THE HYPERSPHERICAL GEOMETRY OF CLIP LATENT SPACE: A SEMANTIC MIXTURE MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Contrastive Language–Image Pretraining (CLIP) representations form a semantic embedding space governed by cosine similarity, reflecting an intrinsic hyperspherical geometry. However, existing probabilistic interpretations typically rely on Gaussian assumptions, which fail to capture this directional and multimodal structure. We propose a principled density model for the CLIP latent space based on Mixtures of von Mises–Fisher (MovMF) distributions defined on the unit hypersphere. Using the Expectation–Maximization (EM) algorithm, we efficiently learn a probabilistic model in which each mixture component corresponds to a coherent semantic concept. This formulation yields a closed-form likelihood naturally aligned with hyperspherical geometry, enabling accurate and interpretable density estimation. Empirically, our model significantly improves long-tailed and out-of-distribution detection and provides a natural semantic decomposition, representing each embedding as a sparse probabilistic combination of interpretable concepts. These results suggest that CLIP latent space is more faithfully characterized as a hyperspherical semantic mixture rather than an isotropic Gaussian, establishing a simple and geometrically consistent probabilistic framework for modeling and understanding multimodal representations.

1 INTRODUCTION

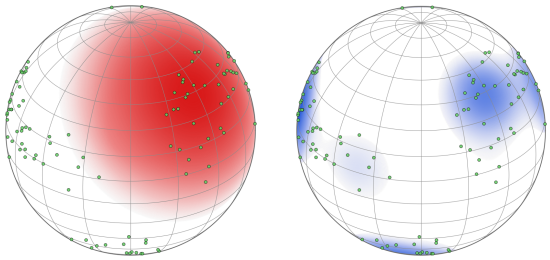
Contrastive Language–Image Pre-training (CLIP) (Radford et al., 2021) has become the foundation of modern vision–language models (VLMs). By aligning visual and textual representations in a shared embedding space, CLIP enables a broad range of downstream tasks, including zero-shot recognition (Radford et al., 2021) and text-to-image generation (Ramesh et al., 2022). Since semantic similarity is measured by cosine distance, the latent space is often idealized as a normalized hypersphere with isotropic geometry.

Despite this theoretical simplicity, the empirical geometry of CLIP representations deviates substantially from the isotropic ideal. Prior studies reveal pronounced structural irregularities, including the *Modality Gap* (Liang et al., 2022) and the so-called *Cone Effect* (Liang et al., 2022), where embeddings concentrate within restricted angular regions. Such anisotropy distorts angular similarity, biases representations toward dominant concepts such as large objects or spurious background features, and degrades robustness across long-tail categories (Abbasi et al., 2025; Wang et al., 2024; Shao et al., 2024).

To enable probabilistic reasoning in this space, recent work has proposed transforming CLIP embeddings into a Gaussianized coordinate system. In particular, Whitened CLIP (W-CLIP) (Betser et al., 2025) applies an invertible whitening transformation and models the resulting features with a single isotropic Gaussian, yielding a tractable likelihood based on Euclidean magnitude. While effective, this unimodal assumption is fundamentally misaligned with the directional and multimodal nature of contrastive embeddings. Natural semantic concepts form multiple structured regions on the hypersphere. Modeling the space with a single Gaussian inevitably conflates *semantic rarity* with *distributional abnormality*, assigning low likelihood to rare but valid concepts and treating them similarly to true out-of-distribution (OOD) samples. Moreover, a global scalar likelihood provides limited interpretability of the underlying semantic structure. It can only indicate how typical a sample is with respect to the entire distribution, but does not reveal which semantic modes contribute to that score or how the representation decomposes across distinct concepts.

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.

In this work, we propose **MovMF-CLIP**, a geometrically consistent probabilistic framework that models CLIP latent space as a *hyperspherical semantic mixture*. We first perform covariance-based geometric calibration via whitening to remove global anisotropy while preserving semantic relationships. Then we model the normalized representations using a Mixture of von Mises–Fisher (MovMF) distributions, estimated efficiently via the Expectation–Maximization (EM) algorithm. Unlike Gaussian approximations that rely on magnitude, MovMF components are defined intrinsically on the unit hypersphere and capture the multimodal directional structures (see Fig. 1). Each mixture component corresponds to a coherent semantic prototype, yielding a closed-form likelihood that respects the geometry of contrastive embeddings. This probabilistic formulation supports two complementary capabilities. First, it provides a principled density estimate for likelihood-based evaluation, decoupling semantic rarity from low probability and improving robustness in long-tailed and OOD settings. Second, the mixture structure enables intrinsic interpretability: embeddings can be decomposed into sparse probabilistic combinations of semantic prototypes.



(a) Global Gaussian in $\mathbb{R}^d$ (b) MovMF-CLIP on $\mathbb{S}^{d-1}$

Figure 1: **Density modeling of CLIP latent space (visualized via dimensionality reduction on real data). (a)** Gaussian-based approaches such as W-CLIP model the latent space with a single global distribution, which can assign low likelihood to valid long-tail concepts due to their distance from the global mean. **(b)** MovMF-CLIP models the space as a hyperspherical semantic mixture, capturing the multimodal structures and providing density estimates aligned with semantic clusters.

Empirically, we validate MovMF-CLIP across long-tailed fairness analysis, OOD detection, semantic decomposition, and semantic stability under iterative generative drift. For OOD detection, MovMF-CLIP reduces FPR95 from 67.76% to **48.00%** on MS-COCO, and from 75.05% to **33.48%** solely on tail concepts. For semantic decomposition, MovMF-CLIP achieves the highest Semantic Relevance of **0.673** while achieving a **13 $\times$ speedup** in inference time compared to the second-best method. Across tasks, modeling the latent space as a hyperspherical semantic mixture consistently improves robustness, interpretability, and stability.

Our main contributions are summarized as follows:

- We offer a geometric re-interpretation of CLIP latent space, arguing that unimodal Gaussian modeling is inherently inconsistent with its directional hypersphere geometry and multimodal semantic structure.
- We propose **MovMF-CLIP**, an elegant and principled density modeling framework that integrates covariance-based geometric calibration with von Mises–Fisher mixture estimation on the unit sphere.
- We show that this unified geometric–probabilistic formulation simultaneously enables (i) calibrated likelihood estimation robust to long-tail and OOD settings, (ii) intrinsic and efficient probabilistic semantic decomposition without auxiliary decoders, and (iii) a lightweight geometric prior for stabilizing multimodal representations under iterative generative drift.

2 MovMF-CLIP: Hyperspherical Density Modeling of CLIP

We model CLIP latent space as a geometrically calibrated hyperspherical density framework. As shown in Fig. 2, our framework proceeds in three stages: (i) metric normalization via whitening to remove global anisotropy (Sec. 2.1); (ii) multimodal directional density modeling using MovMF with EM estimation (Secs. 2.2–2.3); and (iii) probabilistic inference enabling likelihood-based evaluation and semantic attribution (Sec. 2.4). Together, these components form a unified geometric–probabilistic framework for modeling CLIP representations.

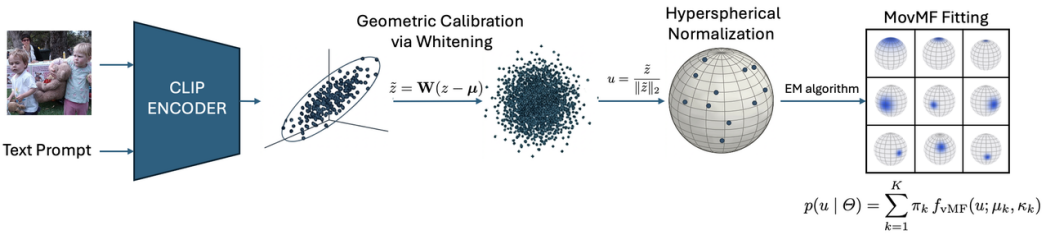


Figure 2: **Overview of the MovMF-CLIP Framework.** We first extract raw embeddings using the CLIP encoder, which exhibit severe anisotropy. To address this, we apply geometric calibration via whitening ($\tilde{z} = \mathbf{W}(z - \boldsymbol{\mu})$) and normalize the features onto a unit hypersphere ($u = \tilde{z}/\|\tilde{z}\|_2$). Finally, we fit an MovMF distribution on the hypersphere using the EM algorithm, yielding a principled multimodal density model.

2.1 Geometric Calibration of CLIP Latent Space

We consider a dataset $\mathcal{D} = \{(x_i^{\text{img}}, x_i^{\text{text}})\}_{i=1}^N$ consisting of image–text pairs. A pre-trained CLIP model provides two encoders, $f_{\text{img}}(\cdot)$ and $f_{\text{text}}(\cdot)$, which map each modality into a shared d-dimensional embedding space. We denote by $z \in \mathbb{R}^d$ the latent representation produced by either encoder. Our geometric calibration and density modeling are applied independently to each modality within this shared space. Although CLIP is trained with cosine similarity and is interpreted as operating on a hypersphere, the raw latent distribution is empirically far from geometrically isotropic. In particular, the embeddings exhibit strong second-order anisotropy and dominant covariance directions, reflecting dataset bias and representation artifacts. As a result, angular distances in the raw space are distorted by global variance structure, potentially obscuring the true semantic organization of the data.

