspondences and the matching constraints.

Prior work has jointly studied set construction and scoring through smooth aggregation and one-to-one matching (Kim et al., 2023; Alomari et al., 2025), and has applied partial transport to image–text matching (Pan et al., 2026). Building on these directions, we investigate how probabilistic compatibility and matching participation can be modeled together when some content may remain unmatched. We distinguish how compatible two matching units are from how much each unit can participate in the allocation, and examine their interaction within a trainable retrieval framework.

Our framework independently represents each image or caption as a set of learned components. Each component pools token information and carries a probability distribution, serving as a cross-modal matching unit without requiring one-to-one correspondence to objects or phrases. Distributional overlap supplies compatibility, learned component masses specify available matching shares, and

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.

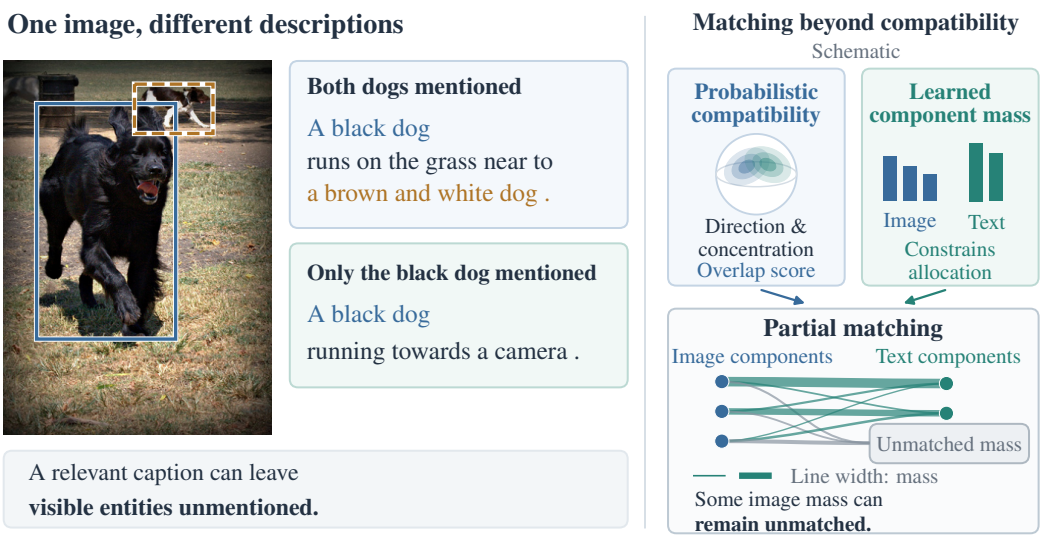


Figure 1: Caption scope and partial-matching design. Left: an image and two original captions from Flickr30k Entities validation, selected because they mention both dogs or only the black dog; colored phrases correspond to annotation boxes of the same color. The example illustrates different caption scopes for the same image. Right: a constructed schematic, independent of the sample, outlines partial matching jointly constrained by probabilistic compatibility and component masses. Nodes represent abstract components, and line widths encode transport mass satisfying the displayed mass constraints. Mass bars follow node order; the unmatched total is prescribed.

one-sided partial transport allocates these shares while permitting unmatched image mass. Figure 1 (right) sketches this construction. The matched-mass ratio is prescribed rather than estimated from each caption's semantic coverage. The core framework learns from image–text pairs; an optional region–phrase supervision extension examines how local content support relates to matching participation, without requiring entity annotations at inference (Plummer et al., 2015).

We contribute this formulation connecting probabilistic component compatibility, learned mass constraints, and partial matching, and evaluate retrieval, structural choices, coverage response, and allocation behavior. The evaluation also identifies boundaries: fixed-representation analysis does not establish a larger benefit for low-coverage captions, and stronger local support under entity supervision can coexist with reduced mass on the corresponding components. These findings distinguish the content captured by a component from its participation in the final matching score.

2 Related Work

2.1 Probabilistic Representations and Cross-Modal Compatibility

Probabilistic representations model the ambiguity of image–text associations by representing samples as probability distributions. PCME estimates matching probabilities by sampling; PCME++ introduces a closed-form distance and pseudo-positives to mitigate false negatives, retaining sample-level distributions while changing matching computation and training (Chun et al., 2021; Chun, 2024). ProbVLM adapts frozen representations probabilistically, while AsymVLM considers hyperspherical geometry and asymmetric cross-modal uncertainty (Upadhyay et al., 2023; Ju et al., 2025). We define distributions at the component level within a set, using distributional overlap to determine compatibility and separate mass constraints to specify matching shares.

2.2 Multiple Embeddings and Local Cross-Modal Interaction

Multiple matching cues can be retained through latent sets or local interactions. PVSE constructs embeddings with multi-head attention; DivE jointly designs competitive aggregation for set diver-

sity and scoring (Song & Soleymani, 2019; Kim et al., 2023). Local interaction approaches differ in where and how cross-modal information is used: SCAN aggregates cross-modal context through cross-attention, FILIP pools local maximum similarities after independent encoding, and SPARC uses sparse local grouping for training alignment (Lee et al., 2018; Yao et al., 2021; Bica et al., 2024). Supervision provides another distinction: MDETR and X-VLM exploit region–phrase or multi-granular correspondence (Kamath et al., 2021; Zeng et al., 2022). Our model learns latent components from paired images and texts through independent encoding, applies cross-modal transport at the scoring stage, and considers entity supervision as an extension.

2.3 Set Scoring and Partial-Matching Constraints

Set scoring determines how local correspondences contribute to overall relevance. PVSE selects the strongest correspondence, DivE aggregates multiple correspondences bidirectionally through Smooth-Chamfer, and MaxMatch restricts component reuse through one-to-one assignment (Song & Soleymani, 2019; Kim et al., 2023; Alomari et al., 2025). OT jointly allocates continuous mass under marginal constraints; partial transport prescribes matched mass, unlike marginal relaxation in unbalanced transport (Chapel et al., 2020; Chizat et al., 2018). UNITER uses OT for pretraining alignment; OMIT directly scores image–text pairs using uniform local masses and global-embedding dustbins on both sides (Chen et al., 2020; Pan et al., 2026). Compared with OMIT, we use probabilistic components and learned masses, with a one-sided dustbin allowing unmatched image mass. Under a prescribed total matched mass, our constraints preserve the relative masses of text components, connecting probabilistic compatibility with constrained matching shares.

