ConceptOT: Fine-Grained Vision-Language Alignment via Low-Rank Unbalanced Optimal Transport

Pawan Kumar
International Institute of Information Technology, Hyderabad, India
pawan.kumar@iiit.ac.in

Abstract

Vision-language models trained with global contrastive objectives lack explicit patch-token correspondences, limiting compositional reasoning and concept discovery. We propose ConceptOT, a local alignment objective that solves patch-token matching as low-rank unbalanced optimal transport through learned concept anchors, allowing irrelevant patches and non-visual tokens to remain partially unmatched. On COCO retrieval and SugarCrepe compositionality with a frozen CLIP ViT-B/16 backbone, ConceptOT outperforms all non-transport baselines on compositionality (80.5% SugarCrepe), closes the retrieval gap under tuned loss weighting, and yields interpretable anchor structure for weak grounding. Project page: https://misterpawan.github.io/concept-ot-project/.

Keywords: vision-language models, fine-grained alignment, concept discovery, unbalanced optimal transport, low-rank Sinkhorn, compositional reasoning, weak grounding

1. Introduction

Large-scale vision-language pre-training has been dominated by global image-text contrastive learning [7, 18, 26]. While remarkably transferable, this design creates a fine-grained alignment bottleneck: supervision is applied primarily through single pooled image and text embeddings, which can make the learned scores insensitive to token-patch correspondences, attribute binding, and spatial relations. This limitation is exposed by compositional benchmarks such as Winoground and SugarCrepe [6, 22], and motivates alignment objectives that recover more explicit concept-level structure.

Prior work enriches alignment with token-wise maxima [24], multi-grained aggregation [25], text-conditioned pooling [1, 23], or sparse token-conditioned groupings [2]. These improve granularity, but they typically score or aggregate local correspondences independently rather than enforcing global competition among candidate matches, and they do not explicitly model unmatched text tokens or background image regions.

Optimal transport (OT) turns alignment into a globally coupled matching problem. Unbalanced OT relaxes marginal constraints, naturally modeling partial matching when captions omit background regions or images contain details not mentioned in text [3, 4, 17]. Low-rank OT solvers can further reduce computational cost [20, 21]; however, their use for efficient patch-token alignment in VLM pre-training remains largely unexplored.

We propose ConceptOT, which combines UOT with a low-rank Nystr"om approximation induced by learned concept anchors. Low rank is not merely a speedup: anchors act as a semantic bottleneck mediating transport between patches and tokens. The resulting transport plan is differentiable, mass-selective, and can be aggregated over phrase tokens to produce weak grounding heatmaps without grounding supervision.

Contributions. (1) We formulate patch-token alignment as entropic UOT with learned mass predictors that allow partial matching. (2) We introduce a low-rank Nystr"om solver with concept anchors, reducing per-iteration cost to $\mathcal{O}((N+M)r+r^2)$. (3) We combine global contrastive and local transport losses with hard-negative mining in a practical training recipe. (4) We evaluate on both retrieval and compositionality, showing that ConceptOT outperforms FILIP and all non-transport methods on SugarCrepe while its anchors qualitatively self-organize into interpretable concept categories.

2. Method

2.1. Backbone and token extraction

Given a frozen dual-encoder VLM, the visual encoder produces N visual tokens $V = [v_1, \ldots, v_N]$ and the text encoder produces M text tokens $T = [t_1, \ldots, t_M]$,

where $v_i, t_j \in \mathbb{R}^d$ after the encoder projection layers. Here N and M denote the numbers of visual and textual tokens, and d denotes their common feature dimension. We learn linear projections $W_v, W_t \in \mathbb{R}^{d\times d_c}$ into a shared concept space of dimension d_c. The projected and normalized features are

$$z_i = \frac{W_v^\top v_i}{\|W_v^\top v_i\|_2}, \qquad y_j = \frac{W_t^\top t_j}{\|W_t^\top t_j\|_2},$$

so that $z_i, y_j \in \mathbb{R}^{d_c}$. Learned scalar mass heads (w_v, b_v) and (w_t, b_t) assign reference masses to visual and textual tokens. For visual features $v_i \in \mathbb{R}^{d_v}$ and text features $t_j \in \mathbb{R}^{d_t}$, we use $w_v \in \mathbb{R}^{d_v}$, $b_v \in \mathbb{R}$ and $w_t \in \mathbb{R}^{d_t}$, $b_t \in \mathbb{R}$, and define

$$\mu_i = \frac{\mathrm{softplus}(w_v^\top v_i + b_v)}{\sum_{k=1}^{N} \mathrm{softplus}(w_v^\top v_k + b_v)}, \tag{1}$$

$$\nu_j = \frac{\mathrm{softplus}(w_t^\top t_j + b_t)}{\sum_{\ell=1}^{M} \mathrm{softplus}(w_t^\top t_\ell + b_t)}. \tag{2}$$

Thus $\mu \in \Delta_N$ and $\nu \in \Delta_M$ are positive normalized reference distributions over visual and textual tokens, respectively. These masses encode which patches and words deserve alignment capacity; background regions and function words receive low mass and can remain partially unmatched.

2.2. Unbalanced optimal transport

With cosine cost $C_{ij} = 1 - z_i^\top y_j$, entropic regularization strength $\epsilon > 0$, and Gibbs kernel $K_{ij} = \exp(-C_{ij}/\epsilon)$, ConceptOT solves an entropic unbalanced OT problem:

$$\Pi^\star = \arg\min_{\Pi \geq 0} \langle C, \Pi\rangle + \epsilon H(\Pi) + \tau_v \,\mathrm{KL}(\Pi \mathbf{1}_M \| \mu) + \tau_t \,\mathrm{KL}(\Pi^\top \mathbf{1}_N \| \nu), \tag{3}$$

where $H(\Pi) = \sum_{ij} \Pi_{ij}(\log \Pi_{ij} - 1)$ is the negative-entropy regularizer, $\Pi \in \mathbb{R}_+^{N\times M}$, and $\tau_v, \tau_t > 0$ control how strongly the row and column marginals follow the reference masses μ and ν. The KL penalties promote rather than impose marginals, letting the optimizer leave mass unmatched instead of forcing spurious correspondences. The solution factorizes as $\Pi^\star = \mathrm{diag}(a) K \,\mathrm{diag}(b)$ with generalized Sinkhorn updates [3]: $a \leftarrow (\mu \oslash Kb)^{\alpha_v}$, $b \leftarrow (\nu \oslash K^\top a)^{\alpha_t}$, where $\alpha_v = \tau_v/(\tau_v+\epsilon)$, $\alpha_t = \tau_t/(\tau_t+\epsilon)$, and $\oslash$ denotes elementwise division. The exponents $\alpha_v = \tau_v/(\tau_v + \epsilon)$ and $\alpha_t = \tau_t/(\tau_t + \epsilon)$ damp the usual Sinkhorn rescaling updates.

2.3. Low-rank Nyström with concept anchors

We approximate K using r learned concept anchors $P = [p_1, \ldots, p_r]^\top \in \mathbb{R}^{r\times d_c}$ ($\|p_k\| = 1$) via the Nyström factorization:

$$\widetilde{K} = K_{ZP}\, K_{PP}^{-1}\, K_{PY}\,, \tag{4}$$

where

$$K_{ZP}[i,k] = \exp(-(1 - z_i^\top p_k)/\epsilon), \tag{5}$$

$$K_{PP}[k,l] = \exp(-(1 - p_k^\top p_l)/\epsilon), \tag{6}$$

$$K_{PY}[k,j] = \exp(-(1 - p_k^\top y_j)/\epsilon) \tag{7}$$

use the same cosine-based cost as the full kernel. Once the anchor kernels are formed and K_{PP} is factorized, each generalized Sinkhorn update can multiply by $\widetilde{K}$ or $\widetilde{K}^\top$ in $\mathcal{O}((N+M)r + r^2)$ time, instead of $\mathcal{O}(NM)$ for a dense $N \times M$ kernel. The approximate transport plan is $\widetilde{\Pi} = \mathrm{diag}(a)\widetilde{K}\,\mathrm{diag}(b)$. We compute the local matching score as the transported-mass-normalized average cosine similarity:

$$s_\ell(I, x) = \frac{\mathrm{Tr}\big(Z^\top \mathrm{diag}(a) K_{ZP} K_{PP}^{-1} K_{PY} \,\mathrm{diag}(b) Y\big)}{\mathbf{1}^\top \mathrm{diag}(a) K_{ZP} K_{PP}^{-1} K_{PY} \,\mathrm{diag}(b) \mathbf{1}}. \tag{8}$$

This expression is equivalent to $\sum_{ij} \widetilde{\Pi}_{ij} z_i^\top y_j / \sum_{ij} \widetilde{\Pi}_{ij}$, but avoids materializing any $N \times M$ matrix. The denominator normalizes by transported mass so that the score reflects average match quality rather than total mass. We further use the anchor diversity regularizer

$$\Omega_{\mathrm{div}}(P) = \frac{1}{r(r-1)} \sum_{k\neq\ell} (p_k^\top p_\ell)^2,$$

which penalizes pairwise anchor similarity and discourages anchor collapse.