Geometric Calibration via Whitening. To remove global anisotropy while preserving semantic structure, we first apply a whitening transformation in $\mathbb{R}^d$. Let $\boldsymbol{\mu}$ denote the empirical mean and $\boldsymbol{\Sigma}$ the covariance matrix computed over the reference dataset (*e.g.* $5k$ validation samples of MS-COCO). We construct a whitening operator $\mathbf{W} = \boldsymbol{\Sigma}^{-1/2}$ and define the calibrated embedding

$$\tilde{z} = \mathbf{W}(z - \boldsymbol{\mu}). \tag{1}$$

This transformation standardizes second-order statistics so that $\text{Cov}(\tilde{z}) \approx \mathbf{I}$. Unlike likelihood-based approaches such as W-CLIP (Betser et al., 2025), we do not assume that the whitened representations follow a Gaussian distribution. Instead, whitening serves purely as a metric normalization step: it removes nuisance covariance structure and **induces an intrinsic Mahalanobis geometry** in the original latent space (see Appendix B). Consequently, directional comparisons are no longer dominated by high-variance axes unrelated to semantic content.

Hyperspherical Geometry. CLIP representations are inherently interpreted through cosine similarity, meaning that their semantics are encoded directionally on the unit hypersphere. Accordingly, after geometric calibration, we operate on normalized embeddings

$$u = \frac{\tilde{z}}{\|\tilde{z}\|_2}, \qquad u \in \mathbb{S}^{d-1}, \tag{2}$$

which respects the intrinsic hyperspherical geometry imposed by contrastive training. This step does not introduce an additional modeling assumption; rather, it aligns our probabilistic framework with the geometry already underlying CLIP. By separating global covariance normalization in $\mathbb{R}^d$ from directional modeling on $\mathbb{S}^{d-1}$, we obtain a geometrically consistent representation that isolates semantic structure from nuisance anisotropy. This calibrated hyperspherical representation forms the foundation for our mixture-based density modeling.

2.2 Modeling Latent Density via MovMF

Following the geometric calibration defined in Sec. 2.1, we operate on whitened and normalized embeddings $\mathcal{U} = \{u_1, \ldots, u_N\} \subset \mathbb{S}^{d-1}$. Since CLIP encodes semantics directionally and compares representations via cosine similarity, the natural probability model on $\mathbb{S}^{d-1}$ is the von Mises–Fisher (vMF) distribution.

Single vMF Distribution. For a unit vector $u \in \mathbb{S}^{d-1}$, the density of a vMF distribution with mean direction $\mu \in \mathbb{S}^{d-1}$ and concentration $\kappa \geq 0$ is

$$f_{\text{vMF}}(u; \mu, \kappa) = C_d(\kappa) \exp(\kappa \mu^\top u), \tag{3}$$

where the normalization constant is

$$C_d(\kappa) = \frac{\kappa^{d/2-1}}{(2\pi)^{d/2} I_{d/2-1}(\kappa)}. \tag{4}$$

$I_\nu(\cdot)$ denotes the modified Bessel function of the first kind. The mean direction μ identifies the center of a semantic cluster on the hypersphere, while κ controls its concentration. Larger κ corresponds to tighter clustering around μ, whereas $\kappa = 0$ yields the uniform distribution on $\mathbb{S}^{d-1}$.

Mixture Model. Natural image and text distributions are inherently multimodal, containing distinct semantic regions (*e.g.,* animals, vehicles, landscapes). To capture this structure, we model the latent density as a mixture of K independent vMF components:

$$p(u \mid \Theta) = \sum_{k=1}^{K} \pi_k \, f_{\text{vMF}}(u; \mu_k, \kappa_k), \tag{5}$$

where $\Theta = \{(\pi_k, \mu_k, \kappa_k)\}_{k=1}^K$ are the set parameters that need to be learned, and the mixing coefficients satisfy $\pi_k \geq 0$ and $\sum_{k=1}^K \pi_k = 1$. Each component corresponds to a coherent semantic mode on the hypersphere, and the resulting density naturally reflects the multimodal organization of the CLIP latent space.

2.3 Parameter Estimation via EM

We estimate the parameters Θ by maximizing the log-likelihood

$$\mathcal{L}(\Theta) = \sum_{i=1}^{N} \log p(u_i \mid \Theta),$$

using the Expectation–Maximization (EM) algorithm for vMF distributions (Banerjee et al., 2005). The EM procedure alternates between computing the vMF posterior assignments (E-step) and updating the model parameters (M-step).

E-Step. Given current parameters, we compute the normalized responsibility

$$\gamma_{ik} = p(k \mid u_i, \Theta) = \frac{\pi_k f_{\text{vMF}}(u_i; \mu_k, \kappa_k)}{\sum_{j=1}^K \pi_j f_{\text{vMF}}(u_i; \mu_j, \kappa_j)}.$$

For numerical stability in high dimensions, we evaluate this expression in the log-domain: $\log \gamma_{ik} \propto \log \pi_k + \kappa_k \mu_k^\top u_i + \log C_d(\kappa_k)$. These responsibilities quantify the soft assignment of each embedding to semantic components.

M-Step. Let $N_k = \sum_i \gamma_{ik}$, $r_k = \sum_i \gamma_{ik} u_i$, and $\bar{R}_k = \|r_k\|_2 / N_k$. The final parameter updates are

$$\pi_k^{\text{new}} = \frac{N_k}{N}, \qquad \mu_k^{\text{new}} = \frac{r_k}{\|r_k\|_2}, \qquad \kappa_k^{\text{new}} \approx \frac{\bar{R}_k (d - \bar{R}_k^2)}{1 - \bar{R}_k^2}. \tag{6}$$

The full derivation and concentration-parameter approximation are provided in Appendix C.2.

2.4 Inference: Density Evaluation and Semantic Decomposition

After estimating the movMF parameters $\Theta = \{(\pi_k, \mu_k, \kappa_k)\}_{k=1}^K$ via EM, the learned model defines an explicit probability density on the hypersphere. Inference therefore reduces to evaluating this density and its associated posterior structure. Importantly, since CLIP maps both images and text into a shared embedding space, the same probabilistic framework applies to either modality.

Likelihood as a Geometric Density Score. For a query sample $u \in \mathbb{S}^{d-1}$, the log-likelihood score $S(u)$ under the learned mixture is

$$S(u) = \log P(u|\Theta) = \log \left(\sum_{k=1}^{K} \pi_k C_d(\kappa_k) \exp(\kappa_k \mu_k^\top u) \right). \tag{7}$$

Table 1: **Long-tail likelihood fairness.** AUROC for separating head and tail categories using likelihood scores (ideal $\rightarrow 0.5$). Lower values indicate reduced bias against rare concepts. MovMF-CLIP substantially improves fairness compared to W-CLIP.

Method	W-CLIP (Betser et al., 2025)	MovMF-CLIP (Ours)
AUROC	0.7826	**0.5819**

This quantity measures how well the sample aligns with the learned semantic modes of the latent space. Unlike Gaussian-based scoring, the likelihood here reflects proximity to multiple semantic directions rather than distance from a single global center. Since the model captures multimodal density structure, low likelihood indicates that a sample lies outside established semantic regions rather than merely being far from an average embedding. *This enables principled evaluation of representation quality, distribution shift in image and text domains, and long-tail behavior within a unified probabilistic framework.*

Posterior Structure and Semantic Attribution. Beyond scalar likelihood scoring, the mixture structure provides a richer, interpretable decomposition. For a given embedding u, we can compute the posterior responsibility

$$\gamma_k(u) = P(k \mid u) = \frac{\pi_k f_{\text{vMF}}(u; \mu_k, \kappa_k)}{\sum_{j=1}^{K} \pi_j f_{\text{vMF}}(u; \mu_j, \kappa_j)}. \tag{8}$$

This posterior distribution quantifies the degree to which the sample is explained by each semantic component. Each mixture component corresponds to a coherent directional mode on the hypersphere. By associating mixture centers with representative samples or textual descriptions from the training data, these components can be interpreted as semantic prototypes. The posterior vector $\{\gamma_k(u)\}_{k=1}^{K}$ therefore provides a soft semantic decomposition of the embedding:

$$u \;\approx\; \sum_{k=1}^{K} \gamma_k(u)\, \mu_k,$$

where we can interpret the representation as a weighted combination of latent semantic modes. This interpretation arises directly from the probabilistic structure of the model and does not require auxiliary decoders or external supervision. Our MovMF-CLIP framework yields both ***accurate density estimation*** and ***intrinsic semantic attribution*** within a unified hyperspherical geometry.

3 Experiments

3.1 Experimental Setup

We evaluate MovMF-CLIP on MS-COCO 2017 for long-tailed likelihood fairness, semantic decomposition, and iterative generative stability, using a cleaned OpenImages subset for OOD detection. All semantic decomposition methods share an OpenCLIP ViT-B/32 backbone, while the other experiments use CLIP ViT-L/14 (full datasets, protocols, and implementation details are provided in Appendix D.1).

3.2 Long-tailed and OOD Detection

A key limitation of unimodal latent density modeling (*e.g.*, W-CLIP) is the conflation of ***semantic rarity*** with ***distributional abnormality***. Under a single isotropic Gaussian assumption, valid but infrequent long-tail concepts tend to receive low likelihood simply due to their distance from the global mean, making them indistinguishable from true OOD samples. In contrast, MovMF-CLIP models the latent space as a hyperspherical mixture, allowing rare but semantically coherent regions to be represented by dedicated mixture components. We evaluate this property through long-tail likelihood fairness and OOD detection.

Long-Tailed Likelihood Fairness (texts). We analyze captions of MS-COCO, partitioning into *head* and *tail* groups based on concept frequency statistics. To quantify bias against rare semantics,

Table 2: **Out-of-distribution detection.** FPR95 (↓) and AUROC (↑) on MS-COCO (full set and tail-only subset) versus OpenImages OOD samples. MovMF-CLIP improves both detection performance and maintains robustness on long-tail subsets.

