ARC-SMOOTHED CAPTION SUPERVISION FOR FINE-GRAINED SPATIAL CLIP

Anonymous authors
Paper under double-blind review

ABSTRACT

CLIP is the image encoder behind zero-shot recognition and many multimodal large language models (MLLMs), yet it is weak at fine-grained spatial perception such as counting objects and judging their positions and relations. Therefore, classifiers and MLLMs that reuse its encoders inherit this weakness. Existing remedies fine-tune the image encoder against a second large model placed in the training loop, such as a frozen diffusion model or a vision-only teacher, and give the text encoder no spatial signal. We propose arc-smoothed caption supervision, a simple alternative whose training targets are all points in CLIP's own image–text embedding space and which uses no other network during training or inference. For each training image, we generate a spatial caption once, offline. During fine-tuning, we move along the great-circle arc between the embeddings of the class template and the spatial caption to a position drawn from a U-shaped Beta law, and use the resulting point as the positive target of CLIP's standard contrastive loss. Targets inside the arc are essential: using the same two captions as separate positives at the same exposure gives lower spatial zero-shot accuracy than template-only fine-tuning. We show that the mixing law trades fidelity to the two caption objectives against an interior term that contrastively supervises the midpoint between each image's template and its own caption, and controlled comparisons are consistent with both sides of this trade-off. On CLIP ViT-L/14, relative to matched template-only fine-tuning and across two captioners, arc smoothing improves spatial zero-shot accuracy by 3.0–3.8 points and text-to-image retrieval R@1 by 1.4–2.1 points, while keeping zero-shot transfer on 11 datasets at the template-only level. Used as the visual encoder of LLaVA-1.5, it improves average VQA accuracy by 3.0 points, including 5.0 points on TallyQA counting, suggesting that spatial information supplied through language during encoder fine-tuning carries over to downstream multimodal reasoning.

1 INTRODUCTION

Contrastive Language–Image Pre-training (CLIP) (Radford et al., 2021) provides one of the most widely used image encoders, both as a zero-shot recognizer and as the visual front-end of multimodal large language models (MLLMs) such as LLaVA (Liu et al., 2023; 2024), BLIP-2 (Li et al., 2023a), and Qwen-VL (Bai et al., 2023). Its image-level contrastive objective captures global semantics well but under-represents fine-grained visual information: CLIP struggles with counting, spatial relations, viewpoint, orientation, and local structure, and MLLMs built on it inherit these failures (Tong et al., 2024b; Kamath et al., 2023; Thrush et al., 2022; Yuksekgonul et al., 2023). Because many MLLMs, including LLaVA-1.5, keep this encoder frozen, its blind spots cannot be trained away downstream and must be addressed in the encoder itself.

Recent work improves CLIP's fine-grained perception by fine-tuning its image encoder against a second model placed *inside* the training loop. Generative-feedback methods back-propagate a reconstruction signal from a frozen text-to-image diffusion model (DIVA; Wang et al., 2025) or a CLIP-conditioned image generator (un^{2}CLIP, Li et al., 2025b; GenHancer, Ma et al., 2025), and KUEA (Gong et al., 2025) aligns CLIP's similarity kernel with the vision-only teacher DINOv2 (Oquab et al., 2024). These methods help, but they share three limitations. First, every training step runs an additional large network, and the outcome depends on that network. Second, their signal is purely visual and lies outside the image–text space in which CLIP is trained and evaluated, so

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

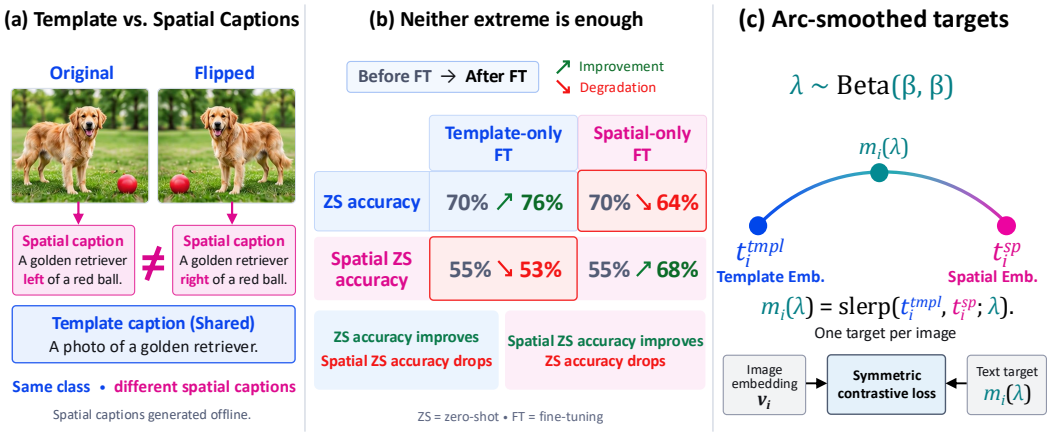


Figure 1: **Arc-smoothed caption supervision. (a)** An image and its horizontally flipped copy share the class template but have different spatial captions, which a multimodal captioner writes once, offline. Both captions are encoded by the same CLIP text encoder. **(b)** Neither extreme is enough. Template-only fine-tuning gives every image of a class the same target, leaving object positions and relations unconstrained. Spatial-only fine-tuning moves the text encoder toward the caption region $\mathcal{M}^{\text{sp}}$, which causes the templates that build the test-time zero-shot classifier to be out of distribution and lowers zero-shot accuracy (Table 1). **(c)** We instead use as the positive target the spherical interpolation $m_i(\lambda)$ between the two caption embeddings, with $\lambda \sim \text{Beta}(\beta, \beta)$. Every target stays on the unit sphere, and the U-shaped law keeps most targets near the two captions while still supervising the interior of the arc.

image–text alignment must be protected separately, for example with a frozen text encoder or a feature regulariser. Third, the signal never states *what* spatial content an image contains, and the text encoder, which builds CLIP's zero-shot classifiers, receives no spatial supervision.

A natural alternative is to supply spatial information in CLIP's own terms: describe each training image in words, which modern MLLM captioners do cheaply offline, and fine-tune with CLIP's contrastive loss. The difficulty is that CLIP's text encoder plays two roles. During fine-tuning it encodes the training targets, and at test time it encodes the class templates ("a photo of a {class}") that form the zero-shot classifier. Fine-tuning on spatial captions alone pulls the text encoder away from templates and lowers zero-shot transfer below that of frozen CLIP. Fine-tuning on templates alone, as in FLYP (Goyal et al., 2023), preserves the classifier but gives every image of a class the same target, leaving object layout unconstrained. The obvious compromise, using both captions as separate positives as in caption-mixing schemes (Fan et al., 2023; Lai et al., 2024; Zhang et al., 2024), does not resolve this tension either: with the same captions at the same exposure, it improves neither spatial zero-shot accuracy nor zero-shot transfer over template-only fine-tuning (Table 1).

We propose *arc-smoothed caption supervision*, which supervises CLIP with targets *between* the two captions rather than at either one (Figure 1). The procedure has four steps. (1) Before training, a frozen MLLM captioner writes one spatial caption per training image, listing its objects with their positions and relations. (2) CLIP's text encoder embeds this caption and a class template. Because CLIP L_2-normalises its embeddings, both are points on the unit hypersphere, separated by an angle that varies considerably across images. (3) For each batch, we draw a mixing weight $\lambda \sim \text{Beta}(\beta, \beta)$ and move from the caption embedding toward the template embedding by the fraction λ of the great-circle arc joining them, using spherical linear interpolation (slerp). (4) The resulting unit vector replaces the template as the positive text target in CLIP's standard contrastive loss, and both encoders are fine-tuned. With $\beta = 0.2$ the law is U-shaped: most targets lie near one of the two captions, and about a third lie in the interior of the arc. No other network runs during training, no loss term is added, and the encoder is used exactly like the original at inference.

Slerp and mixup-style interpolation are established tools (Zhang et al., 2018; Oh et al., 2023); what is new is the supervision signal they create here, and an account of why it works. Mixup combines two different examples and their labels. Our endpoints are two descriptions of the *same* image at different granularities, one of which is the classifier the model is evaluated with. An interior target thus describes the image as its class prototype moved part of the way toward its own layout, which

neither caption provides. Section 5 makes this precise by writing each target in terms of the common and differential components of the two captions. Endpoint mass keeps the objective faithful to the two caption objectives, including the template objective that defines the zero-shot classifier. Interior mass adds a term that, to leading order, contrastively supervises the midpoint of each image's own two captions, with most weight on images whose caption departs most from its template. The mixing law sets the trade-off between the two, and slerp makes it act in the same way on every image despite heterogeneous caption angles. Controlled comparisons are consistent with both sides, and $\mathrm{Beta}(0.2, 0.2)$ is an effective operating point.