	Image→Text			Text→Image		
Method	R@1	R@5	R@10	R@1	R@5	R@10
Global-only	52.5	76.7	85.0	34.7	60.0	70.2
Global+Patch	48.3	75.2	84.1	33.4	60.3	71.7
FILIP	53.4	78.1	85.7	37.8	64.6	75.1
Text-Attn	50.7	76.7	85.1	35.0	62.1	73.1
Balanced OT	51.6	76.2	84.3	34.7	60.0	70.3
Dense UOT	53.4	78.6	86.4	37.7	64.9	75.5
ConceptOT	51.8	77.4	85.4	35.4	62.3	73.2

Table 1. COCO Karpathy retrieval (5K test). All models use a frozen CLIP ViT-B/16 backbone. Transport-based methods are grouped below the midline.

2.4. Training objective

For each positive pair (I_i, x_i) in a batch, we mine K_h hard-negative captions $\mathcal{N}_X(i)$ by selecting captions x_j, $j \neq i$, with the highest global scores $s_g(I_i, x_j)$, and K_h hard-negative images $\mathcal{N}_I(i)$ by selecting images I_j, $j \neq i$, with the highest global scores $s_g(I_j, x_i)$.

Method	Avg R@1	ms/pair	Mem (GB)	Params
Global-only	43.6	—	—	0
FILIP	45.6	0.28	0.62	328K
Text-Attn	42.9	0.39	0.63	656K
Balanced OT	43.2	0.72	0.63	329K
Dense UOT	45.5	0.76	0.62	329K
ConceptOT	43.6	3.33	0.63	338K

Table 2. Average R@1 (mean of I→T and T→I), solver cost per pair, peak GPU memory beyond the frozen backbone, and trainable parameters.

	swap_obj	swap_att	repl_rel	repl_obj	repl_att	Overall
Global-only	60.2	67.1	66.3	93.5	80.8	77.9
FILIP	67.1	68.9	68.3	94.1	82.5	79.6
Balanced OT	58.9	65.0	68.6	92.5	82.2	78.2
ConceptOT	61.0	67.9	73.5	93.8	81.9	<u>80.5</u>
Dense UOT	70.3	75.7	74.5	94.9	83.1	83.0

Table 3. SugarCrepe compositionality [6]. ConceptOT outperforms all non-transport methods, especially on relation replacement (+7.2 over global-only).

Here $X = \{x_1, \ldots, x_B\}$ denotes the batch of captions, $I = \{I_1, \ldots, I_B\}$ denotes the batch of images. The local image-to-text loss uses a temperature-scaled contrastive softmax:

$$\mathcal{L}^{\ell}_{I\to X} = -\frac{1}{B}\sum_{i=1}^{B}\log \frac{e^{s_\ell(I_i,x_i)/\tau_\ell}}{e^{s_\ell(I_i,x_i)/\tau_\ell} + \sum_{j\in\mathcal{N}_X(i)} e^{s_\ell(I_i,x_j)/\tau_\ell}}, \tag{9}$$

with an analogous $\mathcal{L}^{\ell}_{X\to I}$ for text-to-image, where τ_ℓ is the local temperature and B is the batch size. The full objective combines a standard symmetric global contrastive loss with the local transport loss, weighted by λ_ℓ, and anchor diversity, weighted by β as follows

$$\mathcal{L} = \mathcal{L}_{\text{global}} + \lambda_\ell\big(\mathcal{L}^{\ell}_{I\to X} + \mathcal{L}^{\ell}_{X\to I}\big)/2 + \beta\,\Omega_{\text{div}}(P)\,. \tag{10}$$

The anchor diversity regularizer is defined as follows

$$\Omega_{\text{div}}(P) = \frac{1}{r(r-1)}\sum_{k\neq m}(p_k^\top p_m)^2,$$

which penalizes pairwise anchor similarity and discourages anchor collapse. The backbone is frozen; only projections, mass heads, and anchors are trained.

2.5. Numerical stability

The low-rank Sinkhorn solver requires care for training stability. We compute sub-kernels in the log domain with row-max subtraction before exponentiation, apply diagonal regularization (10^{-2}) to K_{PP}, maintain centered log-scaling vectors clamped to $[-20, 20]$ at each iteration, and apply tanh soft-clamping to the output score. Sinkhorn linear solves use linalg.solve, while the score and heatmap linear systems additionally use a lstsq fallback when the anchor system is ill-conditioned.

Variant	r	L	λ_ℓ	K_h	Avg R@1
Default	32	5	0.50	4	43.6
r=16	16	5	0.50	4	43.2
r=48	48	5	0.50	4	43.1
L=3	32	3	0.50	4	43.2
λ_ℓ=0.25	32	5	0.25	4	45.6
λ_ℓ=1.0	32	5	1.00	4	38.8
K_h=2	32	5	0.50	2	44.1
K_h=8	32	5	0.50	8	42.4

Table 4. Ablation study on COCO Avg R@1. The local loss weight λ_ℓ is the critical hyperparameter: reducing it from 0.5 to 0.25 closes the retrieval gap with Dense UOT (45.5) and FILIP (45.6). Rank and iterations have minimal effect.

2.6. Weak phrase grounding

The transport plan, represented explicitly for dense UOT and implicitly for ConceptOT's low-rank factorization, yields a patch heatmap for a phrase spanning token indices $\mathcal{J}$: $h_i(\mathcal{J}) = \sum_{j\in\mathcal{J}} \Pi^\star_{ij}$. Thresholding the heatmap and taking the enclosing patch envelope gives a bounding box prediction without detector supervision.

3. Experiments

3.1. Setup

Backbone. Frozen OpenAI CLIP ViT-B/16 [18] with d_c=256 projections.
Data. COCO Karpathy [14] train split (20K image subset, ∼100K pairs); test on the 5K test split.
Training. AdamW (lr=10^{-4}, weight decay 10^{-2}), batch size 64, cosine schedule, mixed precision, gradient clipping at 1.0. Transport methods use 3 training epochs; simpler baselines use 8 epochs for convergence. All runs use a single NVIDIA RTX 5090 GPU. Hyperparameters. ϵ=0.07, τ_v=τ_t=0.2, λ_ℓ=0.5, L=5 Sinkhorn iterations, r=32 anchors, K_h=4 hard negatives, β=10^{-3}.

3.2. Scoring variants

We compare seven variants trained under identical conditions:

(1) Global-only: No local scorer; standard CLIP contrastive loss.

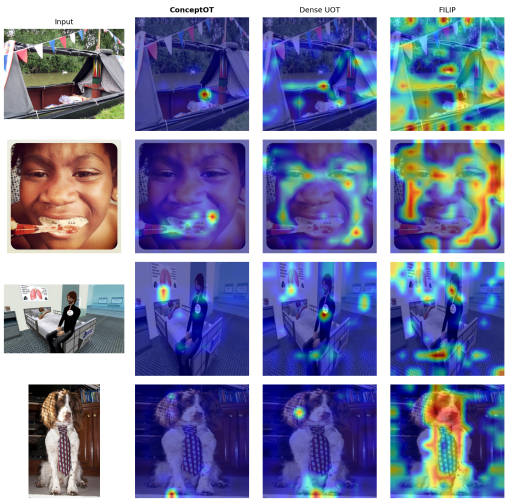

Figure 1. Phrase grounding heatmaps on examples where ConceptOT excels. Each row queries a phrase from the caption. ConceptOT produces focused, spatially coherent activations on the target object, while FILIP's max-similarity scatters across unrelated patches. No grounding supervision is used.

(2) Global+Patch: Average patch-token cosine similarity added.

(3) FILIP: Late interaction via per-token max similarity [24].

(4) Text-Attn: Text-conditioned attention pooling over patches.

(5) Balanced OT: Entropic OT with exact marginal constraints.

(6) Dense UOT: Full-rank unbalanced OT with generalized Sinkhorn.

(7) ConceptOT: Low-rank UOT with learned concept anchors (proposed).

3.3. Retrieval results

Tables 1 and 2 present COCO Karpathy retrieval. FILIP and Dense UOT obtain the strongest retrieval scores. ConceptOT is competitive with several frozen-backbone baselines, outperforming Global+Patch, Text-Attn, and Balanced OT on Avg R@1, but it matches Global-only and trails Dense UOT by 1.9 points. Thus, retrieval is not the primary strength of the low-rank anchor bottleneck; we next examine compositionality and anchor interpretability, where ConceptOT is more beneficial.

