

PROTOlip: FROM SENTENCE-LEVEL TO OBJECT-LEVEL EVIDENCE DISENTANGLEMENT

Yan Zhu

Department of Computer Science
Tulane University
New Orleans, LA, USA
yzhu27@tulane.edu

Y. Chen

Department of Computer Science
Tulane University
New Orleans, LA, USA
ychen88@tulane.edu

Rebecca J. Faust

Department of Computer Science
Tulane University
New Orleans, LA, USA
rfaust1@tulane.edu

Zhengming Ding

Department of Computer Science
Tulane University
New Orleans, LA, USA
zding1@tulane.edu

ABSTRACT

Query-conditioned vision–language models enable fine-grained interpretation by revealing how visual evidence changes with textual queries. However, evidence conditioned on complete descriptions does not necessarily resolve into object-specific evidence, nor does an exposed evidence map necessarily identify the evidence that constitutes the model’s prediction. Across multiple VLM architectures and independent benchmarks, we find that object-level queries often retain evidence from co-occurring objects and shared context. In this paper, we introduce **ProtoLIP**, a lightweight prototype-mediated evidence layer that organizes reusable visual prototypes into text-derived semantic families and uses query-dependent family routing to constrain which prototypes may provide evidence. ProtoLIP requires no spatial annotations and does not retrain the backbone. It improves evidence localization and separation across query granularities, and its localization gains also transfer to independently pretrained VLMs when their patch and text representations are sufficiently aligned. It remains competitive with a spatially supervised grounding model while improving image–text matching. Crucially, ProtoLIP constructs its matching score directly from localized prototype evidence, enabling the score to be exactly decomposed into semantic-family and prototype contributions.

1 INTRODUCTION

Vision–language models (VLMs) learn transferable visual representations through image–text alignment (Radford et al., 2021; Zhai et al., 2023). Recent work has moved beyond global alignment toward finer-grained and text-conditioned visual representations (Zheng et al., 2024; Xiao et al., 2025; Lyu et al., 2026), enabling the same image to be interpreted at different semantic granularities. As queries become more specific, their visual evidence should ideally become correspondingly precise, from complete descriptions to individual objects.

However, evidence conditioned on complete descriptions does not necessarily become object-specific when the query is narrowed. Existing objectives may separate rich textual descriptions without requiring their constituent objects or corresponding visual evidence to be disentangled. Consequently, object-level queries can remain tied to co-occurring objects or shared context. In Figure 1, both models produce reasonable evidence for the full caption “a woman riding a bicycle,” yet ItemizedCLIP (Lyu et al., 2026) retains evidence from the woman and background even when queried specifically about the bicycle. Thus, description-level conditioning does not necessarily extend to object-level evidence separation.

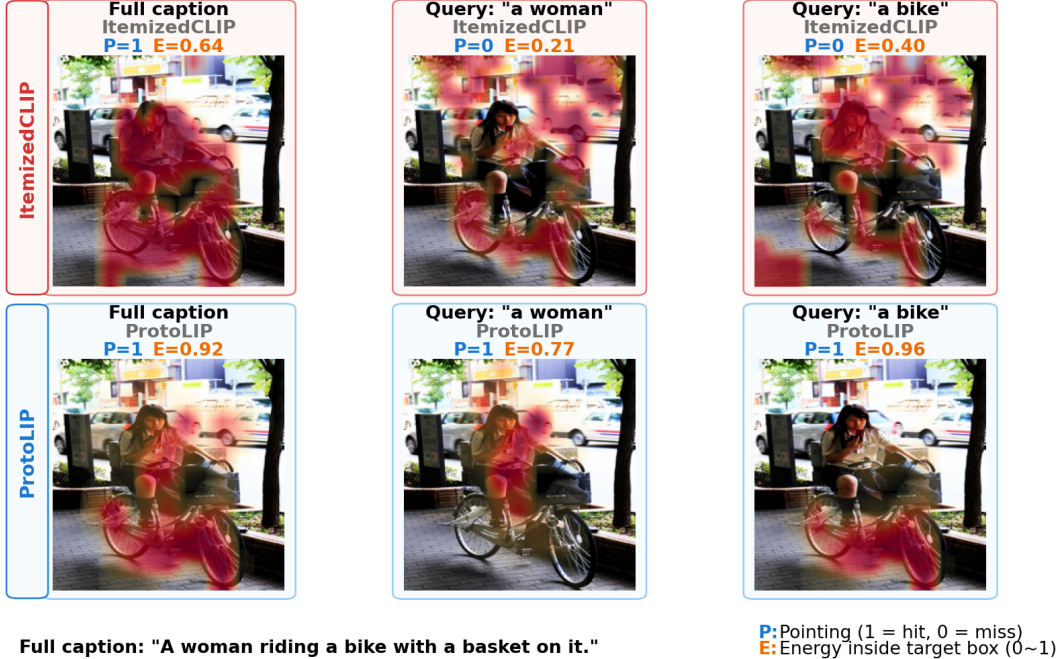


Figure 1: Object-level evidence disentanglement. ItemizedCLIP retains evidence from the woman and background when queried about the bicycle, whereas ProtoLIP separates object-specific evidence. **P** indicates whether the peak evidence falls inside the target box, and **E** measures the fraction of evidence within it.

DreamLIP (Zheng et al., 2024) exploits long captions and sub-caption supervision, FLAIR (Xiao et al., 2025) constructs language-informed visual representations through text-conditioned aggregation, and ItemizedCLIP (Lyu et al., 2026) introduces objectives for item independence and representation completeness. These advances improve correspondence between rich descriptions and visual content, but the unit of textual supervision is not necessarily an individual object. In particular, an ItemizedCLIP item may contain multiple objects and relations, so separation between items need not extend to the constituent objects within an item.

Related studies report complementary symptoms of this limitation. SWIM (Sun et al., 2026) finds that object nouns can produce diffuse visual activations, while analyses of CLIP models reveal reliance on background regions and spurious contextual correlations (Agarwal et al., 2026). These observations motivate our study of evidence disentanglement: whether description-level conditioning yields spatially separated, object-specific evidence.

We use *evidence disentanglement* to denote the structured separation of predictive evidence across semantic and spatial dimensions. We examine two complementary aspects: separating evidence for a sentence into contributions associated with its constituent objects, and spatially separating object-specific evidence from co-occurring entities and context. For this separation to explain a prediction, the evidence must also be explicitly connected to the matching score. Localization alone does not establish this connection: attention weights need not reflect predictive influence (Wu et al., 2024b;a), and gradient-based attribution methods also face faithfulness and completeness limitations (Mehri et al., 2025).

To address these requirements, we introduce ProtoLIP, a lightweight prototype-mediated evidence layer on top of a frozen VLM. Its central design separates coarse semantic routing from fine-grained query scoring: text-derived semantic families constrain which prototypes may contribute, while the complete query evaluates the compatibility of their localized visual evidence. This organization supports object-, phrase-, and sentence-level queries without restricting queries to fixed class labels. The resulting prototype contributions directly and additively constitute the image-text matching score, enabling exact decomposition through semantic families and prototypes without spatial supervision or backbone retraining.