3 Method

3.1 Problem Formulation

Given paired images and texts $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^{N}$, we learn a relevance function $s_\theta(x, y)$ for bidirectional retrieval. Each modality is encoded independently as a set of latent components,

$$\mathcal{Z}_\theta^I(x) = \{z_i^I\}_{i=1}^{K_I}, \qquad \mathcal{Z}_\theta^T(y) = \{z_j^T\}_{j=1}^{K_T}, \tag{1}$$

with normalized component masses $\pi^I \in \Delta^{K_I-1}$ and $\pi^T \in \Delta^{K_T-1}$. Here Δ^{K-1} is the probability simplex, and mass specifies a component's share of the matching allocation.

A caption may describe only part of an image. We therefore seek a matching rule that combines component compatibility with mass constraints while allowing some image content to remain unmatched. Our construction uses probabilistic components to define compatibility and one-sided partial transport to aggregate their correspondences. Figure 2 summarizes the scoring path. The core method uses image–text pair supervision; entity annotations provide an optional training extension.

3.2 Probabilistic Set Representations and Component Compatibility

Component construction Let $\{v_t^m\}_{t=1}^{L_m}$ be projected token features in a common d-dimensional space for modality $m \in \{I, T\}$. A learned routing network produces logits ℓ_{tk}^m, which are normalized over valid tokens to form component features:

$$A_{kt}^m = \frac{\exp(\ell_{tk}^m)}{\sum_{u=1}^{L_m} \exp(\ell_{uk}^m)}, \qquad u_k^m = \sum_{t=1}^{L_m} A_{kt}^m v_t^m. \tag{2}$$

Components can attend to overlapping content. Routing weights determine what each component pools; separate learned masses determine its allocation during matching. The image routing network also uses a multilayer feature residual, detailed in Appendix A.1.

Probabilistic components Each feature u_k^m produces a unit direction μ_k^m, a positive concentration κ_k^m, and a mass π_k^m normalized across components. Direction normalization, a softplus concentration head, and a softmax mass head enforce these constraints. The component $z_k^m = (\mu_k^m, \kappa_k^m, \pi_k^m)$ carries a von Mises–Fisher distribution on $\mathbb{S}^{d-1}$ (Banerjee et al., 2005):

$$p_k^m(z) = C_d(\kappa_k^m) \exp\left(\kappa_k^m (\mu_k^m)^\top z\right). \tag{3}$$

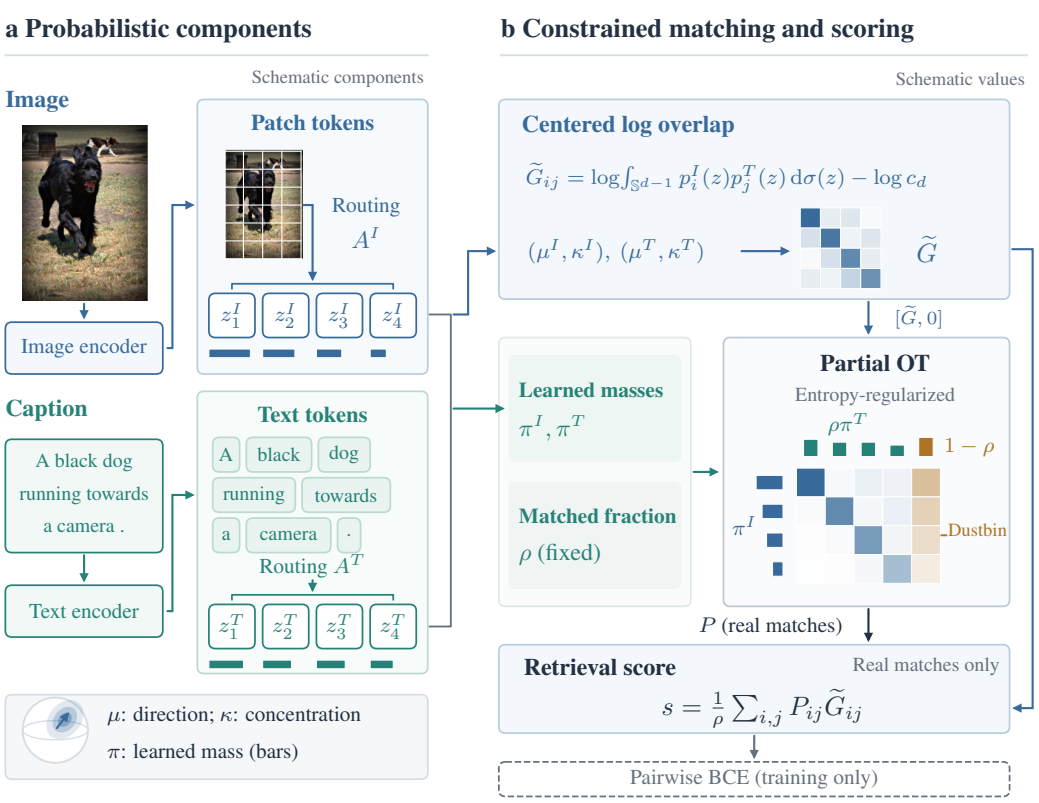


Figure 2: Overview of the proposed image–text retrieval framework. (a) Image and text inputs are independently encoded and routed into weighted sets of probabilistic components. (b) Component compatibility and learned masses jointly determine one-sided partial transport under a fixed matched-mass ratio, yielding a retrieval score. Transport rows and columns correspond to image and text components, respectively; an additional dustbin column receives unmatched image mass. The displayed input pair is illustrative, and component representations and matching values are schematic. Color intensity encodes transport mass on a shared scale across real-match and dustbin columns.

Here C_d is the vMF normalizing constant. Direction specifies the distribution's center, concentration controls its spread, and mass specifies its share of transport.

Compatibility We score an image component and a text component by their log distributional overlap, following the distributional similarity principle of Vilnis & McCallum (2015):

$$\begin{aligned} G_{ij} &= \log \int_{\mathbb{S}^{d-1}} p_i^I(z) p_j^T(z)\, d\sigma(z) \\ &= \log C_d(\kappa_i^I) + \log C_d(\kappa_j^T) \\ &\quad - \log C_d \left(\left\| \kappa_i^I \mu_i^I + \kappa_j^T \mu_j^T \right\|_2 \right). \end{aligned} \tag{4}$$

The integral uses surface-area measure σ. Compatibility depends on both relative directions and concentrations. We center it as $\widetilde{G}_{ij} = G_{ij} - \log c_d$, where $c_d = \Gamma(d/2)/(2\pi^{d/2})$ is the uniform density on the sphere. Zero thus corresponds to the overlap of a uniform density with any normalized density. Numerical evaluation of the normalizer is described in Appendix A.2.