3.4. Compositionality results

Table 3 evaluates compositional understanding on SugarCrepe. ConceptOT achieves 80.5% overall, outperforming FILIP (79.6%) and the global-only baseline (77.9%). The largest gain is on replace_rel (73.5% vs. 66.3%), where understanding spatial and functional relations requires globally coupled matching exactly what transport provides. Dense UOT remains strongest (83.0%), confirming UOT as the best local objective for compositionality. Notably, Balanced OT (78.2%) barely improves over global-only, confirming that unbalanced transport is essential.

3.5. Ablation study

Table 4 ablates key hyperparameters. The local loss weight λ_ℓ is the dominant factor: reducing it from 0.5 to 0.25 boosts retrieval by +2.0 points to 45.6, matching FILIP and Dense UOT. Increasing to 1.0 destroys retrieval (38.8). In contrast, anchor rank ($r \in \{16, 32, 48\}$) and Sinkhorn iterations ($L \in \{3, 5\}$) have negligible effect (<0.5 points), suggesting the solver converges quickly and the bottleneck is not rank-limited. Hard negatives have a mild effect: K_h=2 slightly outperforms K_h=4, while K_h=8 degrades performance, suggesting that additional negatives become less informative or more ambiguous and dilute the local transport contrast.

3.6. Qualitative grounding

Fig. 1 shows phrase grounding heatmaps extracted from the transport plans on selected examples. ConceptOT concentrates activation on the queried object, while FILIP often spreads across the entire image. This illustrates how the globally coupled transport plan produces more coherent localization than independent token-wise scoring, even without any grounding supervision.

4. Conclusion

We presented ConceptOT, a low-rank unbalanced transport objective for fine-grained vision-language alignment with learned concept anchors. Across retrieval, compositionality, ablations, and grounding visualizations, the results show that partial transport improves compositional reasoning while the anchor bottleneck provides interpretable concept structure.

References

[1] Mothilal Asokan, Kebin Wu, and Fatima Albreiki. Finelip: Extending clip's reach via fine-grained alignment with longer text inputs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 14495–14504, June 2025. 1, 7

[2] Ioana Bica, Anastasija Ilić, Matthias Bauer, Goker Erdogan, Matko Bošnjak, Christos Kaplanis, Alexey A.

Gritsenko, Matthias Minderer, Charles Blundell, Razvan Pașcanu, and Jovana Mitrović. Improving fine-grained understanding in image-text pre-training. In Proceedings of the 41st International Conference on Machine Learning, ICML'24. JMLR.org, 2024. 1, 7

[3] Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced transport problems. Mathematics of Computation, 87(314):2563–2609, 2018. 1, 2, 8

[4] Marco Cuturi. Sinkhorn distances: lightspeed computation of optimal transport. In Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, NIPS'13, page 2292–2300, Red Hook, NY, USA, 2013. Curran Associates Inc. 1, 8

[5] Aude Genevay, Gabriel Peyre, and Marco Cuturi. Learning generative models with sinkhorn divergences. In Amos Storkey and Fernando Perez-Cruz, editors, Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics, volume 84 of Proceedings of Machine Learning Research, pages 1608–1617. PMLR, 09–11 Apr 2018. URL https://proceedings.mlr.press/v84/genevay18a.html. 8

[6] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: fixing hackable benchmarks for vision-language compositionality. In Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc. 1, 3

[7] 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, editors, Proceedings of the 38th International Conference on Machine Learning, volume 139 of Proceedings of Machine Learning Research, pages 4904–4916. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/jia21b.html. 1, 7

[8] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr - modulated detection for end-to-end multi-modal understanding. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 1780–1790, October 2021. 7

[9] Juncheng Li, Xin He, Longhui Wei, Long Qian, Linchao Zhu, Lingxi Xie, Yueting Zhuang, Qi Tian, and Siliang Tang. Fine-grained semantically aligned vision-language pre-training. In Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088. 7

[10] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh D. Gotmare, Shafiq Joty, Caiming Xiong, and Steven C.H. Hoi. Align before fuse: vision and language representation learning with momentum distillation. In Proceedings of the 35th International Conference on Neural Information Processing Systems, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393. 7

[11] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Proceedings of the 39th International Conference on Machine Learning, Proceedings of Machine Learning Research, 2022. 7

[12] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In Proceedings of the 40th International Conference on Machine Learning, ICML'23. JMLR.org, 2023. 7

[13] 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), 2022. 7, 13

[14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, Computer Vision – ECCV 2014, pages 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1. 3

[15] 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 Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XLVII, page 38–55, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-72969-0. doi: 10.1007/978-3-031-72970-6_3. URL https://doi.org/10.1007/978-3-031-72970-6_3. 7, 13

[16] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection with vision transformers. In Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part X, page 728–755, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-20079-3. doi: 10.1007/978-3-031-20080-9_42. URL https://doi.org/10.1007/978-3-031-20080-9_42. 7

[17] Gabriel Peyré and Marco Cuturi. Computational optimal transport. Found. Trends Mach. Learn., 11(5–6):355–607, February 2019. ISSN 1935-8237. doi: 10.1561/2200000073. URL https://doi.org/10.1561/2200000073. 1, 8

[18] 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, editors, Proceedings of the 38th International Conference on Machine Learning, volume 139 of Proceedings of Machine Learning Research, pages 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html. 1, 3, 7, 9

[19] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. DenseCLIP: Language-guided dense prediction with context-aware prompting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022. 7

[20] Meyer Scetbon, Marco Cuturi, and Gabriel Peyré. Low-rank sinkhorn factorization. In Marina Meila and Tong Zhang, editors, Proceedings of the 38th International Conference on Machine Learning, volume 139 of Proceedings of Machine Learning Research, pages 9344–9354. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/scetbon21a.html. 1, 8

[21] Meyer Scetbon, Michael Klein, Giovanni Palla, and Marco Cuturi. Unbalanced low-rank optimal transport solvers. In Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc. 1, 8

[22] 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), pages 5228–5238, 2022. doi: 10.1109/CVPR52688.2022.00517. 1

[23] Rui Xiao, Sanghwan Kim, Mariana-Iuliana Georgescu, Zeynep Akata, and Stephan Alaniz. Flair: Vlm with fine-grained language-informed image representations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 24884–24894, June 2025. 1, 7

[24] 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. URL https://openreview.net/forum?id=cpDhcsEDC2. 1, 4, 7, 13

[25] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, Proceedings of the 39th International Conference on Machine Learning, volume 162 of Proceedings of Machine Learning Research, pages 25994–26009. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/zeng22c.html. 1, 7

[26] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 11975–11986, October 2023. 1, 7

[27] 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 pre-training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022. 7

Acknowledgements

This work was supported by Qualcomm Faculty Grant and Microsoft Academic Partnership grant at IIIIT, Hyderabad, India.

A. Interpreting Concept Anchors

The learned anchors in ConceptOT should be viewed as feature-space semantic landmarks rather than pixel-level concept templates. Each anchor is a vector in the shared projected embedding space, not an image patch, detector prototype, or explicit visual part. Its role is to provide a reusable basis through which image patches and text tokens can interact. Consequently, an anchor may correspond to a broad semantic tendency for example people, animals, vehicles, indoor scenes, or object-like regions without being tied to a single spatial pattern or category label.

This interpretation is important because the low-rank factorization intentionally trades some dense patch-token expressivity for computational efficiency and an interpretable bottleneck. A dense transport kernel can represent every patch-token interaction directly. ConceptOT instead routes these interactions through r learned anchors, approximating the full kernel as $K_{ZP}K_{PP}^{-1}K_{PY}$. This reduces the per-iteration cost of generalized Sinkhorn updates and encourages the model to reuse a compact set of semantic directions. The anchors therefore serve two purposes simultaneously: they make transport more scalable, and they expose a basis that can be inspected through anchor-anchor similarity and anchor-word affinity visualizations.

Fine-grained localization is still obtained at the patch level through the induced transport plan. The heatmap for a phrase is computed by aggregating transported mass over the phrase tokens, so the final localization remains a distribution over image patches. What

changes is the route by which patch-token correspondences are formed: instead of all patches and tokens interacting through an unconstrained dense matrix, their correspondences are mediated by the learned anchor basis. Thus, anchors should not be read as bounding boxes or fixed visual templates; they are semantic coordinates that shape how patch-level transport mass is assigned.

This also explains why anchor interpretability is qualitative rather than categorical. An anchor can participate in multiple related matches, and a phrase can use several anchors at once. The semantic organization observed in Figs. 4 and 5 should therefore be interpreted as evidence that the bottleneck has learned meaningful directions in the multimodal feature space, not as a claim that each anchor is a one-to-one detector for a named concept.