Our experiments show improved object-specific localization across datasets and query granularities, alongside preserved or improved full-caption localization. On the frozen ItemizedCLIP backbone, ProtoLIP adds only 1.48% trainable parameters and improves Pointing and Energy by up to 42.2% and 55.1%, respectively. Localization gains also transfer to independently pretrained VLM backbones. Despite using only text-derived weak supervision and no spatial annotations, ProtoLIP remains competitive with a dedicated spatially supervised grounding model. Direct inference improves matching separability, with reductions in strict top-1 retrieval; score-level fusion can partially recover retrieval performance while retaining the same spatial evidence maps.

Our contributions are:

- ProtoLIP, a prototype-mediated evidence layer on top of a frozen vision–language backbone. Text-derived semantic families provide coarse routing over query-relevant visual prototypes, while the complete query performs fine-grained scoring over the routed evidence. The resulting prototype contributions directly and additively constitute the matching score, yielding an exactly decomposable prediction path from localized prototype evidence to the final score.
- Improved evidence disentanglement across query granularities, datasets, and VLM backbones, with stronger localization from object-level queries to complete sentence descriptions. Despite using only text-derived weak supervision and no spatial annotations, ProtoLIP remains competitive with spatially supervised grounding while improving image–text matching.

2 RELATED WORK

Fine-grained vision–language representations. Contrastive vision–language models such as CLIP (Radford et al., 2021) learn transferable representations through global image–text alignment, while SigLIP (Zhai et al., 2023) improves vision–language pretraining with pairwise sigmoid supervision. Subsequent work has pursued finer correspondence between visual and textual content. FILIP (Yao et al., 2022) introduces token-wise interactions between visual and textual representations, while RegionCLIP (Zhong et al., 2022) extends language–image alignment to localized image regions. Other approaches enrich the textual supervision itself: DreamLIP (Zheng et al., 2024) leverages long captions and sub-captions to learn more detailed visual–language correspondences.

More recent work makes the visual representation explicitly dependent on the textual description. FLAIR (Xiao et al., 2025) uses text-conditioned attention pooling over local visual tokens to produce language-informed representations for fine-grained retrieval and localization. Most closely related to our setting, ItemizedCLIP (Lyu et al., 2026) introduces item-independence and representation-completeness objectives to learn distinct visual representations for multiple semantically independent textual items describing the same image.

Together, these approaches progressively refine the granularity of vision–language alignment and textual conditioning. However, the unit of textual supervision is not necessarily an individual visual object. Long captions, sub-captions, and textual items may each contain multiple objects, attributes, and relations; separating such descriptions therefore does not require their constituent objects to have correspondingly separated visual evidence. ProtoLIP addresses this remaining granularity gap by resolving rich textual descriptions into localized evidence for their individual objects.

Vision–language localization and grounding. A parallel line of work seeks to recover object-level spatial evidence from pretrained vision–language models. Training-free methods such as NA-CLIP (Hajimiri et al., 2025) derive dense localization maps directly from pretrained CLIP representations without task-specific grounding training. Dedicated grounding models such as Grounding DINO (Liu et al., 2024) and MM Grounding DINO (Zhao et al., 2024), in contrast, use explicit detection and grounding supervision to learn language-conditioned object localization.

Our setting differs primarily in supervision. Dedicated grounding models rely on explicit detection or grounding annotations, whereas ProtoLIP uses no spatial annotations and keeps the pretrained VLM backbone frozen. Despite this substantially weaker spatial supervision, ProtoLIP achieves competitive localization against supervised grounding models, while outperforming training-free CLIP localization methods under our shared evaluation protocol.

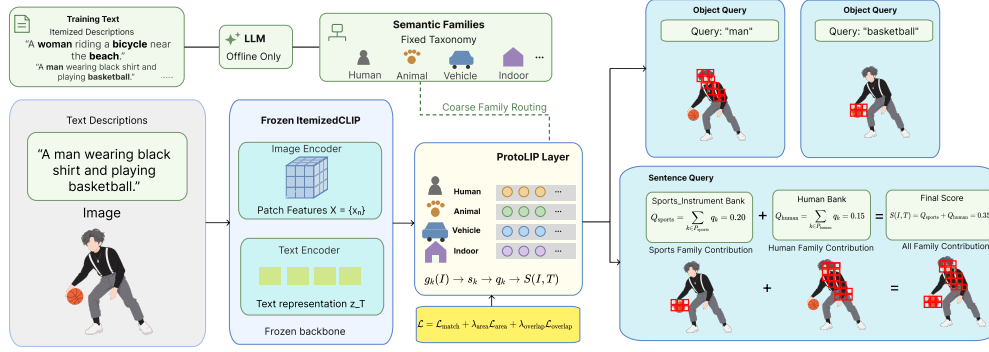


Figure 2: Overview of ProtoLIP. Training text is organized offline into coarse semantic families that constrain prototype eligibility. Given an image–text pair, a frozen vision–language backbone extracts patch features and a complete text representation. Eligible prototypes extract localized visual evidence, whose compatibility with the complete query determines their contribution to the matching score and query-specific spatial evidence.

Prototype-based visual evidence and prediction decomposition. Prototype-based models explain predictions through similarities to learned representative patterns (Chen et al., 2019). Vision–language extensions include class-specific visual prototypes for weakly supervised segmentation (VPL (Xu et al., 2025)) and language-guided part prototypes for interpretable image classification (PRISM (Yu & Yang, 2026)). PRISM uses category-name-conditioned CLIP attribution maps to guide part-prototype learning and aggregates their activations through a fixed-label classifier. Its goal is to explain category predictions, whereas ours is to separate object-specific evidence and decompose image–text matching scores across object, phrase, and sentence queries.

To this end, ProtoLIP combines offline LLM-assisted concept extraction with coarse semantic families that constrain prototype eligibility for each query. The complete query scores the localized evidence pooled by eligible prototypes, whose contributions sum exactly to the matching score. Thus, language controls both evidence eligibility and compatibility during inference, rather than serving only as training guidance. This design connects object-specific evidence learning with explicit decomposition of sentence-level matching predictions.

3 METHOD

Given an image–text pair (I, T) , where I is an image and T is a text query, a frozen VLM encodes I into patch features $X(I) = \{x_n\}_{n=1}^N$ and T into a complete-query embedding z_T . ProtoLIP adds a prototype-mediated layer to these frozen representations (Fig. 2) to produce a matching score $S(I, T)$ and query-specific spatial evidence. Text-derived semantic families constrain prototype eligibility, while the full query determines their evidence compatibility. Family-level regularization encourages compact and separated evidence.

3.1 TEXT-DERIVED SEMANTIC FAMILIES

We construct our main semantic taxonomy from approximately 50% of Itemized-CC0.3M (Lyu et al., 2026). An offline LLM-assisted procedure extracts localizable objects, groups them into 39 semantic families, and normalizes their lexical aliases. The taxonomy organizes 512 prototypes and is frozen before training. Each prototype belongs to exactly one family, so the families partition the prototype bank into sub-banks of 13–14 prototypes. A small number of ambiguous concepts are associated with more than one family, so a single query may activate multiple families, but the sub-banks themselves remain disjoint. The complete prototype bank used by unmatched queries is therefore exactly the union over all families.

During training and inference, deterministic lexical matching selects eligible families, while unmatched queries use the complete prototype bank. No LLM is invoked at runtime, and the complete text representation z_T retains attributes, actions, modifiers, and relations.

