SPACEVLM: SUB-SPACE MODELING OF NEGATION IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent works show that Vision-Language Models (VLMs) like CLIP struggle with logical reasoning, which can contribute to unreliable model behavior and hallucinations. An example of such logical operators is *negation*. Given a prompt like "*retrieve (or generate) a street scene without pedestrians*", VLMs often fail to respect the "*not*". Existing methods try to address the logical blind spots of VLMs by fine-tuning models on larger and richer datasets. However, this fine-tuning can compromise zero-shot generalization and often fails to fully eliminate these blind spots. In this paper, we use the *negation* operator as a case study and show that representing logical operators as traditional point embeddings in a joint embedding space can be limiting. Instead, we propose subspace modeling of negation, where text representation is a subspace in the embedding space rather than a traditional vector (point) embedding. Then we derive a closed-form Mahalanobis distance on a curved surface from an image embedding (point) to a text representation (subspace), replacing the conventional cosine similarity between image and text embeddings. Across retrieval and multiple-choice-question (MCQ) tasks, our new modeling improves negation understanding in VLMs including CLIP, SigLIP, and AIMV2 by about 30% on average over the baseline, without any additional retraining of the VLM, while preserving the model's zero-shot performance. More broadly, our framework could serve as a starting point for research on subspace modeling of other logical operators, not only at inference time, but also during model pre-training. Code is included in the supplementary material.

1 INTRODUCTION

Joint embedding-based Vision-Language Models (VLMs) Radford et al. (2021); Li et al. (2022); Zhai et al. (2023), such as CLIP Radford et al. (2021), have become strong foundations for visual understanding. These models consist of an image encoder and a text encoder that map visual and textual inputs into a shared embedding space, where similarity is measured by dot product. When pretrained on massive image–text datasets, they exhibit strong generalization and are widely used for classification, retrieval, and text-to-image generation Zhang et al. (2024a); Zhao et al. (2023); Klemmer et al. (2025), with successful applications in specialized domains such as healthcare Lu et al. (2024). However, they struggle with inputs that require logical reasoning Kamath et al. (2023); Lewis et al. (2022); Rasekh et al. (2025), particularly those involving negation Alhamoud et al. (2025); Singh et al. (2025). This lack of logical understanding can propagate to unreliable behavior in downstream applications. For instance, Park et al. (2025) showed that CLIP's poor understanding of negation carries over to most Text-to-Image(T2I) models that utilize the CLIP text encoder.

Previous studies attributed this weakness to the lack of logically rich captions in the training data; to address this, they generated large synthetic datasets and fine-tuned VLMs on them Yuksekgonul et al. (2022); Singh et al. (2025); Park et al. (2025). Yet, these fine-tuning methods often reduce the model's zero-shot generalization. Another line of research uses heuristic inference-time steering to improve negation understanding without additional training Aggarwal et al. (2026), but these methods lack a generalizable framework. Despite their differences, both lines of work, fine-tuning approaches and heuristic steering methods, assume that logical operators can be represented using point embeddings in a joint embedding space. In this paper, we challenge this assumption and propose subspace modeling of the text modality in the joint-embedding space.

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.

To develop this subspace framework, we use *negation* as a case study. Beyond being a persistent failure mode of VLMs, negation is critical in high-stakes domains such as healthcare, where about 39.45% of diagnostic keywords in medical reports appear in negated form Weng et al. (2023).

Consider the query "*retrieve an image with a dog but not a cat.*" A model processing this input must correctly exclude images containing cats, while retaining valid alternatives that include dogs. The key challenge is representational. "*Not a cat*" excludes cat, but leaves open many alternatives, such as dog or apple. Representing such a query with a single point embedding, following the dot-product scoring used in joint embedding-based VLMs such as CLIP Radford et al. (2021), SigLIP Zhai et al. (2023), and LiT-tuned AIMV2 Zhai et al. (2022); Fini et al. (2025), is limiting because the valid alternatives can occupy semantically distant regions of the embedding space (Figure 1). We therefore model the text representation as a subspace rather than a point. To this end, we first verify that CLIP's embedding space can be divided into semantically consistent subspaces Bhalla et al. (2024); Zhao et al. (2025). We then model negation as the intersection between an affirmative and a complementary *subspace*, and derive an inference-time scoring rule that requires no VLM retraining. For a caption "$< A >$ *but not* $< N >$", obtained by rephrasing the original caption, we construct subspaces around the embeddings of A and N and identify the region that is close to A but far from N. We then derive a closed-form Mahalanobis distance on a curved surface between the image embedding and this target subspace, replacing conventional point-to-point cosine similarity. Because this scoring operates purely at inference time, it leaves the underlying VLM and its pretrained parameters unchanged. Importantly, our SpaceVLM framework is model-agnostic and applicable to any joint embedding-based VLM. More broadly, our framework (including the scoring rule) opens the door to subspace modeling of other logical operators, including *intersection*, *union*, and *inclusion*.

We validate SpaceVLM across more than 40 experimental settings spanning combinations of VLM backbones, image and video datasets, and diverse negation tasks including multimodal retrieval and Multiple Choice Question (MCQ) (Figure 2). Following NegBench Alhamoud et al. (2025), we use the COCO Lin et al. (2014), VOC-2007 Everingham et al. (2010), and MSR-VTT Xu et al. (2016) datasets for general-domain retrieval and MCQ, and CheXpert Irvin et al. (2019) for medical diagnostics with negation. Our inference-time framework consistently improves negation understanding across all tested joint embedding-based models, including CLIP Radford et al. (2021), SigLIP Tschannen et al. (2025), NegCLIP Yuksekgonul et al. (2022), ConCLIP Singh et al. (2025), AIMV2 Fini et al. (2025), and BiomedCLIP Zhang et al. (2023). Despite requiring no re-training of the VLM, SpaceVLM outperforms fine-tuned baselines such as CLIP-NegFull Alhamoud et al. (2025), ConCLIP Singh et al. (2025), NegCLIP Yuksekgonul et al. (2022), and NegationCLIP Park et al. (2025), and it even surpasses the recent geometric approach DCSM Kang et al. (2025).

In the Appendix, we show that the cosine-similarity threshold, the primary hyperparameter in SpaceVLM, remains robust across a wide practical range (Appendix D). This stability facilitates direct deployment of SpaceVLM to downstream tasks without the need for extensive cross-validation. Additionally, we provide a qualitative visual inspection to confirm that SpaceVLM retrieves and generates diverse images that are semantically consistent with negated prompts (Appendix H). We also discuss extensions of SpaceVLM to multi-negation scenarios and other logical operators in Appendix B. We hope that the efficacy of this subspace perspective encourages future research into geometric frameworks for broader logical reasoning in Vision-Language Models.

2 Related Work

Joint Embedding-based Vision-Language Models align visual and textual representations in a shared embedding space. A representative example is CLIP Radford et al. (2021), which trains an image encoder $\mathcal{I} : x \to \mathbb{R}^d$ and a text encoder $\mathcal{T} : y \to \mathbb{R}^d$ on 400 million image-caption pairs using a contrastive objective. The two encoders map inputs to the surface of a unit sphere, and image-text similarity is measured by what we call the *CLIP dot-product scoring* $\mathcal{I}(x) \odot \mathcal{T}(y)$. Given a caption y, the corresponding image is then retrieved by: $\hat{x} = \text{argmax}_x \ \mathcal{I}(x) \odot \mathcal{T}(y)$.

CLIP's pretrained encoders are widely used across tasks, from multimodal retrieval Luo et al. (2021); Alpay et al. (2023); Lülf et al. (2024) to multimodal LLMs Liu et al. (2023); Wang et al. (2022) and text-to-image generation Tao et al. (2023); Rombach et al. (2022). Several follow-up variants adopt similar principles: SigLIP Zhai et al. (2023) replaces the softmax contrastive loss with a sigmoid

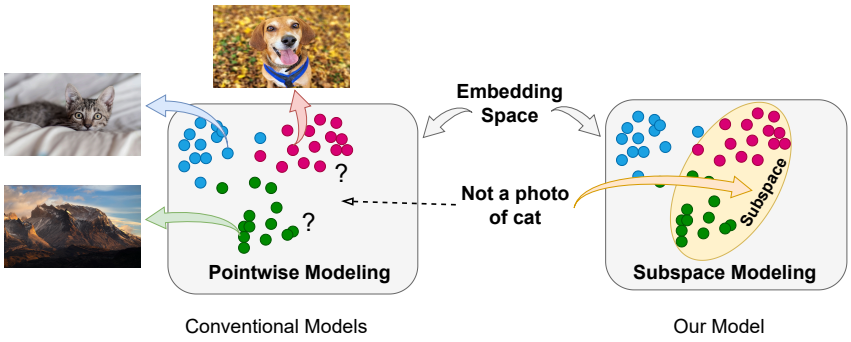


Figure 1: Given a caption such as "*Not a photo of a cat*", standard VLM approaches attempt to map this negative caption to a single point in the embedding space, which makes it ambiguous where the correct destination should be. In contrast, our approach maps the negative caption to a subspace rather than a point, enhancing the model's ability to handle negation effectively.

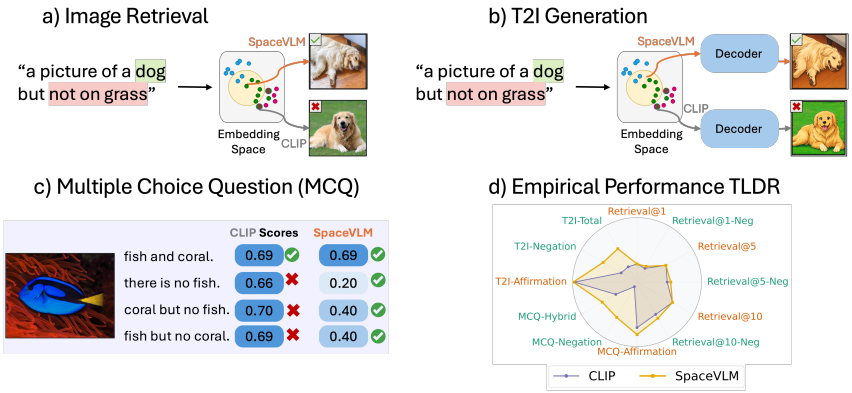
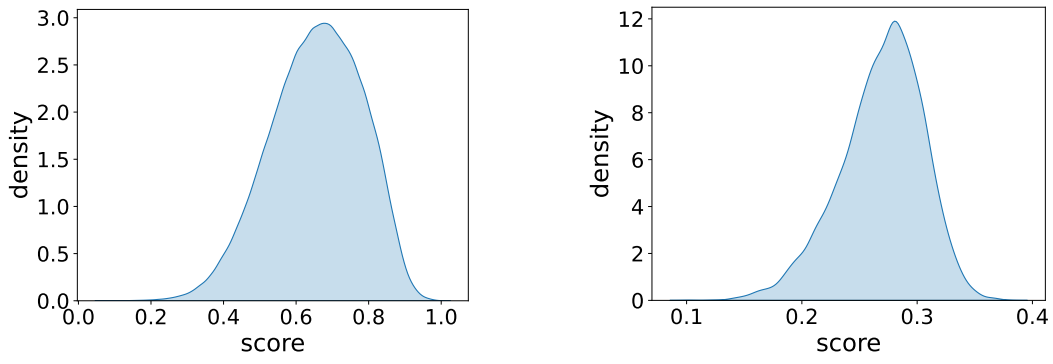


Figure 2: In both (a) Image Retrieval and (b) Text-to-Image (T2I) Generation, **CLIP** embeds the input prompt "a picture of a dog but not on grass" near images that include both dog and grass, leading to incorrect retrievals or generations. In (c) MCQ, **CLIP** assigns similar similarity scores to all captions mentioning "fish" and "coral," regardless of whether they include or exclude a concept, leading to incorrect image-text matching. By modeling negation as a subspace, **SpaceVLM** fixes all these issues. As summarized in (d), this geometric modeling of **SpaceVLM** empirically improves negation understanding across these tasks, while preserving performance on affirmative prompts.

loss; AIMV2 Fini et al. (2025) replaces the contrastive loss with a multimodal autoregressive loss, but its vision and text encoders can be aligned via Locked-Image Text Tuning Zhai et al. (2022), making it applicable to the *CLIP dot-product scoring*. We build on this family of models, improving their handling of negation at inference time without modifying their pretrained parameters.