3.3 One-Sided Partial Transport with Mass Constraints

Partial matching Following the prescribed-mass setting of partial transport (Chapel et al., 2020), a fixed coverage parameter $\rho \in (0, 1]$ specifies the fraction of image mass assigned to text compo-

nents. For $\rho < 1$, a text-side dustbin $\varnothing$ receives the remaining mass with zero centered compatibility. We augment the text marginal and score matrix as

$$b = [\rho\pi^T; 1-\rho], \qquad \overline{G} = [\widetilde{G}, \mathbf{0}]. \tag{5}$$

The augmented plan $\overline{P} \in \mathbb{R}_+^{K_I \times (K_T+1)}$ has feasible set

$$\mathcal{U}(\pi^I, b) = \left\{ \overline{P} \geq 0 : \overline{P}\mathbf{1}_{K_T+1} = \pi^I, \overline{P}^\top \mathbf{1}_{K_I} = b \right\}. \tag{6}$$

Its component-to-component portion P satisfies $P\mathbf{1} \leq \pi^I$ and $P^\top\mathbf{1} = \rho\pi^T$, so exactly ρ mass is matched. Thus ρ is a prescribed modeling parameter, not a pair-specific estimate of content coverage. At $\rho = 1$, the dustbin is removed and all mass is matched.

Plan and score We obtain the plan with entropy regularization (Cuturi, 2013):

$$\begin{aligned} \overline{P}^* &= \arg\max_{\overline{P} \in \mathcal{U}(\pi^I, b)} \left\{ \langle \overline{P}, \overline{G} \rangle + \varepsilon H(\overline{P}) \right\}, \\ H(\overline{P}) &= -\sum_{i,j} \overline{P}_{ij} \log \overline{P}_{ij}, \end{aligned} \tag{7}$$

The entropy includes the dustbin column, with $0 \log 0 = 0$. Learned masses constrain component allocations, compatibility guides their distribution, and ρ fixes their total matched mass. A component may split its mass across several counterparts; dustbin allocation is determined jointly under these constraints and does not itself identify background.

The retrieval score averages centered compatibility over the matched mass:

$$s_\theta(x, y) = \frac{1}{\rho} \sum_{i=1}^{K_I} \sum_{j=1}^{K_T} P^*_{ij} \widetilde{G}_{ij}. \tag{8}$$

Entropy affects the plan but is excluded from the returned score. Consequently, directions and concentrations affect relevance through compatibility, while learned masses affect it through the transport allocation. The plan is computed differentiably; Appendix A.2 gives the numerical procedure.

3.4 Training and Inference

For a batch of B pairs, let $s_{ab} = s_\theta(x_a, y_b)$ and let M_{ab} indicate known image–text associations. All captions of the same image appearing in the batch are positives; the remaining combinations receive zero labels, which need not imply semantic irrelevance. With a learned positive scale $\alpha = \exp(\eta)$ and bias β, we form $q_{ab} = \alpha s_{ab} + \beta$ and optimize

$$\begin{aligned} \mathcal{L}_{\text{ret}} = -\frac{2}{B^2} \sum_{a=1}^{B} \sum_{b=1}^{B} \Big[& M_{ab} \log \operatorname{sigmoid}(q_{ab}) \\ & + (1 - M_{ab}) \log\left(1 - \operatorname{sigmoid}(q_{ab})\right) \Big]. \end{aligned} \tag{9}$$

The core model uses $\mathcal{L}_{\text{ret}}$ alone. Pairwise sigmoid training is also used in SigLIP (Zhai et al., 2023); here labels and normalization follow (9). Gradients propagate through both compatibility computation and the transport plan to learn component representations and masses. Encoder freezing or fine-tuning is specified by the training configuration.

At inference, component sets can be independently precomputed. Each candidate pair still requires compatibility computation and a transport solve. Both retrieval directions rank by the same score in (8), keeping the dustbin on the text side. The learned positive scale and shared bias do not change this ranking.

3.5 Extension with Entity Supervision

Entity boxes and phrase spans define token support targets. We assign entities to distinct components within each modality by routing cross-entropy. The extension uses a bidirectional multi-positive ranking loss $\mathcal{L}_{\text{rank}}$, with spatial support supervision and optional identity supervision on pre-transport component compatibility:

$$\mathcal{L}_{\text{ext}} = \mathcal{L}_{\text{rank}} + \lambda_{\text{spatial}} \mathcal{L}_{\text{spatial}} + \lambda_{\text{id}} \mathcal{L}_{\text{id}}. \tag{10}$$

Table 1: Local evaluation using common manifests and metric implementations. Mean R@1 values and ECCV Caption metrics are percentages. Training protocols differ between the two models.

Model	COCO 5K R@1	COCO 1K R@1	COCO 1K RSUM	CxC R@1	EC mAP@R	EC R-P	Flickr RSUM
PCME++ SWA	55.200	75.530	537.520	56.792	40.043	49.727	514.260
Complete OT	50.372	71.338	524.680	52.107	36.367	46.346	510.260

Retrieval-only training sets both auxiliary weights to zero; spatial supervision sets $\lambda_{\mathrm{id}} = 0$. Appendix B defines the losses and assignments. Entity labels are used only in training. Neither auxiliary loss directly supervises transport mass, and inference remains unchanged.

4 Experiments

4.1 Experimental Setup

We train and select the core model on COCO Captions (Chen et al., 2015), then evaluate the fixed checkpoint on COCO test, CxC, ECCV Caption, and Flickr30k transfer (Parekh et al., 2021; Chun et al., 2022; Young et al., 2014). CxC and ECCV Caption use expanded relevance annotations and their own valid-query sets. Entity extensions train on Flickr30k region–phrase annotations (Plummer et al., 2015) and report validation results separately from transfer.

The complete configuration uses CLIP ViT-B/32, four components per modality, multilayer image routing, learned masses, and partial OT with $\rho = .75$ and $\varepsilon = .2$. Training lasts 25 epochs, freezing encoders in epoch 1 before fine-tuning, with selection by COCO validation RSUM. Structural and parameter ablations freeze CLIP throughout; coverage analysis switches scores on fixed representations. All training comparisons use seed 0. Appendix C details protocols, including the entity extension's bidirectional multi-positive ranking objective.