B. Detailed Discussion

Unbalanced transport is essential for compositionality. The clearest signal in our experiments is the gap between balanced and unbalanced transport. Balanced OT does not improve retrieval over the global-only baseline (43.2 vs. 43.6 Avg R@1) and yields only a marginal compositionality gain (78.2 vs. 77.9), while Dense UOT reaches 83.0 on SugarCrepe (+5.1). This gap suggests that exact mass matching is too restrictive for local vision-language alignment: background patches and function words often need the freedom to remain weakly matched or unmatched.

ConceptOT excels at compositional reasoning. ConceptOT outperforms FILIP on compositionality (80.5 vs. 79.6) and achieves the strongest gains on relation replacement (+7.2 over global-only). This suggests the globally coupled transport plan captures relational structure that token-wise max-similarity misses. Moreover, ablations show the default retrieval gap is not fundamental: reducing λ_ℓ from 0.5 to 0.25 closes it entirely (45.6 Avg R@1, matching FILIP and Dense UOT), indicating the issue was over-weighting the stabilized local loss rather than a limitation of the low-rank formulation itself.

Concept anchors self-organize semantically. The learned anchor bank shows interpretable specialization: anchors respond to people, animals, objects, or scenes (Appendix J), confirming the low-rank bottleneck induces meaningful concept decomposition.

Not all local objectives help. Global+Patch hurts retrieval (40.9 Avg R@1), and Text-Attn (42.9) also underperforms. FILIP matches Dense UOT on retrieval but falls behind on compositionality, highlighting that globally coupled partial matching outperforms independent token-wise scoring on compositional tasks.

C. Extended Related Work

This section provides a more comprehensive survey of related work than space permits in the main paper.

C.1. Global vision-language alignment

The dominant line of VLM pre-training learns a shared space between global image and text representations. CLIP [18] and ALIGN [7] established the modern dual-encoder recipe, and SigLIP [26] further simplified the objective with a pairwise sigmoid loss. ALBEF [10], BLIP [11], and BLIP-2 [12] enlarged the design space by combining alignment with fusion or query bottlenecks. These models are powerful and efficient, but their local semantics is largely emergent rather than explicitly optimized. For concept discovery, that distinction matters: one may obtain good zero-shot classification while still lacking faithful patch-token grounding.

C.2. Fine-grained alignment in VLMs

A second family of methods explicitly pursues fine-grained correspondence. FILIP [24] uses late interaction based on token-wise maximum similarity, avoiding expensive fusion encoders. X-VLM [25] and LOUPE [9] learn multi-grained or semantically aggregated alignments, with the latter using Shapley-style interaction modeling. SPARC [2] learns sparse token-conditioned patch groupings, while FLAIR [23] and FineLIP [1] leverage richer captions and text-conditioned pooling to emphasize localized details. These methods strongly motivate local objectives, but they do not directly solve a globally coupled partial-matching problem. In particular, they do not endow alignment with explicit row-column competition and unmatched mass handling, which is where transport geometry becomes useful.

C.3. Grounding and open-vocabulary localization

Grounding-oriented models such as MDETR [8], GLIP [13], RegionCLIP [27], OWL-ViT [16], DenseCLIP [19], and Grounding DINO [15] demonstrate that language-aware localization can be extremely effective when detector architectures, region proposals, grounding data, or detection supervision are available. These systems should be treated as strong references and, in some cases, upper bounds for localization quality. Our target regime is different: ConceptOT keeps the dual-encoder structure and learns from image-text pairs without box supervision. The relevant question is not whether a weakly supervised dual encoder can

surpass a detector, but whether transport can make a retrieval-style VLM substantially more compositional, interpretable, and localizable than other weakly supervised alternatives.

C.4. Optimal transport and scalable OT

Entropic OT and Sinkhorn scaling made transport practical for machine learning [4]. UOT further relaxed hard marginal constraints through divergence penalties, essential when source and target supports only partially overlap [3, 17]. Sinkhorn-based losses have been used to make OT differentiable in deep learning pipelines [5]. On the scalability side, low-rank OT and low-rank UOT solvers show that transport structure can be compressed without abandoning the optimization viewpoint [20, 21]. However, the specific problem of patch-token concept alignment in dual-encoder VLMs has different needs: one wants an anchor-mediated low-rank approximation that is differentiable, numerically stable, and semantically interpretable.

C.5. What is genuinely new in ConceptOT

The novelty is not simply "use OT in VLMs." The proposal combines four ingredients that, to our knowledge, are not jointly instantiated in prior work: (1) unbalanced transport to model inherent partiality of image-caption correspondence; (2) learned concept anchors for a Nyström approximation whose rank has a semantic interpretation; (3) implicit score evaluation without materializing dense patch-token similarities; (4) the transport plan as a weak grounding signal and concept-discovery object.

D. Theoretical Properties

Theorem 1 (Existence and uniqueness). Assume $\epsilon > 0$, $\tau_v > 0$, $\tau_t > 0$, and strictly positive reference masses $\mu \in \mathbb{R}^N_{++}$, $\nu \in \mathbb{R}^M_{++}$. Then the UOT objective in Eq. 3 admits a unique minimizer $\Pi^\star \in \mathbb{R}^{N\times M}_+$.

Proof sketch. The transport cost is linear in Π, the entropic term is strictly convex on the positive orthant, and the KL penalties are convex in the marginals. Together they make the functional coercive and strictly convex over $\mathbb{R}^{N\times M}_+$. Lower semicontinuity and coercivity guarantee existence; strict convexity guarantees uniqueness [3, 17].

Proposition 1 (Limiting regimes). ConceptOT interpolates several familiar alignment mechanisms:

(i) If $\tau_v, \tau_t \to \infty$, the formulation approaches balanced entropic OT.

(ii) If $\epsilon \to 0$ while τ_v, τ_t remain fixed, the plan becomes sparse and concentrates on low-cost matches.

(iii) If $\epsilon \to \infty$, the kernel flattens and the plan approaches a diffuse coupling dominated by the marginal penalties.

Theorem 2 (Per-iteration complexity with concept anchors). Suppose the Gibbs kernel is approximated by $\widetilde{K} = K_{ZP}K_{PP}^{-1}K_{PY}$. Then each generalized Sinkhorn iteration can be computed in $\mathcal{O}((N+M)r + r^2)$ time and $\mathcal{O}((N+M)r)$ memory, instead of $\mathcal{O}(NM)$ for a dense kernel.

Proof sketch. A multiplication by $\widetilde{K}$ decomposes as: $u \mapsto K_{PY}u$ at $\mathcal{O}(Mr)$, $v \mapsto K_{PP}^{-1}v$ at $\mathcal{O}(r^2)$ with precomputed factorization, and $w \mapsto K_{ZP}w$ at $\mathcal{O}(Nr)$.

Proposition 2 (Finite-iteration stability). Fix L iterations. If all intermediate vectors remain in a compact positive interval, then the scaling vectors satisfy $\|a^{(L)} - \widetilde{a}^{(L)}\|_\infty + \|b^{(L)} - \widetilde{b}^{(L)}\|_\infty \le C_L\|K - \widetilde{K}\|_\infty$, and the local scores satisfy $|s_\ell^{(L)}(K) - s_\ell^{(L)}(\widetilde{K})| \le C'_L\|K - \widetilde{K}\|_\infty$ for constants $C_L, C'_L > 0$.

E. Algorithm Details

This section spells out the two computational routines used by ConceptOT. Algorithm 1 describes the low-rank generalized Sinkhorn solver for a single image-caption pair. The key point is that the dense patch-token Gibbs kernel is never materialized. Instead, matrix-vector products with the approximate kernel are decomposed into three smaller operations: token-to-anchor multiplication through K_{PY}, an anchor-space linear solve with $K_{PP,\lambda}$, and anchor-to-patch multiplication through K_{ZP}. The same sequence is then applied in reverse for the token-side scaling update. This realizes the UOT updates with cost $\mathcal{O}((N+M)r+r^2)$ per iteration, as discussed in Theorem 2.

Algorithm 2 shows how the local transport solver is embedded into training. For each mini-batch, we first compute the standard global contrastive loss using pooled CLIP embeddings. The global similarity matrix is also used to mine hard negatives in both directions: hard captions for each image and hard images for each caption. ConceptOT then runs Algorithm 1 for the positive pair and for each selected hard-negative pair, evaluates local scores with the implicit trace formula, and forms the bidirectional local contrastive loss from these scores. Only the projection layers, mass heads, and anchor bank receive gradient updates; the CLIP backbone remains frozen. The symbols used in the algorithms are summarized in Table 5.