Addressing the Logical Blindspots in VLMs. VLMs struggle with logical reasoning in prompts involving conjunction, disjunction, negation, contrast, comparison, condition, causality, and temporality Ma et al. (2023); Zhou et al. (2025); Li et al. (2024); Kang et al. (2025); Li et al. (2025). Related to our case study is NegBench Alhamoud et al. (2025), which evaluates negation understanding via text-to-image retrieval and image multiple-choice (MCQ) tasks. Most proposed solutions Yuksekgonul et al. (2022); Singh et al. (2025); Park et al. (2025) address these problems by constructing logically rich datasets and fine-tuning VLMs on them. NegCLIP Yuksekgonul et al. (2022) fine-tunes CLIP to improve sensitivity to logical structure, including negation. Singh *et al.* Singh et al. (2025) propose ConCLIP, a CLIP model fine-tuned for negation understanding, and Alhamoud *et al.* Alhamoud et al. (2025) extend this research by fine-tuning CLIP and NegCLIP on CC12M-NegFull, an extension of CC12M Changpinyo et al. (2021) with synthetically augmented negated captions.



(a) (b)

Figure 3: (a) Cosine similarity score distribution of images within the same category. (b) Cosine similarity score distribution between the textual prompt "*A photo of a <category>*" and images belonging to that category.

While these methods improve negation performance, their reliance on fine-tuning has two drawbacks: (i) degraded zero-shot generalization, and (ii) the fundamental inability of joint embedding–based models to represent negation with a single embedding vector, regardless of the scale of fine-tuning data. Another line of research Kang et al. (2025); Aggarwal et al. (2026) has explored inference-time approaches that require less post-training. Kang et al. (2025) note that joint embedding–based models cannot geometrically represent negation. They propose DCSM, a modification to the CLIP scoring function that retains all image patch embeddings and text token embeddings, computes cosine similarities across all pairs, and trains a convolutional projection head to aggregate this information. DCSM differs from our approach in two aspects. First, SpaceVLM explicitly models negation as a logical operation through the intersection of subspaces, whereas DCSM does not directly encode logical operators. Second, DCSM trains a lightweight scoring network on top of the frozen CLIP features for each dataset, while our method requires no additional training and operates entirely at inference time. Aggarwal et al. (2026) proposed a heuristic steering approach to find a new text embedding that also encodes negation information. While it achieves good performance, their framework is not generalizable to other operators by design, as it still relies on a point-based embedding space, which has architectural limitations for logical operations.

Few works address negation in text-to-image generation without fine-tuning Lian et al. (2024); Wu et al. (2024). They use large language models to parse the negated prompt, construct an intermediate image layout, and feed it to Stable Diffusion with negative prompts to suppress excluded concepts. While effective for generative pipelines, these methods are orthogonal to our goal: they target a specific application and do not improve logical understanding in the underlying vision–language encoder from an architectural perspective. Other general work Park et al. (2025) also incorporates negation into text–image modeling, but it does so by fine-tuning the CLIP text encoder on specific datasets, which can introduce task-specific specialization and may affect generalization across other negation tasks, as we show in our experiments.

3 Method

3.1 Empirical Divisibility of the Embedding Space

To model negation more effectively, we first examine the geometric structure of the CLIP embedding space. CLIP aligns images and captions by maximizing cosine similarity, and with ℓ_2-normalized embeddings, all representations lie on the surface of a d-dimensional unit sphere. Empirically, embeddings that refer to the same visual concept (e.g., "dog") occupy compact, well-separated regions on this sphere Bhalla et al. (2024); Zhao et al. (2025).

Figure 3 demonstrates this structure. Figure 3a shows the distribution of pairwise cosine similarities between images within the same CIFAR-100 class, while Figure 3b shows similarities between the

textual prompt "*A photo of a <category>*" and images of that class. Both histograms indicate that intra-class samples form high-similarity clusters that are distinct from other concepts. When such clusters are tight and sufficiently separated, we say that the space is *divisible*: a single cosine-similarity threshold can determine whether a new embedding belongs to a concept region or lies outside it. This divisibility property provides the geometric basis for our approach.

3.2 Modelling Negation

We state the key premise that motivates our approach.

Statement 1. *Negation cannot be consistently modeled by a single point (vector) in a general joint embedding VLM space. Instead, it must be modeled as a subspace or a region.*

Proof. An affirmative query (e.g., "*A photo of a cat*") is semantically *convergent*, targeting a specific cluster of inputs that share common visual features; such a cluster is effectively approximated by a single centroid vector. In contrast, a negated query (e.g., "*Not a photo of a cat*") is *divergent*, encompassing the entire complement of that set. This includes semantically different concepts such as *dogs*, *cars*, and *landscapes*. If we assign a single vector v to represent this negation, geometric consistency requires v to be similar to all valid non-cat instances. By the properties of the continuous metric space, this would imply that all distinct non-cat concepts (e.g., *a car* and *a flower*) must be similar to each other, which contradicts the structure of the semantic manifold. Therefore, negation cannot be mapped to a single point but must be modeled as a subspace to encompass all non-cat concepts. □

3.3 Problem Formulation

Given the divisibility property, we can represent complex textual compositions by reasoning over the regions induced by their constituent concepts. Consider a caption P that we split into an affirmative part "$P_a =< A >$" and a negated part "$P_n =< N >$" where P = "$< A >$ but not $< N >$" [1]. Let

$$e_a = \mathcal{T}(P_a), \qquad e_n = \mathcal{T}(P_n) \tag{1}$$

be the corresponding normalized text embeddings, and let e_I be the normalized image embedding, all produced by CLIP. In standard CLIP inference, the image-text similarity is computed as the dot product $s = e_I \odot e_P$, where e_P is the text embedding of the full caption, P. However, when P contains negation (and possibly other logical operators), this score becomes unreliable (Statement 1). *Instead, our goal is to define a scoring rule that leverages the compact regions (around e_a and e_n) to compute a more faithful score.*

3.4 SpaceVLM: Sub-Space Modeling of Negation

We now define the scoring rule that models negation as a subspace. We start with the affirmative and negated embeddings $e_a = \mathcal{T}(P_a)$ and $e_n = \mathcal{T}(P_n)$. We denote the neighborhood (spherical cap) of a normalized point x in the VLM space as

$$\mathcal{N}(x) = \{\, z \in \mathbb{R}^d \mid x \odot z \geq t \,\}, \qquad t \in [-1, 1],$$

where t is a cosine-similarity threshold. We associate P_a and P_n with their subspaces $\mathcal{N}(e_a)$ and $\mathcal{N}(e_n)$. The target subspace for P is the region that is in the affirmative region but outside the negated one:

$$\mathcal{N}(P) \;=\; \mathcal{N}(e_a) \cap \mathcal{N}^c(e_n),$$

where

$$\mathcal{N}^c(e_n) \;=\; \{\, z \in \mathbb{R}^d \mid z \notin \mathcal{N}(e_n) \,\}.$$

To perform image-text matching, we need a similarity score between an image embedding e_I and this region $\mathcal{N}(P)$. To account for symmetric properties of the $\mathcal{N}(P)$, we need to find the center of this region:

[1] In practice, the splitter can be a rule-based splitter, an LLM, or any kind of splitter that decomposes a caption into its affirmation and negation parts.

Table 1: Image/Video retrieval. R@K uses affirmative queries; R@K–Neg uses negated queries. Our approach narrows (and often closes) the gap between affirmative and negated retrieval when added on top of any baseline.

Dataset	Metric	CLIP		CLIP-NegFull		ConCLIP		NegCLIP		NegCLIP -NegFull		AIMV2		SigLIP-2	
		Base	+ **Ours**	Base	+ **Ours**	Base	+ **Ours**	Base	+ **Ours**	Base	+ **Ours**	Base	+ **Ours**	Base	+ **Ours**
COCO	R@1 ↑	29.8		32.6		28.8		46.1		44.3		41.0		43.4	
	R@1–Neg ↑	25.0	**30.8**	30.4	**33.1**	25.7	**30.1**	41.0	**45.5**	41.3	**44.1**	35.8	**41.7**	32.0	**48.3**
	R@5 ↑	54.8		57.8		54.7		74.0		72.4		66.5		68.7	
	R@5–Neg ↑	47.9	**56.1**	55.0	**58.5**	50.1	**55.9**	68.6	**73.6**	69.0	**71.8**	61.3	**67.4**	57.2	**73.2**
	R@10 ↑	66.0		68.9		66.4		83.0		81.7		76.2		77.8	
	R@10–Neg ↑	59.1	**67.4**	66.5	**69.5**	61.8	**67.6**	79.0	**82.8**	79.1	**81.3**	71.9	**76.8**	67.9	**81.5**
	Avg Δ (pp)		**+7.4**		**+3.1**		**+5.3**		**+4.4**		**+2.6**		**+5.6**		**+15.3**
MSR-VTT	R@1 ↑	26.4		28.4		26.4		30.1		30.9		31.1		33.3	
	R@1–Neg ↑	23.8	**29.0**	28.4	**29.7**	23.3	**28.4**	28.0	**32.4**	29.9	**33.3**	28.2	**31.5**	22.4	**30.8**
	R@5 ↑	48.7		51.7		48.5		51.5		53.9		52.8		55.3	
	R@5–Neg ↑	45.9	**52.4**	51.6	**52.4**	45.4	**52.0**	50.2	**55.7**	51.5	**55.6**	48.8	**56.7**	41.4	**54.5**
	R@10 ↑	60.7		62.8		61.6		62.1		64.3		62.9		65.7	
	R@10–Neg ↑	56.6	**63.1**	62.9	**63.2**	56.4	**63.4**	59.7	**66.9**	63.7	**66.5**	59.4	**66.6**	51.6	**64.0**
	Avg Δ (pp)		**+6.1**		**+0.8**		**+6.2**		**+5.7**		**+3.4**		**+6.1**		**+11.3**

$$\tilde{d} = \frac{\sin(\alpha + \frac{\theta}{2})}{\sin(\theta)} e_a - \frac{\sin(\alpha - \frac{\theta}{2})}{\sin(\theta)} e_n \,, \qquad \hat{d} = \frac{\tilde{d}}{\left\|\tilde{d}\right\|} \tag{2}$$

where

$$\alpha = \arccos(t), \qquad \theta = \arccos(e_a \odot e_n). \tag{3}$$

Then we need to account for different spreads across dimensions, which is important for representing $\mathcal{N}(P)$ as a subspace rather than a single embedding $\hat{d}$. Let $e_I^{\parallel}$ denote the component of the image embedding on the 2D plane spanned by e_n and e_a, and let $e_I^{\perp}$ denote the component of e_I perpendicular to this plane. Borrowing concepts from Riemannian geometry, we come up with a closed form of the Mahalanobis distance on a curved surface as follows:

$$D^2 = \frac{\gamma^2}{\sin^2(\gamma)} \left(\frac{\cos(\phi)^2 \sin^2(\psi)}{\sigma_1^2} + \frac{\sin^2(\phi)}{\sigma_2^2} \right) \tag{4}$$

where

$$\gamma = \arccos\left(e_I \odot \hat{d}\right), \quad \phi = \arccos\left(\left\|e_I^{\parallel}\right\|\right), \quad \psi = \arccos\left(\frac{e_I^{\parallel} \odot \hat{d}}{\left\|e_I^{\parallel}\right\|}\right), \quad \sigma_1 = \frac{\theta}{2}, \quad \sigma_2 = \alpha \tag{5}$$

Intuitively, we measure the Mahalanobis distance of a point (image embedding) to a subspace, noting that the spread of that subspace in the e_a-e_n plane (i.e., σ_1) differs from other dimensions, $\sigma_2 = \alpha$. *The key point is that this subspace could be any compact set resulting from the intersection, union, or inclusion of multiple caps.* That is why this scoring role can be generalized to other logical operators. We include a detailed proof and derivation of $\hat{d}$ and D in Appendix A. We will include the proof in the main text for the camera-ready version. Algorithm 1 provides the pseudocode of our subspace method. When $\theta > 2\alpha$, the affirmative and negative regions are already separated, so $\hat{d} = e_a$ and $\sigma_1 = \alpha$. If we don't have any negated concepts (pure affirmation), $\hat{d} = e_a$, which is the vanilla CLIP inference setup. If we don't have any affirmative concept (pure negation), we found that using a general template for A like "*A photo of something*" yields more robust results due to concentration of meaningful concepts in CLIP embedding space. Also, when $\theta \ll 1$, we approximate $\hat{d}$ and D^2 via a first-order Taylor expansion for numerical stability, as shown in Appendix A.