We report bidirectional recall, RSUM (the sum of R@1, R@5, and R@10 in both directions), mean bidirectional R@1, and ECCV Caption mAP@R and R-Precision. Local metrics distinguish content support, correspondence under ground-truth-assisted assignment, and participation through component or correct-edge mass. Paired image-bootstrap intervals reflect fixed-checkpoint sample variation, excluding training-seed and selection uncertainty.

4.2 Overall Retrieval Performance

Table 1 compares our complete configuration with an independently re-evaluated PCME++ SWA checkpoint (Chun, 2024), using common manifests, preprocessing, caption ordering, and metrics. PCME++ uses Gaussian CSD, pseudo-positives, mixed-sample augmentation, VIB, and SWA, with a different optimizer and schedule.

The complete configuration is lower on every reported metric, with mean R@1 gaps of 4.828 and 4.685 points on COCO 5K and CxC. Published ViT-B/32 references likewise place it above zero-shot CLIP but below trained baselines (Appendix C.1). These comparisons locate system performance; training differences prevent attribution to representation or matching alone.

4.3 Ablating the OT Framework

The frozen-CLIP 2×2 study toggles multilayer routing and a 25% unmatched channel, selecting each configuration independently on validation.

Multilayer routing alone lowers every metric in Table 2. With the unmatched channel, its effect becomes positive on COCO, CxC, and mAP@R but remains negative on Flickr30k. Relative to base OT, the combination slightly improves COCO 5K and CxC R@1 but reduces ECCV Caption mAP@R by 0.155 points, indicating a metric-dependent interaction.

Figure 3 and Appendix C.2 report parameter sweeps. Increasing K did not improve these runs; excluding $K = 1$ prevents establishing a multi-component advantage. Unmatched-mass response

Table 2: Structural ablation of multilayer routing and the unmatched channel. All models freeze CLIP and use one seed.

Multilayer	Unmatched	COCO 5K R@1	COCO 1K RSUM	CxC R@1	EC mAP@R	Flickr RSUM
No	No	40.446	491.888	42.286	30.076	471.620
Yes	No	39.384	489.832	41.210	29.666	469.200
No	Yes	40.408	492.716	42.118	29.609	473.080
Yes	Yes	40.730	494.440	42.496	29.921	472.020

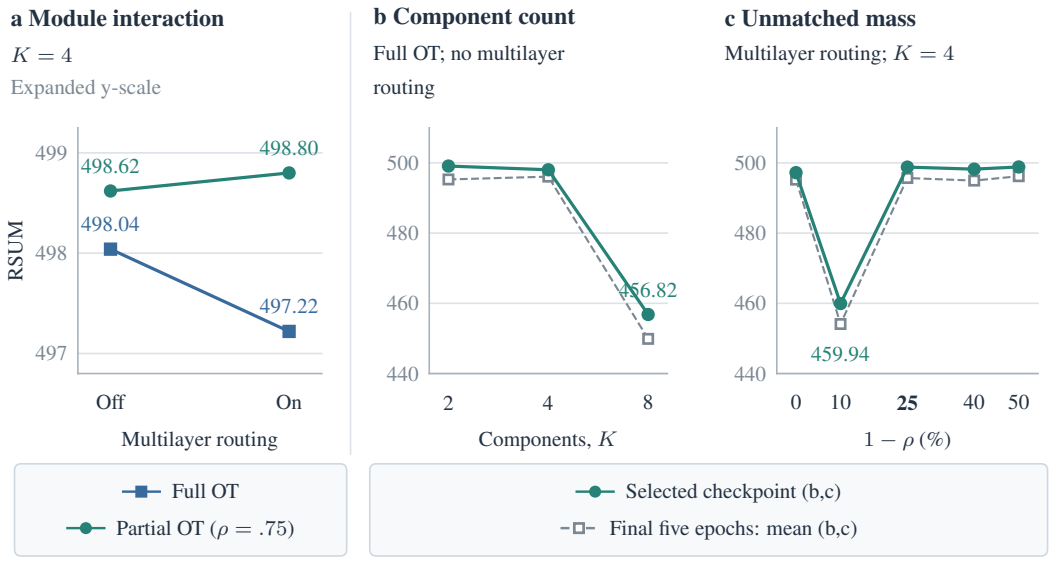


Figure 3: Module interaction and parameter sensitivity. (a) The effect of multilayer routing under full and partial OT. (b) Component count. (c) Unmatched-mass ratio. All vertical axes report COCO validation RSUM; (a) expands the vertical scale to show small differences, and bold 25% in (c) marks the prescribed setting. All runs freeze CLIP and train for 25 epochs with seed 0. In (b,c), solid curves show validation-selected checkpoints and dashed curves show final-five-epoch means. Lines connect evaluated configurations; epoch means are not independent training repeats.

was non-monotonic, and the prescribed 25% capacity was not grid-optimal. The plotted validation RSUM is distinct from Table 2's evaluation metrics.

4.4 Partial Matching under Different Caption Coverage

On Flickr30k Entities validation, we fix a historically validation-selected representation and switch from partial ($\rho = .75$) to full OT ($\rho = 1$). Coverage is the fraction of boxed entities mentioned by a caption relative to the union across all five captions. This tests immediate scoring responses without independent training or testing.

Low-coverage captions are harder under both scores, without an established larger benefit from partial OT. The prespecified low-minus-full contrast is -2.08 points (95% CI $[-6.26, +2.14]$). An auxiliary box-area proxy gives $+1.85$ ($[-0.68, +4.41]$), also inconclusive and affected by scene-level and nested boxes. Low-coverage-specific benefits remain unestablished.

4.5 Analysis of Transport Allocation

For the complete configuration, normalized target-box mass is 0.18604 on real edges versus 0.13225 for dustbin allocation, a difference of 0.05379 (95% CI $[0.04030, 0.06795]$). This indicates regional allocation preferences; dustbin mass does not directly represent segmented background. Numerical feasibility checks appear in Appendix C.

Table 3: Partial versus full OT on fixed representations, stratified by caption coverage. R@1 measures text-to-image retrieval; intervals use paired image bootstrap.

Coverage	Captions / images	Partial R@1	Full R@1	Partial − full (95% CI)
Low	3056 / 907	42.51	42.24	$+0.26\ [-0.79, +1.36]$
Mid	1602 / 783	54.43	54.68	$-0.25\ [-1.58, +1.10]$
Full	342 / 173	59.36	57.02	$+2.34\ [-1.82, +6.42]$
All	5000 / 1000	47.48	47.24	$+0.24\ [-0.64, +1.14]$

Table 4: Entity-supervision results from validation-selected single-seed fine-tuned checkpoints. Text support measures routing probability within GT spans; entity OT measures correspondence under GT-assisted assignment; component and correct-edge masses measure participation.