Method	ID: Full COCO		ID: Tail Subset Only	
	FPR95 (↓)	AUROC (↑)	FPR95 (↓)	AUROC (↑)
MCM (Ming et al., 2022)	84.50%	0.8204	62.90%	0.8754
EOE (Cao et al., 2024a)	83.91%	0.7926	75.54%	0.8364
NegLabel (Jiang et al., 2024)	56.48%	0.8564	38.77%	0.9135
W-CLIP (Betser et al., 2025)	67.76%	0.8600	75.05%	0.7842
MovMF-CLIP (Ours)	**48.00%**	**0.8952**	**33.48%**	**0.9157**

Table 3: **OOD detection across VLMs.** AUROC (↑) and FPR95 (↓) on BLIP-2 and CoCa using OpenImages as the OOD dataset. MovMF-CLIP consistently improves detection performance and significantly reduces false rejection of tail samples.

Model	Method	ID: Full COCO		ID: Tail Subset Only	
		AUROC (↑)	FPR95 (↓)	AUROC (↑)	FPR95 (↓)
BLIP-2	W-CLIP	0.9352	32.02%	0.8473	46.65%
	MovMF-CLIP	**0.9522**	**25.32%**	**0.9197**	**45.57%**
CoCa	W-CLIP	0.8195	86.23%	0.6588	92.12%
	MovMF-CLIP	**0.8586**	**66.14%**	**0.8068**	**80.02%**

we measure how well likelihood scores separate head from tail samples using AUROC. A value closer to 0.5 indicates fairness, meaning likelihood does not systematically penalize rare concepts.

As shown in Table 1, W-CLIP exhibits strong discrimination against tail samples (AUROC > 0.78), confirming that a global Gaussian prior treats long-tail concepts as outliers. In contrast, MovMF-CLIP substantially reduces this bias (AUROC closer to 0.5), indicating that multimodal hyperspherical modeling assigns calibrated likelihoods across semantic frequency groups.

Out-of-Distribution Detection (images). Table 2 shows that MovMF-CLIP consistently outperforms all competing approaches. On the full ID set, FPR95 is reduced from 67.76% to 48.00% compared with W-CLIP. When restricting ID samples to the pure tail subset, the degradation observed in W-CLIP becomes pronounced (FPR95 increases from 67.76% to 75.05%), while MovMF-CLIP remains stable and improves FPR95 to 33.48%. This demonstrates that separating semantic multimodality from global density structure improves both fairness to long-tail concepts and robustness to distributional shifts.

Generalization across VLMs. We further evaluate BLIP-2 (Li et al., 2023) and CoCa (Yu et al., 2022) using the same MS-COCO/OpenImages protocol. Table 3 shows that MovMF-CLIP improves both AUROC and FPR95 over W-CLIP on the full and tail-only ID subsets for both backbones. On CoCa, tail-only FPR95 decreases from 92.12% to 80.02%. These results suggest that the benefits of hyperspherical mixture modeling extend beyond CLIP. Additional long-tailed fairness results are provided in Appendix D.3.

3.3 Interpretable Semantic Decomposition

We evaluate whether the learned mixture structure yields semantically meaningful and human-aligned decompositions of visual representations.

Evaluation Metric. Following SPLICE, we adopt the **Semantic Relevance** metric to quantify alignment between model-derived concepts and ground-truth captions. Higher values indicate stronger semantic agreement. Detailed formulations are kindly referred to Appendix D.2.

Table 4: **Semantic Relevance and inference efficiency.** Comparison with concept-based decomposition methods using the same OpenCLIP ViT-B/32 backbone. MovMF-CLIP achieves higher semantic alignment with ground-truth captions while reducing per-image inference time through its closed-form posterior computation.

Method	Semantic Relevance (↑)	Inference Time (ms) (↓)
MSAE (Zaigrajew et al., 2025)	0.530	14.6
TextSpan (Gandelsman et al., 2024)	0.538	29.4
CLIP + Sparse Decomposition (Bhalla et al., 2024)	0.585	112.2
Negative Concept Weights (Bhalla et al., 2024)	0.635	102.8
SPLICE (Bhalla et al., 2024)	0.655	132.7
MovMF-CLIP (Ours)	**0.673**	**9.8**

Experimental Setup. For all projection-based baselines, we report results using their officially recommended optimal sparsity settings. As for our MovMF-CLIP, we use $K = 500$ mixture components and retain the top $N = 10$ keywords per component. Ablation studies over K and N are provided in Appendix E.

Results and Efficiency. Table 4 shows that MovMF-CLIP achieves the highest Semantic Relevance score, outperforming projection-based baselines. Beyond accuracy, MovMF-CLIP offers a substantial computational advantage. Compared to the second best method, MovMF-CLIP significantly reduces per-image inference time from 132.7 ms to **9.8 ms**, achieving over a **13× speedup**. This is because projection-based methods require solving a sparse regression problem over large concept dictionaries for each image, typically involving iterative CPU-based solvers. In contrast, semantic decomposition in MovMF-CLIP reduces to evaluating posterior responsibilities $\gamma_k(u) = P(k \mid u)$ which requires only a matrix multiplication with the learned cluster centers.

Qualitative Analysis. To provide an intuitive illustration of the learned semantic structure, we visualize and compare the decompositions produced by MovMF-CLIP and SPLICE on several complex real-world scenes (see fig. 4). Projection-based methods often yield broad or loosely related concepts when representing images as sparse combinations over large dictionaries. For example, in the fruit bowl scene, SPLICE primarily identifies generic terms such as "fruit" or "vitamin", without resolving more specific semantics. Our MovMF-CLIP decomposes each image into a sparse probabilistic combination of hyperspherical semantic prototypes. The resulting components tend to align with more precise and fine-grained visual concepts in the scene. This behavior is consistent with the multimodal structure induced by hyperspherical mixture modeling. Fig. 3 further visualizes the concept-specific heatmaps produced by our model, where the highlighted regions closely correspond to the underlying objects associated with each semantic component.

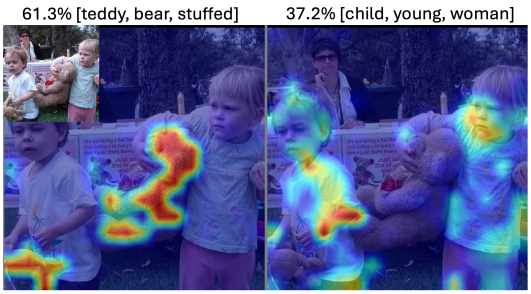


Figure 3: **Concept localization.** MovMF-CLIP represents each image as a sparse mixture of semantic prototypes. Concept-specific heatmaps highlight image regions associated with each semantic component.

Semantic Direction Intervention. Beyond decomposing embeddings into interpretable concepts, the learned vMF components can also be used as explicit semantic directions. Given a target component center μ_k, we manipulate the whitened embedding by subtracting or amplifying its projection along μ_k, while keeping the overall embedding norm fixed before mapping it back for UnCLIP decoding. As shown in Fig. 5, removing the cat-associated direction suppresses the cat semantics, whereas amplifying the same direction progressively strengthens them. This provides qualitative evidence that the mixture components capture controllable semantic factors rather than merely improving aggregate likelihood scores.

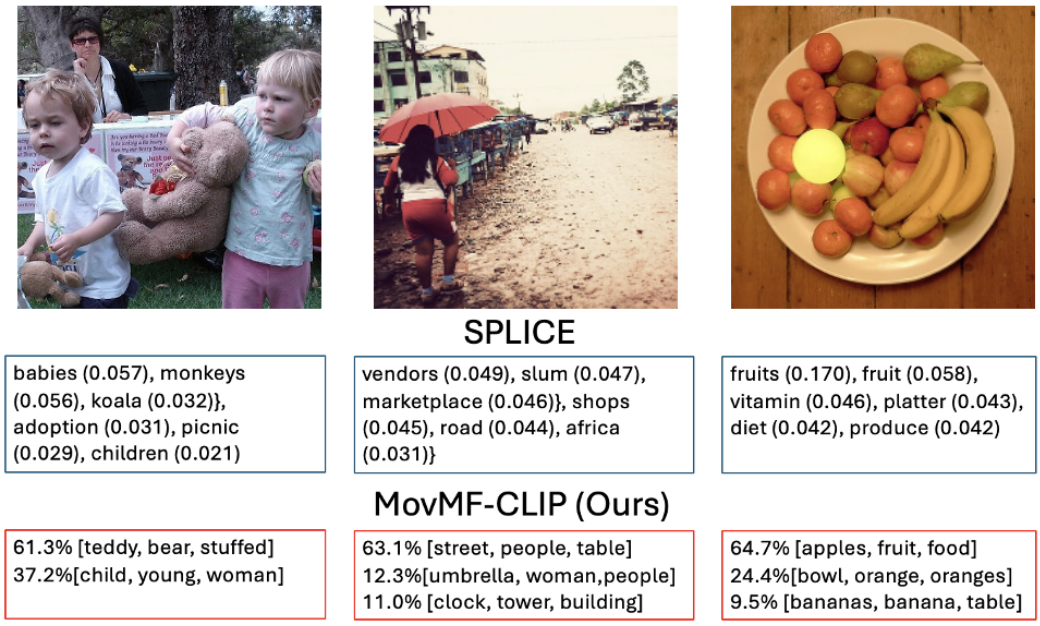


Figure 4: **Qualitative comparison of semantic decomposition.** Concept extraction results for SPLICE and MovMF-CLIP on representative scenes. Projection-based methods may emphasize generic or context-dependent terms, whereas our MovMF-CLIP produces a sparse probabilistic combination of semantically coherent cluster prototypes aligned with object-centric content.