Algorithm 1: PyTorch-style pseudocode of SpaceVLM. Part A computes the center $\hat{d}$ of the target subspace $\mathcal{N}(P)$ (Eqs. 2). Part B scores an image by its anisotropic geodesic (Mahalanobis) distance to $\hat{d}$ (Eq. 4).

```
# Input : caption, image, Splitter, text_encoder, image_encoder,
#         cosine threshold t in (0, 1)
# Output: score, negation-aware image-caption score (higher is better)

# ------ Part A: center of the target subspace N(P) ------

# 1. Split the caption into its affirmative and negated parts
aff_cap, neg_cap = Splitter(caption)            # either part may be None
if aff_cap is None:                          # pure negation
    aff_cap = "A photo of something"

# 2. Encode with the frozen VLM text encoder  (Eq. 1)
e_a = normalize(text_encoder(aff_cap))
alpha = arccos(t)                            # angular radius of each cap
if neg_cap is not None:
    e_n = normalize(text_encoder(neg_cap))
    theta = arccos(e_a . e_n)

# 3. Center d_hat   (Eqs. 2, 5)
overlap = neg_cap is not None and theta < 2 * alpha
if not overlap:                              # no overlap: vanilla center
    d_hat, sigma_1 = e_a, alpha
else:                                        # for theta << 1, see App. A
    d_hat = sin(alpha + theta/2) * e_a - sin(alpha - theta/2) * e_n
    d_hat = normalize(d_hat / sin(theta))
    sigma_1 = theta / 2                      # half-length of in-plane arc
sigma_2 = alpha                              # off-plane spread

# ------------- Part B: Mathcing score of the image -------------

# 4. Encode with the frozen VLM image encoder
e_I = normalize(image_encoder(image))

# 5. Geodesic angle to d_hat and log-map factor  (Eq. 5)
cos_g = dot(e_I, d_hat)
gamma = arccos(cos_g)
k = gamma / sin(gamma)

# 6. Mahalanobis distance in the tangent space at d_hat  (Eq. 4)
if not overlap:                              # isotropic: vanilla VLM ranking
    D2 = (gamma / sigma_2)**2
else:
    e_n_perp = normalize(e_a - dot(e_a, e_n) * e_n) # basis {e_n, e_n_perp}
    e_par  = dot(e_I, e_n) * e_n + dot(e_I, e_n_perp) * e_n_perp
    e_perp = e_I - e_par                    # in-plane / off-plane parts of e_I
    v_1 = k * norm(e_par - cos_g * d_hat)   # = k cos(phi) sin(psi), in-plane
    v_2 = k * norm(e_perp)                  # = k sin(phi), off-plane
    D2 = (v_1 / sigma_1)**2 + (v_2 / sigma_2)**2

# 7. Score
score = -D2
```

4 EXPERIMENTS

We evaluate the effectiveness of our approach for enhancing negation understanding across multiple VLMs.

4.1 EVALUATION PROTOCOL

Tasks. Following the *NegBench* Alhamoud et al. (2025), we assess negation understanding on two tasks: (i) **Image/Video Retrieval** with *negated* queries, and (ii) **Text Retrieval (MCQ)** with *negated* captions. The negated retrieval task measures coarse-grained negation understanding: given a negated query such as, "*A photo of a dog not on grass*", the model must retrieve relevant images or videos (Figure 2a). The MCQ task measures fine-grained negation understanding: given an image, the model selects the correct caption among four closely related candidates drawn from *Affirmation*, *Negation*, and *Hybrid* templates. An example is visualized in Figure 2c.

Table 2: MCQ results showing the effect of our approach on each VLM across Affirmative, Negation, and Hybrid templates. Our method achieves an average improvement of over 30% across all datasets.

Dataset	Metric ↑	CLIP		CLIP-NegFull		ConCLIP		NegCLIP		NegCLIP-NegFull		AIMV2		SigLIP-2	
		Base	+ **Ours**	Base	+ **Ours**	Base	+ **Ours**	Base	+ **Ours**	Base	+ **Ours**	Base	+ **Ours**	Base	+ **Ours**
COCO	Affirmative	70.0	**78.4**	73.1	**84.2**	15.6	**62.7**	49.2	**69.6**	81.0	**80.0**	52.1	**78.5**	45.8	**72.4**
	Negation	6.6	**69.1**	33.2	**68.0**	32.9	**74.1**	13.9	**75.3**	25.9	**72.6**	16.6	**66.6**	9.2	**82.7**
	Hybrid	38.4	**52.6**	54.7	**58.6**	25.3	**52.3**	16.3	**54.8**	60.1	**56.3**	30.5	**46.8**	32.8	**65.8**
	AVG	39.2	(+27.5%) **66.7**	54.2	(+16.2) **70.4**	24.4	(+38.3) **62.7**	26.8	(+39.6) **66.4**	56.5	(+13.1) **69.6**	33.5	(+30.5) **64.0**	29.8	(+43.6) **73.4**
VOC2007	Affirmative	80.9	**83.4**	85.0	**90.8**	24.8	**71.4**	70.5	**79.8**	81.0	**89.1**	64.5	**87.1**	55.9	**86.4**
	Negation	3.0	**83.4**	31.7	**78.8**	23.2	**81.4**	4.6	**90.0**	21.1	**83.2**	9.1	**72.4**	3.8	**89.3**
	Hybrid	58.0	**77.8**	79.5	**81.6**	56.7	**82.0**	42.3	**82.4**	83.7	**83.6**	42.6	**69.7**	39.9	**90.5**
	AVG	37.9	(+43.0%) **80.9**	60.1	(+21.6) **81.7**	38.2	(+42.1) **80.3**	30.2	(+55.0) **85.2**	58.2	(+26.0) **84.2**	31.4	(+41.8) **73.2**	26.8	(+62.7) **89.5**
MSR-VTT	Affirmative	60.9	**82.1**	81.8	**87.5**	57.3	**80.9**	47.5	**78.2**	75.8	**85.4**	57.3	**82.7**	52.2	**79.1**
	Negation	15.6	**34.0**	21.2	**26.1**	30.3	**41.1**	13.6	**51.0**	22.1	**38.2**	6.0	**41.6**	8.8	**48.4**
	Hybrid	18.3	**58.0**	35.6	**62.2**	38.8	**70.8**	16.3	**66.0**	35.3	**65.1**	22.8	**60.9**	21.2	**69.9**
	AVG	31.6	(+26.0%) **57.6**	46.0	(+11.9) **57.9**	42.0	(+21.7) **63.7**	25.8	(+39.0) **64.8**	44.2	(+18.2) **62.4**	28.4	(+33.0) **61.4**	27.2	(+38.2) **65.4**

For medical VLMs, NegBench includes a simplified MCQ task providing a binary choice between negated and affirmative captions: for instance, "*This image shows Lung Opacity*" vs. "*This image does not show Lung Opacity.*"

Datasets. For negated Image/Video Retrieval, we use the negated extensions of COCO Lin et al. (2014) and MSR-VTT Xu et al. (2016) provided by NegBench Alhamoud et al. (2025). For MCQ, samples are drawn from COCO, VOC-2007 Everingham et al. (2010), and MSR-VTT. For MCQ in the medical domain, we use negated CheXpert Irvin et al. (2019); Alhamoud et al. (2025). One interesting point regarding NegBench is that it paraphrases negated captions in a natural tone using Llama-3 Grattafiori et al. (2024), thereby transforming many negated captions in both retrieval and MCQ tasks into different negation forms (Appendix F). So, the Splitter's role as a rephraser is critical, as accuracy depends on correctly extracting the hidden affirmative and negative parts.

Metrics. For retrieval, we report Recall@K (R@K for $K \in \{1, 5, 10\}$), measuring the fraction of queries where at least one relevant image or video appears in the top-K results. We report performance for both standard (affirmative) and negated queries. For MCQ, we report accuracy, decomposed by the template of the correct answer (*Affirmation*, *Negation*, *Hybrid*), to expose performance gaps between affirmative and negated captions. For binary MCQs on medical VLMs, we report accuracy only, since there are only two possible options for each image.

Hyperparameters. The similarity threshold t is tuned per dataset on validation splits. As shown in Appendix D, optimal t values lie in $[0.90, 0.95]$, and performance is robust within this range (negligible degradation), enabling simple hand-set choices of t in new downstream applications. Since the splitter task is not hard, we use Mistral-7B-v0.3 Jiang et al. (2023) as a cost-effective, accurate choice, but anyone can use any other splitter (as our contribution is subspace modeling, not the optimal choice of splitter), including a rule-based one. But we found that rule-based splitters are often unable to detect some kinds of negation, such as implicit negation. In Appendix E.1, we compare lightweight LLMs of different sizes in terms of final performance and inference time. We also include a complementary experiment evaluating the performance of LLMs versus manual splitting of captions in Appendix E.2.

Baselines. We evaluate our method added to nine models spanning both pretrained and fine-tuned VLMs. Pretrained VLMs include CLIP Radford et al. (2021), AIMV2 Fini et al. (2025), and SigLIP-2 Tschannen et al. (2025). Fine-tuned variants include **CLIP-NegFull** (fine-tuned on CC12M-NegFull Alhamoud et al. (2025)), **ConCLIP** (fine-tuned on CC-Neg Singh et al. (2025)), **NegCLIP** (fine-tuned on COCO with hard negative captions), and **NegCLIP-NegFull** (fine-tuned on CC12M-NegFull). For the medical MCQ task, we use **BiomedCLIP** Zhang et al. (2023). We apply our method directly to each baseline and report results with and without our modification to isolate its effect. Unless otherwise stated, all models use the ViT-B/32 backbone for consistency.

4.2 Evaluation on NegBench

Image/Video Retrieval. Table 1 reports results for standard retrieval (R@K; non-negated queries) and negated retrieval (R@K–Neg; negated queries). The R@K accuracy of each model serves as an approximate upper bound for its negated counterpart R@K–Neg.

SpaceVLM improves retrieval across all baselines and datasets, substantially closing the gap between affirmative and negated queries. In some cases, retrieval performance on negated queries even exceeds that of the base model on standard queries, as the additional negation information helps disambiguate similar images and pick the most accurate one. Importantly, performance on affirmative queries remains unchanged, confirming that our scoring rule preserves the original model behavior on non-negated prompts.

MCQ. Table 2 shows MCQ results. Across all models and datasets, SpaceVLM has large gains, especially when the correct caption follows a *Negation* template. More surprisingly, it also improves accuracy when the correct caption follows an *Affirmation* template. This is because it reduces confusion with other templates. For example, vanilla CLIP maps both captions "*a photo of a fish and coral*" and "*a photo of a fish but not coral*" to nearly identical embeddings, which causes the model to select them interchangeably (Figure 2c). With our geometric scoring, these captions become clearly separable. Notably, when applied to vanilla CLIP, our method outperforms several fine-tuned baselines trained specifically for negation understanding, such as CLIP-NegFull. *Since MCQ is the most fine-grained and challenging diagnostic test in NegBench Alhamoud et al. (2025), this result demonstrates that our approach resolves the core failure mode of VLMs without fine-tuning.*

Binary MCQs and medical VLMs. We apply our method to improve the accuracy of BiomedCLIP given medical negations in the CheXpert MCQ task Alhamoud et al. (2025). The control task includes affirmative captions only, whereas the negated dataset includes both an affirmative and a negated caption for each image. The results below suggest that SpaceVLM readily generalizes to specialized domains, such as healthcare.

