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

Anonymous authors
Paper under double-blind review

ABSTRACT

Query-conditioned vision–language models enable fine-grained interpretation by revealing which visual content supports a given textual query and how this evidence changes across queries. However, semantically, sentence-level evidence does not necessarily decompose into object-specific contributions, while spatially, object-level evidence can remain entangled with co-occurring objects and surrounding scene context. Across multiple VLM architectures and independent benchmarks, we observe persistent object-level evidence entanglement. Moreover, exposed evidence maps do not necessarily correspond to the evidence that directly constitutes the model's prediction. To disentangle visual evidence at both semantic and spatial levels, we introduce **ProtoLIP**, a lightweight prototype-mediated evidence layer that organizes reusable visual prototypes into text-derived semantic families and uses *coarse-to-fine evidence routing*, where semantic families constrain prototype eligibility and the complete query determines fine-grained prototype contributions. Our studies show that ProtoLIP improves evidence localization and separation across query granularities, achieving average relative gains of 29% in Pointing and 43% in Energy across four object- and phrase-level OOD benchmarks. Its localization gains also transfer to independently pretrained VLMs, with larger improvements observed in several transfer settings. On the primary backbone, ProtoLIP also improves image–text matching discrimination while remaining competitive with a spatially supervised grounding model in object-level localization. Crucially, ProtoLIP constructs its image–text matching score directly from localized prototype evidence, enabling exact decomposition across prototypes, semantic families, and spatial evidence without spatial annotations or backbone retraining.

1 INTRODUCTION

Vision–language models (VLMs) learn transferable visual representations through image–text alignment (Radford et al., 2021; Zhai et al., 2023). Recent models increasingly support fine-grained and text-conditioned visual representations, from long-caption and sub-caption supervision in DreamLIP (Zheng et al., 2024), to text-conditioned visual aggregation in FLAIR (Xiao et al., 2025), and further to item-level independence and representation completeness in ItemizedCLIP (Lyu et al., 2026). These advances enable the same image to be interpreted at different semantic granularities. Ideally, increasingly specific queries should yield correspondingly precise visual evidence, from complete descriptions to individual objects. However, the objectives of existing fine-grained VLMs do not explicitly require precise object-level visual evidence. ItemizedCLIP, for example, encourages distinct visual evidence for sentence-level items containing multiple objects and relations, but does not require object-level evidence within each sentence to be disentangled. As illustrated in Figure 1, evidence for the woman and bicycle can therefore remain entangled with each other and the surrounding context, even under object-specific queries.

Recent studies report complementary evidence of this limitation. SWIM (Sun et al., 2026) shows that object nouns can produce diffuse visual activations, while analyses of CLIP models identify reliance on background regions and spurious contextual correlations (Agarwal et al., 2026). Related studies further identify limitations in compositional concept binding and object-level localization in pretrained VLMs (Lewis et al., 2024; Bousselham et al., 2024; Assouel et al., 2025). Together, these findings suggest that fine-grained textual conditioning alone does not ensure object-specific

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

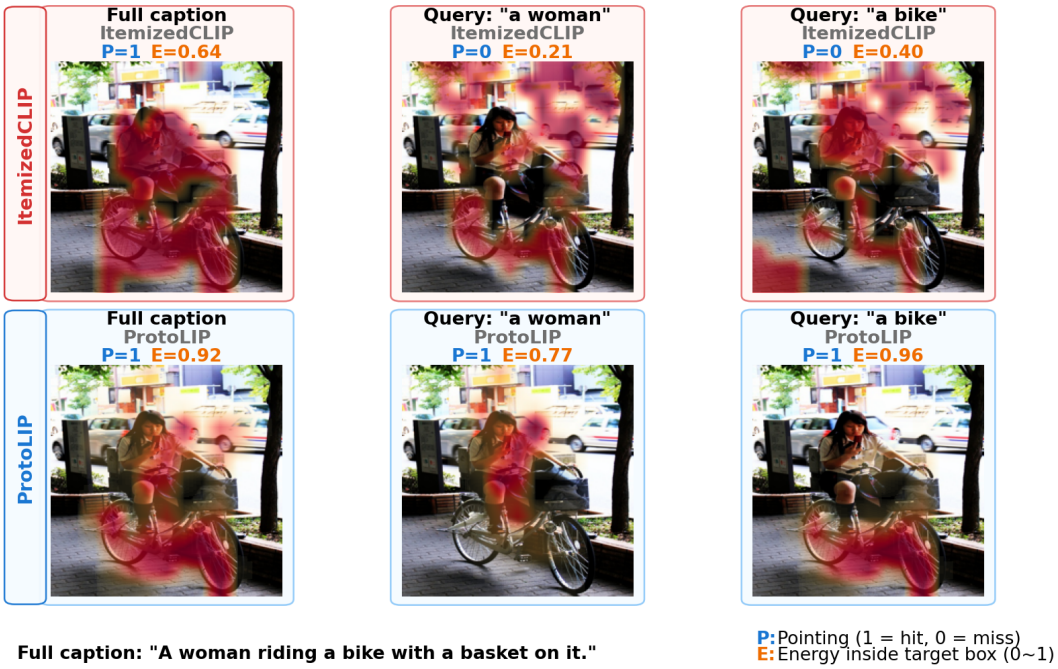


Figure 1: Object-level evidence disentanglement. ItemizedCLIP retains evidence from the woman and background when queried about objects, whereas ProtoLIP separates object-specific evidence. **P** indicates whether the peak evidence falls inside the target box, and **E** measures the fraction of evidence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

visual evidence. We study this limitation as an evidence-disentanglement problem, asking whether sentence-level evidence can be decomposed into distinct object-specific contributions and spatially separated from co-occurring entities and context.