Algorithm 1 Low-rank generalized Sinkhorn with learned concept anchors

Require: Projected patches $Z=\{z_i\}_{i=1}^{N}$, tokens $Y=\{y_j\}_{j=1}^{M}$, reference masses μ, ν, anchors $P=\{p_k\}_{k=1}^{r}$, ϵ, τ_v, τ_t, iterations L

Compute the cosine Gibbs sub-kernels K_{ZP}, K_{PP}, and K_{PY}
Form $K_{PP,\lambda} \leftarrow K_{PP} + \lambda I_r$ and factorize it once for repeated solves
Initialize $a \leftarrow \mathbf{1}_N$, $b \leftarrow \mathbf{1}_M$
$\alpha_v \leftarrow \tau_v/(\tau_v+\epsilon)$, $\alpha_t \leftarrow \tau_t/(\tau_t+\epsilon)$
for $\ell = 1$ to L **do**
  $q \leftarrow K_{PY} b$ ▷ token-to-anchor
  Solve $K_{PP,\lambda} u = q$ for u
  $q \leftarrow K_{ZP} u$ ▷ anchor-to-patch
  $a \leftarrow (\mu \oslash (q + 10^{-8}))^{\alpha_v}$
  $q \leftarrow K_{ZP}^{\top} a$ ▷ patch-to-anchor
  Solve $K_{PP,\lambda}^{\top} u = q$ for u
  $q \leftarrow K_{PY}^{\top} u$ ▷ anchor-to-token
  $b \leftarrow (\nu \oslash (q + 10^{-8}))^{\alpha_t}$
end for
return Scalings a, b and implicit plan $\widetilde{\Pi}(\cdot) = \mathrm{diag}(a) K_{ZP} K_{PP,\lambda}^{-1} K_{PY} \,\mathrm{diag}(b)(\cdot)$

F. Notation Reference

Table 5 collects the main symbols used in the method, algorithms, and theoretical discussion. We use N for the number of visual patch tokens after discarding the CLIP CLS token, M for the number of valid text tokens, and r for the number of learned concept anchors. The notation separates encoder outputs (V, T) from their projected concept-space representations (Z, Y) because ConceptOT trains only the lightweight projections, mass heads, and anchors while keeping the backbone fixed. The transport quantities (C, K, Π) are defined at the patch-token level; in the low-rank implementation, Π is represented implicitly through anchor kernels rather than stored as a dense $N \times M$ matrix.

G. Extended Experimental Setup

G.1. Backbone details

We use the publicly available OpenAI CLIP ViT-B/16 checkpoint [18]. The visual encoder internally produces 197 tokens (196 spatial patches + 1 CLS token), but we discard the CLS token and use the N=196 spatial patch tokens of dimension 768 for local alignment. The text encoder produces up to M=48 tokens of dimension 512. The backbone is kept fully frozen; only the following lightweight components are trained:

- Two linear projections $W_v \in \mathbb{R}^{768\times256}$ and $W_t \in \mathbb{R}^{512\times256}$ mapping visual and textual token features to the shared concept space.
- Two scalar mass heads, implemented as linear layers from the corresponding encoder dimension to one scalar mass logit.
- For ConceptOT: the anchor bank $P \in \mathbb{R}^{32\times256}$, initialized from $\mathcal{N}(0, 0.02)$.

Algorithm 2 One training step for ConceptOT

Require: Mini-batch $\{(I_i, x_i)\}_{i=1}^{B}$, hard-negative count K_h

Encode all images and captions with the frozen VLM backbone
Compute pooled global embeddings and global similarities $g_{ij} = s_g(I_i, x_j)$
Compute $\mathcal{L}_{\text{global}}$ using the standard symmetric contrastive loss
Project visual and text tokens to the concept space and compute learned masses for all batch items
for $i = 1$ to B **do**
  Mine hard-negative captions $\mathcal{N}_X(i) = \mathrm{TopK}_{j\neq i}\, g_{ij}$ and hard-negative images $\mathcal{N}_I(i) = \mathrm{TopK}_{j\neq i}\, g_{ji}$
  Run Algorithm 1 and compute $s_\ell(I_i, x_i)$ for the positive pair
  for $j \in \mathcal{N}_X(i)$ **do**
    Run Algorithm 1 and compute $s_\ell(I_i, x_j)$
  end for
  for $j \in \mathcal{N}_I(i)$ **do**
    Run Algorithm 1 and compute $s_\ell(I_j, x_i)$
  end for
end for
Form $\mathcal{L}_{I\to X}^{\ell}$ and $\mathcal{L}_{X\to I}^{\ell}$ from the positive and hard-negative local scores
Form the full objective $\mathcal{L}$ using Eq. 10
Update only the projection layers, mass heads, and anchors by backpropagation

G.2. Training details

Tables 6 and 7 list the complete hyperparameters. Transport-based methods (Balanced OT, Dense UOT, ConceptOT) are trained for fewer epochs (3) because they converge faster due to the richer local signal; we observed instability in ConceptOT beyond epoch 3 (see Section H). Non-transport methods (Global+Patch, FILIP, Text-Attn) are trained for 8 epochs. Global-only uses no training as the backbone is frozen. Table 5 in the previous section provides a notation reference.

Symbol	Meaning
I, x	Input image and caption.
$V{=}[v_i]_{i=1}^{N}$	Visual patch embeddings from the image encoder.
$T{=}[t_j]_{j=1}^{M}$	Text-token embeddings from the text encoder.
z_i, y_j	Projected, ℓ_2-normalized patch/token embeddings.
$\mu \in \mathbb{R}_+^N$, $\nu \in \mathbb{R}_+^M$	Reference masses for patches and text tokens.
$C \in \mathbb{R}^{N\times M}$	Transport cost matrix: $C_{ij}{=}1{-}z_i^\top y_j$.
$K \in \mathbb{R}_+^{N\times M}$	Gibbs kernel: $K_{ij}{=}\exp(-C_{ij}/\epsilon)$.
$\Pi \in \mathbb{R}_+^{N\times M}$	Transport plan; $\Pi^\star$ is the optimum.
ϵ	Entropic regularization coefficient.
τ_v, τ_t	KL penalties for patch and token marginals.
r	Anchor rank for the Nyström approximation.
$P{=}[p_k]_{k=1}^{r}$	Learned concept anchors.
L	Number of generalized Sinkhorn iterations.
K_h	Number of hard negatives per positive pair.

Table 5. Main notation used throughout the paper.

Hyperparameter	Value
Optimizer	AdamW
Learning rate	10^{-4}
Weight decay	10^{-2}
LR schedule	Cosine decay
Warmup	5% of total steps
Batch size	64
Mixed precision	Yes (FP16)
Gradient clipping	1.0
Training images	20K (COCO Karpathy subset)
Training pairs	∼100K after expanding image-caption pairs
Epochs	3 for ConceptOT/transport runs; 8 in the default full configuration for non-transport scorers
Eval images	5K (COCO Karpathy test)
Rerank top-k	128
GPU	Single RTX 5090 used in our runs

Table 6. Complete training hyperparameters.

G.3. Transport-specific hyperparameters

Table 7 lists the hyperparameters that control the transport solver and local alignment loss. The regularization ϵ determines how diffuse the transport plan is, while τ_v and τ_t determine how strongly the optimized marginals are encouraged to follow the learned patch and token masses. The anchor rank r controls the size of the low-rank concept bottleneck, and L controls the number of generalized Sinkhorn updates used per pair. The local loss weight λ_ℓ, local temperature τ_ℓ, and hard-negative count K_h determine how strongly the transport score influences training relative to the global contrastive objective.

Hyperparameter	Value
Entropic regularization ϵ	0.07
Patch marginal penalty τ_v	0.20
Token marginal penalty τ_t	0.20
Local loss weight λ_ℓ	0.50
Anchor diversity weight β	10^{-3}
Local temperature τ_ℓ	0.07
Sinkhorn iterations L	5
Anchor rank r (ConceptOT)	32
Hard negatives K_h	4
Projection dimension d_c	256

Table 7. Transport-specific hyperparameters.

G.4. Evaluation protocol

Retrieval evaluation uses a two-stage process: (1) encode all images and texts with the frozen backbone to get global embeddings; (2) compute global cosine similarities, select top-128 candidates, then rerank using the combined global + local score: $s(I,x) = g(I,x) + \lambda_\ell \cdot s_\ell(I,x)$. Recall@$K$ is computed for $K \in \{1,5,10\}$ in both I→T and T→I directions.

H. Numerical Stability Details

The low-rank generalized Sinkhorn solver encountered training instability that the dense UOT solver did not. Without stabilization, the ConceptOT loss exploded at epoch 2, with local loss jumping from ∼0.5 to ∼972,000. The root causes were:

1. Unbounded scaling vectors. The generalized Sinkhorn updates $a \leftarrow (\mu/q)^{\alpha_v}$ can produce extreme values when q is near zero, and these are amplified through subsequent matrix multiplications with the Nyström sub-kernels.

2. Ill-conditioned K_{PP}. As concept anchors evolve during training, they can become nearly collinear,

making K_{PP} near-singular. Cholesky decomposition and subsequent solves then amplify numerical errors catastrophically.

3. Unstable score computation. The implicit score in Eq. 8 involves a ratio of quantities computed through multiple linear solves, compounding any conditioning issues.