We also constructed a smaller, 23-family taxonomy (K23) from an earlier 30K-description sample of the training corpus. The two taxonomies yield similar global matching and retrieval performance, with differences within one point on most metrics (Appendix B), indicating that our results are not sensitive to the specific number of semantic families. We adopt K39 as our main taxonomy for its modestly stronger spatial localization and broader semantic coverage.

3.2 FAMILY-CONSTRAINED PROTOTYPE EVIDENCE

Let p_k denote the k -th learned prototype, $\text{sim}(\cdot, \cdot)$ the feature-similarity function, and τ_p the prototype temperature. Each prototype defines a normalized spatial assignment

$$\beta_{k,n}(I) = \text{softmax}_n \left(\frac{\text{sim}(p_k, x_n)}{\tau_p} \right), \quad \sum_{n=1}^N \beta_{k,n}(I) = 1. \quad (1)$$

The assignment $\beta_{k,n}$ determines where prototype k extracts visual evidence from the image. We pool the assigned patch features and refine them into prototype-specific image evidence:

$$\tilde{g}_k(I) = \sum_{n=1}^N \beta_{k,n}(I) x_n, \quad g_k(I) = \phi_\theta(\tilde{g}_k(I), p_k), \quad (2)$$

where $\tilde{g}_k(I)$ is the pooled evidence and ϕ_θ is a lightweight gated residual MLP with learnable parameters θ . Thus, $g_k(I)$ represents the image-specific evidence captured by prototype k .

Query-gated prototype scoring. Let $\mathcal{E}_T$ denote the prototypes belonging to the semantic families matched by query T . Each prototype has a learnable scalar ρ_k , with positive weight $r_k = \text{softplus}(\rho_k)$. Its normalized query-gated weight is

$$\alpha_k(T) = \frac{\mathbf{1}[k \in \mathcal{E}_T] r_k}{\sum_j \mathbf{1}[j \in \mathcal{E}_T] r_j}, \quad \sum_k \alpha_k(T) = 1, \quad (3)$$

where $\mathbf{1}[\cdot]$ is the indicator function, so prototypes outside the routed semantic families receive zero weight.

Prototype compatibility s_k , additive contribution q_k , and the final matching score are

$$s_k(I, T) = \text{sim}(g_k(I), z_T), \quad q_k(I, T) = \alpha_k(T) s_k(I, T), \quad S(I, T) = \sum_k q_k(I, T). \quad (4)$$

Routing therefore determines *which* prototypes may provide evidence, while the complete query embedding z_T determines *how strongly* each eligible prototype supports the query. Because $S(I, T)$ is the sum of the resulting q_k , the prediction is directly decomposable into prototype contributions.

3.3 QUERY-SPECIFIC PREDICTION EVIDENCE

Because each contribution $q_k(I, T)$ is associated with the spatial assignment $\beta_{k,n}(I)$, it can be projected back onto the image:

$$C(n; I, T) = \sum_k \alpha_k(T) s_k(I, T) \beta_{k,n}(I), \quad \sum_n C(n; I, T) = S(I, T). \quad (5)$$

The equality follows from the normalization in Eq. 1. Hence, C is not a post-hoc attribution: it is a score-preserving spatial projection of the prototype-mediated prediction, with positive and negative values supporting and opposing the match, respectively.

For localization and visualization, we use normalized positive support:

$$H(n; I, T) = \frac{[C(n; I, T)]_+}{\sum_j [C(j; I, T)]_+ + \epsilon}, \quad [u]_+ = \max(0, u), \quad (6)$$

where $\epsilon > 0$ ensures numerical stability. C is the signed spatial evidence exposed by the model: by Eq. 5 it is the same set of prototype contributions that constitutes $S(I, T)$, redistributed over patches. H is its normalized positive part, used only for localization and visualization, and is not itself a decomposition of the matching score.

3.4 SPATIAL EVIDENCE REGULARIZATION

Let $\mathcal{F}_T = \{F_c\}$ denote the active semantic families, $\mathcal{P}_T$ the unordered pairs of distinct active families, and P_c the prototypes assigned to family F_c . We aggregate prototype assignments within each family into the normalized family map

$$M_c(n; I, T) = \frac{\sum_{k \in P_c} \alpha_k(T) \beta_{k,n}(I)}{\sum_{k \in P_c} \alpha_k(T)}. \quad (7)$$

This map captures where the evidence associated with an active semantic family is spatially concentrated. We measure its spatial spread and the overlap between two active families as

$$A(M_c) = \frac{1}{N} \exp \left(- \sum_n M_c(n) \log(M_c(n) + \epsilon) \right), \quad (8)$$

$$O(M_c, M_{c'}) = \sum_n \min(M_c(n), M_{c'}(n)). \quad (9)$$

Here, $A(M_c)$ is an entropy-based effective-area measure, while $O(M_c, M_{c'})$ measures shared spatial mass. We penalize only excessive area and overlap:

$$\mathcal{L}_{\text{area}} = \frac{1}{|\mathcal{F}_T|} \sum_{c \in \mathcal{F}_T} [A(M_c) - \tau_a]_+^2, \quad (10)$$

$$\mathcal{L}_{\text{overlap}} = \frac{1}{|\mathcal{P}_T|} \sum_{(c, c') \in \mathcal{P}_T} [O(M_c, M_{c'}) - \tau_o]_+^2, \quad (11)$$

where τ_a and τ_o are tolerated area and overlap thresholds. These losses encourage compact and separated family evidence without forcing semantically related families to occupy disjoint regions.

3.5 LEARNING OBJECTIVE

For a batch of B corresponding image-text pairs, let $\mathbf{S} \in \mathbb{R}^{B \times B}$ with $\mathbf{S}_{ij} = S(I_i, T_j)$ denote the pairwise score matrix and $\mathbf{y} = (1, \dots, B)$ the diagonal matching targets. We train the prototype layer with the symmetric contrastive matching loss

$$\mathcal{L}_{\text{match}} = \frac{1}{2} [\text{CE}(\mathbf{S}/\tau_s, \mathbf{y}) + \text{CE}(\mathbf{S}^\top/\tau_s, \mathbf{y})], \quad (12)$$

where $\text{CE}(\cdot, \cdot)$ denotes cross-entropy over text candidates in the first term and image candidates in the second, and $\tau_s > 0$ is the score temperature. The complete objective is

$$\mathcal{L} = \mathcal{L}_{\text{match}} + \lambda_{\text{area}} \mathcal{L}_{\text{area}} + \lambda_{\text{overlap}} \mathcal{L}_{\text{overlap}}, \quad (13)$$

where λ_{area} and λ_{overlap} weight the two spatial regularizers. The same prototype-mediated score $S(I, T)$ is used during training and inference.

3.6 DATASETS AND IMPLEMENTATION DETAILS

Training and evaluation. We train ProtoLIP on Itemized-CC0.3M (Lyu et al., 2026) with a frozen ItemizedCLIP backbone, optimizing only the prototype evidence layer. All external benchmarks are evaluated zero-shot. We evaluate object queries on COCO (Lin et al., 2014), VOC20 (Everingham et al., 2010), and the complete ADE20K-150 validation split (Zhou et al., 2017); phrase queries on Flickr30K Entities (Plummer et al., 2015); and caption queries on COCO and Flickr30K Entities. Since COCO does not provide phrase-to-box links, we construct a caption-linked object-query benchmark by matching explicit category mentions in captions to COCO object categories and using all annotated boxes of the matched category as the target region. All methods use the same fixed manifest. Details are in Appendix G.