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. ProtoLIP uses semantic families to select relevant prototypes and the full query to determine their contributions, constructing the image–text matching score directly from localized prototype evidence. This enables object-, phrase-, and sentence-level evidence while making the resulting prediction exactly decomposable into prototype and semantic-family contributions: these contributions are the additive terms of the matching score itself, rather than a separate post-hoc explanatory layer.

We evaluate ProtoLIP across query granularities, datasets, and vision–language backbones, examining both evidence localization and image–text matching. Our results show improved object-specific localization and evidence separation across these settings, with localization gains transferring to several independently pretrained VLM backbones. These gains are achieved with a lightweight trainable layer on a frozen backbone. ProtoLIP remains competitive with a spatially supervised grounding model in object-level localization while improving image–text matching on the primary backbone.

Our contributions are as follows.

- ProtoLIP, a lightweight prototype-mediated evidence layer that disentangles visual evidence at semantic and spatial levels while constructing image–text matching predictions

directly from localized prototype evidence, enabling exact decomposition of the resulting prediction.

- A comprehensive evaluation across datasets and VLM backbones, showing improved object-specific localization, evidence separation, and image–text matching on the primary backbone, with localization gains transferring across independently pretrained VLMs.

2 Related Work

Fine-grained vision–language representations. Vision–language models such as CLIP (Radford et al., 2021) learn transferable representations through contrastive image–text alignment, while SigLIP (Zhai et al., 2023) replaces the softmax contrastive objective 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 VLM. Training-free methods such as NACLIP (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.

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)). These methods primarily operate on predefined categories, whereas ProtoLIP targets compositional queries containing multiple constituent objects.

More broadly, visually plausible attribution maps do not necessarily reflect the evidence that actually determines a model's prediction (Adebayo et al., 2018). This issue also arises in prototype-based models, where prototype similarities may directly contribute to a prediction while their spatial explanations are not necessarily faithful to the underlying prototype computation (Wolf et al., 2024). In contrast, ProtoLIP constructs the matching score directly from localized prototype contributions, so that the resulting prediction is exactly decomposable across prototypes and spatial evidence.

3 Method

3.1 Text-Derived Semantic Families

Inspired by the class-specific prototype organization of ProtoPNet (Chen et al., 2019), we introduce **coarse-to-fine evidence routing**. Coarse semantic families first constrain prototype eligibility,

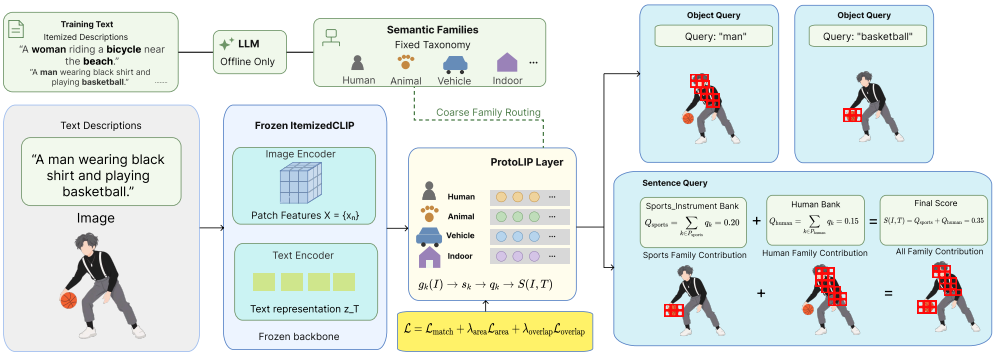


Figure 2: Overview of ProtoLIP. Training text is organized offline into coarse semantic families that constrain prototype eligibility. Given an image–text pair, a frozen vision–language backbone extracts patch features and a complete text representation. Eligible prototypes extract localized visual evidence,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

while the complete query then determines fine-grained prototype contributions. This design provides semantic specialization without requiring the taxonomy itself to encode fine-grained object distinctions. For example, *man*, *woman*, and *child* share the *human* prototype bank but can produce different contributions and spatial evidence.

We construct the taxonomy from training descriptions only. An offline LLM extracts localizable visual concepts and aliases and groups them into coarse families such as *human*, *vehicle*, and *animal*, producing a fixed dictionary $\mathcal{D} = \{(F_c, \mathcal{A}_c)\}_{c=1}^{K}$, where $\mathcal{A}_c$ contains the concepts and aliases associated with family F_c. The LLM is used only for this offline construction.

We use M learnable visual prototypes $\mathcal{P} = \{p_k\}_{k=1}^{M}$, where each prototype is a trainable vector in the frozen backbone's patch-embedding space that learns to capture a reusable visual pattern. The prototypes are learned during training by matching against image patches using the objectives described in Section 3.5.