Our stabilization strategy applies the following mitigations:

- Log-domain sub-kernels: Compute sub-kernels with row-max subtraction before exponentiation to prevent overflow.
- Log-domain Sinkhorn: Work with $\log a$ and $\log b$ instead of a and b, with centering (subtract mean) and clamping to $[-20, 20]$ each iteration.
- Stronger regularization: Increase K_{PP} diagonal regularization from 10^{-4} to 10^{-2}.
- Robust linear solves: Replace Cholesky with linalg.solve, with lstsq fallback.
- Output clamping: Apply tanh to the final score.

These stabilizations successfully prevent loss explosion but likely over-regularize the transport plan, contributing to ConceptOT's retrieval gap relative to Dense UOT. Developing inherently stable low-rank UOT solvers is an important direction for future work.

I. Efficiency Analysis

Method	r / L	ms/pair	Mem (GB)
Dense UOT	— / 5	0.77	0.61
Balanced OT	— / 5	0.73	0.61
ConceptOT	16 / 3	2.37	0.61
ConceptOT	16 / 5	3.14	0.61
ConceptOT	32 / 3	2.49	0.61
ConceptOT	32 / 5	3.31	0.61
ConceptOT	32 / 7	4.13	0.61
ConceptOT	48 / 5	3.65	0.61

Table 8. Solver efficiency across rank r and iterations L. Memory is constant across configurations; cost scales linearly with L and sublinearly with r. At current VLM scales (N=197, $M \leq 48$), log-domain stabilization overhead dominates rank savings, making ConceptOT 3–5× slower than dense solvers.

J. Additional Figures

This section presents supplementary visualizations: a retrieval comparison bar chart (Fig. 2), training convergence curves (Fig. 3), concept anchor analysis (Figs. 4 and 5), and a multi-phrase grounding example (Fig. 6).

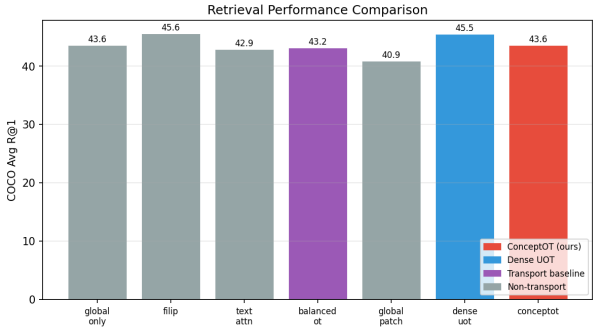

Figure 2. COCO Avg R@1 across all seven scoring variants. Dense UOT and FILIP lead; ConceptOT matches the global-only baseline while providing transport-based interpretability.

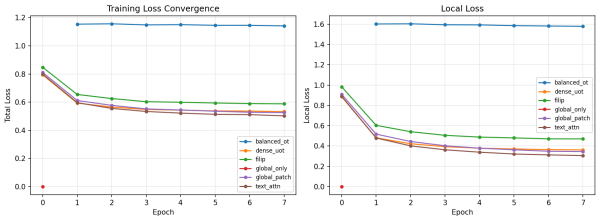

Figure 3. Training convergence for all methods. Left: total loss. Right: local (alignment) loss. Transport-based methods converge within 3 epochs; non-transport methods use 8 epochs. The global-only baseline has zero local loss (no local scorer).

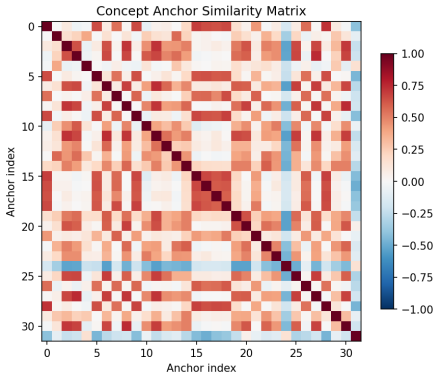

Figure 4. Cosine similarity matrix between the 32 learned concept anchors. The block structure suggests clusters of related concepts. Mean off-diagonal similarity is 0.19, indicating reasonable diversity. Some anchor pairs (e.g., indices 2–3, 24–25) are near-orthogonal (blue), while others share semantic overlap (red).

K. Extended Grounding Examples

Figures 7–10 show multi-phrase grounding heatmaps across 12 COCO test images. Each group shows an

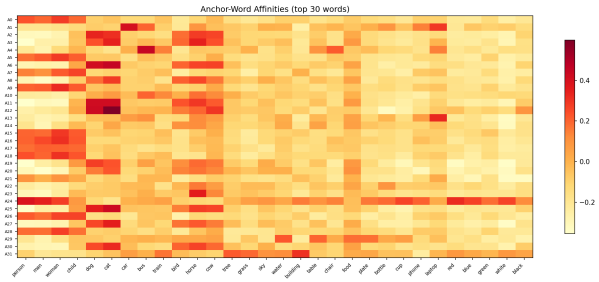

Figure 5. Anchor-word affinity heatmap for 30 common visual concepts. Each row is a learned anchor; each column is a word projected through the text encoder and scorer projection. Anchors specialize to semantic categories: some respond strongly to animals (dog, cat, horse), others to people (woman, man, child), and others to objects or scenes (bus, kitchen, room). The heatmap suggests that anchors behave differently for object and attribute words. Object nouns such as dog, cat, and horse activate multiple anchors, indicating distributed semantic support across animal/foreground-object directions. In contrast, color words such as red, blue, and green are concentrated on fewer anchors, consistent with colors acting as more compact attribute-like directions in the learned concept space.

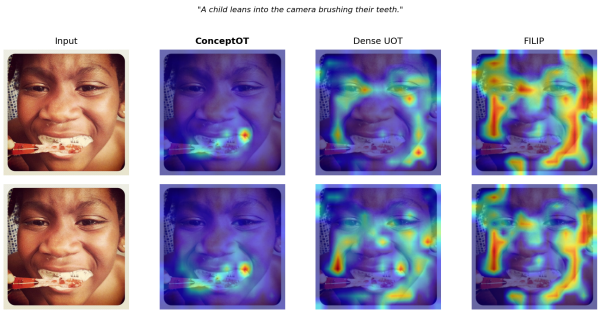

Figure 6. Multi-phrase grounding for "A child leans into the camera brushing their teeth." ConceptOT concentrates activation on the visually diagnostic mouth/toothbrush region and nearby face area, which supports the action phrase "brushing their teeth." FILIP produces more scattered activations, while Dense UOT focuses more narrowly on a small high-confidence region. The example illustrates that ConceptOT yields coherent phrase-level evidence, though the heatmaps should be interpreted as weak grounding rather than full object segmentation.

image with 3 query phrases extracted from the caption, comparing ConceptOT, Dense UOT, and FILIP. ConceptOT consistently produces spatially coherent heatmaps that localize on the queried concept. Dense UOT is often equally focused but occasionally spreads more diffusely. FILIP's max-similarity approach tends to produce scattered or edge-biased activations, especially for abstract or relational phrases. No grounding supervision is used in any method.

L. Limitations and Future Work

- Solver overhead. ConceptOT's solver is 4× slower than Dense UOT due to log-domain stabilization and linear solves required for numerical safety. Developing inherently stable low-rank UOT solvers—perhaps via native log-domain Nyström iterations or adaptive ϵ scheduling—is the most pressing engineering challenge.
- Retrieval gap. ConceptOT trails Dense UOT by 1.9 points on COCO retrieval Avg R@1, suggesting that the heavy stabilization over-regularizes the transport plan. Data-adaptive anchors predicted per pair, rather than globally shared, could close this gap while preserving interpretability.
- Limited evaluation scope. We evaluate on COCO retrieval and SugarCrepe compositionality. Phrase grounding benchmarks (Flickr30k Entities, RefCOCO/+/g) would more directly test ConceptOT's transport plans. Winoground evaluation was not possible due to dataset access restrictions.
- Single backbone and scale. All experiments use a frozen CLIP ViT-B/16. The low-rank advantage of $\mathcal{O}((N+M)r)$ over $\mathcal{O}(NM)$ would be more pronounced at higher resolutions (ViT-L, ViT-H) or with longer token sequences, but this remains untested.
- Anchor interpretability. The claim that anchors self-organize semantically is supported qualitatively (Figs. 4, 5) but not by a quantitative interpretability metric. Developing such metrics is an open problem.
- Training instability. ConceptOT requires careful stabilization that Dense UOT does not. The root cause—ill-conditioning of K_{PP} as anchors co-adapt—suggests that alternative parameterizations (e.g., orthogonality constraints on P) could help.