Model	CheXpert-Control	CheXpert-Negation
BiomedCLIP	67.9	40.6
$\rightarrow$ **+ Ours**	**67.9**	**67.4** (+26.8)

Comparison with Recent Works DCSM Kang et al. (2025) is a recent method that also targets the inability of joint embedding-based VLMs, such as CLIP, to represent negation using a single vector. An important step in DCSM is learning a projection layer specific to each model–dataset pair, requiring additional training for every new domain. Another recent work, NegationCLIP Park et al. (2025), introduces a training dataset for negation and fine-tunes CLIP on it. Also, Aggarwal et al. (2026) is another recent work worth comparing, but we found a huge incompatibility between their reported results and our implementation, and we cannot reproduce it. Accordingly, we omit their results here.

We evaluate NegationCLIP on NegBench MCQ tasks, along with reported DCSM results on COCO and VOC-2007 MCQ tasks. Because the reported DCSM results use CLIP ViT-B/16, we use the same ViT-B/16 backbone across all models in this experiment. As shown below, our SpaceVLM scoring achieves substantially higher accuracy on both datasets with no training required.

Method	COCO	VOC2007
DCSM Kang et al. (2025)	48.6	49.0
NegationCLIP Park et al. (2025)	29.8	38.8
SpaceVLM (ours)	**68.6**	**77.6**

5 Conclusions and Limitations (Future Directions)

We proposed a subspace modeling framework within the joint embedding space for Vision-Language Models (VLMs) to address their architectural limitations in understanding logical operators, particularly focusing on *negation* as our case study. By drawing on concepts from Riemannian geometry, we theoretically derived a closed-form Mahalanobis distance on a curved surface, which forms the foundation of our framework. In addition to the performance improvements achieved by SpaceVLM, we see potential for two future directions that we encourage researchers to explore: (i) benchmarking our framework against other logical operators, including but not limited to *intersection*, *union*, and *inclusion*, and (ii) incorporating subspace modeling during the pretraining stage, since our distance metric could also be utilized as a loss function for training the model as well.

AI USE STATEMENT

In this paper, we used AI tools, Claude and Grammarly, to polish the draft text and proofs. However, all content in the text and proofs is original, as the AI only polished grammar, spelling, and readability. We did not use AI for any other aspects.

REPRODUCIBILITY STATEMENT

We release a self-contained implementation of the proposed scoring rule and of every experiment in the paper as supplementary material. All experiments are produced by a single command (bash scripts/reproduce.sh). Note that datasets and model weights are not redistributed; the provided scripts download them from their original sources and verify them against the included checksums.

REFERENCES

Bhuvan Aggarwal, Amit More, Mudit Soni, and Srinivasa Divakar Bhat. Seeing what's not there: negation understanding needs more than training. In *International Conference on Learning Representations*, volume 2026, pp. 69200–69217, 2026.

Kumail Alhamoud, Shaden Alshammari, Yonglong Tian, Guohao Li, Philip HS Torr, Yoon Kim, and Marzyeh Ghassemi. Vision-language models do not understand negation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29612–29622, 2025.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Leandro von Werra, and Thomas Wolf. Smollm - blazingly fast and remarkably powerful, 2024.

Tayfun Alpay, Sven Magg, Philipp Broze, and Daniel Speck. Multimodal video retrieval with clip: a user study. *Information Retrieval Journal*, 26(1):6, 2023.

Usha Bhalla, Alexander X. Oesterling, Suraj Srinivas, Flávio du Pin Calmon, and Himabindu Lakkaraju. Interpreting clip with sparse linear concept embeddings (splice). *ArXiv*, abs/2402.10376, 2024. URL https://api.semanticscholar.org/CorpusID:267740469.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3558–3568, 2021.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338, 2010.

Enrico Fini, Mustafa Shukor, Xiujun Li, Philipp Dufter, Michal Klein, David Haldimann, Sai Aitharaju, Victor G Turrisi da Costa, Louis Béthune, Zhe Gan, et al. Multimodal autoregressive pre-training of large vision encoders. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9641–9654, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 590–597, 2019.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. Text encoders bottleneck compositionality in contrastive vision-language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4933–4944, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.301. URL https://aclanthology.org/2023.emnlp-main.301/.

Raphi Kang, Yue Song, Georgia Gkioxari, and Pietro Perona. Is clip ideal? no. can we fix it? yes! *ICCV*, 2025.

Konstantin Klemmer, Esther Rolf, Caleb Robinson, Lester Mackey, and Marc Rußwurm. Satclip: Global, general-purpose location embeddings with satellite imagery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 4347–4355, 2025.

Jihoon Kwon, Kyle Min, and Jy-yong Sohn. Enhancing compositional reasoning in clip via reconstruction and alignment of text descriptions. *arXiv preprint arXiv:2510.16540*, 2025.

Martha Lewis, Nihal V Nayak, Peilin Yu, Qinan Yu, Jack Merullo, Stephen H Bach, and Ellie Pavlick. Does clip bind concepts? probing compositionality in large image models. *arXiv preprint arXiv:2212.10537*, 2022.

Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, and Deva Ramanan. Genai-bench: Evaluating and improving compositional text-to-visual generation. *ArXiv*, abs/2406.13743, 2024. URL https://api.semanticscholar.org/CorpusID:270619531.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Siting Li, Pang Wei Koh, and Simon Shaolei Du. Exploring how generative mllms perceive more than clip with the same vision encoder. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10101–10119, 2025.

Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. LLM-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=hFALpTb4fR. Featured Certification.

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. Springer, 2014.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature medicine*, 30(3):863–874, 2024.

Christian Lülf, Denis Mayr Lima Martins, Marcos Antonio Vaz Salles, Yongluan Zhou, and Fabian Gieseke. Clip-branches: Interactive fine-tuning for text-image retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2719–2723, 2024.

Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*, 2021.

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. Crepe: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10910–10921, 2023.

Junsung Park, Jungbeom Lee, Jongyoon Song, Sangwon Yu, Dahuin Jung, and Sungroh Yoon. Know" no"better: A data-driven approach for enhancing negation awareness in clip. *arXiv preprint arXiv:2501.10913*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Ali Rasekh, Sepehr Kazemi Ranjbar, and Simon Gottschalk. Multi-rationale explainable object recognition via contrastive conditional inference. *arXiv preprint arXiv:2508.14280*, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Claude E. Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27:623–656, 1948. URL https://api.semanticscholar.org/CorpusID:55379485.

Jaisidh Singh, Ishaan Shrivastava, Mayank Vatsa, Richa Singh, and Aparna Bharati. Learning the power of "no": Foundation models with negations. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 8002–8012. IEEE, 2025.

Ming Tao, Bing-Kun Bao, Hao Tang, and Changsheng Xu. Galip: Generative adversarial clips for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14214–14223, 2023.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Mistral AI Team. Mistral-7b-instruct-v0.3. https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3, 2024a. Instruct-finetuned version of Mistral-7B v0.3.

Qwen Team. Qwen2.5: A party of foundation models, September 2024b. URL https://qwenlm.github.io/blog/qwen2.5/.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022.

Kung-Hsun Weng, Chung-Feng Liu, and Chia-Jung Chen. Deep learning approach for negation and speculation detection for automated important finding flagging and extraction in radiology report: internal validation and technique comparison study. *JMIR Medical Informatics*, 11:e46348, 2023.

Tsung-Han Wu, Long Lian, Joseph E Gonzalez, Boyi Li, and Trevor Darrell. Self-correcting llm-controlled diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6327–6336, 2024.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5288–5296, 2016.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*, 2022.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18123–18133, 2022.

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.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5625–5644, 2024a.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*, 2024b. URL https://arxiv.org/abs/2401.02385.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.

Jitian Zhao, Chenghui Li, Frederic Sala, and Karl Rohe. Quantifying structure in clip embeddings: A statistical framework for concept interpretation. *arXiv preprint arXiv:2506.13831*, 2025.

Zihao Zhao, Yuxiao Liu, Han Wu, Mei Wang, Yonghao Li, Sheng Wang, Lin Teng, Disheng Liu, Zhiming Cui, Qian Wang, et al. Clip in medical imaging: A comprehensive survey. *arXiv preprint arXiv:2312.07353*, 2023.

Yuchen Zhou, Jiayu Tang, Shuo Yang, Xiaoyan Xiao, Yuqin Dai, Wenhao Yang, Chao Gou, Xiaobo Xia, and Tat-Seng Chua. Logic unseen: Revealing the logical blindspots of vision-language models. *arXiv preprint arXiv:2508.11317*, 2025.

A DERIVATION OF $\hat{d}$ AND D

In this section, we provide a detailed theoretical proof of the SpaceVLM scoring rule. We first prove symmetric properties of the target region, then compute the region's center direction and its spread across dimensions. Finally, we prove the Mahalanobis distance we introduced in Eq. 4.

Notation and assumptions. All embeddings are ℓ_2-normalized, so they lie on the unit sphere $\mathbb{S}^{d-1} = \{z \in \mathbb{R}^d : \|z\| = 1\}$, and $x \odot z$ denotes the inner product. We work on the sphere, so $\mathcal{N}(x) = \{z \in \mathbb{S}^{d-1} : x \odot z \geq t\}$ is a spherical cap of angular radius $\alpha = \arccos(t)$. We assume $t \in (0, 1)$, i.e. $\alpha \in (0, \pi/2)$, and $\theta = \arccos(e_a \odot e_n) \in (0, \pi)$, i.e. $e_a \neq \pm e_n$. Let $\mathcal{E} = \operatorname{span}\{e_a, e_n\}$ be the two-dimensional plane spanned by the two text embeddings, $\Pi_{\mathcal{E}}$ the orthogonal projection onto it, and $\mathcal{C} = \mathcal{E} \cap \mathbb{S}^{d-1}$ the great circle through e_a and e_n. As the orthonormal basis of $\mathcal{E}$ we use e_n and the unit vector in $\mathcal{E}$ perpendicular to e_n,

$$\bar{e}_n = \frac{e_a - \cos\theta\, e_n}{\sin\theta}, \qquad \bar{e}_n \odot e_n = 0, \quad \|\bar{e}_n\| = 1, \quad \bar{e}_n \odot e_a = \sin\theta, \tag{6}$$

which points from e_n towards e_a. Hence $\Pi_{\mathcal{E}} z = (z \odot e_n)\, e_n + (z \odot \bar{e}_n)\, \bar{e}_n$. The target region is $\mathcal{N}(P) = \mathcal{N}(e_a) \cap \mathcal{N}^c(e_n)$.