We use $M = 512$ prototypes distributed approximately uniformly across the $K = 39$ semantic-family banks, with 13–14 prototypes per family (13.1 on average). Here, K is determined by the offline LLM-based taxonomy construction. During training and inference, we perform *semantic-family routing*: deterministic lexical matching activates families whose aliases occur in the query, defining the eligible prototype set $\mathcal{E}_T$. For example, "a woman riding a bicycle" activates the *human* and *vehicle* families; if no family is matched, all prototypes remain eligible.

Importantly, fine-grained evidence is determined by scoring eligible prototypes with the complete query representation z_T. Thus, the taxonomy acts only as a coarse semantic inductive bias, while object-specific evidence remains query-dependent. We further analyze semantic-family routing coverage on ADE20K, a challenging long-tail object benchmark, and Flickr30K Entities, an open-vocabulary phrase-level benchmark. The largest localization gains occur on matched queries, where semantic-family routing is active, while the all-prototype fallback largely preserves localization performance on unmatched queries. Full coverage statistics, matched/unmatched results, and additional taxonomy and fallback analyses are provided in Appendices A and C.

3.2 Family-Constrained Prototype Evidence

Given the prototypes selected by semantic-family routing, we next determine where each prototype extracts visual evidence and how strongly that evidence supports the complete query. Each prototype first defines a spatial assignment over the image patches, and the complete query then scores the resulting prototype-specific evidence. Let p_k denote the k-th learned prototype, $\mathrm{sim}(\cdot,\cdot)$ the feature-similarity function, and τ_p the prototype temperature. Each prototype defines a normalized spatial assignment

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

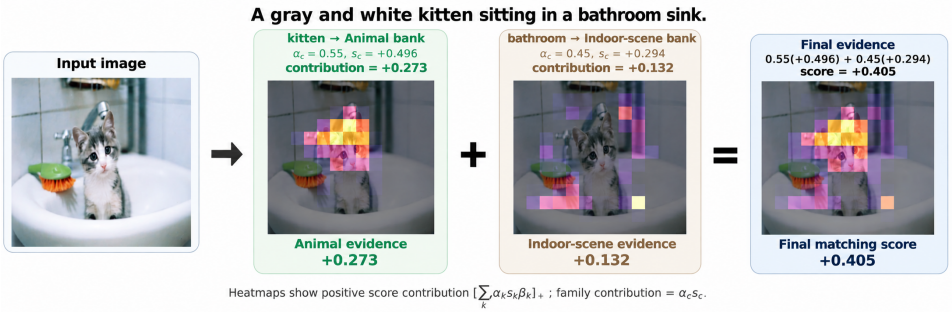


Figure 3: Exact decomposition of prototype evidence for a compositional sentence query.

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

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

where $\widetilde{g}_k(I)$ 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. Implementation details of the prototype evidence layer are provided in Appendix B.

With the prototype evidence extracted, semantic-family routing determines which prototypes are eligible to contribute. Each prototype has a learned positive weight that captures its relative importance within the eligible prototype set $r_k = \mathrm{softplus}(\rho_k)$. For an eligible prototype $k \in \mathcal{E}_T$, we set $\alpha_k(T) = r_k / \sum_{j \in \mathcal{E}_T} r_j$, and set $\alpha_k(T) = 0$ otherwise. The complete query embedding z_T then evaluates the compatibility of each eligible prototype's visual evidence:

$$s_k(I,T) = \mathrm{sim}(g_k(I), z_T), \qquad q_k(I,T) = \alpha_k(T) s_k(I,T), \qquad S(I,T) = \sum_k q_k(I,T). \tag{3}$$

Because the resulting contributions q_k sum directly to $S(I,T)$, the prototype-mediated matching prediction is exactly decomposable into prototype contributions.

3.3 Query-Specific Prediction Evidence

The previous step gives each prototype both a contribution to the matching score and a spatial assignment over the image. We combine these quantities to expose where the prediction is spatially supported while preserving the complete prototype-mediated score:

$$C(n; I, T) = \sum_k q_k(I,T) \beta_{k,n}(I), \qquad \sum_{n=1}^{N} C(n; I, T) = S(I,T). \tag{4}$$

The equality follows from the normalization of β_k. Thus, C provides a score-preserving and exhaustive spatial decomposition of the prototype-mediated prediction: every contribution to $S(I,T)$ is assigned to spatial evidence in C, with no residual score outside the exposed evidence. Positive and negative values respectively support and oppose the image–text match.

For localization and visualization, we normalize the positive part of C to unit mass, denoted H, producing a nonnegative spatial evidence map. Figure 3 illustrates this decomposition for a compositional sentence query, showing distinct contributions from the *animal* and *indoor-scene* families.

3.4 Spatial Evidence Regularization

The prototype assignments above localize evidence, but do not encourage evidence from different semantic families to be compact and spatially separated. We therefore regularize the spatial assignments of active families to discourage diffuse evidence and excessive sharing between families.