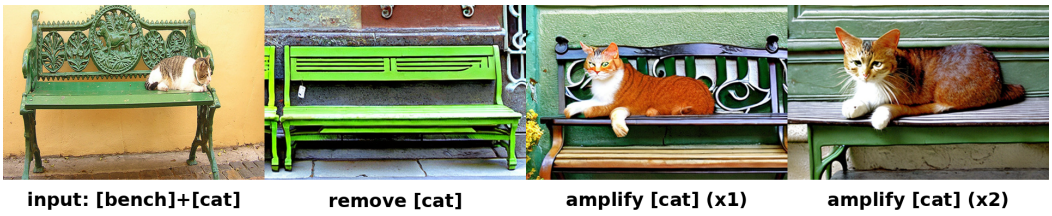


Figure 5: **Directional semantic intervention in CLIP latent space.** We identify a vMF component associated with the cat concept and directly manipulate its direction in the whitened CLIP space before UnCLIP decoding. Removing the component suppresses the cat from the generated image, while amplifying the same direction strengthens the corresponding visual semantics. This illustrates that MovMF-CLIP learns semantically meaningful and controllable directions.

3.4 Semantic Stability Under Iterative Generative Drift

We simulate a process of iterative generative drift by constructing the encoding–decoding pipeline in Fig. 6, which involves conducting CLIP encoding and UnCLIP decoding repeatedly. When an image is projected into the latent space and decoded back to the pixel domain iteratively, small perturbations accumulate, gradually shifting the embedding toward low-density regions between valid semantic modes. This *semantic drift* manifests as progressive degradation in both visual fidelity and semantic consistency. We evaluate whether the hyperspherical mixture structure learned by MovMF-CLIP can serve as a geometric prior to stabilize CLIP latent representations during such iterative processes.

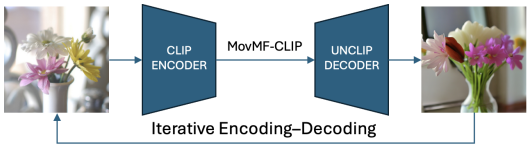


Figure 6: **Pipeline for iterative generative drift analysis.** We insert MovMF-CLIP into the encoding–decoding loop to stabilize the underlying semantics.

Semantic Stabilization via Hyperspherical Mixture Projection. Given a whitened embedding $\tilde{z}$, we first normalize it onto the hypersphere and compute its posterior responsibilities $\{\gamma_k\}$. Components with negligible posterior mass are discarded, and the embedding is reconstructed as a weighted combination of the retained mixture centers. The reconstructed direction is then rescaled to match the original whitened norm and mapped back to the original space via inverse whitening. This operation projects drifting embeddings toward high-density semantic regions while preserving their magnitudes. The resulting embedding is then fed into the next UnCLIP decoding step.

Experimental Setup. We mainly compare MovMF-CLIP against W-CLIP (Betser et al., 2025), which regularizes embeddings by projecting them onto a fixed-radius hypersphere in whitened space without directional correction. To quantify semantic stability, we measure two metrics between the original image and the image obtained after six iterations: LPIPS (AlexNet) (Zhang et al., 2018) for perceptual distortion (↓ better) and CLIP cosine similarity for semantic retention (↑ better).

Quantitative Results. As shown in Table 5, MovMF-CLIP consistently improves both perceptual quality and semantic stability. While W-CLIP constrains embedding magnitude, it does not regulate directional drift on the hypersphere, leading to gradual semantic degradation. In contrast, our mixture-based method explicitly anchors embeddings to high-density semantic modes, resulting in improved cosine similarity and reduced perceptual distortion.

Table 5: **Iterative CLIP–UnCLIP generation.** Average LPIPS (↓) and CLIP cosine similarity (↑) over $1,000$ MS-COCO images after $k = 6$ encode–decode iterations. MovMF-CLIP improves both perceptual consistency and semantic retention.

Method	**LPIPS** (↓)	**CLIP Cosine** (↑)
Vanilla CLIP-UnCLIP	0.7733	0.5911
W-CLIP (Betser et al., 2025)	0.7678	0.6171
MovMF-CLIP (Ours)	**0.7357**	**0.6710**

Qualitative Analysis. Fig. 7 visualizes representative iterative trajectories. For W-CLIP, embeddings progressively drift away from coherent semantic regions, leading to noticeable degradation. Instead, our MovMF-CLIP maintains alignment with semantically meaningful prototypes throughout the iterations, yielding more stable visual and semantic evolution.

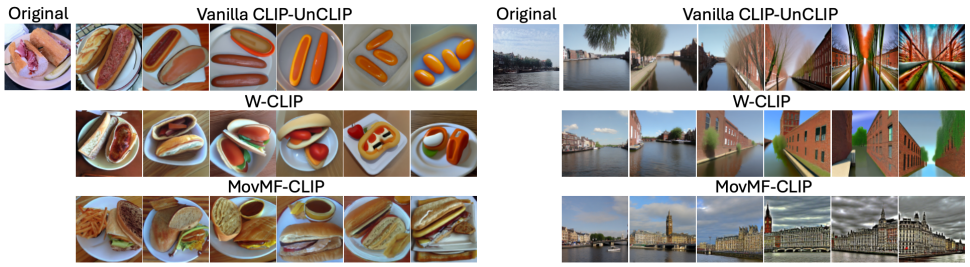


Figure 7: **Qualitative comparison under iterative generative drift.** For each example, the original image is followed by its evolution over six encoding–decoding iterations. The three rows demonstrate the trajectories using vanilla CLIP-UnCLIP (top), W-CLIP (middle), and MovMF-CLIP (bottom). Our hyperspherical mixture aligns well with coherent semantic prototypes, resulting in more stable semantic trajectories.

4 Conclusion

We revisited CLIP latent space through a geometric lens and argued that it is more faithfully modeled as a hyperspherical semantic mixture rather than a unimodal Gaussian. By separating global covariance anisotropy from directional multimodality, MovMF-CLIP provides a simple and principled density model aligned with the intrinsic structure of contrastive embeddings. This unified geometric–probabilistic framework enables calibrated likelihood estimation, intrinsic semantic decomposition, and improved semantic stability under iterative generative drift. Our findings underscore the central role of representation geometry in probabilistic modeling of foundation models and point toward structured, geometry-aware probabilistic frameworks for future multimodal learning.

AI USE STATEMENT

Generative AI tools were used to assist with code generation and to improve the clarity, grammar, and readability of the manuscript. All AI-assisted code was reviewed and tested by the authors, and all AI-assisted edits to the manuscript were carefully verified. The authors take full responsibility for the final content of this work, including all text, claims, results, and artifacts produced with the assistance of generative AI.

REFERENCES

Reza Abbasi, Ali Nazari, Aminreza Sefid, Mohammadali Banayeeanzade, Mohammad Hossein Rohban, and Mahdieh Soleymani Baghshah. Clip under the microscope: A fine-grained analysis of multi-object representation. *CVPR*, 2025. URL https://arxiv.org/abs/2502.19842.

Arindam Banerjee, Inderjit S. Dhillon, Joydeep Ghosh, and Suvrit Sra. Clustering on the unit hypersphere using von mises-fisher distributions. *JMLR*, 2005. URL http://jmlr.org/papers/v6/banerjee05a.html.

Roy Betser, Meir Yossef Levi, and Guy Gilboa. Whitened clip as a likelihood surrogate of images and captions. *ICML*, 2025. URL https://arxiv.org/abs/2505.06934.

Usha Bhalla, Alex Oesterling, Suraj Srinivas, Flavio P. Calmon, and Himabindu Lakkaraju. Interpreting clip with sparse linear concept embeddings (splice). *Neurips*, 2024. URL https://arxiv.org/abs/2402.10376.

Valentin De Bortoli, Emile Mathieu, Michael Hutchinson, James Thornton, Yee Whye Teh, and Arnaud Doucet. Riemannian score-based generative modelling. *NeurIPS*, 2022. URL https://arxiv.org/abs/2202.02763.

Chentao Cao, Zhun Zhong, Zhanke Zhou, Yang Liu, Tongliang Liu, and Bo Han. Envisioning outlier exposure by large language models for out-of-distribution detection. *ICML*, 2024a. URL https://arxiv.org/abs/2406.00806.

Yunkang Cao, Jiangning Zhang, Luca Frittoli, Yuqi Cheng, Weiming Shen, and Giacomo Boracchi. Adaclip: Adapting clip with hybrid learnable prompts for zero-shot anomaly detection. *ECCV*, 2024b. URL http://dx.doi.org/10.1007/978-3-031-72761-0_4.

Jiwan Chung and Seon Joo Kim. Global geometry is not enough for vision representations, 2026. URL https://arxiv.org/abs/2602.03282.

Grégoire Dhimoïla, Thomas Fel, Victor Boutin, and Agustin Picard. Cross-modal redundancy and the geometry of vision-language embeddings, 2026. URL https://arxiv.org/abs/2602.06218.

Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis. *ICLR*, 2022. URL https://arxiv.org/abs/2202.01197.

Yossi Gandelsman, Alexei A. Efros, and Jacob Steinhardt. Interpreting clip's image representation via text-based decomposition. *ICLR*, 2024. URL https://arxiv.org/abs/2310.05916.

Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan A. Rossi, Vishwa Vinay, and Aditya Grover. Cyclip: Cyclic contrastive language-image pretraining. *Neurips*, 2022. URL https://arxiv.org/abs/2205.14459.

Tao Gong, Qi Chu, Bin Liu, Zhou Wei, and Nenghai Yu. Fe-clip: Frequency enhanced clip model for zero-shot anomaly detection and segmentation. *ICCV*, 2025. URL https://iccv.thecvf.com/virtual/2025/poster/2571.

Jianfang He, Min Cao, Silong Peng, and Qiong Xie. Rareclip: Rarity-aware online zero-shot industrial anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24478–24487, October 2025.

Nhat Hoang-Xuan, Xiyuan Wei, Wanli Xing, Tianbao Yang, and My T. Thai. Advancing interpretability of CLIP representations with concept surrogate model. *Neurips*, 2025. URL https://openreview.net/forum?id=KxoPiQ03BT.