A.1 THE CENTER $\hat{d}$ OF $\mathcal{N}(P)$

Lemma 1 (Symmetry). *Let $R \in \mathbb{R}^{d\times d}$ be orthogonal with $Rx = x$ for all $x \in \mathcal{E}$. Then $R\mathcal{N}(P) = \mathcal{N}(P)$. Moreover, the only points of $\mathbb{S}^{d-1}$ that are fixed by every such R are the points of $\mathcal{C}$.*

Proof. Since $R^\top = R^{-1}$ also fixes $\mathcal{E}$, $(Rz) \odot e_a = z \odot R^\top e_a = z \odot e_a$ and likewise $(Rz) \odot e_n = z \odot e_n$. Membership in $\mathcal{N}(e_a)$ and in $\mathcal{N}^c(e_n)$ depends only on these two inner products, so $z \in \mathcal{N}(P) \iff Rz \in \mathcal{N}(P)$. Points of $\mathcal{C}$ are fixed by definition. If $z \in \mathbb{S}^{d-1} \setminus \mathcal{C}$, then $z_\perp = z - \Pi_{\mathcal{E}} z \neq 0$, and the reflection $R = I - 2ww^\top$ with $w = z_\perp/\|z_\perp\| \in \mathcal{E}^\perp$ fixes $\mathcal{E}$ but sends z to $z - 2z_\perp \neq z$. □

Lemma 1 says that $\mathcal{N}(P)$ looks the same in every direction orthogonal to $\mathcal{E}$. Hence any center that is defined by the region alone must be fixed by all these symmetries, and so it must lie on the great circle $\mathcal{C}$. In the basis $\{e_n, \bar{e}_n\}$ of equation 6, we parameterize $\mathcal{C}$ by the signed angle s from e_n, positive towards e_a:

$$c(s) = \cos s\; e_n + \sin s\; \bar{e}_n, \qquad s \in (\theta - \pi,\, \theta + \pi]. \tag{7}$$

Then $\|c(s)\| = 1$, $c(0) = e_n$ and $c(\theta) = e_a$. Substituting equation 6 gives

$$c(s) = \frac{\sin(\theta - s)}{\sin\theta}\, e_n + \frac{\sin s}{\sin\theta}\, e_a, \qquad c(s) \odot e_n = \cos s, \qquad c(s) \odot e_a = \cos(\theta - s). \tag{8}$$

Lemma 2 (The region on the great circle). *Let $s_{\min} = \max(\alpha,\ \theta - \alpha)$. Up to its end points, $\mathcal{N}(P) \cap \mathcal{C} = \{c(s) : s_{\min} < s \leq \theta + \alpha\}$, a single non-empty arc.*

Proof. Since $s - \theta \in (-\pi, \pi]$, equation 8 gives $c(s) \in \mathcal{N}(e_a) \iff \cos(\theta - s) \geq \cos\alpha \iff |s - \theta| \leq \alpha$. For such s, $s \in [\theta - \alpha, \theta + \alpha]$, and this interval lies in $(-\alpha,\ 2\pi - \alpha)$ because $\theta > 0$ and $\theta + 2\alpha < 2\pi$ (as $\theta < \pi$, $\alpha < \pi/2$). On $(-\alpha, 2\pi - \alpha)$ we have $\cos s < \cos\alpha \iff s > \alpha$. Hence $c(s) \notin \mathcal{N}(e_n) \iff s > \alpha$. Combining both conditions gives $\max(\alpha, \theta - \alpha) < s \leq \theta + \alpha$. The arc is non-empty because $\theta + \alpha > \alpha$ and $\theta + \alpha > \theta - \alpha$. □

Proposition 3 (Center and in-plane spread). *Define $\hat{d}$ as the midpoint of the arc in Lemma 2, and σ_1 as its half-length. Then:*

1. *If $\theta < 2\alpha$ (overlapping caps), $\hat{d}$ is given by equation 2, with $\|\tilde{d}\| = 1$, and $\sigma_1 = \theta/2$.*
2. *If $\theta \geq 2\alpha$ (separated caps), $\hat{d} = e_a$ and $\sigma_1 = \alpha$.*

The two cases agree at $\theta = 2\alpha$, so $\hat{d}$ and σ_1 are continuous in θ.

Proof. The arc runs from $s = s_{\min}$ to $s = \theta + \alpha$, so its midpoint is $s^* = (s_{\min} + \theta + \alpha)/2$ and its half-length is $\sigma_1 = (\theta + \alpha - s_{\min})/2$. (1) If $\theta < 2\alpha$, then $s_{\min} = \alpha$, so $s^* = \theta/2 + \alpha$ and $\sigma_1 = \theta/2$. By equation 8,

$$c(s^*) = \frac{\sin(\frac{\theta}{2} - \alpha)}{\sin\theta}\, e_n + \frac{\sin(\frac{\theta}{2} + \alpha)}{\sin\theta}\, e_a = \tilde{d}.$$

Since $c(s^*) \in \mathbb{S}^{d-1}$, $\|\tilde{d}\| = 1$ and $\hat{d} = \tilde{d}$. The normalization in equation 2 therefore only removes round-off error. (2) If $\theta \geq 2\alpha$, then $s_{\min} = \theta - \alpha$, so $s^* = \theta$, giving $\hat{d} = c(\theta) = e_a$, and $\sigma_1 = \alpha$. At $\theta = 2\alpha$ both cases give $s^* = \theta$ and $\sigma_1 = \alpha$. □

Corollary 4 (Bisector form of $\hat{d}$)**.** *For $\theta < 2\alpha$, let $\hat{m} = (e_a + e_n)/\|e_a + e_n\|$ (the bisector of e_a and e_n) and $\hat{h} = (e_a - e_n)/\|e_a - e_n\|$. Then $\hat{m} \perp \hat{h}$ and*

$$\hat{d} = \cos\alpha\ \hat{m} + \sin\alpha\ \hat{h}. \tag{9}$$

That is, $\hat{d}$ lies on $\mathcal{C}$ at angle α from the bisector, on the side of e_a.

Proof. Expand $\sin(\alpha \pm \frac{\theta}{2}) = \sin\alpha\cos\frac{\theta}{2} \pm \cos\alpha\sin\frac{\theta}{2}$ in equation 2 and use $\sin\theta = 2\sin\frac{\theta}{2}\cos\frac{\theta}{2}$:

$$\tilde{d} = \sin\alpha\, \frac{e_a - e_n}{2\sin\frac{\theta}{2}} + \cos\alpha\, \frac{e_a + e_n}{2\cos\frac{\theta}{2}}.$$

Finally, $\|e_a - e_n\|^2 = 2 - 2\cos\theta = 4\sin^2\frac{\theta}{2}$, $\|e_a + e_n\|^2 = 4\cos^2\frac{\theta}{2}$, and $(e_a + e_n) \odot (e_a - e_n) = 0$. □

Remark 5 (Small θ). For $\theta \ll 1$, both coefficients in equation 2 are close to $0/0$, and $\theta = \arccos(e_a \odot e_n)$ is inaccurate because $\arccos$ is steep near 1. The first-order Taylor expansion $\sin\frac{\theta}{2} \approx \frac{\theta}{2}$, $\cos\frac{\theta}{2} \approx 1$ in the proof of Corollary 4 gives

$$\tilde{d} \approx \frac{\sin\alpha}{\theta}\,(e_a - e_n) + \frac{\cos\alpha}{2}\,(e_a + e_n), \tag{10}$$

with error $O(\theta^2)$, followed by $\hat{d} = \tilde{d}/\|\tilde{d}\|$. Computing $\theta = 2\arcsin(\|e_a - e_n\|/2)$ instead of via $\arccos$ keeps θ accurate for small angles. The bisector form equation 9 is exact and also stable, since it only uses $e_a \pm e_n$ and their norms.

A.2 The distance D

Let $e_I \in \mathbb{S}^{d-1}$ be an image embedding with $e_I \neq -\hat{d}$. We split it as

$$e_I^{\parallel} = \Pi_{\mathcal{E}}\, e_I = (e_I \odot e_n)\, e_n + (e_I \odot \bar{e}_n)\, \bar{e}_n, \qquad e_I^{\perp} = e_I - e_I^{\parallel}, \qquad \|e_I^{\parallel}\|^2 + \|e_I^{\perp}\|^2 = 1, \tag{11}$$

and use $\gamma \in [0, \pi)$, $\phi \in [0, \pi/2]$ and $\psi \in [0, \pi]$ from equation 5. Thus $\sin\phi = \|e_I^{\perp}\|$, ϕ is the angle between e_I and the plane $\mathcal{E}$, and ψ is the angle inside $\mathcal{E}$ between the shadow $e_I^{\parallel}$ and $\hat{d}$. If $e_I^{\parallel} = 0$ ($\phi = \pi/2$), ψ is undefined; every statement below still holds for any value of ψ, since ψ only appears multiplied by $\cos\phi = 0$.

Lemma 6 (Spherical Pythagorean theorem)**.** $\cos\gamma = \cos\psi\ \cos\phi$.

Proof. $\hat{d} \in \mathcal{E}$ and $e_I^{\perp} \perp \mathcal{E}$, so $e_I^{\perp} \odot \hat{d} = 0$. Hence $\cos\gamma = e_I \odot \hat{d} = e_I^{\parallel} \odot \hat{d} = \|e_I^{\parallel}\| \cos\psi = \cos\phi\cos\psi$. □

Lemma 7 (Local coordinates around $\hat{d}$)**.** *Let $\hat{t} \in \mathcal{E}$ be the unit vector with $\hat{t} \perp \hat{d}$ and $e_I^{\parallel} \odot \hat{t} \geq 0$. Let $\hat{w} = e_I^{\perp}/\|e_I^{\perp}\|$ if $\phi > 0$, and any unit vector in $\mathcal{E}^{\perp}$ otherwise. Then $\{\hat{d}, \hat{t}, \hat{w}\}$ is orthonormal and*

$$e_I = \cos\phi\,\big(\cos\psi\,\hat{d} + \sin\psi\,\hat{t}\big) + \sin\phi\,\hat{w}. \tag{12}$$

Proof. $\{\hat{d}, \hat{t}\}$ is an orthonormal basis of $\mathcal{E}$ and $\hat{w} \in \mathcal{E}^{\perp}$. In this basis, $e_I^{\parallel} = (e_I^{\parallel} \odot \hat{d})\,\hat{d} + (e_I^{\parallel} \odot \hat{t})\,\hat{t}$ with $e_I^{\parallel} \odot \hat{d} = \cos\phi\cos\psi$ and $(e_I^{\parallel} \odot \hat{t})^2 = \|e_I^{\parallel}\|^2 - (e_I^{\parallel} \odot \hat{d})^2 = \cos^2\phi\sin^2\psi$. Both $e_I^{\parallel} \odot \hat{t}$ and $\cos\phi\sin\psi$ are non-negative, so they are equal. Adding $e_I^{\perp} = \sin\phi\,\hat{w}$ gives equation 12. □

Definition 8 (Tangent space, exponential and logarithmic maps)**.** The tangent space of the sphere at $\hat{d}$ is the flat $(d-1)$-dimensional space $T_{\hat{d}} = \{v \in \mathbb{R}^d : v \odot \hat{d} = 0\}$. The exponential map follows the great circle that leaves $\hat{d}$ in direction v for an arc length $\|v\|$:

$$\mathrm{Exp}_{\hat{d}}(v) = \cos\|v\|\ \hat{d} + \sin\|v\|\ \frac{v}{\|v\|}, \qquad \mathrm{Exp}_{\hat{d}}(0) = \hat{d}.$$

Its inverse, the logarithmic map, is defined for $z \neq -\hat{d}$ by

$$\mathrm{Log}_{\hat{d}}(z) = \kappa(\gamma_z)\left(z - \cos\gamma_z\,\hat{d}\right), \qquad \gamma_z = \arccos(z \odot \hat{d}), \qquad \kappa(\gamma) = \frac{\gamma}{\sin\gamma},\ \kappa(0) = 1. \quad (13)$$

Lemma 9 (Properties of $\mathrm{Log}_{\hat{d}}$)**.** *For $z \neq -\hat{d}$: (i) $\mathrm{Log}_{\hat{d}}(z) \in T_{\hat{d}}$; (ii) $\|\mathrm{Log}_{\hat{d}}(z)\| = \gamma_z$, the geodesic distance from $\hat{d}$ to z; (iii) $\mathrm{Exp}_{\hat{d}}(\mathrm{Log}_{\hat{d}}(z)) = z$.*

Proof. (i) $(z - \cos\gamma_z\hat{d}) \odot \hat{d} = \cos\gamma_z - \cos\gamma_z = 0$. (ii) $\|z - \cos\gamma_z\hat{d}\|^2 = 1 - 2\cos^2\gamma_z + \cos^2\gamma_z = \sin^2\gamma_z$, and $\sin\gamma_z \geq 0$ for $\gamma_z \in [0, \pi)$, so the norm is $\kappa(\gamma_z)\sin\gamma_z = \gamma_z$ (for $\gamma_z = 0$, $z = \hat{d}$ and both sides are 0). (iii) For $\gamma_z > 0$ the unit direction of $\mathrm{Log}_{\hat{d}}(z)$ is $(z - \cos\gamma_z\hat{d})/\sin\gamma_z$, so $\mathrm{Exp}_{\hat{d}}(\mathrm{Log}_{\hat{d}}(z)) = \cos\gamma_z\hat{d} + (z - \cos\gamma_z\hat{d}) = z$. □

So $\mathrm{Log}_{\hat{d}}$ flattens the sphere around $\hat{d}$ while keeping every geodesic distance to $\hat{d}$ exact. The plain projection $z - \cos\gamma_z\hat{d}$ has length $\sin\gamma_z$, which is not monotone on $[0, \pi)$ and would shrink large angles. This is why the distance is defined in the tangent space.