Baselines and metrics. We compare against the frozen ItemizedCLIP backbone, released VLMs, training-free localization methods, and spatially supervised grounding models (Figure 6). We quantify spatial evidence quality with Pointing (Zhang et al., 2016) and Energy, matching with AUC and BAcc, and retrieval with I@k and T@k.

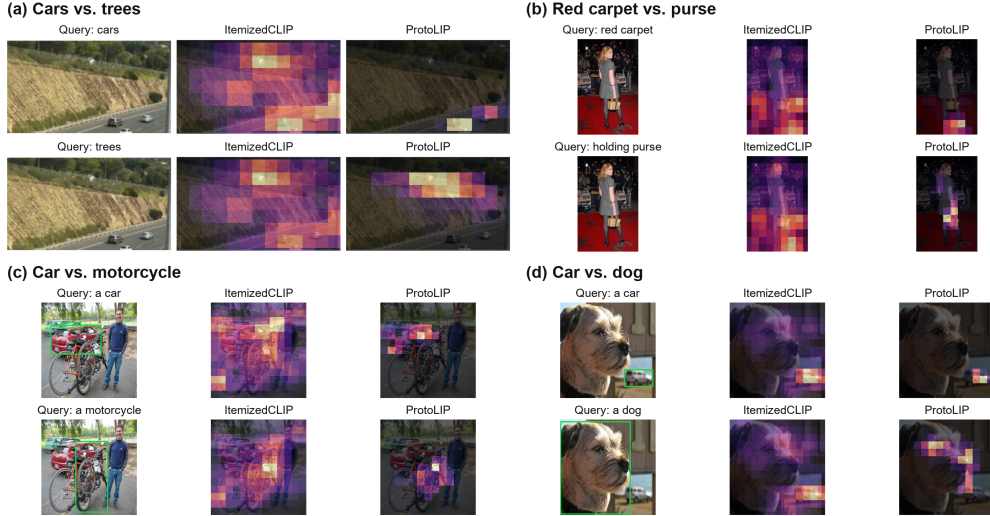


Figure 3: **Object-level evidence disentanglement.** ItemizedCLIP entangles evidence across object queries, whereas ProtoLIP localizes the queried object. Green boxes denote ground-truth regions.

4 RESULTS

Our evaluation examines whether ProtoLIP improves evidence separation across query granularities and whether these gains generalize across datasets and vision–language backbones. We first evaluate evidence disentanglement from object-level to full-caption queries (Section 4.1), and then show how the same prototype evidence exactly decomposes the model’s predictions (Section 4.2). We next test transfer to frozen public backbones and compare against public localization and supervised grounding methods (Section 4.3), followed by component ablations (Section 4.4). Additional results and controlled comparisons are provided in the appendix.

4.1 EVIDENCE DISENTANGLEMENT ACROSS QUERY GRANULARITIES

We first test whether ProtoLIP can mitigate the object-level evidence entanglement illustrated in Figures 1 and 3 on OOD benchmarks. As shown in Figure 4, ProtoLIP improves both Pointing and Energy on VOC20 object queries and ADE20K class queries. Gains on ADE20K are more modest, consistent with its broader long-tail vocabulary and greater semantic mismatch with the 0.3M text-only training corpus. Nevertheless, ProtoLIP also improves AUC and BAcc on both benchmarks, showing that the localization gains are accompanied by stronger matching performance.

We next test whether these gains extend to richer compositional queries. On COCO and Flickr30K, ProtoLIP further improves Pointing, Energy, AUC, and BAcc for object queries, while preserving or modestly improving full-caption localization (Figure 4), which shows that stronger object-level separation does not compromise evidence quality for complete sentence descriptions. Direct prototype-mediated inference improves matching separability but shows a moderate decrease in caption retrieval. Simple score-level fusion with the frozen backbone largely recovers this gap while retaining ProtoLIP’s spatial evidence (Table 1).

Table 1: Flickr30K Entities object-query score fusion. Fusion combines 0.75 ProtoLIP and 0.25 frozen-backbone scores.

Metric pair	Backbone	ProtoLIP	Fusion
Object P/E	0.378/0.308	0.454/0.427	0.454/0.427
AUC/BAcc	0.881/0.812	0.925/ 0.845	0.927/0.845
I@1/T@1	0.148/0.165	0.111/0.152	0.140/0.160

We thus provide two inference modes reflecting an interpretability–utility trade-off. *Direct ProtoLIP* preserves a fully decomposable prediction path in which the exposed prototype evidence directly constitutes the model score, whereas *Fused ProtoLIP* incorporates the backbone’s global matching signal to improve retrieval and matching utility while preserving the spatial evidence maps. We use Direct ProtoLIP as the primary model.

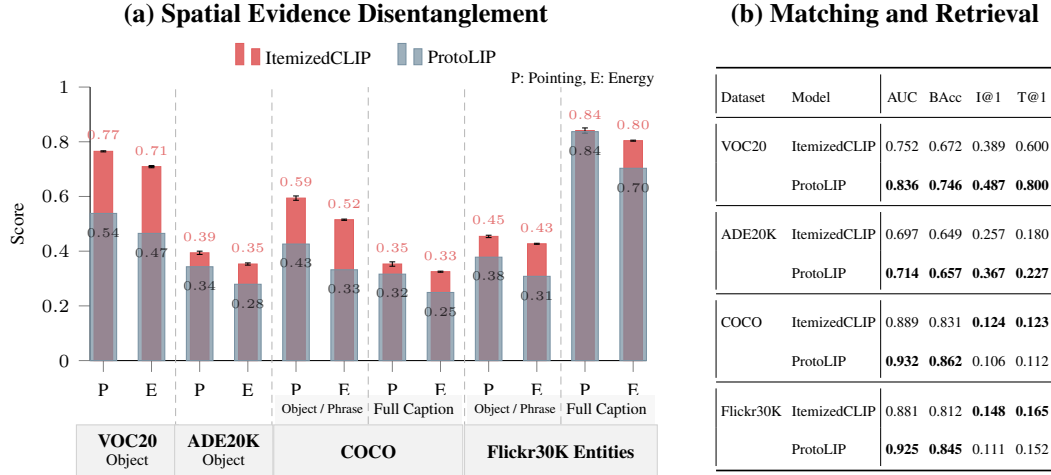


Figure 4: **Spatial evidence disentanglement and matching across OOD benchmarks.** (a) ProtoLIP improves Pointing (P) and Energy (E); error bars show $\pm$ standard deviation over three seeds. (b) ProtoLIP improves AUC and BAcc.

4.2 TRANSPARENT EVIDENCE COMPOSITION

The same prototype pathway that separates object-level evidence also exactly decomposes the matching score of a complete sentence query. Figure 5 illustrates this for the caption “a gray and white kitten sitting in a bathroom sink.” The complete text representation evaluates the image-specific evidence of all eligible prototypes through $s_k(I, T)$, and each prototype contributes $q_k(I, T) = \alpha_k(T)s_k(I, T)$ to the final score. Aggregating these prototype contributions within semantic families reveals distinct support from the animal and indoor-scene evidence, and the resulting family contributions sum exactly to the final prototype-mediated matching score.