Chin-Wei Huang, Milad Aghajohari, Avishek Joey Bose, Prakash Panangaden, and Aaron Courville. Riemannian diffusion models. *NeurIPS*, 2022. URL https://arxiv.org/abs/2208.07949.

Linlan Huang, Xusheng Cao, Haori Lu, Yifan Meng, Fei Yang, and Xialei Liu. Mind the gap: Preserving and compensating for the modality gap in clip-based continual learning. *ICCV*, 2025. URL https://arxiv.org/abs/2507.09118.

Xue Jiang, Feng Liu, Zhen Fang, Hong Chen, Tongliang Liu, Feng Zheng, and Bo Han. Negative label guided ood detection with pretrained vision-language models. *ICLR*, 2024. URL https://arxiv.org/abs/2403.20078.

Raphi Kang, Yue Song, Georgia Gkioxari, and Pietro Perona. Is clip ideal? no. can we fix it? yes! *ICCV*, 2025. URL https://arxiv.org/abs/2503.08723.

Elias Kempf, Simon Schrodi, Max Argus, and Thomas Brox. When and how does clip enable domain and compositional generalization? *ICML*, 2025. URL https://arxiv.org/abs/2502.09507.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, March 2020. ISSN 1573-1405. doi: 10.1007/s11263-020-01316-z. URL http://dx.doi.org/10.1007/s11263-020-01316-z.

Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Clearclip: Decomposing clip representations for dense vision-language inference. *ECCV*, 2024. URL https://arxiv.org/abs/2407.12442.

Meir Yossef Levi and Guy Gilboa. The double-ellipsoid geometry of clip. *ICML*, 2025. URL https://arxiv.org/abs/2411.14517.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *ICML*, 2022. URL https://arxiv.org/abs/2201.12086.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ICML*, 2023. URL https://arxiv.org/abs/2301.12597.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Neurips*, 2022. URL https://arxiv.org/abs/2203.02053.

Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. *Advances in neural information processing systems*, 35:35087–35102, 2022.

Jayneel Parekh, Pegah Khayatan, Mustafa Shukor, Alasdair Newson, and Matthieu Cord. A concept-based explainability framework for large multimodal models. *Neurips*, 2024. URL https://arxiv.org/abs/2406.08074.

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. *ICML*, 2021. URL https://arxiv.org/abs/2103.00020.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. URL https://arxiv.org/abs/2204.06125.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *CVPR*, 2022. URL https://arxiv.org/abs/2112.10752.

Jie-Jing Shao, Jiang-Xin Shi, Xiao-Wen Yang, Lan-Zhe Guo, and Yu-Feng Li. Examining the achilles' heel of CLIP models: The worst-performing categories, 2024. URL https://openreview.net/forum?id=0S0CgZEYxR.

Liangliang Shi, Jack Fan, and Junchi Yan. OT-CLIP: Understanding and generalizing CLIP via optimal transport. *ICML*, 2024. URL https://openreview.net/forum?id=X8uQ1TslUc.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *ICLR*, 2021. URL https://arxiv.org/abs/2011.13456.

Harald Steck, Chaitanya Ekanadham, and Nathan Kallus. Is cosine-similarity of embeddings really about similarity? *WWW*, 2024. URL http://dx.doi.org/10.1145/3589335.3651526.

Weijie Tu, Weijian Deng, and Tom Gedeon. A closer look at the robustness of contrastive language-image pre-training (clip). *Neurips*, 2023. URL https://arxiv.org/abs/2402.07410.

Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. *NeurIPS*, 2021. URL https://arxiv.org/abs/2106.05931.

Qizhou Wang, Yong Lin, Yongqiang Chen, Ludwig Schmidt, Bo Han, and Tong Zhang. A sober look at the robustness of clips to spurious features. *Neurips*, 2024. URL https://arxiv.org/abs/2403.11497.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. *ICML*, 2022. URL https://arxiv.org/abs/2005.10242.

Xin Wen, Bingchen Zhao, Yilun Chen, Jiangmiao Pang, and Xiaojuan Qi. What makes clip more robust to long-tailed pre-training data? a controlled study for transferable insights. *Neurips*, 2024. URL https://arxiv.org/abs/2405.21070.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *TMLR*, 2022. URL https://arxiv.org/abs/2205.01917.

Vladimir Zaigrajew, Hubert Baniecki, and Przemyslaw Biecek. Interpreting clip with hierarchical sparse autoencoders. *ICML*, 2025. URL https://arxiv.org/abs/2502.20578.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. *CVPR*, 2018. URL https://arxiv.org/abs/1801.03924.

Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. *ICLR*, 2024. URL https://arxiv.org/abs/2310.18961.

APPENDIX CONTENTS

A Related Work **14**

A.1 Geometry and Anisotropy in CLIP 14

A.2 Likelihood Estimation in Latent Space 14

A.3 Interpretability of Multimodal Representations 15

B Future Work **15**

B.1 Connection with Score-based Generative Models 15

B.2 Hierarchical Modeling of Semantic Structure 15

C Additional Methodological Details **15**

C.1 Why Whitening *Before* Hyperspherical Mixture Modeling? 15

C.2 Detailed EM Updates 16

D Extended Experimental Results **17**

D.1 Experimental Setup 17

D.2 Implementation Details. 17

D.3 Long-tailed and OOD Detection for Other VLMs 17

D.4 Visualization of Likelihood Distributions for OOD Detection 18

E Ablation Studies of Hyperparameters **18**

F Extended Qualitative Analyses **18**

F.1 Semantic Decomposition Comparison 18

F.2 Concept Localization 20

G Extended Analysis Across Iterations **24**

A Related Work

A.1 Geometry and Anisotropy in CLIP

CLIP and related multimodal contrastive models (Radford et al., 2021; Li et al., 2022; Yu et al., 2022) embed images and text into a shared latent space optimized under cosine similarity. This objective implicitly normalizes embeddings onto a hypersphere and is often interpreted as encouraging an isotropic and uniformly distributed representation geometry (Wang & Isola, 2022). However, a growing body of empirical evidence demonstrates that the actual geometry of CLIP latent space deviates substantially from this idealized picture.

One prominent deviation is the *Modality Gap*, wherein image and text embeddings occupy distinct regions of the hypersphere rather than forming a fully overlapping distribution (Liang et al., 2022; Levi & Gilboa, 2025). While initially viewed as a deficiency in cross-modal alignment, subsequent studies suggest that such separation may serve functional roles, including improved robustness and mitigation of catastrophic forgetting (Huang et al., 2025). Beyond modality-level separation, CLIP representations also exhibit significant anisotropy and uneven semantic coverage. Embeddings tend to concentrate around dominant visual patterns, such as large foreground objects or frequently co-occurring features, while underrepresenting rare or fine-grained concepts (Abbasi et al., 2025; Wang et al., 2024; Lan et al., 2024). The anisotropic structure leads to substantial performance variability across semantics, particularly on long-tailed and rare concepts (Tu et al., 2023; Shao et al., 2024).

These observations indicate that CLIP latent space possesses structured and non-uniform density rather than a single homogeneous distribution. Consequently, cosine similarity alone cannot fully capture semantic relationships, as it ignores variations in representation density and concentration (Steck et al., 2024; Kang et al., 2025). Several works have therefore explored alternative geometric perspectives, including cycle-consistency constraints (Goel et al., 2022) and optimal transport formulations (Shi et al., 2024). More recent analyses explicitly characterize CLIP geometry as structured anisotropy, such as the "double-ellipsoid" structure separating common and rare concepts (Levi & Gilboa, 2025; Wen et al., 2024). Collectively, these findings suggest that CLIP latent space exhibits an inherently multimodal semantic organization.

To enable likelihood-based reasoning and improve calibration, recent work has proposed geometric normalization techniques. W-CLIP (Betser et al., 2025), for example, applies a whitening transformation to normalize second-order statistics and estimate likelihood under a Gaussian assumption. Related approaches similarly attempt to regularize representation geometry through linear normalization (Chung & Kim, 2026). While effective for calibration, such methods rely on unimodal approximations of the latent space. In contrast, our work directly models CLIP latent space as a mixture of hyperspherical semantic components, providing a probabilistic formulation that better captures its intrinsic multimodal structure.

A.2 Likelihood Estimation in Latent Space

Estimating likelihood directly in pixel space is computationally challenging and often poorly aligned with semantic similarity, motivating likelihood estimation in learned representation spaces. Recent work has explored using pretrained multimodal encoders such as CLIP to define semantically meaningful likelihood surrogates. In particular, W-CLIP (Betser et al., 2025) projects embeddings into a whitened space and estimates likelihood under a Gaussian assumption, providing a tractable likelihood proxy. However, modeling the latent space with a single Gaussian imposes a unimodal assumption that conflicts with the inherently multimodal structure of semantic representations. Natural image distributions consist of multiple semantic modes corresponding to different objects, attributes, and compositions. Under a unimodal Gaussian model, rare but valid concepts may be assigned artificially low likelihood, conflating semantic rarity with distributional abnormality. Alternative approaches, such as retrieval-based similarity (He et al., 2025), prompt adaptation (Zhou et al., 2024; Cao et al., 2024b), or architectural modifications (Gong et al., 2025), can improve robustness but do not provide a unified probabilistic formulation.

In contrast, we model the CLIP latent distribution using a Mixture of von Mises–Fisher distributions, which naturally respects the hyperspherical geometry of CLIP embeddings. This formulation captures multiple semantic modes while retaining analytical tractability for likelihood estimation.

A.3 Interpretability of Multimodal Representations

Understanding and interpreting dense representations learned by multimodal models has become an important area of research. Prior work has explored sparse linear decompositions that express embeddings as combinations of interpretable concepts drawn from predefined vocabularies or learned dictionaries (Bhalla et al., 2024; Hoang-Xuan et al., 2025; Parekh et al., 2024). Other approaches investigate internal model structure, including attention mechanisms and sparse autoencoders, to identify latent features and disentangle semantic factors within CLIP representations (Gandelsman et al., 2024; Kempf et al., 2025; Zaigrajew et al., 2025; Dhimoïla et al., 2026). While these methods provide valuable insights, they often require training auxiliary models, introducing additional complexity and dependence on external supervision or architectural modifications.