Supervision	RSUM	Text support	Entity OT	Text eff. K	GT-slot mass	Correct-edge mass
Retrieval only	493.18	4.07%	28.40%	3.87	63.67%	0.0490
Spatial support	494.60	64.31%	24.38%	1.36	6.55%	0.0042
Spatial + vMF identity	494.82	40.40%	29.95%	1.53	8.71%	0.0067

Figure 4 examines whether allocation yields distinguishable phrase evidence. Component contributions are $c_i(q) = \rho^{-1} \sum_j P^*_{ij}(q) \widetilde{G}_{ij}(q)$. For $d_i = c_i(q_0) - c_i(q_1)$, removing the mass-proportional shift gives $\delta_i = d_i - \pi^I_i \sum_k d_k$. Its normalized positive and negative parts weight image routing maps to form relative evidence. Food/hand shows alignment; player/fence exposes localization failure. Across 100 images, the area-adjusted target-versus-other-region grounding gap is 0.02286 (95% CI $[-0.00071, 0.04545]$), leaving reliable entity separation unestablished.

Deleting components selected by the original contribution contrast, renormalizing masses, and recomputing OT yields selectivity win rates of 58.5% ($[51.0\%, 65.5\%]$) against random deletion and 50.0% ($[42.5\%, 57.5\%]$) against mass-matched deletion. Selectivity measures the difference between phrase-score drops; an advantage over the mass-matched control remains unestablished. We next examine whether entity supervision improves local correspondence and translates into matching participation and retrieval gains.

4.6 Entity-Supervised Extension

We compare retrieval-only, spatial-support, and spatial-plus-identity supervision under a common Flickr30k fine-tuning and ranking protocol (Section 3.5). Relative to their frozen counterparts, RSUM increases by 10.16, 9.38, and 10.54, while entity OT correspondence falls from 37.76%, 34.31%, and 34.34% to 28.40%, 24.38%, and 29.95%, respectively. Paired bootstrap 95% intervals exclude zero for all six changes: retrieval improvement coexists with correspondence deterioration.

Table 4 separates supervision effects. Spatial supervision raises text support from 4.07% to 64.31%, but correspondence declines, GT-slot mass falls from 63.67% to 6.55%, and correct-edge mass from 0.0490 to 0.0042; RSUM changes by $+1.42$ (95% CI $[-1.44, 4.08]$). Identity supervision improves correspondence by 5.57 points ($[4.36, 6.73]$), while text support falls and GT-slot mass remains 8.71%; additional RSUM gain is $+0.22$ ($[-2.54, 2.98]$).

Content support, identity correspondence, and participation need not improve together, and local improvements have not established additional retrieval gains. Neither auxiliary loss directly constrains transport mass, allowing well-supported components to receive little matching mass. Evidence comes from single-seed, validation-selected checkpoints; bootstrap intervals capture only sample variation. Without mass interventions holding other factors fixed, downweighting's causal role in retrieval changes remains unresolved. These observations establish a dissociation in the evaluated configurations, not an inevitable optimization conflict.

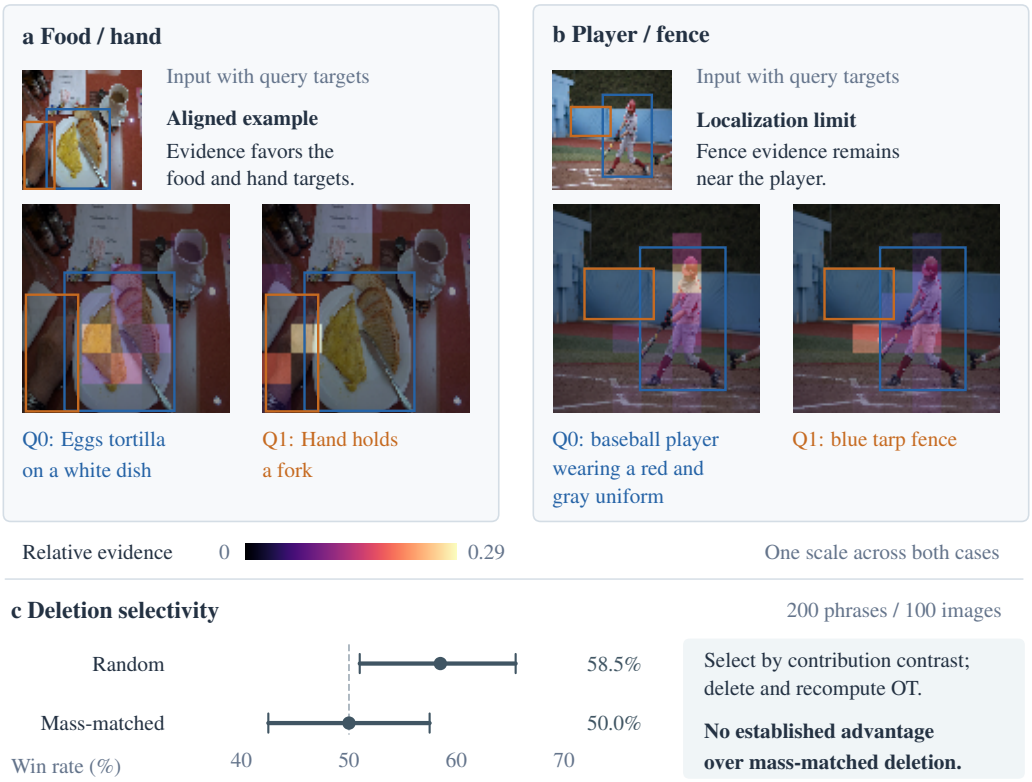


Figure 4: Relative phrase evidence and deletion selectivity. (a,b) An aligned example and a localization failure. Maps represent relative score evidence from mass-centered component-contribution contrasts, at native 7×7 resolution on a shared color scale. Food/hand was selected post hoc from five cases passing grounding and both deletion-control checks for both phrases; player/fence was a previously inspected failure. (c) Components selected by the original contribution contrast are deleted, remaining masses are renormalized, and OT is recomputed. Selectivity is the queried phrase's score drop minus the other's; win rates are measured against random and mass-matched controls. Error bars are 95% image-bootstrap intervals over 100 development images at the fixed seed-0 checkpoint.