Contributions.

- **A text-side route to spatial CLIP.** We introduce arc-smoothed caption supervision: same-image geodesic targets between a class template and an offline spatial caption, trained with CLIP's own contrastive loss and no additional network, loss term, or inference-time module (Section 4). We shot that the interior of the arc, not the captions alone, drives the gains.
- **An exact account of the mixing law.** Using identities we show that the law splits supervision between the common and differential modes of the two captions and that its interior budget weights an endpoint-slope term, absent from endpoint sampling, that contrastively supervises each image's template–caption midpoint. This defines a trade-off between endpoint fidelity and interior supervision that acts alike on every example under slerp (Section 5).
- **Consistent empirical gains.** Under a matched recipe on ViT-L/14 with two captioners, arc smoothing improves spatial zero-shot accuracy ($+3.0$ to $+3.8$), spatial linear probing ($+1.6$ to $+2.1$), and text-to-image R@1 on Flickr30K and MSCOCO ($+1.4$ to $+2.1$) over template-only fine-tuning, keeps 11-dataset zero-shot transfer at the template-only level, and improves LLaVA-1.5 VQA by 3.0 points on average (Section 7).

2 Related Work

Fine-tuning CLIP for visual detail. Beyond the generative-feedback and teacher-alignment methods above, CLIPSelf (Wu et al., 2024), RegionCLIP (Zhong et al., 2022), and locality alignment (Covert et al., 2025) use other vision-side targets; Qiu et al. (2025) preserve spatial correlations under region–language alignment for dense prediction; and mixture-of-encoders MLLMs (Shi et al., 2025; Tong et al., 2024a; Zong et al., 2024) add a second visual encoder. All rely on visual supervision outside CLIP's text space; in our method, gradients flow only through CLIP, and spatial information enters through the text branch.

Generated and long captions for CLIP. Menon & Vondrick (2023); Pratt et al. (2023) build CLIP classifier heads from LLM-generated descriptive prompts at inference, without fine-tuning. LaCLIP (Fan et al., 2023) replaces web captions with LLM rewrites during pretraining, and VeCLIP (Lai et al., 2024) enriches captions with visual concepts, including spatial relations, and mixes them with the original alt-text. Long-CLIP (Zhang et al., 2024) and FLAIR (Xiao et al., 2025) fine-tune CLIP with long or dense captions alongside short ones as separate positives. DAC (Doveh et al., 2023) fine-tunes CLIP on captioner-generated dense captions, and Paiss et al. (2023) add counterfactual counting captions; these change the captions themselves. Within our recipe, caption mixing corresponds to our balanced-endpoint baseline; we instead use image-grounded spatial captions as one endpoint of an on-sphere target whose other endpoint is the evaluation template.

Caption-conditioned fine-tuning. FLYP (Goyal et al., 2023) fine-tunes CLIP with its pretraining loss on class templates and outperforms LP-FT, robust fine-tuning (Wortsman et al., 2022), and L2-SP under distribution shift; we extend it by interpolating each template embedding with an image-specific caption embedding. Feature distortion under fine-tuning (Kumar et al., 2022) motivates our spatial-only baseline.

Mixup, label smoothing, and slerp. Neither mixup (Zhang et al., 2018) nor manifold mixup (Verma et al., 2019) stays on $\mathcal{S}^{d-1}$, and Geodesic Multi-Modal Mixup (Oh et al., 2023) applies slerp *across* the modality gap to create hard negatives. We apply slerp to two *text* embeddings of the same image, which relates our targets to label smoothing (Szegedy et al., 2016; Müller et al., 2019). Expansions of the mixup objective in the mixing strength reveal data-dependent regularisers (Zhang et al., 2021;

Carratino et al., 2022); Proposition 2 is an exact analogue for the arc. Mistretta et al. (2025) document CLIP's intra-modal misalignment, which is consistent with our measurements of the template and caption regions (Section 4.1).

3 PRELIMINARIES

CLIP. CLIP (Radford et al., 2021) consists of a visual encoder f_ψ and a text encoder g_ω that produce L_2-normalised embeddings $\bar{f}(\cdot), \bar{g}(\cdot) \in \mathcal{S}^{d-1}$. Pretraining uses a symmetric InfoNCE loss (van den Oord et al., 2018) on web-scale image–text pairs. For zero-shot classification, the classifier weight of class c is $\bar{g}(T_c)$ for a template caption T_c, and the logit of image I is $\bar{f}(I) \cdot \bar{g}(T_c)/\tau$.

FLYP. FLYP (Goyal et al., 2023) fine-tunes CLIP on labelled data with its pretraining loss. For batch size B,

$$\mathcal{L}_{\text{FLYP}} = \frac{1}{2B} \sum_{i=1}^{B} \left[-\log \frac{\exp(\bar{f}(I_i) \cdot \bar{g}(T_i)/\tau)}{\sum_j \exp(\bar{f}(I_i) \cdot \bar{g}(T_j)/\tau)} - \log \frac{\exp(\bar{f}(I_i) \cdot \bar{g}(T_i)/\tau)}{\sum_j \exp(\bar{f}(I_j) \cdot \bar{g}(T_i)/\tau)} \right], \tag{1}$$

where T_i is drawn uniformly from the 80 OpenAI templates for the class c_i of image I_i. Our method changes only the text targets supplied to this loss.

Spherical linear interpolation. For unit vectors $\vec{a}, \vec{b} \in \mathcal{S}^{d-1}$ separated by the angle $\theta = \cos^{-1}(\vec{a} \cdot \vec{b}) \in (0, \pi)$, slerp, used for multimodal mixup by Oh et al. (2023), is

$$m_\lambda(\vec{a}, \vec{b}) \;=\; \frac{\sin(\lambda\theta)}{\sin\theta}\,\vec{a} \;+\; \frac{\sin((1-\lambda)\theta)}{\sin\theta}\,\vec{b}, \qquad \lambda \in [0, 1]. \tag{2}$$

It satisfies $m_0 = \vec{b}$ and $m_1 = \vec{a}$, and traces the shortest great-circle arc from $\vec{b}$ to $\vec{a}$. For coincident endpoints we use the constant path; antipodal endpoints do not occur in our data.

4 ARC-SMOOTHED CAPTION SUPERVISION

4.1 TWO CAPTIONS PER TRAINING IMAGE

Each ImageNet-1K image I_i with class label c_i has a template caption T_i^{tmpl} and a spatial caption T_i^{sp} that lists the objects in the image with their positions and relations (Section 6). Frozen CLIP places these two kinds of caption in different regions of the hypersphere (Appendix Table 6 and Figure 3). Templates of a class share their syntax and class token, so they cluster tightly; we call the union of these clusters the *template region* $\mathcal{M}^{\text{tmpl}}$. Spatial captions are far more varied (the mean cosine between random pairs is 0.21, compared with 0.53 for templates) and occupy a broader *caption region* $\mathcal{M}^{\text{sp}}$ whose centroid is displaced from that of the templates. Matched spatial captions are slightly closer to their images than templates are (mean cosine 0.28 versus 0.26).

Two properties shape our design. First, the text encoder must remain compatible with $\mathcal{M}^{\text{tmpl}}$, because the fine-tuned encoder rebuilds the zero-shot classifier from templates. Second, the angle θ_i between an image's two captions varies widely: the same-image cosine averages 0.35 with standard deviation 0.15, four times that of matched image–text pairs. Spatial captions are thus not uniform rephrasings of templates, and a rule for mixing them must behave consistently across very different angles.

4.2 ENDPOINT TARGETS AND WHY THEY FALL SHORT

Three simple choices use only the two captions themselves as targets.

Template-only ($T_i = T_i^{\text{tmpl}}$, i.e. FLYP) trains exactly the text regime used at test time, but gives every image of a class the same target, so object positions and relations remain unconstrained.

Spatial-only ($T_i = T_i^{\text{sp}}$) supplies layout information but removes direct supervision of templates. We expect the text encoder to drift toward $\mathcal{M}^{\text{sp}}$, which causes the templates that build the zero-shot classifier to be out of distribution, an instance of the fine-tuning distortion studied by Kumar et al. (2022).