M. Extended Discussion

1. Why unbalanced transport matters. The patch-token alignment problem is intrinsically partial. Background clutter, occlusion, and abstract language make exact mass preservation undesirable. ConceptOT formalizes this observation instead of treating unmatched content as optimization noise. Our experiments confirm this: Balanced OT (43.2 Avg R@1) barely exceeds the global-only baseline (43.6), while Dense UOT reaches 45.5.

2. Why low rank is more than an acceleration trick. In many efficient OT pipelines, low rank is introduced only for complexity reduction. Here it also induces a reusable concept basis. If the anchor bank becomes semantically coherent, then efficiency and interpretability are obtained from the same modeling choice.

3. Relation to fine-grained late interaction. FILIP-style maxima [24] are effective for selective local matches but do not couple all rows and columns. ConceptOT is a globally coupled alternative that handles competition and partial matching. Interestingly, FILIP matches Dense UOT on retrieval (45.6 vs. 45.5), suggesting that coupling may not be necessary for pure retrieval but could matter for compositional tasks.

4. Relation to detector-style grounding. Detector-based approaches such as GLIP [13] and Grounding DINO [15] should remain stronger on localization benchmarks. Our contribution targets a different regime: improving locality and concept faithfulness while remaining in the weakly supervised dual-encoder framework.

5. Expected failure modes. The method may struggle with non-visual words, negation, long-range discourse, and phrases requiring external knowledge. If the caption omits the crucial object, no alignment mechanism can recover it from paired supervision alone.

6. Retrieval-versus-grounding trade-off. A stronger local branch might overfit fine details and disturb global retrieval geometry. Our recipe keeps the global branch intact and uses transport as a complementary objective.

7. Downstream implications. If transport plans become reliable and sparse, they can serve as supervision for open-vocabulary segmentation, grounded captioning, or controllable generation.

8. Scientific value. A useful weakly supervised method need not outperform specialized detectors. It is valuable if it improves compositionality, reveals interpretable concept structure, and offers a principled local objective for compact VLM training.

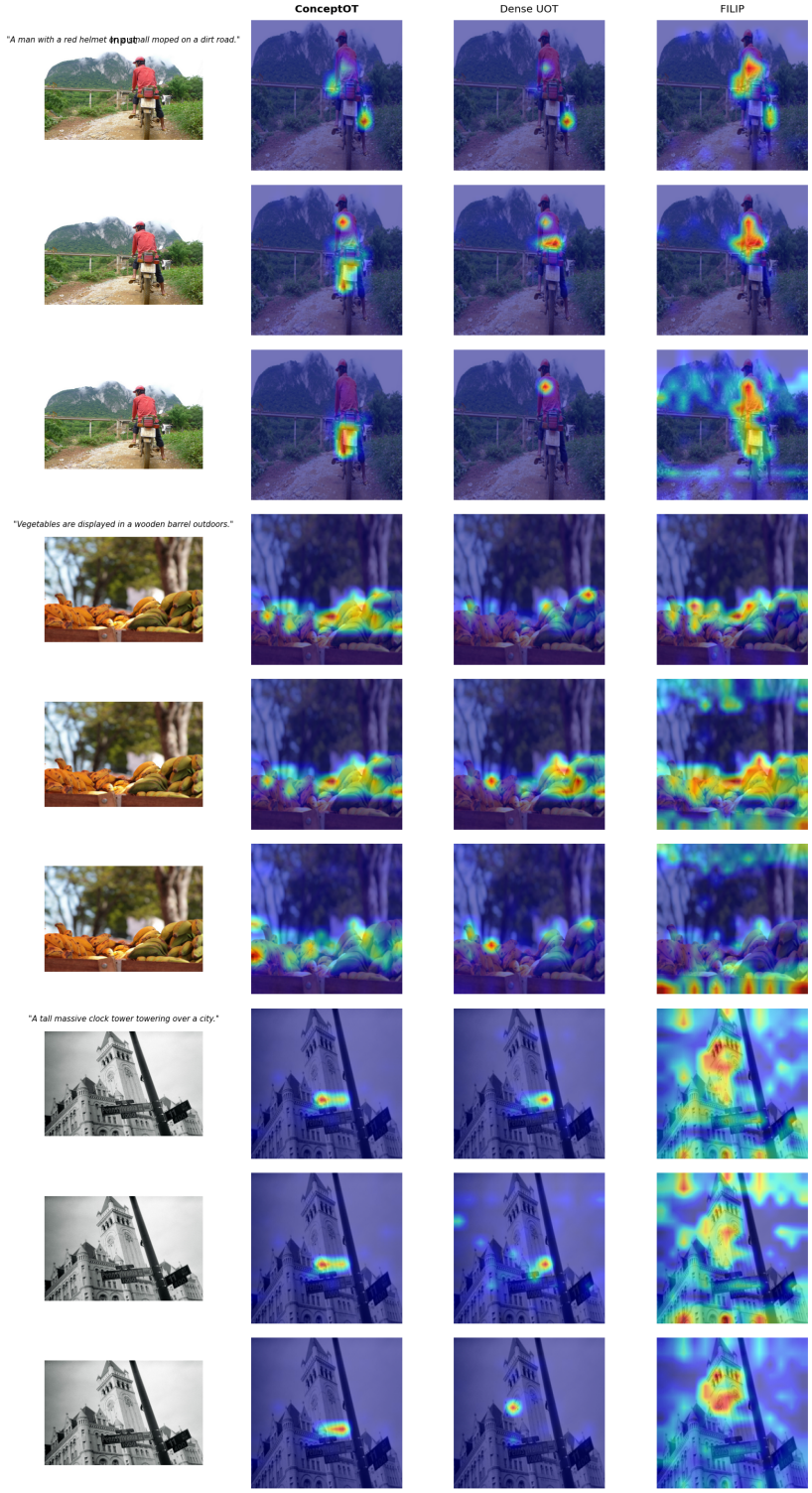

Figure 7. Multi-phrase grounding examples (1/4). Three images with 3 phrases each. Captions shown above each group.

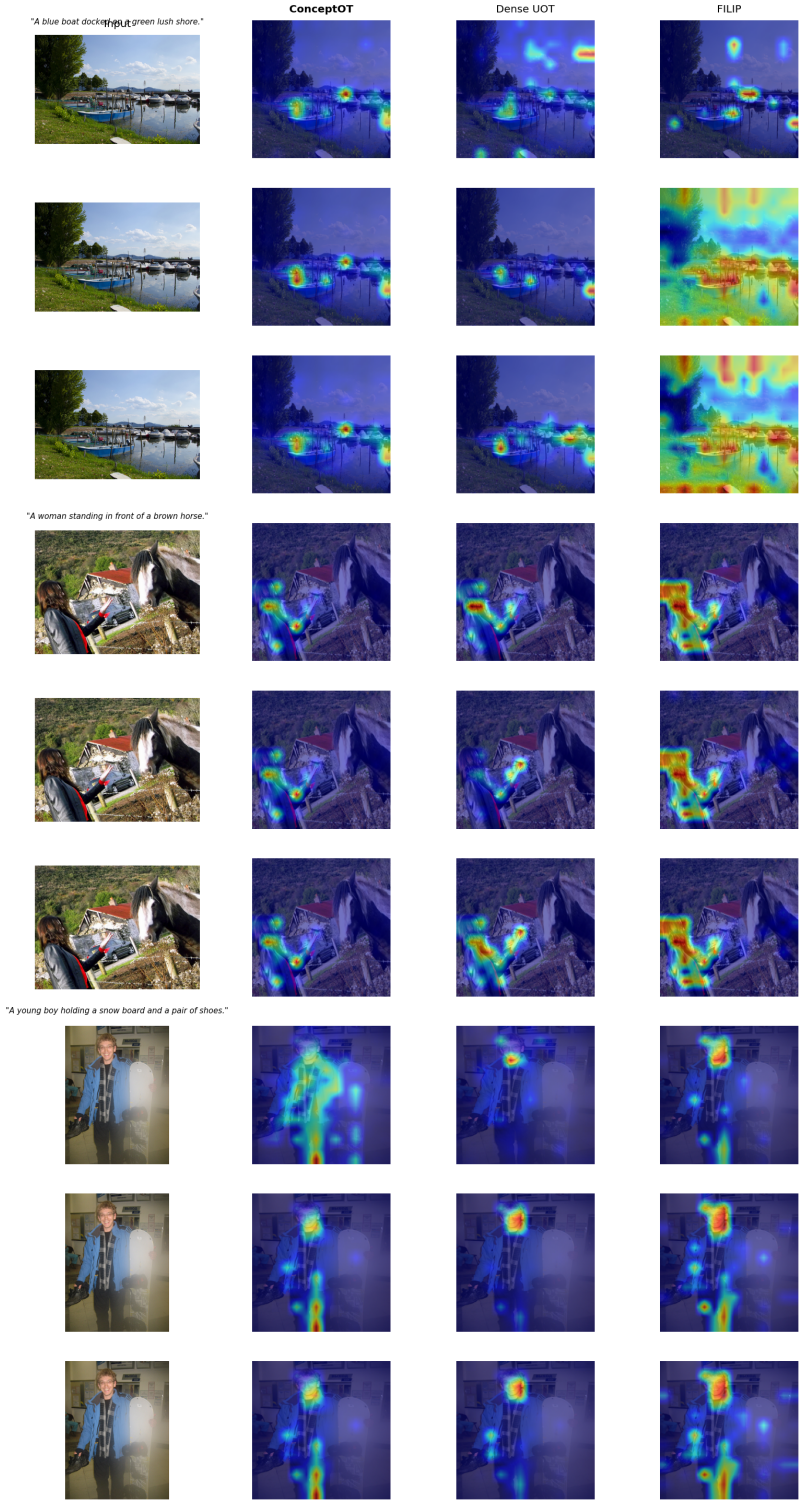

Figure 8. Multi-phrase grounding examples (2/4).