5 Conclusion

We presented an image–text retrieval framework combining probabilistic component representations with one-sided partial transport to model differences between image content and caption scope. The framework uses vMF distribution overlap to measure component compatibility and combines learned masses with partial matching constraints to determine component contributions to retrieval scores, allowing some image mass to remain unmatched to real text components. Experiments show that entity content support, cross-modal correspondence, and matching participation do not always improve together in the evaluated configurations, motivating joint consideration of local representation quality and its role in scoring. Future work will explore adaptive matched-mass ratios and learning mechanisms that coordinate entity supervision with scoring participation.

AI Use Statement

Generative AI was used primarily to improve the language and clarity of the manuscript and to assist with translation and LaTeX editing. It also supported research record organization, implementation checks, literature review, and preliminary checks of supporting material.

REFERENCES

Hani Alomari, Anushka Sivakumar, Andrew Zhang, and Chris Thomas. Maximal matching matters: Preventing representation collapse for robust cross-modal retrieval. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31769–31785, 2025. doi: 10.18653/v1/2025.acl-long.1533. URL https://aclanthology.org/2025.acl-long.1533/.

Arindam Banerjee, Inderjit S. Dhillon, Joydeep Ghosh, and Suvrit Sra. Clustering on the unit hypersphere using von Mises–Fisher distributions. *Journal of Machine Learning Research*, 6:1345–1382, 2005. URL https://www.jmlr.org/papers/v6/banerjee05a.html.

Ioana Bica, Anastasija Ilic, Matthias Bauer, Goker Erdogan, Matko Bošnjak, Christos Kaplanis, Alexey A. Gritsenko, Matthias Minderer, Charles Blundell, Razvan Pascanu, and Jovana Mitrovic. Improving fine-grained understanding in image-text pre-training. In *International Conference on Machine Learning*, volume 235, pp. 3974–3995, 2024. URL https://proceedings.mlr.press/v235/bica24a.html.

Laetitia Chapel, Mokhtar Z. Alaya, and Gilles Gasso. Partial optimal transport with applications on positive-unlabeled learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://arxiv.org/abs/2002.08276.

Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. Learning the best pooling strategy for visual semantic embedding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15789–15798, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Chen_Learning_the_Best_Pooling_Strategy_for_Visual_Semantic_Embedding_CVPR_2021_paper.html.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. URL https://arxiv.org/abs/1504.00325.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: UNiversal Image-TExt representation learning. In *European Conference on Computer Vision*, 2020. URL https://arxiv.org/abs/1909.11740v3.

Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018. doi: 10.1090/mcom/3303. URL https://arxiv.org/abs/1607.05816.

Sanghyuk Chun. Improved probabilistic image-text representations. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/ad9d6ab10446114cf5482d5e1f971a84-Abstract-Conference.html.

Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio de Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. URL https://arxiv.org/abs/2101.05068v2.

Sanghyuk Chun, Wonjae Kim, Song Park, Minsuk Chang, and Seong Joon Oh. ECCV caption: Correcting false negatives by collecting machine-and-human-verified image-caption associations for MS-COCO. In *European Conference on Computer Vision*, 2022. URL https://www.ecva.net/papers/eccv_2022/papers_ECCV/papers/136680001.pdf.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transportation distances. In *Advances in Neural Information Processing Systems*, volume 26, 2013. URL https://arxiv.org/abs/1306.0895.

Li Ju, Max Andersson, Stina Fredriksson, Edward Glöckner, Andreas Hellander, Ekta Vats, and Prashant Singh. Exploiting the asymmetric uncertainty structure of pre-trained VLMs on the unit hypersphere. *arXiv preprint arXiv:2505.11029*, 2025. URL https://arxiv.org/abs/2505.11029v2. Version 2.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. MDETR—modulated detection for end-to-end multi-modal understanding. In *IEEE/CVF International Conference on Computer Vision*, 2021. URL https://arxiv.org/abs/2104.12763v2.

Dongwon Kim, Namyup Kim, and Suha Kwak. Improving cross-modal retrieval with set of diverse embeddings. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. URL https://arxiv.org/abs/2211.16761v3.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *European Conference on Computer Vision*, 2018. URL https://arxiv.org/abs/1803.08024v2.

Zhengxin Pan, Haishuai Wang, Fangyu Wu, Bailing Zhang, Jiajun Bu, and Hongyang Chen. Learning image-text matching with optimal partial transport. *arXiv preprint arXiv:2603.14349*, 2026. URL https://arxiv.org/abs/2603.14349v1. Version 1.

Zarana Parekh, Jason Baldridge, Daniel Cer, Austin Waters, and Yinfei Yang. Crisscrossed captions: Extended intramodal and intermodal semantic similarity judgments for MS-COCO. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 2855–2870, 2021. doi: 10.18653/v1/2021.eacl-main.249. URL https://aclanthology.org/2021.eacl-main.249/.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k Entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *IEEE International Conference on Computer Vision*, 2015. URL https://www.cv-foundation.org/openaccess/content_iccv_2015/papers/Plummer_Flickr30k_Entities_Collecting_ICCV_2015_paper.pdf.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Yale Song and Mohammad Soleymani. Polysemous visual-semantic embedding for cross-modal retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. URL https://arxiv.org/abs/1906.04402v2.

Uddeshya Upadhyay, Shyamgopal Karthik, Massimiliano Mancini, and Zeynep Akata. ProbVLM: Probabilistic adapter for frozen vision-language models. In *IEEE/CVF International Conference on Computer Vision*, 2023. URL https://arxiv.org/abs/2307.00398v3.

Luke Vilnis and Andrew McCallum. Word representations via Gaussian embedding. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6623.

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. URL https://arxiv.org/abs/2111.07783v1. Version 1.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. doi: 10.1162/tacl_a_00166. URL https://aclanthology.org/Q14-1006/.

Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. In *International Conference on Machine Learning*, 2022. URL https://arxiv.org/abs/2111.08276v3.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Zhai_Sigmoid_Loss_for_Language_Image_Pre-Training_ICCV_2023_paper.html.