For a query with at least one matched family, let $\mathcal{F}_T$ denote the active semantic families and $\mathcal{P}_c$ the prototypes assigned to family F_c. We aggregate the prototype assignments within each active family into the normalized family map

$$M_c(n; I, T) = \frac{\sum_{k \in \mathcal{P}_c} \alpha_k(T) \beta_{k,n}(I)}{\sum_{k \in \mathcal{P}_c} \alpha_k(T)}. \tag{5}$$

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

$$A(M_c) = \frac{1}{N} \exp\left(- \sum_{n=1}^{N} M_c(n) \log(M_c(n) + \epsilon) \right), \tag{6}$$

$$O(M_c, M_{c'}) = \sum_{n=1}^{N} \min(M_c(n), M_{c'}(n)), \tag{7}$$

where $\epsilon > 0$ ensures numerical stability. $A(M_c)$ is an entropy-based effective-area measure, while $O(M_c, M_{c'})$ measures shared spatial mass.

Let $\mathcal{R}_T$ denote the unordered pairs of distinct active families. We penalize only excessive area:

$$\mathcal{L}_{\text{area}} = \frac{1}{|\mathcal{F}_T|} \sum_{c \in \mathcal{F}_T} [A(M_c) - \tau_a]_+^2. \tag{8}$$

When $\mathcal{R}_T$ is nonempty, we additionally penalize excessive pairwise overlap:

$$\mathcal{L}_{\text{overlap}} = \frac{1}{|\mathcal{R}_T|} \sum_{(c,c') \in \mathcal{R}_T} [O(M_c, M_{c'}) - \tau_o]_+^2. \tag{9}$$

Here, $[u]_+ = \max(0, u)$, and τ_a and τ_o are the tolerated area and overlap thresholds. These losses encourage compact and separated family evidence without forcing semantically related families to occupy disjoint regions.

For single-family queries, $\mathcal{L}_{\text{overlap}} = 0$; unmatched queries use all prototypes and are excluded from spatial regularization. Within-family distinctions instead rely on query-conditioned prototype scoring.

3.5 Learning Objective

We jointly optimize image–text matching and the spatial organization of the prototype evidence defined above. The objective combines a symmetric contrastive matching loss with the two spatial regularizers.

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 $\mathbf{y} = (1, \ldots, B)$ the diagonal matching targets. The matching loss is

$$\mathcal{L}_{\text{match}} = \frac{1}{2} \left[\text{CE}(\mathbf{S}/\tau_s, \mathbf{y}) + \text{CE}(\mathbf{S}^\top/\tau_s, \mathbf{y}) \right], \tag{10}$$

where $\text{CE}(\cdot, \cdot)$ denotes cross-entropy and $\tau_s > 0$ is the score temperature.

The complete objective is

$$\mathcal{L} = \mathcal{L}_{\text{match}} + \lambda_{\text{area}} \mathcal{L}_{\text{area}} + \lambda_{\text{overlap}} \mathcal{L}_{\text{overlap}}, \tag{11}$$

where λ_{area} and λ_{overlap} weight the two spatial regularizers.

Spatial regularization is applied only to positive diagonal pairs, while the same score $S(I, T)$ is used for training and inference.

4 Experiments

4.1 Experimental Setup

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-level evidence on COCO (Lin et al., 2014), VOC20 (Everingham et al., 2010), and ADE20K-150 validation split (Zhou et al., 2017); phrase-level evidence on Flickr30K Entities (Plummer et al., 2015); and caption-level evidence on COCO and Flickr30K Entities. Because COCO lacks phrase-to-box links, we construct a caption-linked object-query benchmark by matching explicit category mentions in captions to COCO categories and treating all annotated instances of the matched category as the target region. All methods use the same fixed manifest. Additional full COCO-80 evaluation is provided in Appendix B.1; full implementation, benchmark, and evaluation details are provided in Appendix B.

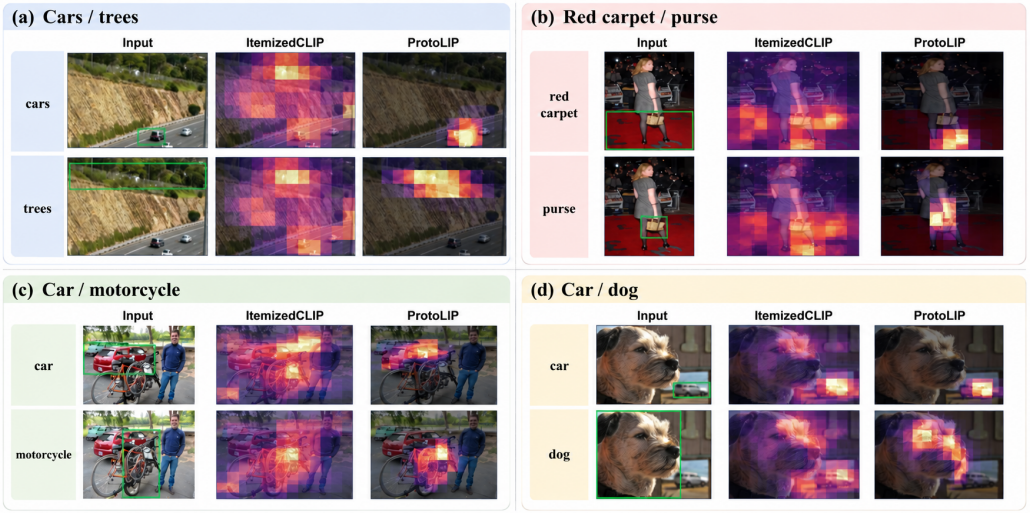


Figure 4: **Object-level evidence disentanglement.** ItemizedCLIP entangles evidence across object queries, whereas ProtoLIP localizes the queried object. Green boxes denote ground-truth regions.