Balanced endpoint sampling uses, for each batch, the template with probability $\frac{1}{2}$ and the spatial caption otherwise. It is the within-recipe analogue of caption mixing (Fan et al., 2023; Lai et al., 2024; Zhang et al., 2024), and it is also what a symmetric Beta draw becomes when rounded to its nearest endpoint. It therefore matches our method's exposure to each caption but never uses a target between them, which makes it the direct control for the value of interior targets.

Table 1 shows that neither alternative to template-only training improves on it in spatial zero-shot accuracy, and both lose zero-shot transfer. The missing ingredient is a target *between* the two captions.

4.3 Geodesic targets between the two captions

We interpolate the two text embeddings of the same image along their shortest arc (Figure 1c):

$$\tilde{T}_i \;=\; m_\lambda\big(\bar{g}(T_i^{\text{tmpl}}),\, \bar{g}(T_i^{\text{sp}})\big), \qquad \lambda \sim \nu_\beta = \text{Beta}(\beta, \beta), \tag{3}$$

and fine-tune with the FLYP loss (1) in which $\bar{g}(T_i)$ is replaced by $\tilde{T}_i$:

$$\mathcal{L}_{\text{arc}} \;=\; \frac{1}{2B}\sum_{i=1}^{B}\left[-\log\frac{\exp(\bar{f}(I_i)\cdot\tilde{T}_i/\tau)}{\sum_j \exp(\bar{f}(I_i)\cdot\tilde{T}_j/\tau)} - \log\frac{\exp(\bar{f}(I_i)\cdot\tilde{T}_i/\tau)}{\sum_j \exp(\bar{f}(I_j)\cdot\tilde{T}_i/\tau)}\right]. \tag{4}$$

Gradients flow through both text embeddings and the image embedding, so both encoders are fine-tuned. We draw one λ per batch and share it across the batch. This makes the training objective an average over a one-parameter family of losses, which Section 5 analyses; Appendix D (Proposition 4) compares it with per-example draws. By default $\beta = 0.2$, under which 32.7% of draws fall in $(0.1, 0.9)$. Interior targets are synthetic unit vectors; they are not assumed to correspond to literal captions. The method adds no loss term and no inference-time module; its only training overhead is encoding a second caption per image.

Why slerp. Slerp's coefficient has the same geometric meaning on every arc; that of linear interpolation does not.

Lemma 1 (Calibration of the arc parameter)**.** *Let $t, s \in \mathcal{S}^{d-1}$ with $\theta = \arccos(t^\top s) \in (0, \pi)$. For every $\lambda \in [0, 1]$, $\|m_\lambda(t, s)\|_2 = 1$ and*

$$d_{\mathcal{S}}\big(s, m_\lambda(t, s)\big) = \lambda\theta, \qquad d_{\mathcal{S}}\big(t, m_\lambda(t, s)\big) = (1-\lambda)\theta. \tag{5}$$

In contrast, the normalised chord $n_w = c_w/\|c_w\|_2$, with $c_w = wt + (1-w)s$, lies at the fractional position $h_\theta(w) = d_{\mathcal{S}}(s, n_w)/\theta = w + \frac{1}{6}w(1-w)(2w-1)\,\theta^2 + O(\theta^4)$ along the same arc.

Under slerp, the law of λ is the law of the fractional position along *every* arc. A fixed coefficient law for the normalised chord instead induces a different position law for each pair, with a distortion that grows quadratically in the angle; since θ_i varies widely (Section 4.1), this is a heterogeneity across examples, not a global reparameterisation. The unnormalised chord also leaves the sphere, rescaling the positive logit by an angle-dependent factor (Appendix A). Calibration is what lets Section 5 describe the law by scalars shared by all examples.

5 What the Mixing Law Controls

Section 4 fixed the path of the targets. This section characterises the remaining design choice, the law ν of the position Λ on that path. Proposition 1 characterises the laws themselves, and Proposition 2 is an exact identity for the training objective at fixed parameters; proofs and further results are in Appendix D. They describe what each law optimises. They do not bound downstream accuracy and do not select β, which we choose empirically (Section 7.2).

Notation. For a batch at fixed parameters, let $u_i = \bar{f}(I_i)$, $t_i = \bar{g}(T_i^{\text{tmpl}})$, $s_i = \bar{g}(T_i^{\text{sp}})$, $\theta_i = \arccos(t_i^\top s_i) \in (0, \pi)$, and $z_i(\lambda) = m_\lambda(t_i, s_i)$, the target $\tilde{T}_i$ of Eq. (3). Let $\mathcal{L}(\lambda)$ denote the loss (4) with one shared λ, so that $\mathcal{L}(1)$ is the template (FLYP) objective and $\mathcal{L}(0)$ is the spatial-caption objective. For a law ν on $[0, 1]$ that is symmetric about $\frac{1}{2}$, the training objective is $\mathcal{L}_\nu = \mathbb{E}_{\Lambda\sim\nu}[\mathcal{L}(\Lambda)]$, and the balanced endpoint objective is $\mathcal{L}_{\text{end}} = \frac{1}{2}[\mathcal{L}(0) + \mathcal{L}(1)]$. Let $D(\lambda) = \min(\lambda, 1-\lambda)$ be the fractional distance to the nearer caption, which by Lemma 1 has the same meaning for every pair,

and let $S_\nu(\xi) = \Pr_\nu(D(\Lambda) > \xi)$. The *interior budget* of ν is

$$\mathcal{I}(\nu) = \mathbb{E}_\nu[D(\Lambda)] = \int_0^{1/2} S_\nu(\xi)\, d\xi \in [0, \tfrac{1}{2}], \tag{6}$$

which is zero exactly for laws supported on the two endpoints.

5.1 Endpoint fidelity: common and differential modes

Each pair j spans a plane with the orthonormal basis

$$e_j^+ = \frac{t_j + s_j}{2\cos(\theta_j/2)}, \qquad e_j^- = \frac{t_j - s_j}{2\sin(\theta_j/2)},$$
$$z_j(\lambda) = \cos\varphi_j(\lambda)\, e_j^+ + \sin\varphi_j(\lambda)\, e_j^-, \qquad \varphi_j(\lambda) = (\lambda - \tfrac{1}{2})\theta_j. \tag{7}$$

The *common mode* e_j^+ is the direction the two captions share, and the *differential mode* e_j^- is what distinguishes them: $t_j = \cos(\frac{\theta_j}{2})e_j^+ + \sin(\frac{\theta_j}{2})e_j^-$, so the template-specific part of t_j lies along e_j^-. Writing $c_{ij} = u_i^\top e_j^+$ and $d_{ij} = u_i^\top e_j^-$, every logit takes the form

$$a_{ij}(\lambda) = \frac{u_i^\top z_j(\lambda)}{\tau} = \frac{\cos\varphi_j(\lambda)\, c_{ij} + \sin\varphi_j(\lambda)\, d_{ij}}{\tau}, \tag{8}$$

so the law acts on the objective only through the gains $\cos\varphi_j$ and $\sin\varphi_j$.

Proposition 1 (Mode gains). *Let ν be symmetric about $\frac{1}{2}$ and $0 < \theta < \pi$, and write $\varphi = (\Lambda - \frac{1}{2})\theta$.*

(i) $\mathbb{E}_\nu \sin\varphi = 0$, $r_\nu(\theta) := \mathbb{E}_\nu \cos\varphi \in [\cos(\theta/2), 1]$, *and* $v_\nu(\theta) := \mathbb{E}_\nu \sin^2\varphi \in [0, \sin^2(\theta/2)]$.

(ii) The balanced endpoint law $\frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$ is the unique symmetric law with $v_\nu(\theta) = \sin^2(\theta/2)$, and the midpoint law $\delta_{1/2}$ is the unique one with $v_\nu(\theta) = 0$.

(iii) For $\nu_\beta = \mathrm{Beta}(\beta, \beta)$, $v_{\nu_\beta}(\theta)$ is strictly decreasing in β, whereas $r_{\nu_\beta}(\theta)$, $S_{\nu_\beta}(\xi)$ for every $\xi \in (0, \frac{1}{2})$, and hence $\mathcal{I}(\nu_\beta)$, are strictly increasing in β.

Every symmetric law supervises the common mode on average, since $\mathbb{E}_\nu z_j(\Lambda) = r_\nu(\theta_j)e_j^+$. Laws differ in how strongly they supervise the differential mode, i.e. how fully each step commits to one actual caption: fully for the endpoint law, never for the midpoint law. We call v_ν the *endpoint fidelity* of ν, a descriptive name: a law with high v_ν trains close to the two caption objectives themselves, including the template objective from which the zero-shot classifier is built. Training on the mean target is not equivalent: the expected loss adds a nonnegative prediction-variation term (Proposition 3, Appendix D).