Proposition 10 (In-plane and off-plane parts)**.** *The tangent space splits orthogonally as $T_{\hat{d}} = L \oplus \mathcal{E}^\perp$, where $L = \mathcal{E} \cap T_{\hat{d}} = \mathrm{span}\{\hat{t}\}$ is the in-plane tangent line and $\mathcal{E}^\perp$ (dimension $d-2$) contains all off-plane directions. Moreover,*

$$\mathrm{Log}_{\hat{d}}(e_I) = v_1\,\hat{t} + v_2\,\hat{w}, \qquad v_1 = \kappa(\gamma)\cos\phi\,\sin\psi, \qquad v_2 = \kappa(\gamma)\sin\phi, \quad (14)$$

where $v_1\hat{t} = \kappa(\gamma)\left(e_I^{\parallel} - \cos\gamma\,\hat{d}\right)$ is the part in L, $v_2\hat{w} = \kappa(\gamma)\,e_I^{\perp}$ is the part in $\mathcal{E}^\perp$, and $v_1^2 + v_2^2 = \gamma^2$.

Proof. For $v \in T_{\hat{d}}$, $\Pi_{\mathcal{E}} v \odot \hat{d} = v \odot \hat{d} = 0$ (as $\hat{d} \in \mathcal{E}$), so $\Pi_{\mathcal{E}} v \in L$. This gives the orthogonal splitting. By Lemma 6, $\cos\gamma = \cos\psi\cos\phi$, so subtracting $\cos\gamma\,\hat{d}$ from equation 12 gives $e_I - \cos\gamma\,\hat{d} = \cos\phi\sin\psi\,\hat{t} + \sin\phi\,\hat{w}$. Multiplying by $\kappa(\gamma)$ gives equation 14. Finally, $v_1^2 + v_2^2 = \|\mathrm{Log}_{\hat{d}}(e_I)\|^2 = \gamma^2$ by Lemma 9(ii). □

Choice of σ_1 and σ_2. *In-plane:* for $|s - s^*| < \pi$, $\mathrm{Log}_{\hat{d}}(c(s)) = (s - s^*)\,\hat{t}$ (up to the sign of $\hat{t}$), because $c(s)$ lies on the great circle through $\hat{d}$ at geodesic distance $|s - s^*|$. Hence the arc of Lemma 2 maps exactly onto the segment $[-\sigma_1, \sigma_1]\,\hat{t}$ of L, and σ_1 from Proposition 3 is the half-width of the region along L. *Off-plane:* every $z \in \mathcal{N}(P)$ satisfies $z \in \mathcal{N}(e_a)$, so the angle between z and $\mathcal{E}$ is at most the angle between z and $e_a \in \mathcal{E}$, which is at most α. The region therefore extends at most α away from the plane in every direction of $\mathcal{E}^\perp$, and we set $\sigma_2 = \alpha$, the angular radius of the affirmative cap.

Definition 11 (Mahalanobis distance on the sphere)**.** With $\sigma_1, \sigma_2 > 0$, define the covariance operator on $T_{\hat{d}}$

$$\Sigma = \sigma_1^2\,\Pi_L + \sigma_2^2\,\Pi_{\mathcal{E}^\perp},$$

and $D^2(e_I) = \mathrm{Log}_{\hat{d}}(e_I)^\top\,\Sigma^{-1}\,\mathrm{Log}_{\hat{d}}(e_I)$, with Σ^{-1} the inverse of Σ on $T_{\hat{d}}$.

Theorem 12 (Closed form of D)**.** *For every $e_I \in \mathbb{S}^{d-1}$ with $e_I \neq -\hat{d}$,*

$$D^2(e_I) = \frac{v_1^2}{\sigma_1^2} + \frac{v_2^2}{\sigma_2^2} = \frac{\gamma^2}{\sin^2\gamma}\left(\frac{\cos^2\phi\,\sin^2\psi}{\sigma_1^2} + \frac{\sin^2\phi}{\sigma_2^2}\right),$$

which is equation 4.

Proof. Π_L and $\Pi_{\mathcal{E}^\perp}$ are complementary orthogonal projections on $T_{\hat{d}}$, so $\Sigma^{-1} = \sigma_1^{-2}\Pi_L + \sigma_2^{-2}\Pi_{\mathcal{E}^\perp}$. By Proposition 10, $\Pi_L\,\mathrm{Log}_{\hat{d}}(e_I) = v_1\hat{t}$ and $\Pi_{\mathcal{E}^\perp}\,\mathrm{Log}_{\hat{d}}(e_I) = v_2\hat{w}$ with unit $\hat{t}, \hat{w}$, so $D^2 = v_1^2/\sigma_1^2 + v_2^2/\sigma_2^2$. Substituting equation 14 gives equation 4. □

Corollary 13 (Consistency checks)**.** *Under the assumptions of Theorem 12:*

1. Isotropic spread. *If $\sigma_1 = \sigma_2 = \sigma$, then $D = \gamma/\sigma$ exactly.*
2. Pure in-plane deviation. *If $\phi = 0$, then $\gamma = \psi$ and $D = \psi/\sigma_1$ exactly.*
3. Pure off-plane deviation. *If $\psi = 0$, then $\gamma = \phi$ and $D = \phi/\sigma_2$ exactly.*
4. Small angles. *With $\varepsilon^2 = \psi^2 + \phi^2 \to 0$, $v_1 = \psi\big(1 - \frac{1}{3}\phi^2 + O(\varepsilon^4)\big)$ and $v_2 = \phi\big(1 + \frac{1}{6}\psi^2 + O(\varepsilon^4)\big)$, so $D^2 = \psi^2/\sigma_1^2 + \phi^2/\sigma_2^2$ up to a relative error of order ε^2.*

Proof. (1) $D^2 = (v_1^2 + v_2^2)/\sigma^2 = \gamma^2/\sigma^2$ by Proposition 10. (2) $\phi = 0$ gives $\cos\gamma = \cos\psi$, so $\gamma = \psi$, $v_2 = 0$ and $v_1 = \kappa(\psi)\sin\psi = \psi$. (3) $\psi = 0$ gives $\gamma = \phi$, $v_1 = 0$ and $v_2 = \kappa(\phi)\sin\phi = \phi$. (4) From $\cos\gamma = \cos\psi\cos\phi$, $\gamma^2 = \psi^2 + \phi^2 + O(\varepsilon^4)$, so $\kappa(\gamma) = 1 + \frac{1}{6}(\psi^2 + \phi^2) + O(\varepsilon^4)$. Multiply by $\cos\phi\sin\psi = \psi\big(1 - \frac{1}{6}\psi^2 - \frac{1}{2}\phi^2\big) + O(\varepsilon^5)$ and by $\sin\phi = \phi\big(1 - \frac{1}{6}\phi^2\big) + O(\varepsilon^5)$. □

Remark 14 (Special cases in Section 3.4). (a) *Separated caps* ($\theta \geq 2\alpha$): by Proposition 3, $\hat{d} = e_a$ and $\sigma_1 = \sigma_2 = \alpha$, so by Corollary 13(1), $D = \arccos(e_I \odot e_a)/\alpha$. The ranking of images is then the same as ranking by $e_I \odot e_a$. (b) *Pure affirmation* (no e_n): with $\hat{d} = e_a$ and $\sigma_1 = \sigma_2 = \alpha$, the same argument gives the vanilla CLIP ranking, and no plane $\mathcal{E}$ is needed. (c) D is undefined only at $e_I = -\hat{d}$, where every direction is a shortest path. $D = 0$ at $e_I = \hat{d}$. D does not depend on the sign of ψ or the orientation of $\hat{t}$. (d) One scalar σ_2 is shared by all $d - 2$ off-plane directions, so the model is isotropic outside $\mathcal{E}$.

Remark 15 (Numerically stable evaluation). Proposition 10 gives an equivalent form of equation 4 that needs neither ϕ nor ψ:

$$D^2 = \kappa(\gamma)^2 \left(\frac{\|e_I^{\parallel} - (e_I \odot \hat{d})\,\hat{d}\|^2}{\sigma_1^2} + \frac{\|e_I^{\perp}\|^2}{\sigma_2^2} \right), \qquad \gamma = \operatorname{atan2}\big(\|e_I - (e_I \odot \hat{d})\,\hat{d}\|,\ e_I \odot \hat{d}\big). \tag{15}$$

This avoids $\arccos$ near ± 1 and the undefined ψ at $e_I^{\parallel} = 0$. For $\gamma \ll 1$, use $\kappa(\gamma) = 1 + \gamma^2/6 + O(\gamma^4)$ instead of $\gamma/\sin\gamma$.

B Discussion on Multi-Negation Scenarios

The SpaceVLM framework naturally extends to scenarios involving multiple concepts within the affirmative set A and the negative set N. However, evaluating SpaceVLM on such cases requires a concrete benchmark for multi-negation alongside other compositional operators, which is currently lacking in the literature. Nevertheless, we provide a theoretical discussion on the SpaceVLM formulation for multiple-concept scenarios.

Consider a general case where a caption P contains K affirmative concepts, $\{A_i\}_{i\in[K]}$, and L negative concepts, $\{N_j\}_{j\in[L]}$. The semantic meaning of P can be formalized as the set difference between the aggregated affirmative and negative sets:

$$(A_1 \diamond A_2 \diamond \cdots \diamond A_K) \setminus (N_1 \diamond N_2 \diamond \cdots \diamond N_L), \tag{16}$$

where $\diamond \in \{\cup, \cap\}$ represents the requisite logical aggregation (union or intersection). Since standard VLMs like CLIP struggle with compositional reasoning Kwon et al. (2025), combining all concepts into a single caption often results in semantic collapse. Extending our subspace methodology to this setup offers a robust generalization for arbitrary compositional operators.

Recalling the notation from Section 3.4, the geometric regions for the affirmative and negative sets can be defined as the aggregation of their individual concept neighbors:

$$\mathcal{N}(A) = \mathcal{N}(A_1) \diamond \mathcal{N}(A_2) \diamond \cdots \diamond \mathcal{N}(A_K)$$
$$\mathcal{N}(N) = \mathcal{N}(N_1) \diamond \mathcal{N}(N_2) \diamond \cdots \diamond \mathcal{N}(N_L)$$

The valid semantic region for the full caption $\mathcal{N}(P)$ is then the intersection of the affirmative region and the complement of the negative region:

$$\mathcal{N}(P) = \mathcal{N}(A) \cap \mathcal{N}^c(N).$$

Table 3: The effectiveness of SpaceVLM on various model sizes.

Model	ViT-B/32	ViT-B/16	ViT-L/14
CLIP	39.2	41.4	38.5
→ + **Ours**	**66.7** (+27.5)	**68.6** (+27.2)	**70.8** (+32.3)

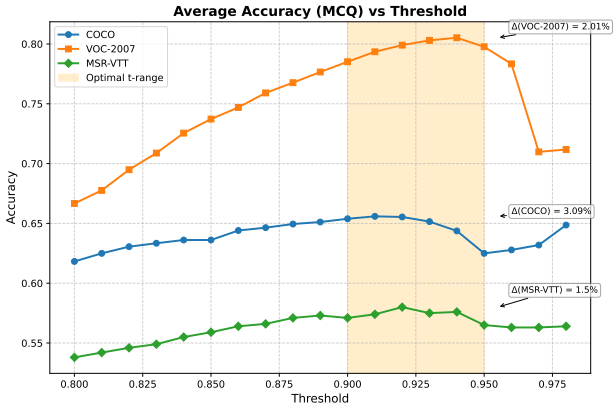


Figure 4: Effect of varying threshold t on MCQ average accuracy. The range $[0.90, 0.95]$ is near-optimal with at most a 3.09% drop on COCO. For best results, perform cross-validation on the target task.