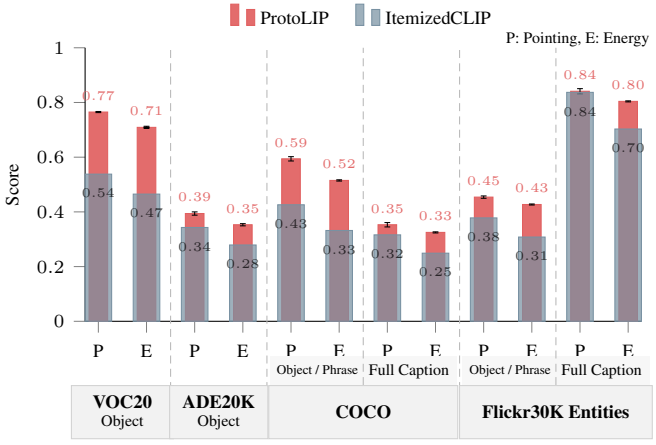
Baselines and metrics. We compare against the frozen ItemizedCLIP backbone, released VLMs, training-free localization methods, and spatially supervised grounding models (Figure 6). We evaluate spatial evidence using Pointing Game Accuracy (Zhang et al., 2016), which measures whether the maximum-evidence location falls inside the ground-truth region, and the Energy-Based Pointing Game (Energy) (Wang et al., 2020), which measures the fraction of total positive evidence contained within the ground-truth region. We evaluate image–text matching using area under the ROC curve (AUC) and balanced accuracy (BAcc), and retrieval using Recall@k (I@k/T@k).

4.2 Evidence Disentanglement Across Query Granularities

We first test whether ProtoLIP can mitigate the object-level evidence entanglement illustrated in Figures 1 and 4 on OOD benchmarks. We evaluate whether evidence for an object-level query is better localized to the annotated region of the queried object, while also measuring whether this improved localization is accompanied by stronger image–text matching.

As shown in Figure 5, 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.

(a) Spatial Evidence Disentanglement



(b) Matching and Retrieval

Dataset	Model	AUC	BAcc	I@1	T@1
VOC20	ItemizedCLIP	0.752	0.672	0.389	0.600
	ProtoLIP	**0.836**	**0.746**	**0.487**	**0.800**
ADE20K	ItemizedCLIP	0.697	0.649	0.257	0.180
	ProtoLIP	**0.714**	**0.657**	**0.367**	**0.227**
COCO	ItemizedCLIP	0.889	0.831	**0.124**	**0.123**
	ProtoLIP	**0.932**	**0.862**	0.106	0.112
Flickr30K	ItemizedCLIP	0.881	0.812	**0.148**	**0.165**
	ProtoLIP	**0.925**	**0.845**	0.111	0.152

Figure 5: **Spatial evidence disentanglement and matching across OOD benchmarks.** (a) ProtoLIP improves Pointing (P) and Energy (E); error bars show ± standard deviation over three seeds. (b) ProtoLIP matching and retrieval.

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/phrase queries, while preserving or modestly improving full-caption localization (Figure 5), which shows that stronger object-level separation does not compromise evidence quality for complete sentence descriptions. Direct prototype-mediated inference improves matching separability but moderately reduces caption retrieval. A simple 50/50 score-level fusion with the frozen backbone not only recovers this gap but improves retrieval over the frozen backbone, while retaining spatial evidence (Table 1). This separates two complementary uses of ProtoLIP: Direct ProtoLIP provides a fully decomposable prototype-mediated prediction, whereas Fused ProtoLIP combines the prototype-mediated and frozen-backbone scores to improve retrieval utility, at the cost of exact decomposition of the fused score. We use Direct ProtoLIP as the primary model because it preserves the direct correspondence between localized prototype evidence and the matching prediction.

Table 1: Fused ProtoLIP on Flickr30K Entities using 50/50 score-level fusion.

Metric pair	Backbone	ProtoLIP	Fusion
Caption P/E	0.8367/0.7030	**0.8437/0.8136**	**0.8437/0.8136**
AUC/BAcc	0.8814/0.8118	0.9249/**0.8487**	**0.9252**/0.8444
I@1/T@1	0.1483/0.1653	0.1112/0.1523	**0.1543/0.1693**

4.3 Generalization across vision–language backbones

We next test whether ProtoLIP generalizes across independently pretrained vision–language backbones. We apply the same prototype evidence layer to SigLIP-WebLI and FLAIR merged-30M, which differ from ItemizedCLIP in both pretraining data and objectives. For each model, the pretrained backbone remains fully frozen, and only a new prototype evidence layer is trained on Itemized-CC0.3M. We then evaluate the resulting models zero-shot on the same OOD benchmarks.