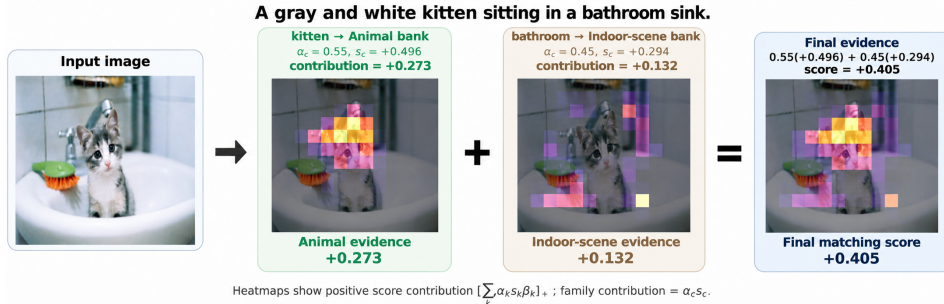


Figure 5: Exact decomposition of prototype evidence for a compositional sentence query.

4.3 GENERALIZATION ACROSS VISION–LANGUAGE BACKBONES

We then transfer ProtoLIP to independently pretrained SigLIP-WebLI and FLAIR merged-30M backbones. Training only a new prototype layer on Itemized-CC0.3M improves Pointing and Energy in all six backbone–dataset combinations (Table 2), showing that the localization benefit is not specific to ItemizedCLIP. Global matching and retrieval are also improved for SigLIP-WebLI and remain broadly stable for FLAIR, with only modest metric-specific variation.

We also compare with complementary external baselines (Figure 6): FLAIR and ItemizedCLIP, NACLIP as a training-free localization reference, and the spatially supervised MM Grounding DINO. All methods are re-evaluated on identical object queries and target regions. Despite using only Itemized-CC0.3M, compared with FLAIR’s 30M pairs and NACLIP’s CLIP-B/16 pretrained on approximately 400M pairs, ProtoLIP outperforms both fine-grained VLM baselines on all eight comparisons and exceeds NACLIP on six of eight, trailing it only on ADE20K. Against MM Grounding DINO, ProtoLIP is stronger on COCO and VOC20 under both metrics, while trailing on ADE20K and Flickr30K Entities. Thus, despite training on only 0.3M image–text pairs, ProtoLIP remains competitive with models benefiting from substantially larger-scale pretraining and, on several benchmarks, with a dedicated grounding model trained using explicit spatial supervision.

Table 2: Transfer of ProtoLIP to frozen public vision–language backbones on OOD Object-level benchmarks. Only the ProtoLIP layer is trained on Itemized-CC0.3M.

Dataset	Backbone	Variant	Pointing	Energy	AUC	BAcc	I1	T1
COCO Object	SigLIP-WebLI	Frozen	0.292	0.245	0.791	0.749	0.906	0.773
		+ ProtoLIP Layer	0.390	0.384	0.833	0.764	0.909	0.813
	FLAIR merged-30M	Frozen	0.245	0.313	0.906	0.827	0.659	0.707
		+ ProtoLIP Layer	0.558	0.483	0.888	0.804	0.812	0.687
ADE20K	SigLIP-WebLI	Frozen	0.226	0.230	0.793	0.730	0.623	0.587
		+ ProtoLIP Layer	0.412	0.363	0.802	0.737	0.643	0.580
	FLAIR merged-30M	Frozen	0.214	0.278	0.764	0.691	0.626	0.560
		+ ProtoLIP Layer	0.436	0.397	0.785	0.719	0.667	0.553
VOC20	SigLIP-WebLI	Frozen	0.397	0.365	0.862	0.828	0.914	1.000
		+ ProtoLIP Layer	0.684	0.625	0.912	0.834	0.908	1.000
	FLAIR merged-30M	Frozen	0.272	0.473	0.899	0.833	0.912	1.000
		+ ProtoLIP Layer	0.802	0.747	0.890	0.814	0.921	0.950

Finally, to isolate backbone effects from pretraining data, we retrain Multi-positive SigLIP, FLAIR, and DreamLIP on Itemized-CC0.3M. ProtoLIP improves localization for ItemizedCLIP, Multi-positive SigLIP, and FLAIR, but not DreamLIP. Global utility is largely preserved for FLAIR but degrades for Multi-positive SigLIP. These findings suggest that well-aligned frozen patch–text representations are important for preserving global matching, whereas localization gains transfer across multiple compatible backbones. See details in Appendix F.

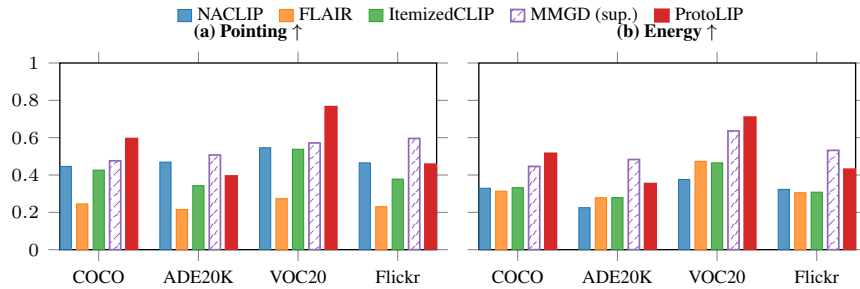


Figure 6: OOD object-level localization across fine-grained VLMs, training-free localization, spatially supervised grounding, and ProtoLIP.

4.4 ABLATION STUDY

We ablate ProtoLIP on COCO to isolate the effects of prototype mediation, semantic organization, and spatial regularization. As shown in Table 3, a flat prototype layer already improves object-level localization, while semantic family organization provides the largest additional gain. The area and overlap constraints further improve evidence quality by discouraging diffuse responses and excessive sharing between active families, with their combination achieving the strongest overall Pointing and Energy.

All prototype variants use the same parameter budget, adding only 2.23M trainable parameters (1.48% of the frozen backbone). Full parameter counts are provided in Appendix D.

5 CONCLUSION

We show that increasingly fine-grained text conditioning does not necessarily yield correspondingly precise object-level evidence. We introduce ProtoLIP, a lightweight prototype-mediated evidence layer that improves query-specific localization and separation without spatial supervision or backbone retraining. Crucially, ProtoLIP constructs its matching score directly from localized prototype evidence, enabling the score of a complete sentence query to be exactly decomposed into semantic-family and prototype contributions.

Table 3: Ablation of ProtoLIP components on COCO object-level.