A Representation and Numerical Details

A.1 Token Routing and Parameter Heads

The representation uses spatial image tokens after removing the global special token. Text padding and special tokens are excluded; when no content token remains, the implementation falls back to the non-padding tokens. Modality-specific projections map encoder outputs to $v_t^m = W_p^m h_t^m + b_p^m$. Routing logits are computed by

$$\ell_t^m = \gamma_m \left[W_r^m \,\mathrm{GELU}\left(W_b^m v_t^m + b_b^m + r_t^m \right) + b_r^m \right]. \tag{11}$$

The image-side residual aggregates features at corresponding spatial positions from selected encoder layers:

$$r_t^I = W_f \,\mathrm{Concat}_{\ell \in \mathcal{L}} \left[\mathrm{LN}(h_t^{I,(\ell)}) \right] + b_f, \qquad r_t^T = 0. \tag{12}$$

Layer normalization has no learned affine parameters, and the residual projection is initialized to zero. Multilayer features affect the routing logits; the pooled values remain the projected final-layer tokens. The text-side residual is zero. The component heads are

$$\begin{aligned} \mu_k^m &= \frac{u_k^m}{\|u_k^m\|_2}, \\ \kappa_k^m &= \mathrm{softplus}\left((w_\kappa^m)^\top u_k^m + b_\kappa^m \right) + \kappa_{\min}, \\ \pi_k^m &= \frac{\exp((w_\pi^m)^\top u_k^m + b_\pi^m)}{\sum_a \exp((w_\pi^m)^\top u_a^m + b_\pi^m)}. \end{aligned} \tag{13}$$

Direction normalization uses a denominator floor of 10^{-8}. The positive offset $\kappa_{\min}$ keeps concentrations away from zero. These heads learn mass independently of token support: a component's routing distribution does not determine its transport marginal.

A.2 Distributional Scores and Transport Solver

The vMF normalizer in (3) is

$$C_d(\kappa) = \frac{\kappa^{d/2-1}}{(2\pi)^{d/2} I_{d/2-1}(\kappa)}, \tag{14}$$

where I_ν is the modified Bessel function of the first kind. The implementation evaluates $\log C_d$ through a second-order Debye approximation. Thus the analytic overlap in (4) is evaluated approximately in the implemented score.

For the augmented transport problem, dual potentials f and g give a plan of the form

$$\overline{P}_{ij} = \exp\left(\frac{\overline{G}_{ij} + f_i + g_j}{\varepsilon} \right). \tag{15}$$

The solver uses a geometric schedule from a larger entropy regularization strength to the target ε, warm-starting the dual potentials at each stage. Stabilized Sinkhorn updates (Cuturi, 2013) and damped Newton updates enforce the prescribed marginals. Newton updates fix the final text potential to zero to remove the additive gauge freedom. The larger-regularization stages provide initialization and are detached; gradients pass through the target-regularization solve.

Convergence is checked using the maximum absolute row and column marginal residual. If the final residual is above the tolerance but no greater than ten times it, the plan is corrected by scaling rows and columns down to avoid exceeding their targets, then adding a nonnegative rank-one correction for the remaining marginal deficits. A larger residual, or a failed correction, is treated as a solver

failure. This procedure approximates the entropy-regularized plan while checking feasibility at the specified tolerance.

For each candidate pair, component compatibility requires $O(K_I K_T d)$ time, and the augmented plan occupies $O(K_I(K_T + 1))$ space. Solver cost additionally depends on its iteration count and the dual linear systems. These quantities exclude encoder computation.

B Entity Supervision

For an annotated sample, let $\mathcal{E}$ be the valid entities and let T_e^m be the nonnegative token support target for entity e, normalized so that $\sum_t T_{et}^m = 1$. Image targets are derived from entity boxes and text targets from associated phrase spans. Each modality assigns entities injectively to components:

$$a_m^* = \arg \min_{a_m : \mathcal{E} \hookrightarrow \{1,\dots,K_m\}} \sum_{e \in \mathcal{E}} \left[-\sum_t T_{et}^m \log A_{a_m(e),t}^m \right]. \tag{16}$$

This requires $|\mathcal{E}| \le K_m$. A linear assignment solver selects the assignment, which is held fixed during backpropagation. The loss for an annotated sample is

$$\begin{aligned} \mathcal{L}_{\text{sup}}^m &= \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \left[-\sum_t T_{et}^m \log A_{a_m^*(e),t}^m \right], \\ \mathcal{L}_{\text{spatial}} &= \frac{\mathcal{L}_{\text{sup}}^I + \mathcal{L}_{\text{sup}}^T}{2}. \end{aligned} \tag{17}$$

Losses are averaged first across valid entities within a sample, then across samples with valid annotations. Samples without valid entities contribute no auxiliary loss. This spatial term does not directly supervise transport mass; the ranking and optional identity terms in (10) are defined below.

B.1 Ranking and Identity Objectives

The entity experiments replace the core BCE objective with bidirectional multi-positive ranking. Let $\mathcal{P}_a = \{b : M_{ab} = 1\}$ and $\mathcal{Q}_b = \{a : M_{ab} = 1\}$ collect captions and images with matching image IDs. With $r_{ab} = \alpha s_{ab}$,

$$\mathcal{L}_{\text{rank}} = -\frac{1}{2B} \left[\sum_a \log \frac{\sum_{b \in \mathcal{P}_a} \exp r_{ab}}{\sum_b \exp r_{ab}} + \sum_b \log \frac{\sum_{a \in \mathcal{Q}_b} \exp r_{ab}}{\sum_a \exp r_{ab}} \right]. \tag{18}$$

For entity e, let $i_e = a_I^*(e)$ and $j_e = a_T^*(e)$. Identity supervision applies bidirectional component cross-entropy to the pre-transport compatibility:

$$\ell_e^{\text{id}} = -\frac{1}{2} \left[\log \frac{\exp(G_{i_e j_e}/\tau)}{\sum_{i \in \mathcal{C}_e^I} \exp(G_{i j_e}/\tau)} + \log \frac{\exp(G_{i_e j_e}/\tau)}{\sum_{j \in \mathcal{C}_e^T} \exp(G_{i_e j}/\tau)} \right]. \tag{19}$$

Each candidate set retains the target and all unassigned components. Another assigned component is excluded when its entity support overlaps the target with coefficient at least 0.5. To compute this coefficient, each support vector is first divided by its maximum; the sum of elementwise minima is then divided by the smaller support sum. We average ℓ_e^{id} over entities and then over samples with at least two valid entities, using $\tau = 1$. Neither compatibility nor this loss is weighted by component mass.

All three entity arms use (18). Spatial supervision uses $\lambda_{\text{spatial}} = .002$; the spatial-plus-identity arm additionally uses $\lambda_{\text{id}} = .002$. Both CLIP encoders are fine-tuned from the first epoch for 25 epochs, with batch size 128 and configured annotation fraction 0.05. Results use single-seed Flickr validation-selected checkpoints. These conditions differ from the core COCO BCE protocol.

C Experimental Protocols and Extended Results

Models and training protocols The complete configuration uses CLIP ViT-B/32 encoders, four vMF components per modality, multilayer image routing, learned component masses, and one-sided

Table 5: Overview of the experimental protocols. Core and entity protocols use different retrieval objectives; see Appendix B.1.