As shown in Table 2, ProtoLIP improves Pointing and Energy in all eight backbone–dataset combinations, indicating that its localization gains are not specific to the ItemizedCLIP backbone. Matching performance also improves for SigLIP-WebLI and remains broadly stable for FLAIR, with modest metric-specific variation. Under the same-training-data in-domain setting, ProtoLIP further achieves the best AUC, BAcc, and Recall@5/10 among the compared baselines (Appendix F).

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

Dataset	Backbone	Variant	Pointing	Energy	AUC	BAcc
COCO Object	SigLIP-WebLI	Frozen	0.292	0.245	0.791	0.749
		+ ProtoLIP Layer	**0.390**	**0.384**	**0.833**	**0.764**
	FLAIR merged-30M	Frozen	0.245	0.313	**0.906**	**0.827**
		+ ProtoLIP Layer	**0.558**	**0.483**	0.888	0.804
ADE20K	SigLIP-WebLI	Frozen	0.226	0.230	0.793	0.730
		+ ProtoLIP Layer	**0.412**	**0.363**	**0.802**	**0.737**
	FLAIR merged-30M	Frozen	0.214	0.278	0.764	0.691
		+ ProtoLIP Layer	**0.436**	**0.397**	**0.785**	**0.719**
VOC20	SigLIP-WebLI	Frozen	0.397	0.365	0.862	0.828
		+ ProtoLIP Layer	**0.684**	**0.625**	**0.912**	**0.834**
	FLAIR merged-30M	Frozen	0.272	0.473	**0.899**	**0.833**
		+ ProtoLIP Layer	**0.802**	**0.747**	0.890	0.814
Flickr30K Entities(phrase)	SigLIP-WebLI	Frozen	0.296	0.279	0.852	0.778
		+ ProtoLIP Layer	**0.503**	**0.451**	**0.870**	**0.791**
	FLAIR merged-30M	Frozen	0.272	0.354	**0.915**	**0.837**
		+ ProtoLIP Layer	**0.521**	**0.484**	0.898	0.814

To further isolate backbone effects from differences in pretraining data, we retrain ItemizedCLIP, Multi-positive SigLIP, FLAIR, and DreamLIP on the same Itemized-CC0.3M data and then freeze their backbones before training ProtoLIP. ProtoLIP improves localization for ItemizedCLIP, Multi-positive SigLIP, and FLAIR. In particular, across the same-data retrained Multi-positive SigLIP and FLAIR backbones, it achieves average relative gains of 66% in Pointing and 37% in Energy across three evaluation datasets. Localization does not improve for DreamLIP, suggesting that transfer also depends on the underlying patch–text representations. Full matched-data results and patch–text alignment diagnostics are provided in Appendix G.

4.4 Comparison with External Localization Methods

We further ask whether the object-level evidence produced by ProtoLIP is competitive with existing localization approaches despite training the prototype evidence layer without explicit spatial supervision. To test this, we compare against complementary external baselines (Figure 6): FLAIR (merged-30M) and ItemizedCLIP as fine-grained VLM baselines, NACLIP with a CLIP backbone pretrained on approximately 400M image–text pairs as a training-free localization baseline, and the spatially supervised MM Grounding DINO. All methods are re-evaluated on identical object queries and target regions. Because MM Grounding DINO outputs scored bounding boxes rather than dense evidence maps, we rasterize its predictions into dense spatial maps before applying the same Pointing and Energy metrics; the rasterization protocol is detailed in Appendix B.

Despite training the prototype evidence layer only on 0.3M image–text pairs, ProtoLIP outperforms both fine-grained VLM baselines on all eight comparisons and exceeds NACLIP on six of eight, with lower Pointing on ADE20K and slightly lower Pointing on Flickr30K. The weaker performance on ADE20K is consistent with its broader 150-class vocabulary and mismatch with the Itemized-CC0.3M training distribution. Against MM Grounding DINO, ProtoLIP is stronger on COCO and VOC20 under both metrics, while trailing on ADE20K and Flickr30K. The MM Grounding DINO Large-All checkpoint is trained with spatial supervision from detection and grounding datasets, including COCO and Flickr30K. Thus, ProtoLIP remains competitive with a grounding model benefiting from spatial supervision and training-data overlap with two of the evaluation benchmarks.

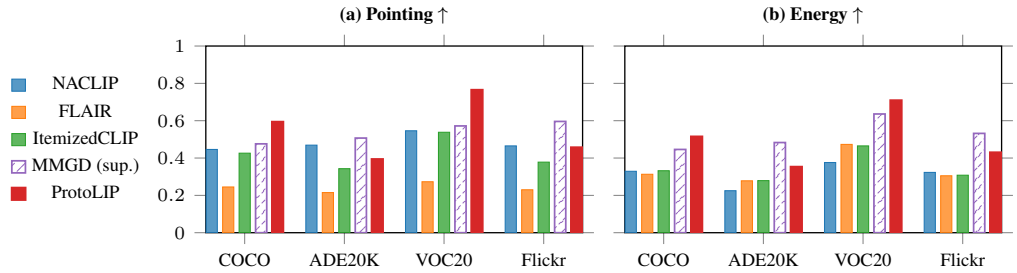


Figure 6: OOD object-level localization across fine-grained VLMs, training-free localization, spatially supervised grounding, and ProtoLIP.

4.5 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 improves object-level localization, while semantic family organization provides the largest additional gain. Combining the area and overlap constraints yields the strongest localization.