In contrast, our approach derives interpretability directly from probabilistic modeling of the latent space. By representing embeddings as a mixture of hyperspherical distributions, each component corresponds to a coherent semantic mode, and individual embeddings can be interpreted through their probabilistic associations with these components. The mixture centers naturally act as semantic prototypes, enabling semantic explanations and likelihood estimation within a unified framework, without auxiliary decoders or external concept dictionaries.

B Future Work

B.1 Connection with Score-based Generative Models

Beyond density estimation, the hyperspherical density learned by MovMF-CLIP may provide a useful geometric prior for generative modeling. Modern latent diffusion and score-based models (Song et al., 2021; Rombach et al., 2022; Vahdat et al., 2021) typically assume simple Gaussian priors in the latent space. In contrast, the movMF mixture defines a directional density on the hypersphere, which naturally yields a closed-form score function on the manifold. This observation suggests a potential connection between hyperspherical density modeling and Riemannian score-based generative models (Bortoli et al., 2022; Huang et al., 2022). In particular, the learned density could provide a geometry-aware guidance signal that encourages latent trajectories to remain within semantically meaningful regions of the representation space. Exploring such integrations with latent diffusion or unCLIP-style generation pipelines is an interesting direction for future work.

B.2 Hierarchical Modeling of Semantic Structure

Our MovMF-CLIP currently models CLIP latent space using a flat mixture with a fixed number of components. While this formulation already captures multimodal structure, semantic concepts in natural data often exhibit hierarchical organization. A natural extension is therefore to explore hierarchical mixture formulations that allow components to be organized across multiple semantic levels. For example, Bayesian nonparametric approaches such as hierarchical Dirichlet processes could allow the number of semantic components to adapt to the data automatically. Such hierarchical modeling may provide a coarse-to-fine representation of semantic structure and could potentially improve robustness to long-tail concepts by allowing rare categories to share statistical strength with related parent clusters. Investigating hierarchical extensions of our hyperspherical mixture models remains an important and interesting future direction.

C Additional Methodological Details

C.1 Why Whitening *Before* Hyperspherical Mixture Modeling?

CLIP embeddings are compared by cosine similarity, which motivates modeling their directions on a hypersphere. However, the raw latent distribution is typically far from isotropic: second-order anisotropy introduces dominant covariance directions that can distort angular geometry and confound hyperspherical clustering. Let $z \in \mathbb{R}^d$ denote CLIP embeddings with mean μ and covariance

$$\Sigma \;=\; \mathbb{E}\big[(z-\mu)(z-\mu)^\top\big]\,.$$

Empirically, $\Sigma \neq I$, reflecting global variations induced by dataset bias, frequent concepts, and representation artifacts.

Whitening as metric normalization. We first apply a whitening transform in $\mathbb{R}^d$:

$$y \;=\; W(z-\mu), \qquad W \approx \Sigma^{-1/2},$$

so that y has approximately standardized second-order statistics:

$$\mathbb{E}[y] \approx 0, \qquad \mathrm{Cov}(y) \approx I.$$

This step removes nuisance anisotropy and defines an **intrinsic Mahalanobis geometry** in the original space. In particular, the angular similarity in whitened space corresponds to a Mahalanobis cosine similarity in the original coordinates:

$$\frac{y_i^\top y_j}{\|y_i\|\|y_j\|} \;=\; \frac{(z_i-\mu)^\top \Sigma^{-1}(z_j-\mu)}{\|z_i-\mu\|_{\Sigma^{-1}}\,\|z_j-\mu\|_{\Sigma^{-1}}}.$$

Thus, whitening calibrates directions according to the covariance structure of the latent distribution rather than letting a few high-variance directions dominate. This ensures that the resultant hypersphere is defined under the intrinsic second-order geometry of the data.

Normalization and hyperspherical density. After whitening, we normalize to the unit hypersphere:

$$u \;=\; \frac{y}{\|y\|} \in \mathbb{S}^{d-1},$$

and model the directional density with a mixture of von Mises–Fisher distributions:

$$p(u) \;=\; \sum_{k=1}^{K} \pi_k\, C_d(\kappa_k)\, \exp\!\big(\kappa_k\, \mu_k^\top u\big), \qquad \|\mu_k\| = 1.$$

Why not fit movMF directly on $z/\|z\|$? If we normalize first, $u_0 = z/\|z\|$, the subsequent movMF fit must simultaneously explain (i) global anisotropic scaling and (ii) genuine semantic multimodality. In practice, this causes mixture components to align with dominant covariance directions rather than semantic modes, leading to redundant components, unstable concentration estimates, and degraded likelihood calibration. Whitening *before* normalization removes global second-order effects in $\mathbb{R}^d$ and yields a hyperspherical representation whose angular geometry better reflects true density structure.

Importantly, whitening does not impose Gaussianity. Unlike approaches that use whitening to justify a unimodal Gaussian likelihood, we use whitening only to normalize second-order geometry. The distribution on the hypersphere remains explicitly multimodal and is captured by the vMF mixture. This separation between global covariance normalization and multimodal semantic density improves identifiability, statistical efficiency of EM, and interpretability of mixture components as semantic prototypes.

C.2 Detailed EM Updates

Given responsibilities γ_{ik} from the E-step, the M-step maximizes the expected complete-data log-likelihood. Define the effective mass and weighted resultant vector of component k as

$$N_k = \sum_{i=1}^{N} \gamma_{ik}, \qquad r_k = \sum_{i=1}^{N} \gamma_{ik} u_i. \tag{9}$$

The mixing coefficient and mean direction then admit the closed-form updates

$$\pi_k^{\mathrm{new}} = \frac{N_k}{N}, \qquad \mu_k^{\mathrm{new}} = \frac{r_k}{\|r_k\|_2}. \tag{10}$$

Let $\bar{R}_k = \|r_k\|_2/N_k$ denote the mean resultant length. The maximum-likelihood estimate of the concentration parameter satisfies

$$A_d(\kappa_k) = \bar{R}_k, \qquad A_d(\kappa) = \frac{I_{d/2}(\kappa)}{I_{d/2-1}(\kappa)}, \tag{11}$$

where $I_\nu(\cdot)$ is the modified Bessel function of the first kind. Direct inversion is computationally expensive in high dimensions. Following Banerjee et al. (2005), we therefore use

$$\kappa_k^{\text{new}} \approx \frac{\bar{R}_k(d - \bar{R}_k^2)}{1 - \bar{R}_k^2}, \tag{12}$$

which yields stable and efficient updates for high-dimensional CLIP embeddings. These expressions give the compact updates summarized in Eq. 6.

D Extended Experimental Results

D.1 Experimental Setup

Datasets. We adopt MS-COCO 2017 as the primary in-distribution (ID) dataset. Its validation split is used for both density estimation and long-tailed evaluation across semantic categories. For OOD evaluation, we use a cleaned subset of OpenImages (Kuznetsova et al., 2020), following VOS (Du et al., 2022) to remove overlapping categories.

Long-tailed and OOD Detection. We assess likelihood robustness by measuring how well the learned density distinguishes ID samples from OOD data and how it behaves across frequent versus rare semantic categories. This setup allows us to evaluate whether hyperspherical mixture modeling mitigates the common conflation between semantic rarity and distributional abnormality.

Interpretable Semantic Decomposition. To evaluate interpretability, we compare against SPLICE (Bhalla et al., 2024), MSAE (Zaigrajew et al., 2025), and TextSpan (Gandelsman et al., 2024) by measuring alignment between mixture-based semantic decomposition and ground-truth captions. This metric evaluates whether the posterior responsibilities of mixture components correspond to meaningful linguistic concepts in the annotations.

Semantic Stability Under Iterative Generative Drift. To study generative robustness, we simulate semantic drift through iterative encoding–decoding cycles using the pipeline of CLIP + UnCLIP[1]. At each iteration, images are re-encoded into CLIP space, allowing us to track density and semantic changes. We compare against W-CLIP (Betser et al., 2025) in terms of semantic stability.

Implementation Details. For experiments on semantic decomposition, following Bhalla et al. (2024), all methods use the same OpenCLIP ViT-B/32 backbone trained on LAION-400M. By contrast, the pre-trained CLIP ViT-L/14 backbone is used for all the other experiments. Following Betser et al. (2025), we estimate the whitening transformation from MS-COCO validation set. The hyperparameters of our method are fixed across tasks. Additional qualitative analyses and extended results on BLIP-2 (Li et al., 2023) and CoCa (Yu et al., 2022) are provided in Appendices D and F.

D.2 Implementation Details.

Details of the Semantic Relevance Metric. For a given image, let W_{cluster} denote the set of keywords extracted from the most activated mixture components, and let W_{caption} denote the content words (nouns and verbs) in the ground-truth caption. We compute the Hausdorff distance d_H between the two embedding sets and define

$$\text{Semantic Relevance} = 1 - d_H\big(E(W_{\text{cluster}}), E(W_{\text{caption}})\big), \tag{13}$$

where $E(\cdot)$ denotes the text encoder. Higher values indicate stronger semantic agreement between the decomposed representation and the annotations.

Inference Measurement Details. All inference speeds are measured as the average per-image inference time on a workstation equipped with a single NVIDIA A100 GPU and an Intel Xeon Platinum 8468 CPU.

D.3 Long-tailed and OOD Detection for Other VLMs

Complementing the cross-VLM OOD detection results in Section 3.2 (Table 3), we report additional long-tailed likelihood fairness results for BLIP-2 and CoCa using the same evaluation protocol.