5.2 Interior supervision: the endpoint-slope term

Let $p_{ij}(\lambda)$ and $q_{ij}(\lambda)$ be the row (image-to-text) and column (text-to-image) softmax probabilities of the logit matrix, and let $P_{ij}^\lambda = p_{ij}(\lambda) + q_{ij}(\lambda)$.

Proposition 2 (Profile identity). *For every law ν symmetric about $\frac{1}{2}$,*

$$\mathcal{L}_\nu - \mathcal{L}_{\mathrm{end}} = \int_0^{1/2} \rho(\xi)\, S_\nu(\xi)\, d\xi, \qquad \rho(\xi) = \tfrac{1}{2}\big[\mathcal{L}'(\xi) - \mathcal{L}'(1-\xi)\big]. \tag{9}$$

At the endpoints, the profile is the endpoint-slope term

$$\rho(0) = \frac{1}{4B\tau} \sum_{i,j} \theta_j \Big[\sin(\tfrac{\theta_j}{2})\big(P_{ij}^0 + P_{ij}^1 - 4\delta_{ij}\big)\, c_{ij} + \cos(\tfrac{\theta_j}{2})\big(P_{ij}^0 - P_{ij}^1\big)\, d_{ij} \Big], \tag{10}$$

and $\mathcal{L}_\nu = \mathcal{L}_{\mathrm{end}} + \mathcal{I}(\nu)\,\rho(0) + \frac{1}{2}\,\mathbb{E}_\nu[D(\Lambda)^2]\,\rho'(0) + R_\nu$, with $|R_\nu| \le \frac{1}{6}\mathbb{E}_\nu[D(\Lambda)^3]\sup_{[0,1/2]}|\rho''|$.

The endpoint law ($S_\nu \equiv 0$) optimises $\mathcal{L}_{\mathrm{end}}$ alone; interior mass adds the weighted profile, the slope of the loss as the target moves inward from each caption. Every target lies in the plane of its two captions, so this term adds no new directions but reweights supervision within each plane (Eq. (10)). Its first part is contrastive on the bisectors $e_j^+ \propto t_j + s_j$: with the softmax weights held fixed, lowering it raises each image's score against its own bisector (the coefficient is negative for $i = j$) and lowers

its scores against other images' bisectors. Because s_j is image-specific, so is e_j^+, and the weight $\theta_j \sin(\theta_j/2)$ grows with the template–caption angle. The second part acts on the differential mode only in proportion to how differently the two endpoint objectives fit each pair, vanishing when they fit it equally. To leading order, interior mass thus emphasises agreement with the midpoint of each image's two captions, most where they differ most. This reading is descriptive: at CLIP's temperature the leading term need not dominate, and only Eq. (9) is exact (Appendix D).

5.3 The trade-off and the operating point

Propositions 1 and 2 identify two competing properties of a law: endpoint fidelity v_ν, a descriptive measure of how directly the objective trains the two caption objectives, and the interior budget $\mathcal{I}(\nu)$ (more generally S_ν), the weight on the interior term. Maximal fidelity is attained only by the endpoint law, where $\mathcal{I}(\nu) = 0$ (Proposition 1(ii)). Within the Beta family, increasing β lowers v_ν and raises S_ν (Proposition 1(iii)), in the same direction for every example because the law of D is the same on every arc (Lemma 1). The two scalars do not rank all laws (Appendix D), so we treat β as a one-parameter path through the trade-off. $\mathrm{Beta}(0.2, 0.2)$ retains 73% of the endpoint law's fidelity at $\bar{\theta}$ $(0.236/0.325)$ while giving the interior term 40% of the uniform law's budget $(0.101/0.250$; Appendix Table 7), within the endpoint-concentrated range common for mixup (Zhang et al., 2018). This suggests two hypotheses, examined in Section 7.2: in-distribution ImageNet accuracy, which the template objective trains directly, should fall as β grows; and any gain over balanced endpoint sampling must come from the interior term. Which β is best remains an empirical question.

6 Experimental Setup

Training. We fine-tune OpenAI CLIP ViT-L/14 on ImageNet-1K (Deng et al., 2009) with AdamW (Loshchilov & Hutter, 2019), a learning rate of 8×10^{-6}, weight decay 0.05, a cosine schedule with 500 warmup steps, batch size 224, and 10 epochs. All label-supervised rows use this recipe and differ only in the text target. Templates are sampled uniformly from the 80 OpenAI templates. Spatial captions are generated once, offline, by CapRL (Xing et al., 2026) as comma-separated object–relation phrases; a second caption set is generated by Qwen2.5-VL-7B-Instruct (Bai et al., 2025) with the same prompt (Appendix C). We use $\mathrm{Beta}(0.2, 0.2)$ with one draw per batch unless stated otherwise. Each fine-tuned row is a single training run.

Evaluation. Following KUEA (Gong et al., 2025), we use CLIP-Benchmark (LAION-AI, 2022) for ImageNet and 11 zero-shot transfer datasets (ZS-11), Flickr30K and MSCOCO retrieval (Young et al., 2014; Lin et al., 2014), and zero-shot and linear-probe evaluation on the four "spatial" tasks of Gong et al. (2025): SVHN, GTSRB, CLEVR-Distance, and CLEVR-Counts (Netzer et al., 2011; Stallkamp et al., 2012; Johnson et al., 2017). SVHN and GTSRB test fine-grained local recognition, while the CLEVR tasks directly test counting and distance. ImageNet is in distribution for all label-supervised models. For LLaVA-1.5-7B (Liu et al., 2024), we replace the visual encoder and LoRA-tune the LLM (Hu et al., 2022) on the 665k instruction mixture with LLaVA's defaults (rank 128, $\alpha = 256$, learning rate 2×10^{-4}, 1 epoch, batch size 128), using the same protocol for the CLIP, KUEA, and arc-smoothed encoders. We report TextVQA (Singh et al., 2019), POPE (Li et al., 2023b), TallyQA (Acharya et al., 2019), and VQA-v2 (Goyal et al., 2017).

Baselines. Label-free results for CLIP, Projection, Feature, DIVA, and KUEA are taken from Gong et al. (2025); DIVA and KUEA take no text input, so captions cannot be supplied to them. Label-supervised baselines (Section 4.2) use our recipe; normalised linear interpolation uses the raw $\mathrm{Beta}(0.2, 0.2)$ coefficient law.

7 Results

7.1 Which text target yields spatial gains?

Table 1 isolates the text target under a matched recipe. Relative to template-only fine-tuning, the CapRL arc-smoothed model raises the spatial zero-shot average from 38.39 to 41.36 and the spatial-probe average from 58.91 to 60.46, improves text-to-image R@1 on Flickr30K and MSCOCO by 1.5 and 1.4 points, and keeps ZS-11 transfer at the template-only level (69.96 versus 69.28). The cost is

Table 1: Matched-recipe comparison of text targets on ViT-L/14 (%). All label-supervised rows share the backbone, optimiser, 10-epoch schedule, and learning rate, and each is a single run. The balanced-endpoint ImageNet value is the mean of two runs (86.22 and 86.24). Spatial averages cover SVHN, GTSRB, CLEVR-Distance, and CLEVR-Counts; per-dataset results are in Appendix B.