Variant	Proto.	Family	Area	Overlap	Object-level localization		Matching/retrieval			
					Pointing	Energy	AUC	BAcc	I@1	T@1
Frozen ItemizedCLIP	–	–	–	–	0.426	0.332	0.889	0.831	0.124	0.123
Flat Prototype	✓	–	–	–	0.492	0.407	0.919	0.846	0.081	0.105
Family Prototype	✓	✓	–	–	0.586	0.461	0.934	<u>0.863</u>	<u>0.113</u>	0.111
+ Area	✓	✓	✓	–	<u>0.588</u>	<u>0.494</u>	<u>0.933</u>	<u>0.862</u>	0.103	<u>0.112</u>
+ Overlap	✓	✓	–	✓	0.582	0.459	0.934	0.864	0.101	0.110
Full ProtoLIP	✓	✓	✓	✓	0.594	0.515	0.932	0.862	0.106	<u>0.112</u>

REFERENCES

- Aishwarya Agarwal, Srikrishna Karanam, and Vineet Gandhi. Concept regions matter: Benchmarking CLIP with a new cluster-importance approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2864–2874, 2026.
- Chaofan Chen, Oscar Li, Chaofan Tao, Alina Jade Barnett, Cynthia Rudin, and Jonathan K. Su. This looks like that: Deep learning for interpretable image recognition. In *Advances in Neural Information Processing Systems*, volume 32, pp. 8930–8941, 2019.
- Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.
- Sina Hajimiri, Ismail Ben Ayed, and Jose Dolz. Pay attention to your neighbours: Training-free open-vocabulary semantic segmentation. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 5061–5071, 2025.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755, 2014.
- 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 *European Conference on Computer Vision*, 2024. doi: 10.1007/978-3-031-72970-6_3.
- Yiwei Lyu, Chenhui Zhao, Soumyanil Banerjee, Shixuan Liu, Akshay Rao, Akhil Kondepudi, Honglak Lee, and Todd C. Hollon. Learning complete and explainable visual representations from itemized text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21110–21120, 2026.
- Faridoun Mehri, Mahdieh Soleymani Baghshah, and Mohammad Taher Pilehvar. LibraGrad: Balancing gradient flow for universally better vision transformer attributions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 67–78, 2025. doi: 10.1109/CVPR52734.2025.00016.
- 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 *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2641–2649, 2015.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.
- Boyuan Sun, Bo-Wen Yin, Yuan-Ming Li, Xihan Wei, and Qibin Hou. See what i mean: Aligning vision and language representations for video fine-grained object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36814–36827, 2026.

-
- Junyi Wu, Bin Duan, Weitai Kang, Hao Tang, and Yan Yan. Token transformation matters: Towards faithful post-hoc explanation for vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10926–10935, 2024a.
- Junyi Wu, Weitai Kang, Hao Tang, Yuan Hong, and Yan Yan. On the faithfulness of vision transformer explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10936–10945, 2024b.
- 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*, pp. 24884–24894, 2025.
- Zhongxing Xu, Feilong Tang, Zhe Chen, Yingxue Su, Zhiyi Zhao, Ge Zhang, Jionglong Su, and Zongyuan Ge. Toward modality gap: Vision prototype learning for weakly-supervised semantic segmentation with CLIP. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 9023–9031, 2025. doi: 10.1609/aaai.v39i9.32976.
- 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.
- Anni Yu and Yu-Bin Yang. PRISM: Prototype-based reasoning with inter-modal semantic mining for interpretable image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2853–2863, 2026.
- 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*, pp. 11975–11986, 2023.
- Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. In *European Conference on Computer Vision*, pp. 543–559, 2016.
- Xiangyu Zhao, Yicheng Chen, Shilin Xu, Xiangtai Li, Xinjiang Wang, Yining Li, and Haian Huang. An open and comprehensive pipeline for unified object grounding and detection. *arXiv preprint arXiv:2401.02361*, 2024.
- Kecheng Zheng, Yifei Zhang, Wei Wu, Fan Lu, Shuailei Ma, Xin Jin, Wei Chen, and Yujun Shen. DreamLIP: Language-image pre-training with long captions. In *European Conference on Computer Vision*, pp. 73–90, 2024.
- Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. RegionCLIP: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16793–16803, 2022. doi: 10.1109/CVPR52688.2022.01629.
- Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 633–641, 2017.

A LLM PROMPT TEMPLATES FOR TAXONOMY CONSTRUCTION

We used the following offline prompt templates to construct the semantic family taxonomy. The LLM was given only sampled Itemized-CC0.3M training descriptions; no images, evaluation labels, downstream test data, or model outputs were provided.

Prompt 1: Visual-concept extraction.

You are helping construct a visual concept vocabulary from image-text training descriptions.

Input: a batch of image descriptions.

Your task is to extract only explicitly mentioned visual entities or image regions that can be spatially localized in an image.

Rules:

1. Keep concrete objects, object parts, people/animals, scene regions, and visually identifiable materials or places when they are localizable.
2. Exclude actions, relations, colors, styles, emotions, subjective attributes, quantities alone, and abstract concepts.
3. Normalize obvious lexical variants to one canonical concept.
4. Record aliases only if they appear in the provided descriptions.
5. Do not introduce concepts or aliases unsupported by the input.
6. If a phrase is too vague or not spatially localizable, discard it.

Return JSON:

```
{
  "concepts": [
    {
      "canonical": "canonical concept name",
      "aliases": ["alias 1", "alias 2"],
      "evidence_descriptions": [0, 3]
    }
  ]
}
```

The `evidence_descriptions` field lists the indices of input descriptions where the concept was observed.

Input descriptions:
[BATCH_OF_DESCRIPTIONS]

Prompt 2: Adaptive semantic grouping.

You are helping organize a visual concept vocabulary into coarse semantic families for routing visual prototypes.

Input: a list of canonical visual concepts and their observed aliases.

Your task is to group concepts into semantic families that are visually and semantically coherent.

Rules:

1. Choose the number of families based on the concept inventory.
2. Families should be coarse routing groups, not fine-grained class labels.
3. Separate concepts with substantially different visual structure or role.
4. Merge lexical variants and close synonyms into the same family.
5. Allow multiple-family membership only for genuinely ambiguous concepts.
6. Do not create families for actions, relations, colors, or styles.
7. Do not add concepts or aliases absent from the input inventory.

Return JSON:

```
{
  "families": [
    {
      "family": "family name",
      "description": "short visual definition",
      "concepts": [
        "canonical concept 1",
        "canonical concept 2"
      ],
      "aliases": ["alias 1", "alias 2"]
    }
  ]
}
```

```
}
```

```
Concept inventory:
[CONCEPT_INVENTORY]
```

Prompt 3: Taxonomy audit.

You are auditing a draft semantic taxonomy for visual prototype routing.

Input: a draft family-to-alias taxonomy and the original concept inventory.

Your task is to check whether the taxonomy is complete, consistent, and supported by the training-text inventory.

Audit requirements:

1. Identify missing concepts that should be assigned to a family.
2. Identify duplicate aliases or inconsistent family assignments.
3. Identify visually incoherent families.
4. Identify aliases or concepts unsupported by the input inventory.
5. Identify overly broad or overly narrow families.
6. Preserve coarse semantic routing.
7. Base every finding and revision only on the provided concept inventory; do not introduce unsupported concepts.

Return JSON:

```
{
  "issues": [
    {
      "type": "missing | duplicate | ambiguous | incoherent | unsupported",
      "item": "concept or alias",
      "explanation": "brief reason",
      "suggested_fix": "specific correction"
    }
  ],
  "revised_taxonomy": [
    {
      "family": "family name",
      "concepts": [
        "canonical concept 1",
        "canonical concept 2"
      ],
      "aliases": ["alias 1", "alias 2"]
    }
  ]
}
```

```
Draft taxonomy:
[DRAFT_TAXONOMY]
```

```
Concept inventory:
[CONCEPT_INVENTORY]
```

B TAXONOMY GRANULARITY ANALYSIS

We compare the main 39-family taxonomy (K39), constructed from the expanded extraction, with a smaller 23-family taxonomy (K23) built from 30K sampled training descriptions. Only the ProtoLIP layer is retrained, and both variants are evaluated under the same held-out Itemized-CC0.3M split used throughout the paper, reported here for the full 1,200-pair evaluation set. Results are shown in Table 4.