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

					Object-level localization		**Matching**	
Variant	Proto.	Family	Area	Overlap	Pointing	Energy	AUC	BAcc
Frozen ItemizedCLIP	–	–	–	–	0.426	0.332	0.889	0.831
Flat Prototype	✓	–	–	–	0.492	0.407	0.919	0.846
Family Prototype	✓	✓	–	–	0.586	0.461	**0.934**	0.863
+ Area	✓	✓	✓	–	0.588	0.494	0.933	0.862
+ Overlap	✓	✓	–	✓	0.582	0.459	**0.934**	**0.864**
Full ProtoLIP	✓	✓	✓	✓	**0.594**	**0.515**	0.932	0.862

ProtoLIP adds only 2.23M trainable parameters, equivalent to 1.48% of the frozen backbone. The resulting localization gains therefore require only a small parameter overhead. Full parameter counts are provided in Appendix E.

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.

AI USE STATEMENT

In this work, we used generative AI tools to assist with the offline construction of the semantic-family taxonomy, including extracting and grouping localizable visual concepts from training descriptions. We also used generative AI tools to assist with manuscript editing and improving readability. All AI-assisted outputs were reviewed and verified by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

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.

Rim Assouel, Pietro Astolfi, Florian Bordes, Michal Drozdzal, and Adriana Romero-Soriano. Object-centric binding in contrastive language-image pretraining. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Walid Bousselham, Felix Petersen, Vittorio Ferrari, and Hilde Kuehne. Grounding everything: Emerging localization properties in vision-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3828–3837, 2024.

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.

Martha Lewis, Nihal Nayak, Peilin Yu, Jack Merullo, Qinan Yu, Stephen Bach, and Ellie Pavlick. Does CLIP bind concepts? probing compositionality in large image models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pp. 1487–1500. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-eacl.101.

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.

Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 24–25, 2020.

Tom Nuno Wolf, Fabian Bongratz, Anne-Marie Rickmann, Sebastian Pölsterl, and Christian Wachinger. Keep the faith: Faithful explanations in convolutional neural networks for case-based reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5921–5929, 2024.

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, 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 Evaluation Protocol and Implementation Details

Prototype layer architecture. In our implementation, ϕ_θ is a gated two-layer residual MLP with hidden width $d/2$, where d is the frozen backbone's patch-embedding dimension. We use GELU activation, a residual scale of $\gamma = 0.25$, and final ℓ_2 normalization.

Prototype initialization. Before training, the $M = 512$ prototype vectors are independently sampled from a standard Gaussian distribution and ℓ_2-normalized,

$$p_k \sim \mathcal{N}(0, I), \qquad p_k \leftarrow \frac{p_k}{\|p_k\|_2}.$$

We then project each prototype onto the training patch-feature space. Using patch features from up to 8,192 randomly sampled training image–text feature rows, each prototype is replaced by the ℓ_2-normalized mean of its 128 most similar patch features. The resulting prototypes are subsequently optimized as learnable parameters throughout training. This initialization is independent of the semantic-family construction: the K23/K39 taxonomy determines only prototype-bank membership

and query routing. No downstream evaluation data, spatial annotations, K-means clustering, or text embeddings are used for prototype initialization.

Optimization and hyperparameters. 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.

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

For MM Grounding DINO, each query is evaluated independently using the single-query prompt followed by a period. We use box and text confidence thresholds of 0.25 and retain the bounding boxes and corresponding detection scores returned by the model's post-processing procedure. To obtain a dense map compatible with Pointing and Energy, each retained detection score is added to every cell of a 64×64 grid whose center lies inside the corresponding predicted box. Scores from overlapping boxes are summed, and the resulting heatmap is normalized by its total mass. Pointing is computed from the maximum-mass cell, while Energy is the normalized heatmap mass inside the ground-truth region.

This conversion preserves the detector's predicted regions and returned scores while providing a common spatial representation for evaluation. MM Grounding DINO is spatially supervised with detection and referring-expression annotations, whereas ProtoLIP receives no spatial annotations and keeps its VLM backbone frozen. We therefore treat this comparison as a reference against supervised grounding.

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.

B.1 Additional COCO-80 Evaluation

We additionally evaluate ProtoLIP on the complete set of 80 COCO object categories. This evaluation complements the caption-linked COCO object-query benchmark in the main paper by covering the full COCO-80 category set. As shown in Table 4, ProtoLIP improves both Pointing and Energy over the frozen ItemizedCLIP backbone.

C Natural Fallback and Semantic-Family Coverage

Our semantic-family taxonomy provides a coarse routing prior rather than a closed vocabulary. When a query matches one or more semantic families, only prototypes from the corresponding family banks are eligible for scoring. When no family is matched, ProtoLIP falls back to the full

Table 4: Object-level evidence localization on the complete COCO-80 category set.

Model	Pointing	Energy
ItemizedCLIP	0.313	0.250
ProtoLIP	**0.389**	**0.358**
Improvement	+0.075	+0.108

Table 5: Semantic-family routing coverage and matched/unmatched localization analysis. ΔP and ΔE denote the absolute improvements of ProtoLIP over the frozen ItemizedCLIP backbone in Pointing and Energy, respectively.