Text target	ImageNet	ZS-11	Spat. ZS	Spat. probe	Flickr T→I	COCO T→I
Label-free						
CLIP (frozen)	74.90	65.26	36.80	50.59	63.62	34.51
KUEA (Gong et al., 2025)	75.52	66.54	39.17	56.10	64.78	35.98
Label-supervised (identical recipe)						
Template-only (FLYP-style)	86.40	69.28	38.39	58.91	66.06	39.54
Spatial-caption-only	77.65	62.19	37.27	59.58	67.36	40.81
Balanced endpoint sampling	86.23	66.94	37.14	59.20	66.82	40.12
Normalised linear interpolation	81.02	64.42	36.99	54.78	66.04	40.28
Ours (arc-smoothed, $\mathrm{Beta}(0.2, 0.2)$)						
CapRL captions	85.54	**69.96**	41.36	60.46	**67.54**	40.97
Qwen2.5-VL captions	84.48	69.55	**42.21**	**61.05**	67.43	**41.67**

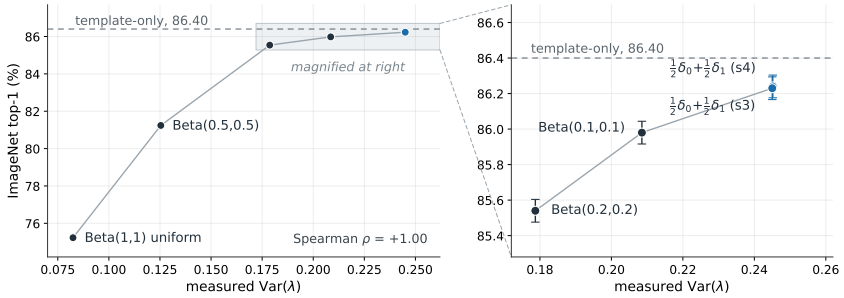


Figure 2: **Controlled law sweep: the fidelity side of the trade-off.** Six runs under an identical recipe differ only in the batch-shared mixing law, and all have measured mean $\lambda \in [0.500, 0.505]$, so exposure to each caption is matched. In-distribution ImageNet top-1 increases with the measured variance of λ (Spearman $\rho = +1.00$ across the five laws), that is, it decreases as β and the interior budget increase. The two balanced-endpoint points (s3, s4) are independent runs. Error bars in the magnified panel show $\pm 2\sigma$ for a common run-to-run standard deviation σ; they are indicative, not per-law confidence intervals. The dashed line is template-only training, which uses no captions.

0.9 points of in-distribution ImageNet accuracy (1.9 with Qwen2.5-VL captions). The baselines locate the source of these gains. Spatial-only training is competitive on retrieval but loses 3.1 transfer points relative to frozen CLIP, so the template regime must be retained. Balanced endpoint sampling uses the same two captions at the same exposure, yet reaches only 37.14 spatial zero-shot accuracy and 66.94 on ZS-11, both below template-only training; the interior of the arc accounts for the difference. Normalised linear interpolation trails slerp on every metric in these single runs; Appendix A compares the two parameterisations, and we do not attribute the gap to a specific mechanism. Qwen2.5-VL captions give the same profile (+3.8 spatial zero-shot, +2.1 probe, +1.4/+2.1 text-to-image R@1).

7.2 The mixing law trades endpoint fidelity against interior supervision

We now examine the two hypotheses of Section 5.3. Figure 2 varies only the law, under the same recipe and batch-shared protocol, with measured mean $\lambda \in [0.500, 0.505]$ so that exposure to each caption is matched. *Fidelity side.* ImageNet top-1 increases with the measured variance of λ, the angle-free form of endpoint fidelity ($v_\nu(\theta) = \theta^2 \operatorname{Var}_\nu(\Lambda) + O(\theta^4)$), i.e. it decreases as β grows: from 86.22–86.24 for balanced endpoints through 85.54 at $\mathrm{Beta}(0.2, 0.2)$ to 75.23 for the uniform law, which gives up almost the entire in-distribution gain of fine-tuning. *Interior side.* Moving from the endpoint law to $\mathrm{Beta}(0.2, 0.2)$ switches on the interior term while changing nothing else, and it raises spatial zero-shot accuracy by 4.2 points, the spatial probe by 1.3, and ZS-11 transfer by 3.0, at a cost of 0.7 ImageNet points (Table 1). Out-of-distribution transfer thus benefits from interior supervision even as in-distribution accuracy declines slightly.

Table 2: LLaVA-1.5-7B VQA (%) with different ViT-L/14 visual encoders. "Released" is the original LLaVA-1.5 model (CLIP encoder, no further tuning). The other rows LoRA-tune the LLM with the same protocol on top of the stated encoder. Avg is the unweighted mean of the five columns.

Encoder	TextVQA-ocr	TextVQA-pure	POPE	TallyQA	VQA-v2	**Avg**
Released (CLIP)	55.37	41.16	84.64	60.39	75.07	63.33
CLIP + LoRA	54.54	41.82	84.35	57.19	74.71	62.52
KUEA + LoRA	55.03	42.13	84.37	59.87	74.87	63.25
Ours + LoRA	**57.49**	**45.88**	**84.83**	**62.20**	**77.28**	**65.54**

Held-out loss along the arc shows the same trade-off (Appendix Figure 4): arc smoothing lowers the loss at the spatial-caption end, where template-only training does not, while template-only training has lower loss over the rest of the arc.

7.3 Spatial perception and retrieval

Relative to template-only fine-tuning, the four-task gains are $+3.0$ zero-shot and $+1.6$ probe (Table 1). Per task (Appendix Table 4), the largest gains over frozen CLIP are on CLEVR-Counts ($+8.5$ zero-shot, $+14.6$ probe) and CLEVR-Distance ($+4.0$, $+15.0$), which most directly test counting and relative position; averaged over the two CLEVR tasks, zero-shot accuracy rises from 19.8 (CLIP) and 22.3 (KUEA) to 26.1. For retrieval (Appendix Table 5), spatial-only training already captures most of the text-to-image gain; arc smoothing retains it while restoring transfer.

7.4 Downstream multimodal reasoning with LLaVA-1.5

Table 2 replaces LLaVA-1.5's visual encoder under a matched LoRA protocol. The arc-smoothed encoder improves the average by 3.0 points over CLIP and by 2.3 points over KUEA under the same protocol, and by 2.2 points over the released LLaVA-1.5 model. The largest gains are on TallyQA counting ($+5.0$ over CLIP with LoRA) and TextVQA without OCR tokens ($+4.1$). Spatial information supplied through captions therefore transfers to an MLLM that consumes the encoder's patch features.

8 Conclusion

We introduced arc-smoothed caption supervision, which improves CLIP's spatial perception by fine-tuning it toward targets on the arc between each image's class template and its own spatial caption. Slerp gives the mixing weight the same geometric meaning for every image; the mixing law trades fidelity to the two caption objectives against an interior term that contrastively supervises each image's template–caption midpoint; and $\mathrm{Beta}(0.2, 0.2)$ is an effective operating point in this trade-off. Arc targets improve spatial perception and retrieval over template-only fine-tuning, keep zero-shot transfer at its level, and carry over to counting and VQA in LLaVA-1.5, with no additional network in training or inference.

Limitations. Spatial content comes from an offline captioner and may contain errors, and our two captioners share a model family. We use CLIP ViT-L/14 only; since the method changes only the text target, it applies unchanged to other dual encoders such as SigLIP, which we have not evaluated. The theory concerns the training objective, and does not select β; the other Beta laws were evaluated on ImageNet only.

AI use statement

Generative AI tools were used for language editing, for restructuring the exposition, and to assist in drafting and checking mathematical derivations. They were not used to generate research ideas, experimental results, or figure data. The authors designed and ran all experiments and verified all mathematical statements, proofs, and reported numbers.

Ethics statement

This work aims to improve the fine-grained spatial perception of CLIP-style vision–language models while preserving zero-shot transfer. Potential benefits include more reliable visual reasoning in downstream multimodal systems, especially for counting, spatial relations, object layout, and local grounding. However, improved spatial perception may also increase the effectiveness of systems used for surveillance, monitoring, or other privacy-sensitive applications. The method relies on captions from an external multimodal captioner and may therefore inherit or amplify its biases, stereotypes, omissions, or hallucinations. Improved spatial reasoning does not guarantee reliability in high-stakes domains. Deployment should include task-specific validation, bias and robustness evaluation, privacy safeguards for image data, and human oversight in high-stakes settings. This paper does not involve human-subjects research beyond the use of publicly released vision benchmarks.

Reproducibility statement

We will release the code, training configurations, and, subject to their licences, the offline caption files used for ImageNet-1K. Section 6 specifies the optimiser, schedule, batch size, backbone, Beta law, and LLaVA LoRA protocol. Appendix C provides the captioner prompt and failure-mode taxonomy. Appendices A and D contain the interpolation identities and all proofs. The label-free baseline results are taken from Gong et al. (2025); all label-supervised baselines use the same recipe (Table 1).

References

Manoj Acharya, Kushal Kafle, and Christopher Kanan. TallyQA: Answering complex counting questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 8076–8084, 2019.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Luigi Carratino, Moustapha Cissé, Rodolphe Jenatton, and Jean-Philippe Vert. On mixup regularization. *J. Mach. Learn. Res.*, 23:325:1–325:31, 2022. URL https://jmlr.org/papers/v23/20-1385.html.