Table 4: In-domain FullPlus comparison between the 23-family taxonomy (K23) and the finer 39-family taxonomy (K39). Both models use direct prototype-mediated inference without knowledge distillation and are evaluated on the same 1,200-pair protocol.

Model	AUC	BAcc	I@1	I@5	I@10	T@1	T@5	T@10
ProtoLIP K23	0.956	0.881	0.340	0.466	0.546	0.323	0.448	0.522
ProtoLIP K39	0.958	0.880	0.347	0.485	0.553	0.337	0.463	0.526

Table 5: Effect of taxonomy granularity. K23 and K39 denote the 23-family and 39-family taxonomies, respectively; both use direct prototype-mediated inference without knowledge distillation (no-KD). Differences between the two taxonomies are within one point on all reported metrics.

Dataset / Query	K23 (no-KD)		K39 (no-KD)	
	Pointing	Energy	Pointing	Energy
COCO object / phrase	.5868	.5187	.5939	.5143
Flickr30K phrase	.4553	.4293	.4571	.4301

Table 4 shows that increasing the number of semantic families from 23 to 39 produces only modest changes in matching and retrieval. K39 slightly improves AUC and several retrieval cutoffs, while BAcc remains essentially unchanged. Table 5 shows that K39 additionally yields modestly stronger spatial localization, with differences from K23 within one point on all reported metrics. We adopt K39 as our main taxonomy for this modest localization benefit and its broader semantic coverage; the small gap to K23 on both matching and localization indicates that our results are not sensitive to this granularity choice.

C BEHAVIOR ON NATURALLY UNMATCHED QUERIES

We next examine whether queries that fail to match any semantic family are systematically degraded. This analysis uses the held-out in-domain Itemized-CC0.3M split and is distinct from the OOD benchmarks evaluated in the main paper.

Reported values are mean $\pm$ sample standard deviation over three independent training runs with different random seeds. The same evaluation protocol is applied to each trained model. Results are reported in Table 6.

Table 6: Performance on naturally matched and unmatched Itemized-CC0.3M queries. Values are mean $\pm$ sample standard deviation over three seeds. Unmatched queries use the complete prototype-bank fallback.

Query subset / routing	AUC	BAcc	I@1	T@1
Matched, standard routing	0.999 $\pm$ 0.000	0.989 $\pm$ 0.001	0.746 $\pm$ 0.010	0.798 $\pm$ 0.012
Matched, forced fallback	0.999 $\pm$ 0.000	0.987 $\pm$ 0.001	0.795 $\pm$ 0.018	0.815 $\pm$ 0.009
Naturally unmatched, fallback	0.998 $\pm$ 0.000	0.986 $\pm$ 0.002	0.782 $\pm$ 0.010	0.777 $\pm$ 0.013

As shown in Table 6, naturally unmatched queries retain AUC and balanced accuracy close to those of matched queries. For these queries, standard inference and explicitly forced fallback are identical because unmatched queries already use the complete prototype bank.

As a controlled intervention, we additionally force matched queries to use the complete bank instead of their routed families. Balanced accuracy remains essentially unchanged, while I@1 increases. Matched and naturally unmatched queries are semantically different populations by construction, so direct comparisons between these subsets should be interpreted descriptively rather than causally.

This same direction is visible under a separate, model-level comparison: relative to the frozen ItemizedCLIP baseline under the identical in-domain FullPlus protocol (Table 8), ProtoLIP’s strict I@1 is also lower (0.340 vs 0.368) despite leading on AUC, BAcc, and broader retrieval cutoffs. Together, these two independent in-domain comparisons suggest that family-constrained routing carries a consistent, modest cost to strict top-1 retrieval, traded for the localization, separation, and matching gains reported throughout the paper.

Overall, the fallback mechanism prevents unmatched queries from being silently discarded or assigned to arbitrary semantic families.

D PARAMETER BUDGET

To verify that the improvements in the main ablation are not caused by differences in trainable capacity, we report the parameter budget of all prototype variants in Table 7.

Table 7: Parameter budget across ablation variants. The ItemizedCLIP backbone remains frozen, and only the ProtoLIP layer is optimized. All prototype variants introduce the same 2.23M trainable parameters (1.48% of the backbone). Family routing uses fixed masks, while the area and overlap objectives are parameter-free.

Model	Frozen Backbone	Trainable	Total	Increase
Frozen ItemizedCLIP	150.686M	0	150.686M	0.00%
Flat Prototype	150.686M	2.230M	152.916M	1.48%
Family Prototype	150.686M	2.230M	152.916M	1.48%
Full ProtoLIP	150.686M	2.230M	152.916M	1.48%

Table 7 confirms that Flat Prototype, Family Prototype, and Full ProtoLIP use identical trainable parameter budgets. Improvements from semantic family routing and spatial regularization therefore cannot be attributed to increased model capacity.

E SAME-DATA IN-DOMAIN CONTROLS

We additionally compare ProtoLIP with alternative vision-language objectives trained using the same Itemized-CC0.3M training manifest. This controls for differences in pretraining data that affect comparisons with released large-scale checkpoints. Results are reported in Table 8.

Table 8: In-domain image-text matching and retrieval on the same-data FullPlus test set with 1,200 image-text pairs. All transferred models are trained on Itemized-CC0.3M. ProtoLIP uses the direct prototype-mediated score without knowledge distillation or inference-time backbone fusion. I@K and T@K denote image-to-text and text-to-image Recall@K, respectively. Best results are shown in **bold**.

Model	AUC	BAcc	I@1	I@5	I@10	T@1	T@5	T@10
ItemizedCLIP	0.845	0.779	0.368	0.406	0.434	0.356	0.408	0.438
Single-positive Sigmoid-0.3M	0.908	0.820	0.263	0.375	0.428	0.223	0.355	0.398
Multi-positive Sigmoid-0.3M	0.905	0.820	0.334	0.422	0.466	0.306	0.403	0.437
FLAIR-0.3M	0.892	0.804	0.323	0.417	0.455	0.290	0.399	0.441
DreamLIP-0.3M	0.829	0.763	0.100	0.203	0.264	0.073	0.168	0.253
ProtoLIP	0.956	0.881	0.340	0.466	0.546	0.323	0.448	0.522

As shown in Table 8, ProtoLIP achieves the strongest AUC and balanced accuracy among the same-data controls. Although ItemizedCLIP retains the highest strict Recall@1, ProtoLIP provides the strongest retrieval performance at several broader cutoffs, including I@5, I@10, T@5, and T@10. These results indicate that the gains in matching quality are not explained solely by differences in pretraining data or scale.

F MATCHED-DATA BACKBONE TRANSFER AND PATCH-TEXT ALIGNMENT

To separate backbone effects from pretraining-data differences, we evaluate ProtoLIP on backbones trained on the same Itemized-CC0.3M corpus. The results in Table 9 show that transfer remains backbone-dependent. We hypothesize that ProtoLIP requires frozen patch tokens that are meaningfully aligned with the text space, since its matching score is constructed entirely from prototype-pooled patch evidence.

We test this property on a shared set of 400 held-out image-text pairs. Patch Top-8 scores are computed by averaging the eight highest patch-text cosine similarities, without prototype pooling or family routing. Patch Top-8 AUC measures pairwise discrimination, while Patch I@1 measures image-to-text retrieval over the 400 candidates.