Dataset	Unique-query cov.	Sample cov.	Matched ΔP / ΔE	Unmatched ΔP / ΔE
ADE20K	50.00%	72.59%	+0.071 / +0.093	−0.007 / +0.016
Flickr30K Entities	63.76%	73.26%	+0.103 / +0.154	+0.013 / +0.033

prototype pool, allowing arbitrary queries to remain supported without requiring an explicit taxonomy assignment.

We further examine the coverage of semantic-family routing and the behavior of this fallback on two broader-vocabulary benchmarks: ADE20K, which contains a challenging long-tail object vocabulary, and Flickr30K Entities, which contains open-vocabulary phrase-level queries. We report both unique-query coverage, which measures the fraction of distinct queries matched by the taxonomy, and sample coverage, which measures the fraction of evaluation samples for which semantic-family routing is active. We additionally split the localization results according to whether each query is matched or unmatched and report the improvement of ProtoLIP over the frozen ItemizedCLIP backbone.

As shown in Table 5, the largest localization gains occur on matched queries, where semantic-family routing constrains prototype eligibility. For unmatched queries, the all-prototype fallback largely preserves localization performance: Pointing changes are close to zero, while Energy shows modest improvements on both benchmarks. These results indicate that semantic-family routing provides the primary localization benefit when applicable, while the fallback maintains support for queries outside the taxonomy without substantial degradation.

C.1 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±0.000	0.989±0.001	0.746±0.010	0.798±0.012
Matched, forced fallback	0.999±0.000	0.987±0.001	0.795±0.018	0.815±0.009
Naturally unmatched, fallback	0.998±0.000	0.986±0.002	0.782±0.010	0.777±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 10), ProtoLIP's strict I@1 is also lower 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 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 7.

Table 7: In-domain 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 8: 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. Differences between the two taxonomies are within one point on all reported metrics.

	K23		K39	
Dataset / Query	Pointing	Energy	Pointing	Energy
COCO object / phrase	.5868	.5187	.5939	.5143
Flickr30K phrase	.4553	.4293	.4571	.4301

Table 7 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 8 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.

E 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 9.

Table 9 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.

Table 9: 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%

F 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 10.

Table 10: 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**; second-best results are underlined.

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 SigLIP-0.3M	0.908	0.820	0.263	0.375	0.428	0.223	0.355	0.398
Multi-positive SigLIP-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	**0.958**	**0.880**	0.347	**0.485**	**0.553**	0.337	**0.463**	**0.526**

As shown in Table 10, 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.

G 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 11 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 information 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 11), suggesting that the missing discriminability is concentrated in their local patch representations.

Table 11: 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
COCO object	ItemizedCLIP-0.3M	Frozen	.426	.332	.889	.831
		+ ProtoLIP	**.594**	**.515**	**.932**	**.862**
	Multi-positive SigLIP-0.3M	Frozen	.185	.210	**.821**	**.750**
		+ ProtoLIP	**.456**	**.374**	.711	.659
	FLAIR-0.3M	Frozen	.243	.260	**.793**	**.729**
		+ ProtoLIP	**.329**	**.315**	**.793**	.721
	DreamLIP-0.3M	Frozen	**.306**	**.298**	**.791**	**.729**
		+ ProtoLIP	.262	.283	.731	.676
VOC20 full val	ItemizedCLIP-0.3M	Frozen	.538	.465	.752	.672
		+ ProtoLIP	**.765**	**.709**	**.836**	**.746**
	Multi-positive SigLIP-0.3M	Frozen	.288	.314	**.872**	**.798**
		+ ProtoLIP	**.654**	**.543**	.801	.734
	FLAIR-0.3M	Frozen	.399	.408	**.850**	**.778**
		+ ProtoLIP	**.522**	**.487**	.845	.758
	DreamLIP-0.3M	Frozen	**.515**	**.440**	**.811**	**.739**
		+ ProtoLIP	.438	.424	.765	.702
ADE20K full val	ItemizedCLIP-0.3M	Frozen	.342	.279	.697	.649
		+ ProtoLIP	**.394**	**.353**	**.714**	**.657**
	Multi-positive SigLIP-0.3M	Frozen	.212	.225	.685	.634
		+ ProtoLIP	**.303**	**.275**	**.685**	**.637**
	FLAIR-0.3M	Frozen	.248	.252	.644	.602
		+ ProtoLIP	**.283**	**.271**	**.695**	**.642**
	DreamLIP-0.3M	Frozen	**.257**	**.244**	.721	.673
		+ ProtoLIP	.253	.237	**.730**	**.676**

Table 12: Patch–text alignment on the shared 400-pair diagnostic set. These diagnostic values are separate from the formal benchmark results in Table 11.

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

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 13 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 12) tracks the matching columns across all three backbones, but does not separate DreamLIP-0.3M from Multi-positive SigLIP-0.3M, whose alignment is comparable (0.513 vs

Table 13: 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 11.

Backbone	ΔPointing	ΔEnergy	ΔAUC	ΔI@1
ItemizedCLIP-0.3M	$+0.149$	$+0.167$	$+0.048$	$+0.059$
Multi-positive SigLIP-0.3M	$+\mathbf{0.243}$	$+0.148$	-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

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.