Ian Covert, Tony Sun, James Zou, and Tatsunori Hashimoto. Locality alignment improves vision-language models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=qssVptHTPN.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, 2009.

Sivan Doveh, Assaf Arbelle, Sivan Harary, Roei Herzig, Donghyun Kim, Paola Cascante-Bonilla, Amit Alfassy, Rameswar Panda, Raja Giryes, Rogério Feris, Shimon Ullman, and Leonid Karlinsky. Dense and aligned captions (DAC) promote compositional reasoning in VL models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/efe406d6d2674d176cdcd958ce605d17-Abstract-Conference.html.

Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving CLIP training with language rewrites. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=SVjDiiVySh.

Shizhan Gong, Yankai Jiang, Qi Dou, and Farzan Farnia. Kernel-based unsupervised embedding alignment for enhanced visual representation in vision-language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 19912–19931. PMLR, 2025.

Sachin Goyal, Ananya Kumar, Sankalp Garg, Zico Kolter, and Aditi Raghunathan. Finetune like you pretrain: Improved finetuning of zero-shot vision models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19338–19347, 2023.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6325–6334, 2017.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1988–1997, 2017.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? Investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9161–9175, 2023.

Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=UYneFzXSJWh.

Zhengfeng Lai, Haotian Zhang, Bowen Zhang, Wentao Wu, Haoping Bai, Aleksei Timofeev, Xianzhi Du, Zhe Gan, Jiulong Shan, Chen-Nee Chuah, Yinfei Yang, and Meng Cao. VeCLIP: Improving CLIP training via visual-enriched captions. In *European Conference on Computer Vision*, pp. 111–127. Springer, 2024.

LAION-AI. CLIP-Benchmark: CLIP-like model evaluation, 2022. URL https://github.com/LAION-AI/CLIP_benchmark.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pp. 19730–19742, 2023a.

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, Vienna, Austria, 2025a. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.499. URL https://aclanthology.org/2025.acl-long.499/.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, 2023b.

Yinqi Li, Jiahe Zhao, Hong Chang, RuiBing Hou, Shiguang Shan, and Xilin Chen. un^{2}CLIP: Improving CLIP's visual detail capturing ability via inverting unCLIP. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b. URL https://openreview.net/forum?id=kpdFjNitGW.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755, 2014.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26286–26296, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Shijie Ma, Yuying Ge, Teng Wang, Yuxin Guo, Yixiao Ge, and Ying Shan. GenHancer: Imperfect generative models are secretly strong vision-centric enhancers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24402–24412, 2025.

Sachit Menon and Carl Vondrick. Visual classification via description from large language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=jlAjNL8z5cs.

Marco Mistretta, Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Andrew D. Bagdanov. Cross the gap: Exposing the intra-modal misalignment in CLIP via modality inversion. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=VVVfuIcmKR.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems*, 2019.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, 2011.

Changdae Oh, Junhyuk So, Hoyoon Byun, YongTaek Lim, Minchul Shin, Jong-June Jeon, and Kyungwoo Song. Geodesic multi-modal mixup for robust fine-tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=iAAXq60Bw1.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching CLIP to count to ten. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 3147–3157. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00294. URL https://doi.org/10.1109/ICCV51070.2023.00294.

Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? Generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15645–15655, 2023.

Congpei Qiu, Yanhao Wu, Wei Ke, Xiuxiu Bai, and Tong Zhang. Refining CLIP's spatial awareness: A visual-centric perspective. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=38No4B8sx6.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, Yilin Zhao, De-An Huang, Hongxu Yin, Karan Sapra, Yaser Yacoob, Humphrey Shi, Bryan Catanzaro, Andrew Tao, Jan Kautz, Zhiding Yu, and Guilin Liu. Eagle: Exploring the design space for multimodal LLMs with mixture of encoders. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Y2RW9EVwhT.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8309–8318, 2019.

Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural Networks*, 32:323–332, 2012.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5228–5238, 2022.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024a.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? Exploring the visual shortcomings of multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9568–9578, 2024b.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning*, pp. 6438–6447, 2019.

Wenxuan Wang, Quan Sun, Fan Zhang, Yepeng Tang, Jing Liu, and Xinlong Wang. Diffusion feedback helps CLIP see better. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=tLFWU6izoA.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7949–7961, 2022.

Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. CLIPSelf: Vision transformer distills itself for open-vocabulary dense prediction. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=DjzvJCRsVf.

Rui Xiao, Sanghwan Kim, Mariana-Iuliana Georgescu, Zeynep Akata, and Stephan Alaniz. FLAIR: VLM with fine-grained language-informed image representations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 24884–24894. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.02317. URL https://openaccess.thecvf.

com/content/CVPR2025/html/Xiao_FLAIR_VLM_with_Fine-grained_Language-informed_Image_Representations_CVPR_2025_paper.html.

Long Xing, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Jianze Liang, Qidong Huang, Jiaqi Wang, Feng Wu, and Dahua Lin. CapRL: Stimulating dense image caption capabilities via reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=JLelnhqXaC.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

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? In *International Conference on Learning Representations*, 2023.

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of CLIP. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LI*, volume 15109 of *Lecture Notes in Computer Science*, pp. 310–325. Springer, 2024. doi: 10.1007/978-3-031-72983-6_18. URL https://doi.org/10.1007/978-3-031-72983-6_18.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

Linjun Zhang, Zhun Deng, Kenji Kawaguchi, Amirata Ghorbani, and James Zou. How does mixup help with robustness and generalization? In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=8yKEo06dKNo.

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. RegionCLIP: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16772–16782, 2022.

Zhuofan Zong, Bingqi Ma, Dazhong Shen, Guanglu Song, Hao Shao, Dongzhi Jiang, Hongsheng Li, and Yu Liu. Mova: Adapting mixture of vision experts to multimodal context. *Advances in Neural Information Processing Systems*, 37:103305–103333, 2024.

A SPHERICAL INTERPOLATION AND ANGULAR CALIBRATION

Proof of Lemma 1 (slerp part). Let t, s be unit vectors with $0 < \theta = \arccos(t^\top s) < \pi$, and set $v = (t - \cos\theta\, s)/\sin\theta$, so that s, v are orthonormal and $t = \cos\theta\, s + \sin\theta\, v$. Substituting into Eq. (2) and using $\sin(\lambda\theta)\cos\theta + \sin((1-\lambda)\theta) = \cos(\lambda\theta)\sin\theta$ gives

$$m_\lambda(t, s) = \cos(\lambda\theta)\, s + \sin(\lambda\theta)\, v. \tag{11}$$

Consequently $\|m_\lambda\|_2 = 1$, $m_\lambda^\top s = \cos(\lambda\theta)$, and $m_\lambda^\top t = \cos\theta\cos(\lambda\theta) + \sin\theta\sin(\lambda\theta) = \cos((1-\lambda)\theta)$. Both angles lie in $[0, \pi)$, so applying $\arccos$ gives Eq. (5), and $\|\partial_\lambda m_\lambda\|_2 = \theta$. These identities concern the shortest arc; unit norm and constant speed alone would also admit other spherical paths, and we make no uniqueness claim. When $\theta = 0$ we set $m_\lambda(t, s) = t = s$.

Unnormalised and normalised linear interpolation. The chord $c_w = wt + (1-w)s$ satisfies $\|c_w\|_2^2 = 1 - 2w(1-w)(1-\cos\theta)$, which is below one for $w \in (0, 1)$ and $\theta > 0$. At $\theta = \arccos(0.35) \approx 69.5°$, the midpoint norm is $\cos(\theta/2) \approx 0.822$, which scales every logit involving that target by 0.822 at fixed temperature. (This angle is computed from the mean cosine; it is not the mean of the individual angles.)