A critical observation is that $\mathcal{N}(P)$ may not be a single continuous subspace. Instead, it may consist of a union of H disjoint continuous subspaces, such that $\mathcal{N}(P) = \bigcup_{h=1}^{H} \mathcal{N}_h(P)$, where each $\mathcal{N}_h(P)$ represents a distinct valid semantic mode. In this multimodal scenario, simply calculating a single global center $\hat{d}$ is insufficient.

Therefore, given an image x with embedding e_I, the matching score $s_{\text{neg}}(x, P)$ should be defined as the maximum similarity to any of the valid disjoint regions:

$$s_{\text{neg}}(x, P) = \max_{h \in [H]} \left(e_I \odot \hat{d}(\mathcal{N}_h(P)) \right),$$

where $\hat{d}(\mathcal{N}_h(P))$ denotes the central direction of the h-th spherical cap. The intuition behind this "*Max-Region*" scoring is that multiple distinct visual realizations may satisfy the logical conditions of the caption. Since any valid realization is acceptable, the model should reward the image for being close to the *nearest* valid subspace, rather than forcing it to align with an arbitrary mean of all possibilities.

C Varying VLM Complexity

We evaluate how VLM complexity affects our method by testing three backbones of increasing size: ViT-B/32, ViT-B/16, and ViT-L/14. Using the MCQ COCO task, the inline table below reports average results. Across all model sizes, our method consistently improves performance, indicating its robustness and applicability to VLMs of varying capacity.

D Sensitivity to the Cosine Threshold t

We analyze performance sensitivity to t on the MCQ benchmark by varying $t \in [0.90, 0.95]$ (Figure 4). The maximum drop is 3.09% (COCO), indicating robustness and enabling practical, hand-set choices of t in many applications. For the highest accuracy, we recommend cross-validation on the target task.

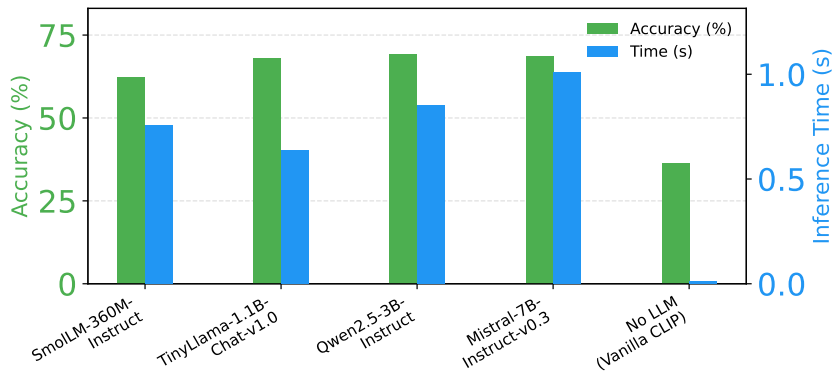


Figure 5: Accuracy vs. inference time trade-off among different LLMs as the splitter for our method.

Table 4: Isolation of the LLM effect as the splitter for Image/Video Retrieval. CLIP ViT-B/32 is used as the base model.

Dataset	Model	R@1 ↑	R@1-Neg ↑	R@5 ↑	R@5-Neg ↑	R@10 ↑	R@10-Neg ↑	Avg Δ (pp)
COCO	CLIP	29.8	25.0	54.8	47.9	66.0	59.1	
	→ + Ours (LLM)		30.8		56.1		67.4	**+7.4**
	→ + Ours (manual)		30.0		55.0		66.2	**+6.4**
MSR-VTT	CLIP	26.4	23.8	48.7	45.9	60.7	56.6	
	→ + Ours (LLM)		29.0		52.4		63.1	**+6.1**
	→ + Ours (manual)		26.4		48.8		60.7	**+3.2**

E LLM AS SPLITTER

E.1 VARYING LLM SIZE

We study how the choice of LLM, used to split captions into affirmative and negative parts, affects computation time and downstream performance. We ablate across several LLMs of varying scale and capability, evaluating SmolLM-360M-Instruct Allal et al. (2024), TinyLlama-1.1B-Chat-v1.0 Zhang et al. (2024b), Qwen2.5-3B-Instruct Team (2024b), and Mistral Team (2024a).

Figure 5 shows this comparison. We evaluate performance on the NegBench MCQ tasks by reporting SpaceVLM's average accuracy across COCO, VOC2007, and MSR-VTT. For inference time (the time required to obtain the text embedding for a given caption), we similarly average across datasets. The inference time is computed for a 32-input batch on a single H100 GPU. TinyLlama-1B provides a favorable balance between accuracy and inference speed relative to the other models, which makes it a practical choice in real-world settings.

Table 5: Isolation of LLM effect as a rephraser for MCQ. CLIP ViT-B/32 is used as the base model.

Dataset	Model	Affirmative ↑	Negation↑	Hybrid ↑	AVG ↑
COCO	CLIP	70.3	6.6	38.1	39.2
	→ + Ours (LLM)	78.4	69.1	52.6	66.7 (**+27.5**)
	→ + Ours (manual)	76.0	52.7	55.4	61.6 (**+22.4**)
VOC2007	CLIP	80.9	3.0	58.0	37.9
	→ + Ours (LLM)	83.4	83.4	77.8	80.9 (**+43.0**)
	→ + Ours (manual)	82.8	70.7	80.3	76.6 (**+38.7**)
MSR-VTT	CLIP	60.9	15.6	18.3	31.6
	→ + Ours (LLM)	82.1	34.0	58.0	57.6 (**+26.0**)
	→ + Ours (manual)	88.2	43.4	71.4	66.7 (**+35.1**)

Table 6: Examples of different negation types extracted from the rephrased COCO and MSR-VTT datasets in NegBench. The examples illustrate the diversity of linguistic structures our framework must handle, ranging from standard explicit negation to complex nested and implicit forms.

Negation Type	Example Caption	Dataset
Explicit Negation	This image features a knife, but **no** car is present.	COCO
	This image features a person, but **not** a truck.	COCO
	Several dogs are playing dead, with **no** running in sight.	MSR-VTT
Implicit Negation	This image features a refrigerator, but **lacks** a bottle.	COCO
	Spices are being mixed in a bowl; cake is noticeably **absent** from the scene.	MSR-VTT
	An animated video shows a bottle rolling across an **empty** hallway.	MSR-VTT
Quantifier Negation	A handbag is visible, but there's **no one** around.	COCO
	No one is running in the video; instead, a cartoon character nods off.	MSR-VTT
	A man is talking, surrounded by people, and **no one** appears to be an actor.	MSR-VTT
Double Negation	**Neither** a school **nor** students are visible in the video.	MSR-VTT
	A man sings **without** accompaniment, notably with **no** piano.	MSR-VTT
	One of two guys is walking on eggs, with **neither** of them running.	MSR-VTT

E.2 Comparison of LLM Splitting vs. Manual Splitting

To compare LLM splitting performance, we compare our automated pipeline against a baseline in which captions are manually separated into affirmative and negative components. Tables 4 and 5 present the results of using an LLM (Mistral-7B-Instruct) versus manual decomposition.

We observe that the LLM-based approach achieves a 6.8% improvement on Image/Video Retrieval and 32.2% on MCQ tasks. Comparatively, manual splitting yields improvements of 4.8% and 32.0%, respectively. These results demonstrate two key findings. First, off-the-shelf LLMs are highly effective at converting natural language captions into the structured "$< A >$ *but not* $< N >$" format. Second, the performance of the SpaceVLM framework is robust to the source of the rephrasing; as long as the affirmative and negative components are correctly isolated, the geometric subspace modeling delivers consistent gains.

F Examples of Different Negation Types

Table 6 presents illustrative examples extracted from the evaluation datasets representing diverse negation categories: (i) *Explicit Negation*, (ii) *Implicit Negation*, (iii) *Quantifier Negation*, and (iv) *Double Negation*.

G Application to Text-to-Image Generation (T2I)

Experimental Setup. We test whether the proposed SpaceVLM scoring improves negation adherence in T2I generation systems. CLIP text encoders are widely used in modern T2I models Rombach et al. (2022); Tao et al. (2023), yet their limited handling of negation often causes generated images to exhibit objects explicitly excluded in the prompt. **An important note: image generator models accept only text embeddings, not subspaces. This is a limitation of these models, so our approach cannot be directly assessed. Yet we still use the center of the target region, $\hat{d}$, as the text representation to perform this experiment. However, the results of this experiment do not capture the full capability of our subspace modeling.** We apply our method to GALIP Tao et al. (2023), a GAN-based generator that uses a CLIP text encoder and produces image quality comparable to Stable Diffusion Rombach et al. (2022) and Matching Flows Lipman et al. (2022). We evaluate on the 107 negated prompts introduced by Park et al. (2025), and use Gemma-3-27B-it Team et al. (2025) as an evaluator for presence/absence checks under human supervision. Table 7 reports the T2I results. Our method substantially improves negation adherence, with up to 37% higher accuracy over baselines. Intuitively, these gains arise because the subspace formulation explicitly removes the negated concept from the CLIP text embedding, enabling the generator to condition on representations that better match the intended prompt semantics (Figure 2b).

Table 7: Text-to-Image generation. *Affirmative Acc* is the mean accuracy of correctly generating the positive concept; *Negation Acc* is the mean accuracy of omitting the negated concept; *Acc* is the mean accuracy of generating and omitting the positive and negative concepts at the same time for each image.

Model	Aff-Acc ↑	Neg-Acc ↑	Acc ↑
CLIP	97.3	28.5	27.3
→ **+ Ours**	**98.8**	**60.9**	**59.7** (+32.4)
CLIP-NegFull	98.1	40.7	39.7
→ **+ Ours**	**97.4**	**64.0**	**61.8** (+22.1)
ConCLIP	27.7	68.3	11.0
→ **+ Ours**	**86.6**	**57.8**	**48.2** (+37.2)
NegCLIP	98.8	24.5	23.7
→ **+ Ours**	**98.9**	**60.6**	**59.8** (+36.1)
NegCLIP-NegFull	98.6	35.5	34.8
→ **+ Ours**	**98.0**	**63.9**	**62.3** (+27.5)
NegationCLIP	98.8	45.2	44.5
→ **+ Ours**	**97.1**	**60.7**	**58.6** (+14.1)

H VISUALIZATION

H.1 TEXT-TO-IMAGE

Figures 6 and 7 show samples of T2I generated by GALIP with negation as evaluated in Table 7. As can be seen, our method generates more aligned images compared to the vanilla baseline.

H.2 IMAGE RETRIEVAL

We compare SpaceVLM with vanilla CLIP by conducting an image-retrieval study on CIFAR-100 to evaluate both *exclusion* (retrieving images outside a negated category) and *diversity* among the retrieved results. We consider two settings: (i) vanilla prompting with "*Not a photo of a <category>*", (ii) our subspace-based negation using an affirmative prompt "*This is a photo*" combined with a negation prompt "*A photo of a <category>*".

Our goal is to retrieve images that do not belong to the specified category and that are semantically diverse. To quantify diversity, we compute the (Shannon) entropy Shannon (1948) over the CIFAR-100 categories of the top-5 retrieved images. Figure 8a reports these entropy values: our method consistently yields higher entropy than vanilla CLIP, indicating more diverse results. Moreover, decreasing the threshold t increases entropy, consistent with a larger feasible region in the embedding space. Figure 8b shows the top-5 results for the prompt "*Not a photo of a mountain*". Our approach retrieves diverse images that are not labeled "*mountain*", whereas vanilla CLIP often fails to account for negation. As t decreases, the retrieved images further diverge from the "*mountain*" category, reflecting the expanded subspace and increased diversity. To retrieve more relevant concepts related to the negated one (e.g., "*sky*" to "*mountain*"), the threshold should be kept in the range of [0.9, 0.95]. A higher threshold leads to the leakage of the negated concept in the retrieved image, while a lower threshold results in totally irrelevant retrieved concepts.