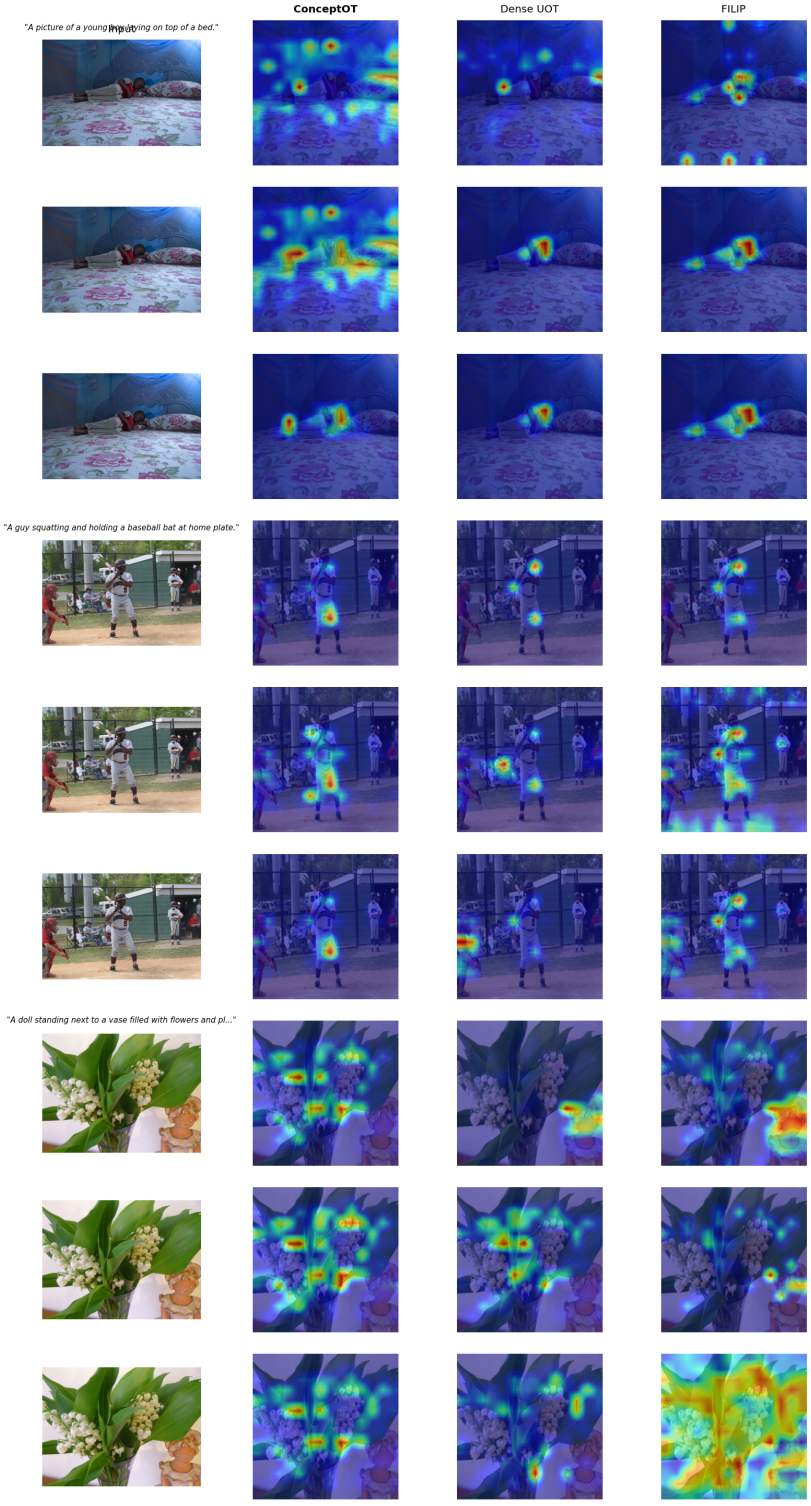

Figure 9. Multi-phrase grounding examples (3/4).

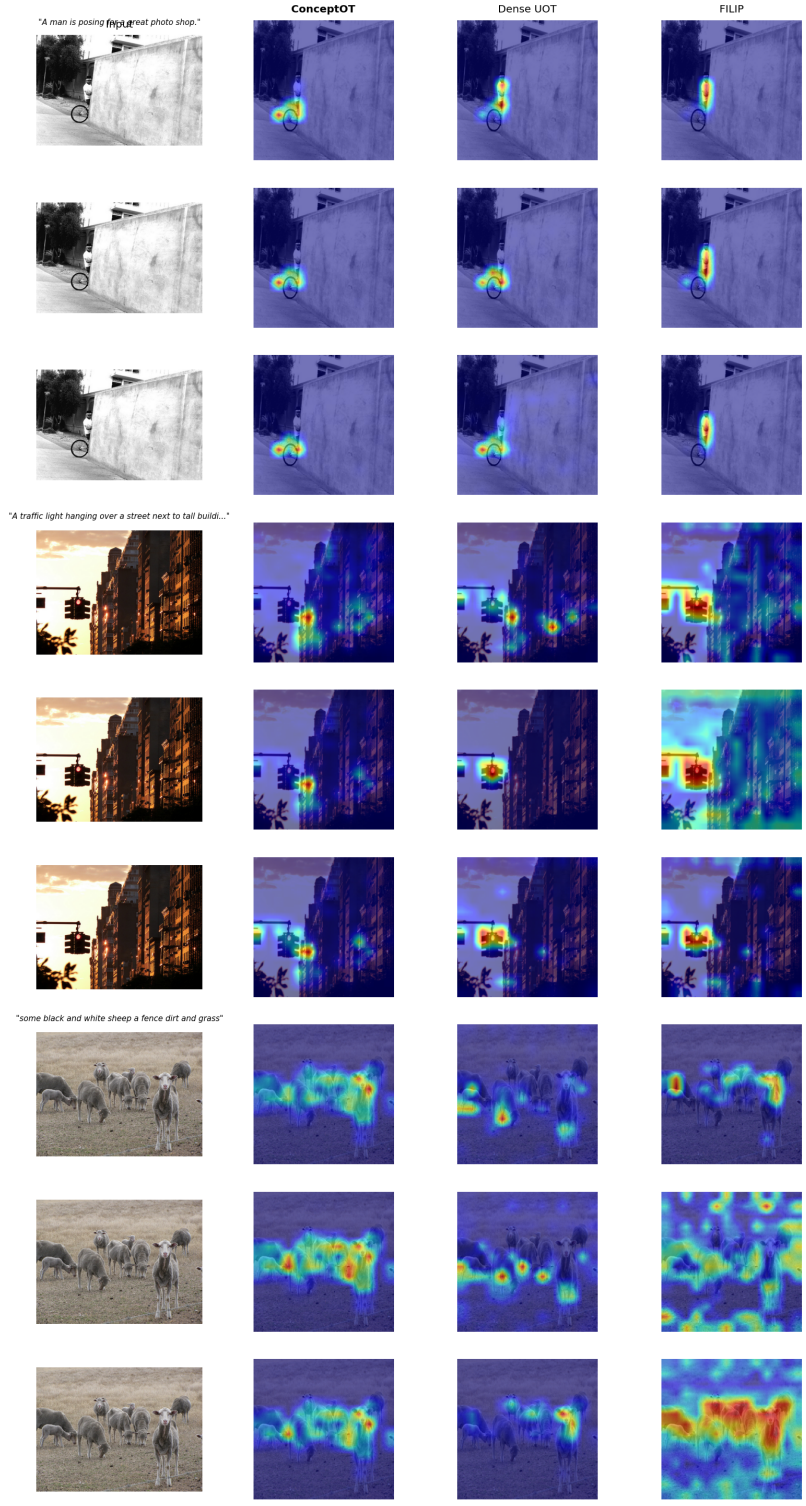

Figure 10. Multi-phrase grounding examples (4/4).