Purpose	Training data	Encoder	K	Matching / supervision	Selection	Reporting	Seed
Complete model	COCO	Freeze epoch 1; tune both	4	Multilayer; $\rho = .75$, $\varepsilon = .2$	COCO val	COCO, CxC, EC, Flickr	0
Structural 2×2	COCO	Frozen	4	Full/partial OT; toggle multilayer	COCO val	COCO, CxC, EC, Flickr	0
Component count	COCO	Frozen	2/4/8	Base full OT; pair supervision	COCO val	COCO val; local probes	0
Unmatched capacity	COCO	Frozen	4	Multilayer; $1 - \rho = 0/.1/.25/.4/.5$	COCO val	COCO val; local probes	0
Coverage scoring	Flickr Entities val	Fixed historical model	4	Fixed features; $\rho = .75/1$	None	Flickr Entities val	–
Entity extension	Flickr30k	Fine-tune both	4	$\rho = .75$; retrieval/spatial/identity	Flickr val	Flickr val; entity probes	0

Table 6: Published CLIP ViT-B/32 reference results from Table 1 of Chun (2024). Trained reference methods report three-run means; the complete OT configuration is a local single run under a different training protocol.

Method	Source / statistic	Prob.	COCO 5K R@1	COCO 1K R@1	COCO 1K RSUM	CxC R@1	EC mAP@R	EC R-P
CLIP zero-shot	Published reference	No	40.3	59.5	471.9	42.0	26.8	36.9
InfoNCE	Published, 3-run mean	No	53.0	74.0	532.6	54.9	39.0	48.7
VSE∞	Published, 3-run mean	No	55.2	75.5	536.5	57.1	40.0	49.5
PCME	Published, 3-run mean	Yes	53.0	73.8	532.0	54.7	39.1	48.9
PCME++	Published, 3-run mean	Yes	55.1	75.4	537.0	56.8	40.1	49.7
Complete OT configuration	Local, single run	Yes	50.372	71.338	524.680	52.107	36.367	46.346

partial OT with matched-mass ratio $\rho = 0.75$ and entropy regularization $\varepsilon = 0.2$. We train for 25 epochs, freezing both CLIP encoders in the first epoch and fine-tuning them with a smaller learning rate thereafter. The checkpoint with the highest COCO validation RSUM is selected. This configuration has currently been trained with one seed, so its results characterize the executed protocol rather than variation across training runs.

The ablations use controlled protocols tailored to each question. The 2×2 structural study fixes $K = 4$ and toggles multilayer routing and the unmatched channel. The component-count study uses base full OT and varies $K \in \{2, 4, 8\}$. The unmatched-capacity study fixes multilayer routing and varies $1 - \rho \in \{0, .1, .25, .4, .5\}$. These experiments freeze CLIP, train for 25 epochs with seed 0, and select checkpoints by COCO validation RSUM. The coverage study performs no retraining and instead changes the scoring rule for fixed representations. The entity extension compares retrieval-only training, spatial support supervision, and spatial support plus vMF identity supervision on Flickr30k.

The two full-OT runs in the structural study were resumed from checkpoints after 12 complete epochs because of a logging I/O failure. Optimizer and loop state were retained, the incomplete epoch was replayed, and the model, data, and solver protocols were unchanged. All four runs completed 25 epochs and the same number of valid updates, although batchwise bit-level equivalence across the resume was not separately established.

The complete configuration reaches 533.36 COCO validation RSUM at its selected epoch 25. In the local allocation analysis, it has no OT fallback; marginal errors before and after deletion are below 10^{-4}, and the maximum recorded training dustbin-mass error is 3.58×10^{-7}. These checks validate numerical feasibility and score reconstruction, not retrieval utility.

C.1 Published Retrieval References

Table 6 reproduces selected ViT-B/32 results from the common evaluation in Chun (2024). Reference methods share CLIP initialization and optimization hyperparameters; trained references report three-run means. VSE∞ is attributed to Chen et al. (2021), while the numbers come from PCME++. The local OT row uses a different objective and selection protocol, so this is a reference band rather than a controlled ranking.

C.2 Parameter Sensitivity

C.2.1 Number of Components

Under the base full-OT protocol, $K = 2$ attains a best validation RSUM of 499.10, slightly above 498.04 for $K = 4$, but its mean over the final five epochs is lower. Increasing the set to $K = 8$

Table 7: Component-count sweep under base full OT.

K	Best epoch	Best RSUM	Last-5 RSUM	Grounding gap	Flip stability	Norm. mass entropy
2	23	499.10	495.28	0.02348	0.95127	0.9922
4	24	498.04	496.08	0.06389	0.93866	0.9948
8	7	456.82	449.88	0.12233	0.54282	0.9680

Table 8: Unmatched-capacity sweep. The complete model retains the prespecified 25% capacity.

Unmatched capacity	0%	10%	25%	40%	50%
Best val RSUM	497.22	459.94	498.80	498.22	498.84
Last-5 RSUM	495.22	454.11	495.69	494.96	496.20

reduces the best RSUM to 456.82 and yields an early selected checkpoint. Retrieval performance is therefore not monotonic in the number of components.

The $K = 8$ model has a larger grounding gap while performing substantially worse in retrieval and perturbation stability. A higher local diagnostic or raw effective rank is thus not a substitute for task utility. This sweep supports empirical selection of K and shows that a larger component set is not automatically more useful. Because the sweep does not include $K = 1$, it does not establish an advantage of multiple components over a single component.

C.2.2 Matched-Mass Ratio

The best validation RSUM values for unmatched capacities of 25%, 40%, and 50% are 498.80, 498.22, and 498.84. The highest point in this single-seed coarse grid, 50%, exceeds 25% by only 0.04 RSUM. The 10% configuration falls to 459.94, showing that the response is non-monotonic.

The evidence therefore supports only a descriptive plateau over the medium-to-high capacity range in this protocol. It does not identify 50% as a generally optimal capacity, nor does the isolated degradation at 10% establish a continuous capacity law. The complete configuration retains the prespecified 25% capacity; this setting is not retrospectively described as the optimum selected from the grid.

C.3 Computational Scope

Component sets can be cached independently, but candidate scoring computes a $K_I \times K_T$ compatibility matrix and a transport plan. Complexity is given in Appendix A.2. Existing end-to-end logs differ in evaluation stages, hardware load, and encoder state and do not constitute a controlled timing benchmark. No wall-clock advantage is established.