In the frame above, $c_w = (1 - w + w\cos\theta)s + w\sin\theta\, v$. Normalisation does not change its direction, so the fractional distance of n_w from s is

$$h_\theta(w) = \frac{\operatorname{atan2}(w\sin\theta,\, 1 - w + w\cos\theta)}{\theta}, \qquad h'_\theta(w) = \frac{\sin\theta}{\theta\,[1 - 2w(1-w)(1-\cos\theta)]} > 0. \tag{12}$$

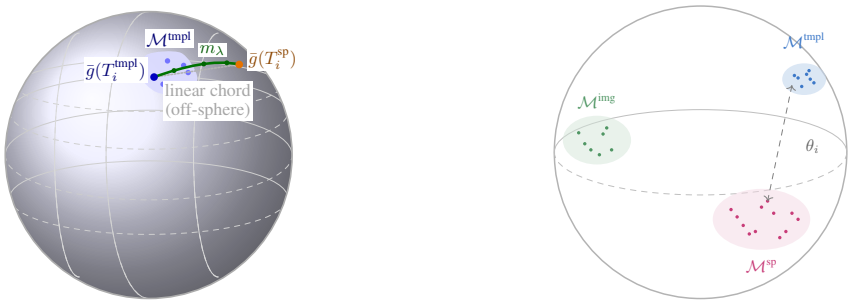
The derivative depends on w whenever $0 < \theta < \pi$, so a fixed coefficient law induces different position laws at different angles. Expanding about $\theta = 0$ gives the series in Lemma 1. For example, $w = 1/4$ is placed at fractional distances 0.246, 0.225, and 0.192 for $\theta = 30°$, $69.5°$, and $100°$, whereas slerp places $\lambda = 1/4$ at 0.25 for every angle.

An angle-matched chord. For $\lambda \in [0, 1]$ and $0 < \theta < \pi$, let

$$w_\theta(\lambda) = \frac{\sin(\lambda\theta)}{\sin(\lambda\theta) + \sin((1-\lambda)\theta)}. \tag{13}$$

Then $c_{w_\theta(\lambda)}$ is a positive multiple of the slerp expression, whose norm is one, so $n_{w_\theta(\lambda)}(t, s) = m_\lambda(t, s)$. Slerp is thus the normalised chord with an angle-dependent coefficient; the two coincide only if the coefficient is recomputed for every pair (and differentiated through θ when the endpoints are trainable). Our normalised-linear baseline instead uses the raw coefficient law, which is the natural implementation of linear interpolation.

The chord's effective law. Under a raw coefficient law $W \sim \nu$, the chord's fractional position is $h_\theta(W)$. Since $h_\theta(1-w) = 1 - h_\theta(w)$, this position law is again symmetric, but it depends on θ. For $W \sim \mathrm{Beta}(0.2, 0.2)$, its interior budget $\mathbb{E}[\min(h_\theta(W), 1 - h_\theta(W))]$ is 0.0997, 0.0926, and 0.0824 at $\theta = 30°$, $69.5°$, and $100°$, compared with 0.101 for slerp at every angle; for a uniform W it is 0.247, 0.233, and 0.213, compared with 0.250. The chord thus applies a slightly more endpoint-concentrated, angle-dependent law to each pair. These differences are modest at the observed angles; we do not know how much a budget shift of this size changes downstream accuracy. The normalised chord also routes gradients to the endpoints through the normalisation rather than through the slerp coefficients, so the two parameterisations can differ in optimisation even where their targets nearly coincide. We therefore report the empirical gap as observed from single runs and do not attribute it to the position law alone.



(a) Geodesic targets and the off-sphere chord.

(b) Template, caption, and image regions.

Figure 3: **Geometry of the two captions.** **(a)** Slerp targets (green) stay on $\mathcal{S}^{d-1}$ along the arc between an image's template and spatial caption, whereas the linear chord (dotted) passes through the interior of the sphere. **(b)** Schematic of the tight template region $\mathcal{M}^{\text{tmpl}}$, the broader caption region $\mathcal{M}^{\text{sp}}$, and the image region $\mathcal{M}^{\text{img}}$, which the modality gap (Liang et al., 2022) separates from both; θ_i is the template–caption angle of image i.

B Additional Results

Table 3: Template-based transfer classification accuracy (%). The 11 datasets other than ImageNet are zero-shot, and "Avg ZS" is their mean. Label-free numbers are from Gong et al. (2025). Template-only FLYP obtains 86.40 on ImageNet and 69.28 Avg ZS (Table 1).

		Zero-shot datasets											
Alignment strategies	ImageNet	CIFAR-10	CIFAR-100	Caltech	FER	Pets	DTD	RESISC	EuroSAT	PCAM	IN-Sketch	IN-O	Avg ZS
Label-free													
CLIP	74.90	95.20	71.08	83.30	50.00	93.21	55.21	63.35	62.65	52.00	59.59	32.25	65.26
Projection	70.42	95.10	74.14	80.31	24.51	80.02	37.02	37.03	32.48	50.02	56.04	28.05	54.07
Feature	75.16	95.77	75.13	83.58	49.62	93.35	55.73	63.13	63.11	51.34	58.08	34.20	65.73
DIVA	74.84	95.10	70.96	84.01	49.47	93.57	55.05	63.48	62.93	53.79	59.53	32.05	65.45
KUEA	75.52	96.27	77.04	84.32	50.31	93.40	55.74	63.83	63.83	52.68	59.67	34.90	66.54
Label-supervised													
Ours	85.54	**96.65**	**79.66**	**85.71**	**53.87**	**93.59**	**56.91**	**64.70**	**68.44**	**58.48**	**63.14**	**48.35**	**69.96**

Table 4: Spatial tasks with ViT-L/14 (%), CapRL captions. Bold marks the best entry per column; note that Ours is label-supervised on ImageNet, whereas CLIP and KUEA are label-free. Template-only and Qwen2.5-VL averages are in Table 1.

Method	SVHN	GTSRB	CLEVR-Dist.	CLEVR-Counts	**Avg**
Zero-shot					
CLIP	57.02	50.55	20.21	19.43	36.80
KUEA	**59.63**	52.53	23.39	21.11	39.17
Ours	58.87	**54.41**	**24.24**	**27.92**	**41.36**
Linear probe					
CLIP	65.20	72.94	22.97	41.25	50.59
KUEA	69.39	**74.51**	30.82	49.67	56.10
Ours	**74.00**	74.02	**37.94**	**55.89**	**60.46**

Held-out loss along the arc. On a held-out pool of 24,000 ImageNet validation images with independently generated CapRL captions, we evaluate $\mathcal{L}(\lambda)$ on a shared grid of λ values over 40 batches of 512 images with distinct classes (20,480 image evaluations). Figure 4 reports the paired change relative to frozen CLIP.

Embedding geometry. We embed ImageNet training images and their template and spatial captions with frozen OpenAI CLIP ViT-L/14. The *cone gap* between modalities a and b is $\|\hat{\mu}_a - \hat{\mu}_b\|_2$, where $\hat{\mu}_a$ is the L_2-normalised mean embedding of modality a (at most 2). *Matched* cosines use co-indexed

Table 5: Zero-shot retrieval R@1/R@5 with ViT-L/14 (%), CapRL captions. Text-to-image R@1 for template-only training and the Qwen2.5-VL variant is in Table 1. Repeated evaluation passes of probes and retrieval agree to within about 0.5 points; this measures evaluation repeatability, not variation across training runs.

	Flickr30K				MSCOCO			
Method	I→T R@1	I→T R@5	T→I R@1	T→I R@5	I→T R@1	I→T R@5	T→I R@1	T→I R@5
CLIP	81.40	96.20	63.62	86.36	50.64	74.20	34.51	59.21
KUEA	**83.00**	**96.90**	64.78	87.20	51.72	75.52	35.98	61.08
Ours	81.80	95.00	**67.54**	**88.58**	**54.12**	**77.10**	**40.97**	**66.24**

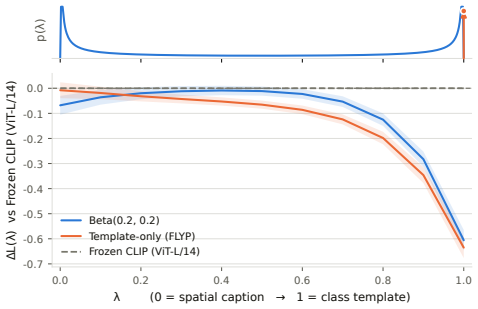


Figure 4: **Held-out loss along the arc.** Paired change in contrastive loss relative to frozen CLIP, on the same held-out batches and grid of λ. $\mathrm{Beta}(0.2, 0.2)$ lowers the loss at the spatial-caption end ($\lambda = 0$), where template-only training barely changes it. Template-only training, which never samples the interior, has lower loss over the rest of the arc, reflecting parameter sharing. Bands show one batch standard deviation. These are loss diagnostics, not downstream bounds.

pairs; *random cross-pair* cosines average over mismatched pairs across the two modalities; *intra-modality* cosines use two distinct samples of one modality. *Centroid* ℓ_2 is the distance between unnormalised means.

Table 6: Frozen ViT-L/14 embedding geometry on a random subset of the ImageNet training split. Cross-modal statistics (*top*) and mean cosine between mismatched pairs within each modality (*bottom*).