[1] https://huggingface.co/sd2-community/stable-diffusion-2-1-unclip-small

Long-Tailed Fairness across VLMs. As shown in table 6, the unimodal Gaussian assumption used in W-CLIP consistently exhibits bias against tail categories across different VLM architectures, producing high AUROC values when separating head and tail samples. In contrast, MovMF-CLIP substantially reduces this bias, bringing the AUROC closer to the ideal value of 0.5. These results suggest that modeling the latent space as a hyperspherical mixture provides a more balanced density estimate across semantic frequencies, and the improvements remain consistent across models such as BLIP-2 and CoCa.

Table 6: **Long-tailed fairness across VLMs.** AUROC ($\rightarrow 0.5$) for distinguishing head and tail categories using likelihood scores. MovMF-CLIP consistently reduces bias against tail concepts compared with W-CLIP across BLIP-2 and CoCa.

Model	BLIP-2		CoCa	
Method	W-CLIP	MovMF-CLIP	W-CLIP	MovMF-CLIP
AUROC	0.6809	**0.5172**	0.8308	**0.5465**

D.4 Visualization of Likelihood Distributions for OOD Detection

To provide further intuition for the OOD detection results, we visualize the empirical log-likelihood distributions of ID and OOD samples in Fig. 8. The top row shows the distributions when the full MS-COCO is used as the ID dataset, while the bottom row restricts the ID samples to the tail subset containing rare concepts. In both settings, MovMF-CLIP produces a clearer separation between ID and OOD samples compared with W-CLIP. The improvement is particularly evident when the ID set contains only tail samples, where unimodal Gaussian modeling tends to assign low likelihood to rare but valid data.

E Ablation Studies of Hyperparameters

In Sec. 3.3 of the main text, we showed that MovMF-CLIP outperforms the SPLICE baseline (Bhalla et al., 2024) in terms of Semantic Relevance. To evaluate the robustness of our approach, we conduct a grid search over two key hyperparameters: the number of mixture components (K) and the number of top-frequency keywords extracted per cluster (N).

The results are summarized in table 7. Overall, MovMF-CLIP demonstrates stable performance across a broad range of configurations. Increasing the number of keywords N consistently improves Semantic Relevance, as richer vocabularies better describe the semantic content of each cluster. Regarding the number of clusters K, performance improves when moving from small values to moderate values, reflecting the benefit of capturing more fine-grained semantic structure. When K becomes very large (e.g., $K \geq 800$), the improvement saturates and slightly decreases, likely due to over-fragmentation of the latent space. The best performance is achieved with $K = 500$ and $N = 10$, reaching a Semantic Relevance of **0.6726**. Importantly, even small configurations (*e.g.,* $K = 70$, $N = 3$) remain competitive, indicating that the learned hyperspherical mixture structure is robust across different sets of hyperparameters.

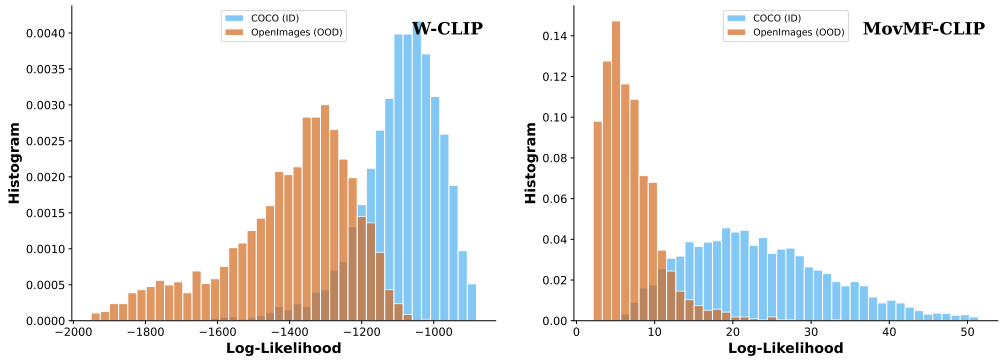
F Extended Qualitative Analyses

We provide extensive results to further demonstrate the robustness, fine-grained interpretability of concept localization, and semantic stability under generative drifts of our MovMF-CLIP framework.

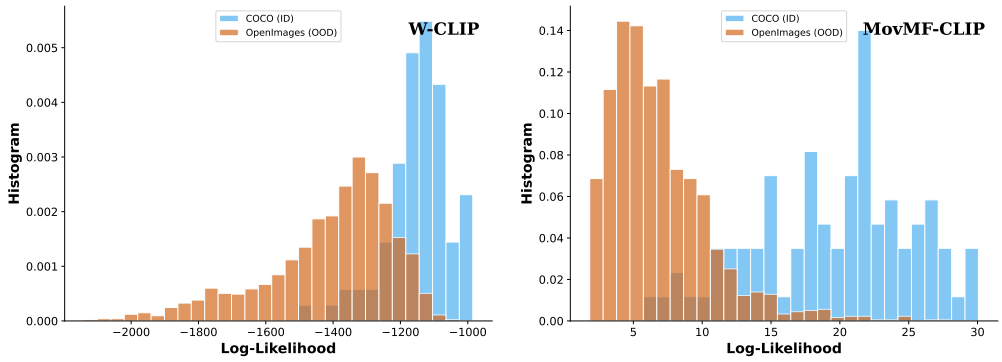
F.1 Semantic Decomposition Comparison

To further validate our semantic decomposition capabilities (extending Fig. 4 of the main text), we present 6 additional comparison cases against the SPLICE baseline in Fig. 9.

The results clearly demonstrate that SPLICE frequently suffers from hallucinated or overly generic descriptions. For instance, given an image of a person holding a laptop (top-right), SPLICE outputs



(a) ID: **Full COCO** vs. OOD: OpenImages



(b) ID: **Tail Subset Only** vs. OOD: OpenImages

Figure 8: **Likelihood distributions for OOD detection.** Histograms of log-likelihood scores for ID and OOD samples. **Left**: W-CLIP exhibits substantial overlap between ID and OOD distributions, particularly when the ID set contains rare concepts. **Right**: MovMF-CLIP produces better separation, reflecting improved robustness to long-tail semantics while maintaining strong OOD discrimination.

Table 7: **Grid Search for Semantic Relevance.** We report the Semantic Relevance ($1 - d_H$) on the MS-COCO validation set using OpenCLIP ViT-B/32, varying the number of clusters (K) and keywords per cluster (N).

Number of Clusters (K)	**Top-N Keywords per Cluster**			
	$N = 3$	$N = 5$	$N = 7$	$N = 10$
70	0.6434	0.6552	0.6612	0.6695
100	0.6443	0.6563	0.6624	0.6700
300	**0.6445**	0.6565	0.6640	0.6722
500	0.6442	**0.6566**	**0.6653**	**0.6726**
800	0.6423	0.6561	0.6644	0.6706
1000	0.6425	0.6565	0.6638	0.6711

irrelevant terms like *"protester"*, *"chimney"*, or *"hvac"*. In contrast, MovMF-CLIP precisely extracts fine-grained semantic prototypes such as *"laptop"*, *"table"*, and *"people"*. This confirms that our hyperspherical mixture successfully maps complex visual scenes to highly accurate linguistic concepts without cross-modal contamination.

F.2 Concept Localization

Fig. 3 provides a representative concept-localization example in the main text, while Fig. 10 presents six additional cases. The concept-specific heatmaps closely correspond to the image regions associated with each semantic component. For example, in the outdoor street scene (Row 4), MovMF-CLIP cleanly separates the *"fire hydrant"* from the *"dog"*. Similarly, it distinguishes a *"bird"* perched on a *"cow"* (Row 5), and localizes specific objects in complex environments, such as isolating *"wine"* from a *"pizza"* on a dining table (Row 3), or a *"man"* from boxes of *"bananas"* (Row 6). This spatial grounding verifies that the learned mixture components capture distinct semantic concepts.