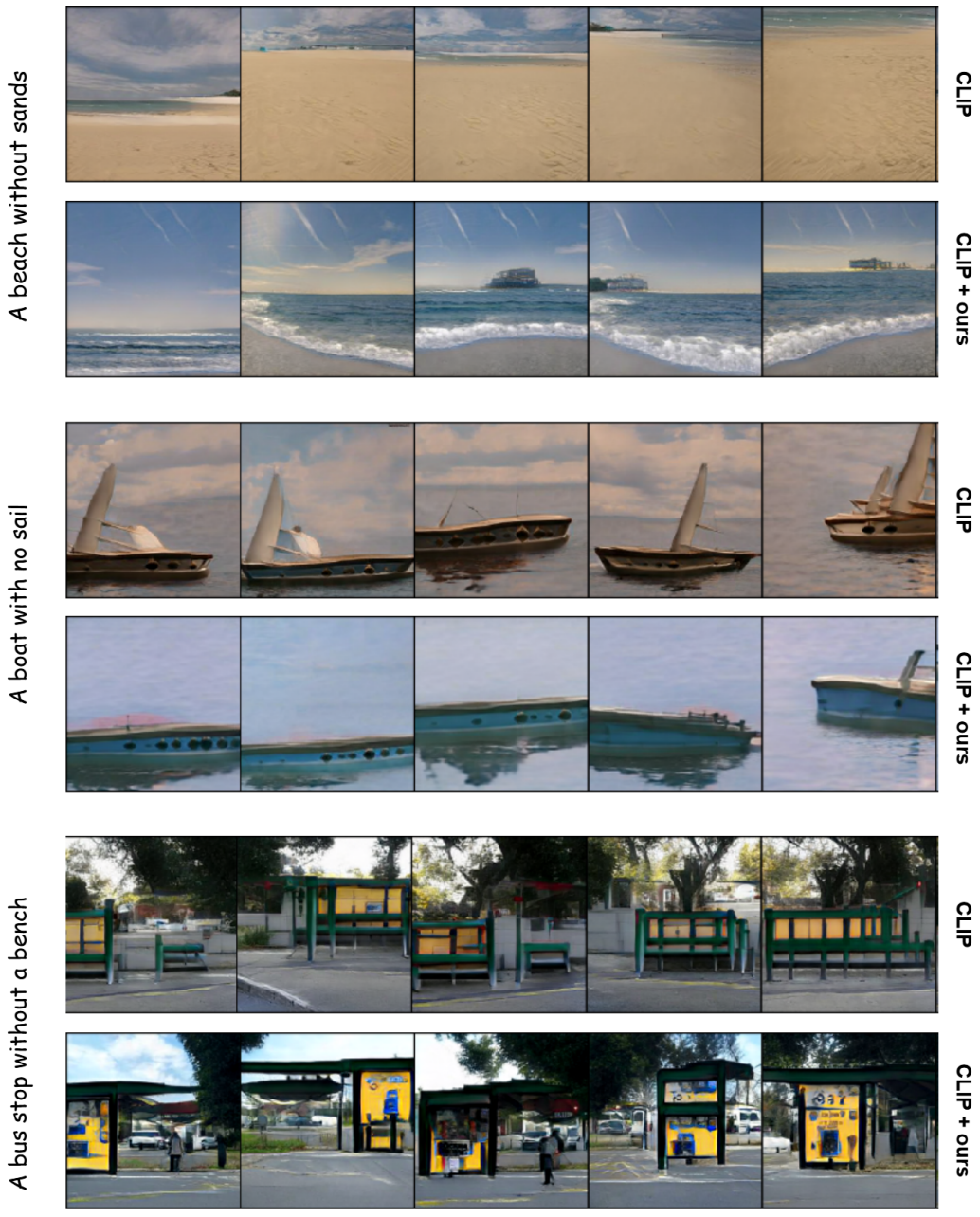


Figure 6: Samples of T2I with negation. As can be seen, our method consistently generates more aligned images with the query, while CLIP ignores negation in the query.

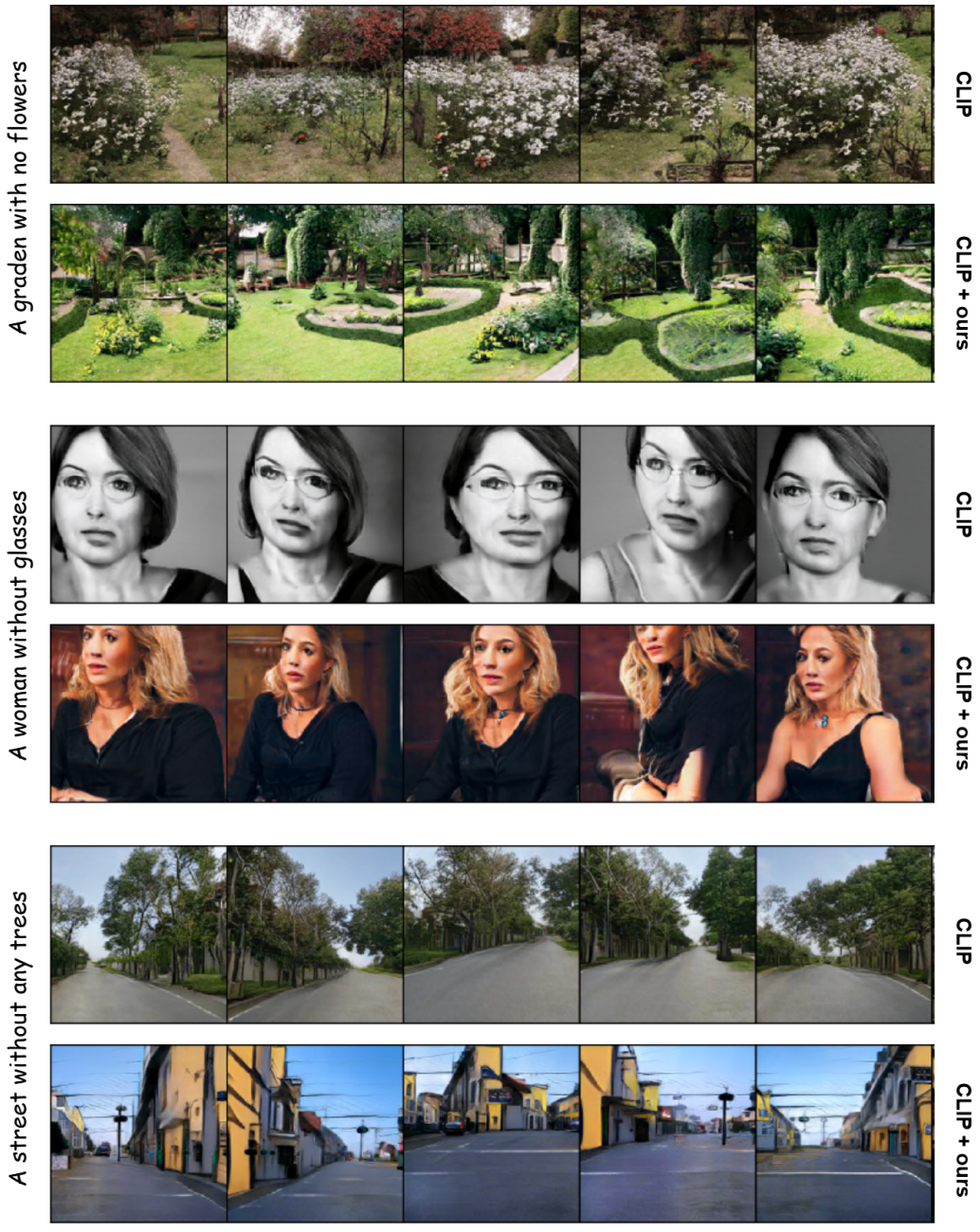
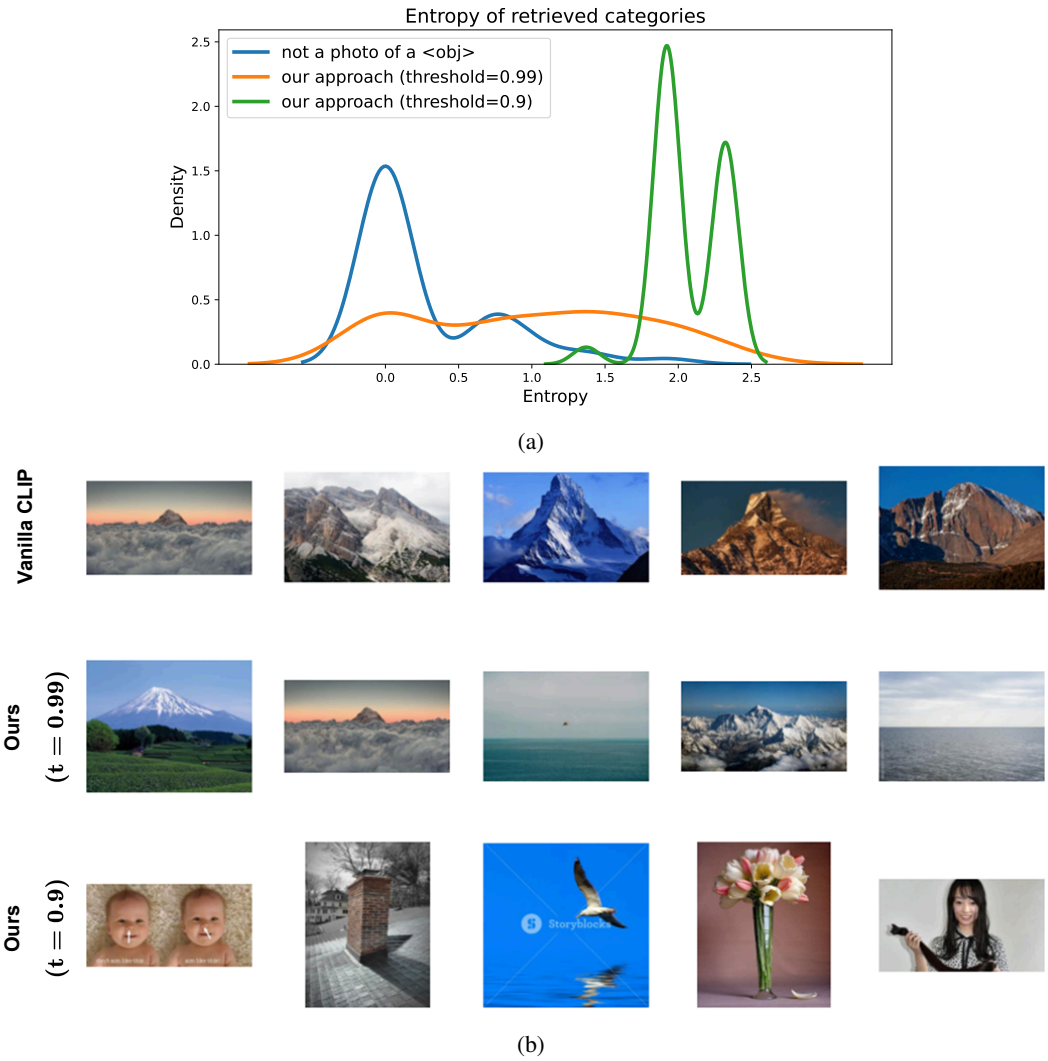


Figure 7: Samples of T2I with negation. As can be seen, our method consistently generates more aligned images with the query, while CLIP ignores negation in the query.



(a)

(b)

Figure 8: (a) Entropy of retrieved labels. (b) Retrieved images.