Patch Top-8 AUC near 0.5 and Patch I@1 near the $1/400 = 0.0025$ chance level indicate that the patch tokens of DreamLIP and Multi-positive SigLIP carry almost no text-discriminative infor-

Table 9: Matched-data transfer across frozen vision–language backbones under the full evaluation protocol. All backbones are trained on Itemized-CC0.3M and then frozen; only the ProtoLIP layer is trained. COCO uses the object-query benchmark, while ADE20K and VOC20 use full validation splits. ItemizedCLIP-0.3M rows report three-seed means (seeds 1713–1715); the remaining backbones report a single run.

Dataset	Frozen backbone	Variant	Pointing	Energy	AUC	BAcc	I@1	T@1
COCO object	ItemizedCLIP-0.3M	Frozen	.426	.332	.889	.831	.124	.123
		+ ProtoLIP	.594	.515	.932	.862	.106	.112
	Multi-positive SigLIP-0.3M	Frozen	.185	.210	.821	.750	.410	.240
		+ ProtoLIP	.456	.374	.711	.659	.199	.213
	FLAIR-0.3M	Frozen	.243	.260	.793	.729	.388	.253
		+ ProtoLIP	.329	.315	.793	.721	.365	.387
	DreamLIP-0.3M	Frozen	.306	.298	.791	.729	.268	.213
		+ ProtoLIP	.262	.283	.731	.676	.099	.133
VOC20 full val	ItemizedCLIP-0.3M	Frozen	.538	.465	.752	.672	.389	.600
		+ ProtoLIP	.765	.709	.836	.746	.487	.800
	Multi-positive SigLIP-0.3M	Frozen	.288	.314	.872	.798	.545	.800
		+ ProtoLIP	.654	.543	.801	.734	.340	.500
	FLAIR-0.3M	Frozen	.399	.408	.850	.778	.524	.900
		+ ProtoLIP	.522	.487	.845	.758	.500	.750
	DreamLIP-0.3M	Frozen	.515	.440	.811	.739	.395	.350
		+ ProtoLIP	.438	.424	.765	.702	.248	.450
ADE20K full val	ItemizedCLIP-0.3M	Frozen	.342	.279	.697	.649	.257	.180
		+ ProtoLIP	.394	.353	.714	.657	.367	.227
	Multi-positive SigLIP-0.3M	Frozen	.212	.225	.685	.634	.289	.213
		+ ProtoLIP	.303	.275	.685	.637	.175	.213
	FLAIR-0.3M	Frozen	.248	.252	.644	.602	.268	.253
		+ ProtoLIP	.283	.271	.695	.642	.247	.247
	DreamLIP-0.3M	Frozen	.257	.244	.721	.673	.228	.127
		+ ProtoLIP	.253	.237	.730	.676	.157	.173

Table 10: Patch–text alignment on the shared 400-pair diagnostic set. These diagnostic values are separate from the formal benchmark results in Table 9.

Backbone	Patch Top-8 AUC	Patch I@1	Routed ProtoLIP AUC
DreamLIP-0.3M	0.513	0.003	0.845
Multi-positive SigLIP-0.3M	0.496	0.000	0.880
FLAIR-0.3M	0.782	0.308	0.953

mation on their own, whereas FLAIR exposes substantially more (0.782 AUC, 0.308 Patch I@1). These backbones remain strong at the image level (Table 9), so the missing discriminability appears to be concentrated in their global representations rather than absent from the model. Since the ProtoLIP score is constructed entirely from prototype-pooled patch evidence, matching performance is retained only where such alignment is present, consistent with the more stable score-level transfer observed on FLAIR. Disabling family routing does not recover the lost matching performance, suggesting that weak patch–text alignment, rather than taxonomy mismatch, is the limiting factor.

Table 11 summarizes the average change from adding a ProtoLIP layer. Localization and score-level matching do not move together: Multi-positive SigLIP-0.3M gains the most in Pointing (+0.243) while losing the most in retrieval (−0.177), whereas DreamLIP-0.3M degrades on both. Patch–text alignment (Table 10) tracks the matching columns across all four backbones, but does not sepa-

Table 11: Average change from adding a ProtoLIP layer to matched-data frozen backbones across COCO object, ADE20K full val, and VOC20 full val. Localization gains and matching retention behave differently. Computed from Table 9.

Backbone	Δ Pointing	Δ Energy	Δ AUC	Δ I@1
ItemizedCLIP-0.3M	+0.149	+0.167	+0.048	+0.059
Multi-positive SigLIP-0.3M	+0.243	+0.148	−0.060	−0.177
FLAIR-0.3M	+0.081	+0.051	+0.015	−0.023
DreamLIP-0.3M	−0.042	−0.013	−0.032	−0.129

rate DreamLIP-0.3M from Multi-positive SigLIP-0.3M, whose alignment is comparable (0.513 vs 0.496) while their localization moves in opposite directions. We therefore report localization transfer and score-level matching transfer as separate empirical behaviors, and leave the determinants of the former to future work.

G EVALUATION PROTOCOL AND IMPLEMENTATION DETAILS

Prototype layer architecture. In our implementation, ϕ_θ is a gated two-layer residual MLP with hidden width $d/2$, GELU activation, residual scale $\gamma = 0.25$, and final ℓ_2 normalization. We optimize the prototype evidence layer with AdamW for 6,000 steps using a learning rate of 1.5×10^{-3} and weight decay of 10^{-3} . The effective batch size is 1,024. Distributed runs use two NVIDIA GPUs with 256 samples per GPU and two gradient-accumulation steps; single-GPU sensitivity experiments use a batch size of 256 with four accumulation steps. Unless otherwise specified, reported ProtoLIP results are averaged over three independent training runs with random seeds 1713, 1714, and 1715.

The prototype and matching temperatures are set to 0.04 and 0.05, respectively. Image patches are pooled with a 2×2 region size. We set $\lambda_{\text{area}} = 1.5$, $\tau_a = 0.14$, $\lambda_{\text{overlap}} = 0.25$, and $\tau_o = 0.20$. Gradient checkpointing and chunked score computation are used to reduce memory consumption without changing the model or objective. The frozen vision–language backbone is shared across all variants, and no KD, teacher branch, or inference-time score fusion is used for the main model.

External baseline checkpoints. We evaluate the public `openmmlab-community/mm_grounding_dino_large_all` checkpoint, whose training corpus includes GoldG, V3Det, COCO2017, LVIS-v1, COCO2014, GRIT, RefCOCO, RefCOCO+, RefCOCOg, and gRefCOCO. NACLIP uses CLIP-B/16, and FLAIR uses the released merged-30M checkpoint.

COCO caption-linked object queries. COCO does not provide phrase–box correspondences, so we construct a fixed caption-linked object-query manifest from validation captions. Before model evaluation, we define a boundary-aware alias list for COCO instance categories, e.g., *bike* for *bicycle* and *woman/man* for *person*, and apply deterministic lexical matching to each caption. We retain only matched categories with at least one ground-truth instance box in the image, discarding unmatched, non-localizable, or unannotated mentions. If multiple instances of a matched category are present, their boxes form a union target: Pointing is correct when the top evidence patch falls inside any instance box, and Energy is the evidence mass within their union. The resulting manifest contains 681 object queries over 557 images and is shared by all methods.