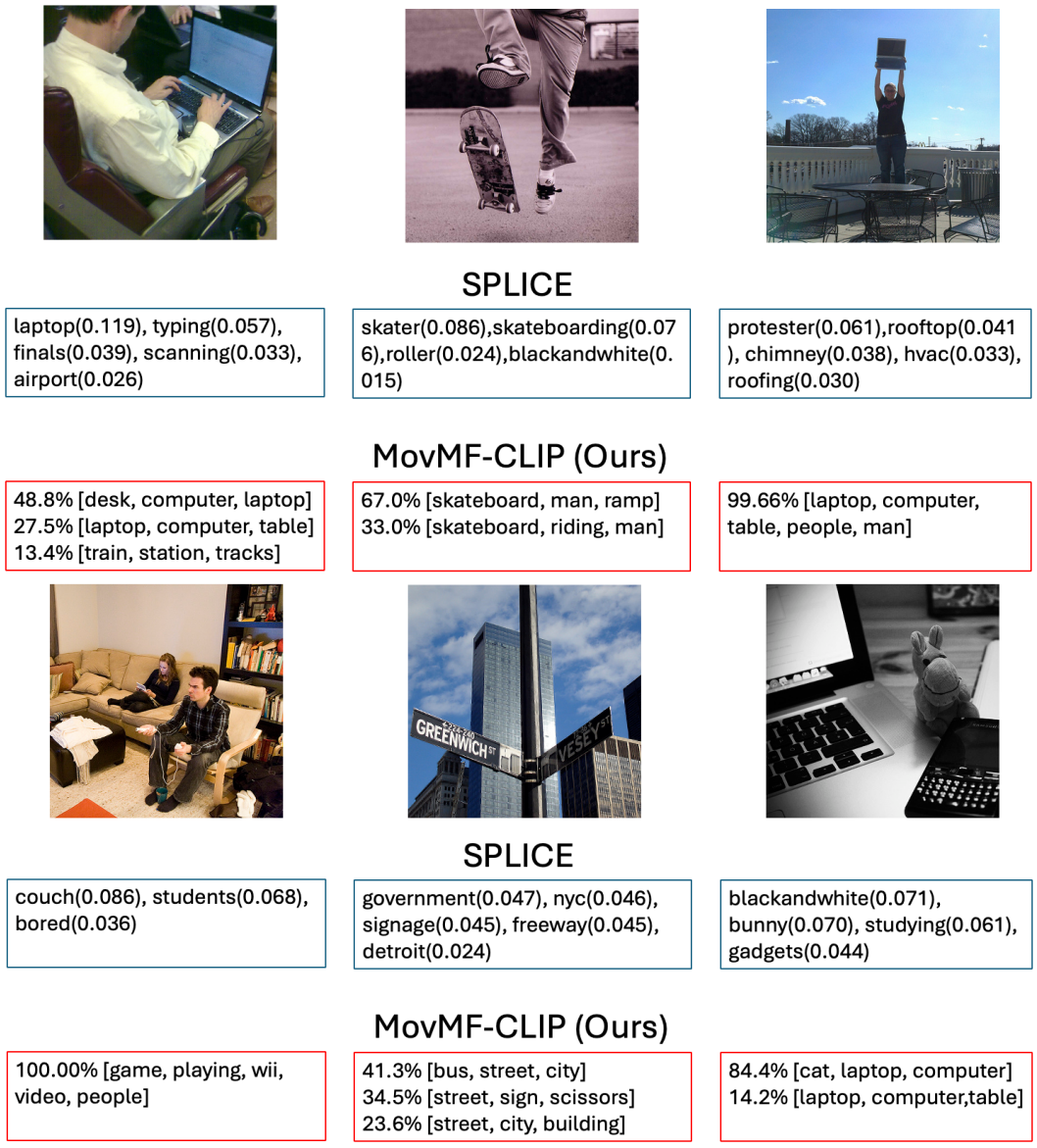


Figure 9: **Extended Results of Semantic Decomposition.** Comparison of text descriptions decoded from the latent embeddings. MovMF-CLIP consistently extracts precise, fine-grained semantic concepts (e.g., “wii”, “skateboard”), while the baseline SPLICE often struggles with hallucinated or coarse terms (e.g., “bored”, “hvac”).

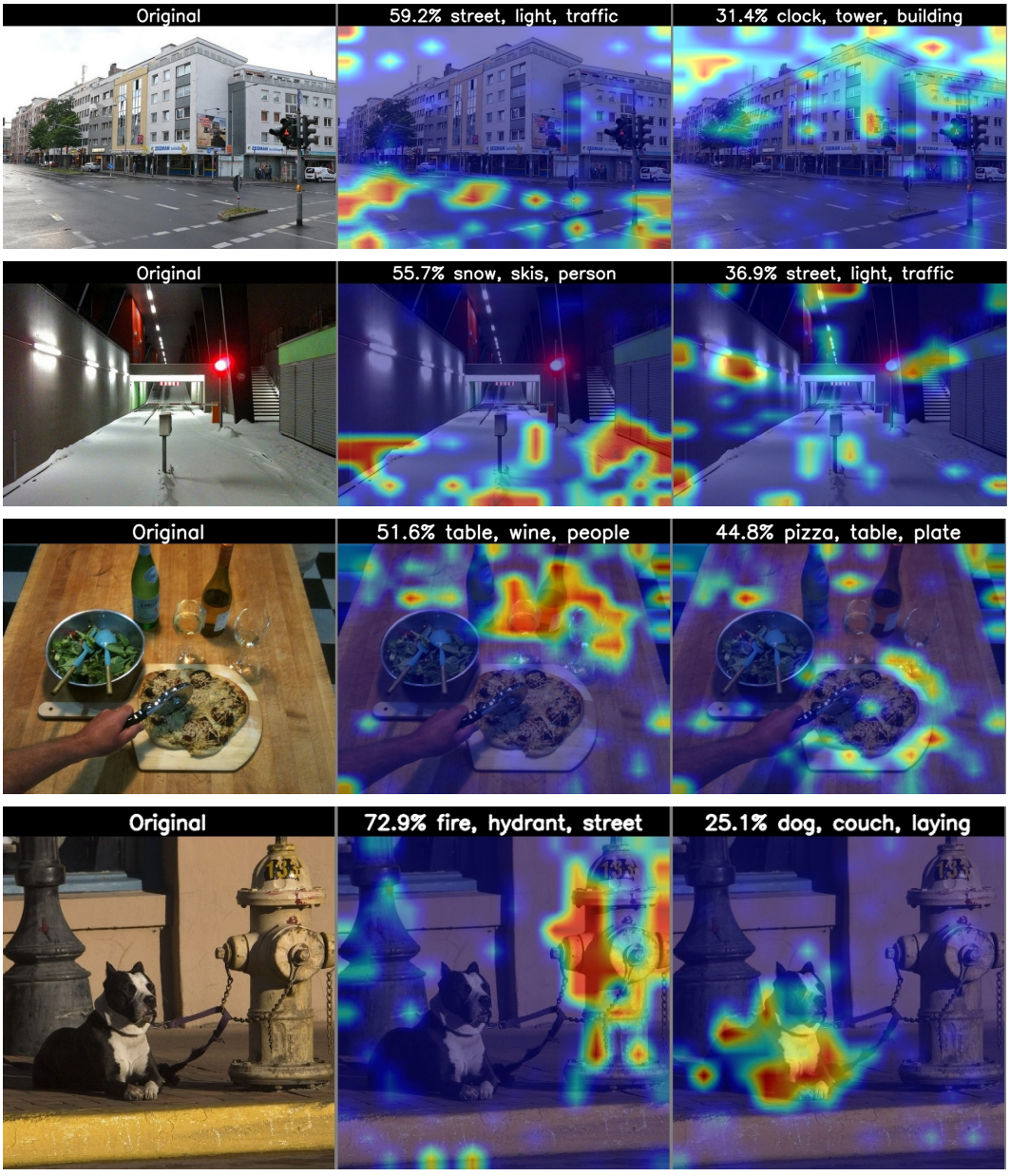


Figure 10: **Extended Results of Concept Localization.** (Continued on next page.)



Figure 10: **Extended Results of Concept Localization (Continued).** MovMF-CLIP successfully disentangles complex visual scenes into orthogonal semantic components. Each heatmap accurately grounds specific keywords derived from our mixture components, demonstrating precise spatial isolation of objects.

G Extended Analysis Across Iterations

We further analyze semantic stability under iterative generative drift across six consecutive steps. We begin by evaluating the quantitative degradation of the semantic representations, followed by a detailed visual analysis.

As shown in Fig. 11, we track the quantitative degradation across 1,000 COCO images. Vanilla CLIP-UnCLIP and W-CLIP initially show slightly lower LPIPS and higher cosine similarity in the first round. However, their representations rapidly deteriorate in subsequent iterations, reflected by a sharp, accelerating drop in CLIP cosine similarity and a continuous spike in LPIPS distance. In contrast, MovMF-CLIP effectively stabilizes the semantic trajectory. By dynamically pulling latent embeddings toward high-density semantic prototypes, MovMF-CLIP significantly outperforms the baselines from round 3 onward.

Fig. 12 displays the generated results across six consecutive steps. Vanilla CLIP-UnCLIP representations quickly fall into low-density semantic voids, causing catastrophic mode collapse (*e.g.,* a surfer turning into colorful horizontal stripes or a pizza collapsing into abstract red and green patterns). While W-CLIP mitigates some noise, it suffers from severe semantic blurring over time. Conversely, MovMF-CLIP leverages the learned semantic prototypes as geometric anchors and preserves the core semantic identities across iterations.

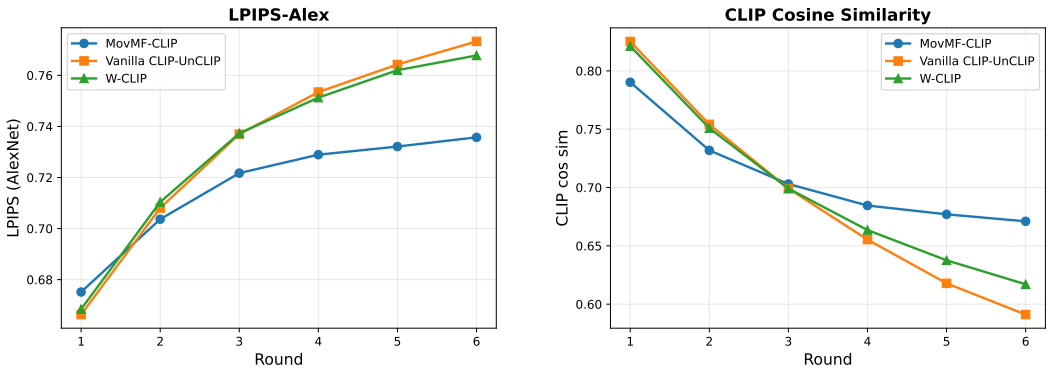


Figure 11: **Quantitative trends across iterations.** We track the average LPIPS (left, ↓) and CLIP cosine similarity (right, ↑) over 1,000 COCO images across six CLIP-UnCLIP iterations. While W-CLIP and Vanilla CLIP-UnCLIP show a slight initial advantage, they suffer from accelerating degradation. In contrast, MovMF-CLIP stabilizes the trajectory, significantly outperforming the baselines from round 3 onward.



Figure 12: **Extended results of iterative generative drift.** Trajectories over six CLIP-UnCLIP iterations. Vanilla CLIP-UnCLIP suffers from severe mode collapse (*e.g.,* generating meaningless stripes). W-CLIP loses semantic details. MovMF-CLIP correctly anchors the embeddings to high-density semantic prototypes, better preserving the original concepts.