	Image–template	**Image–spatial**	**Template–spatial**
Matched cosine mean	0.258	0.281	0.350
Matched cosine std	0.038	0.037	0.153
Cone gap $\|\hat{\mu}_a - \hat{\mu}_b\|_2$	1.239	1.259	1.074
Random cross-pair cosine mean	0.119	0.068	0.142
Centroid ℓ_2 (unnormalised means)	0.888	0.761	0.676
Intra-modality random-pair cosine mean			
Image / template / spatial	0.500 /	0.526 /	0.214

C SPATIAL CAPTION GENERATION

Spatial captions are produced once, offline, before training. We use CapRL (Xing et al., 2026), a dense captioning model trained with reinforcement learning to produce detailed descriptions of natural images. CapRL is invoked once per ImageNet-1K training image; the captions are stored and never regenerated or back-propagated through. At training time they are fed to CLIP's text encoder g_ω exactly like templates.

Prompt. Every image receives the same prompt:

```
List the objects, their spatial positions, and relationships
in this image.
Use short comma-separated phrases. No sentences. No filler
words.
Format: [object] [position/relation], [object]
[position/relation], ...
```

The format constraint concentrates the caption's content in object–position phrases rather than in sentence-level syntax.

Is an MLLM captioner with a CLIP-like vision tower circular? An MLLM captioner may use a CLIP-family vision encoder, so its captions might seem unable to contain information beyond what CLIP already encodes. Li et al. (2025a) show otherwise: generative MLLMs perceive substantially more than CLIP *with the same vision encoder and weights*. They attribute the gap to components absent from CLIP's pooled-similarity inference: (i) the use of all patch tokens rather than only the `[CLS]` token, (ii) positional embeddings that preserve where each patch lies, and (iii) prompt-driven attention that reads selectively from local regions. CLIP's zero-shot inference collapses an image to one global vector before computing similarity. A captioner can therefore read out spatial information that CLIP's pooled embedding does not surface, and arc smoothing routes that information back into CLIP through its text branch.

Caption failure modes. We observe four recurring failure modes. *(i) Class-label leakage.* Some captions are dominated by the class name and nearly paraphrase the template (e.g. for *golden retriever*, "a golden retriever in the centre, fur visible, dog standing"). The angle θ_i is then small, so the interpolated targets stay close to both endpoints; this does not guarantee that the caption adds useful spatial supervision. *(ii) Hallucinated objects.* The captioner occasionally invents objects. Such captions provide misleading supervision; template-side targets retain class information, but neither slerp nor the mixing law identifies or removes hallucinations, and our analysis gives no robustness guarantee against them. *(iii) Format violations and very short captions.* Some outputs are prose despite the prompt, or contain only one or two phrases. We discard these and fall back to template-only supervision for those images. The fallback rate is small on ImageNet but could be larger in domains where the captioner follows the format less reliably. *(iv) Class-conflated captions in cluttered scenes.* With several objects of the same coarse type, the captioner sometimes describes the scene at the level of class plurality rather than the labelled instance; the caption remains informative about layout but is less aligned with the single class label.

Second captioner and compute. Qwen2.5-VL-7B-Instruct (Bai et al., 2025) captions use the same prompt and are likewise generated once and never updated. CapRL is itself trained from a Qwen2.5-VL model, so the two captioners share a model family but differ in scale and training objective. Captioning ImageNet-1K is a one-time cost at dataset preparation; all fine-tuning runs reuse the stored captions.

D Proofs and full-objective characterisation

All statements condition on a fixed batch, its two captions per image, and the current parameters, with a fixed positive temperature. The identities hold at any parameter setting, including when both encoders and the temperature are learned; they do not compare optimisation trajectories.

The midpoint frame and the arc tangents. For $0 < \theta_j < \pi$, $\|t_j + s_j\|_2 = 2\cos(\theta_j/2)$, $\|t_j - s_j\|_2 = 2\sin(\theta_j/2)$, and $(t_j + s_j)^\top (t_j - s_j) = 0$, so $e_j^\pm$ in Eq. (7) are orthonormal, with $t_j = \cos(\frac{\theta_j}{2})e_j^+ + \sin(\frac{\theta_j}{2})e_j^-$ and $s_j = \cos(\frac{\theta_j}{2})e_j^+ - \sin(\frac{\theta_j}{2})e_j^-$. Substituting into Eq. (2) and using $\sin(\lambda\theta) \pm \sin((1-\lambda)\theta)$ sum-to-product identities gives $z_j(\lambda) = \cos\varphi_j\, e_j^+ + \sin\varphi_j\, e_j^-$. Hence

$$\partial_\lambda z_j(\lambda) = \theta_j\big[-\sin\varphi_j\, e_j^+ + \cos\varphi_j\, e_j^-\big], \qquad \partial_\lambda^2 z_j(\lambda) = -\theta_j^2\, z_j(\lambda). \tag{14}$$

In particular $\partial_\lambda z_j(0) = \theta_j w_j^{s\to t}$ and $\partial_\lambda z_j(1) = -\theta_j w_j^{t\to s}$, where

$$w_j^{s\to t} = \sin(\tfrac{\theta_j}{2})\, e_j^+ + \cos(\tfrac{\theta_j}{2})\, e_j^-, \qquad w_j^{t\to s} = \sin(\tfrac{\theta_j}{2})\, e_j^+ - \cos(\tfrac{\theta_j}{2})\, e_j^- \tag{15}$$

are the unit tangents to the arc at s_j pointing toward t_j and at t_j pointing toward s_j. Each is orthogonal to its endpoint, and $w_j^{t\to s} = (s_j - \cos\theta_j\, t_j)/\sin\theta_j$. If $\theta_j = 0$, the target is constant, e_j^- is arbitrary, and pair j does not contribute to any λ-derivative.

Symmetric Beta laws. For $\nu_\beta = \mathrm{Beta}(\beta,\beta)$, the distance $D = \min(\Lambda, 1-\Lambda)$ has density $g_\beta(d) = 2[d(1-d)]^{\beta-1}/\mathrm{B}(\beta,\beta)$ on $(0,\frac{1}{2})$, where B is the beta function. Hence

$$\mathbb{E}_{\nu_\beta}[D^k] = \frac{2}{\mathrm{B}(\beta,\beta)} \int_0^{1/2} x^{\beta+k-1}(1-x)^{\beta-1}\,dx,$$
$$C_\delta(\nu_\beta) = 1 - \frac{2}{\mathrm{B}(\beta,\beta)} \int_0^{\delta} x^{\beta-1}(1-x)^{\beta-1}\,dx, \tag{16}$$

with $C_\delta(\nu) = \Pr_\nu(\delta < \Lambda < 1-\delta) = S_\nu(\delta)$ and $\mathcal{I}(\nu_\beta) = \mathbb{E}_{\nu_\beta}[D]$. For $\beta_2 > \beta_1$, the ratio $g_{\beta_2}(d)/g_{\beta_1}(d) \propto [d(1-d)]^{\beta_2-\beta_1}$ is strictly increasing on $(0,\frac{1}{2})$, because $d(1-d)$ is. This monotone likelihood ratio implies strict first-order stochastic dominance: $S_{\nu_{\beta_2}}(d) > S_{\nu_{\beta_1}}(d)$ for $0 < d < \frac{1}{2}$. Consequently, the expectation of any strictly increasing function of D strictly increases with β. As $\beta \to 0^+$, ν_β converges weakly to the balanced endpoint law, since $\mathbb{E}[\Lambda(1-\Lambda)] = \beta/[2(2\beta+1)] \to 0$ and symmetry fixes the weights. At $\beta = 1$, $\mathcal{I} = \frac{1}{4}$ and $\mathbb{E}[D^2] = \frac{1}{12}$. Table 7 evaluates Eq. (16), $r_\nu(\bar\theta) = \mathbb{E}_\nu \cos((\Lambda - \frac{1}{2})\bar\theta)$, and $v_\nu(\bar\theta) = \mathbb{E}_\nu \sin^2((\Lambda - \frac{1}{2})\bar\theta)$ by quadrature; for the uniform law, $r = 2\sin(\bar\theta/2)/\bar\theta$ and $v = \frac{1}{2}(1 - \sin\bar\theta/\bar\theta)$ in closed form.

Table 7: Analytic statistics of the symmetric laws in our sweep: interior budget $\mathcal{I}$, central mass $C_{0.1} = \Pr(0.1 < \Lambda < 0.9)$, and the mode gains r_ν and v_ν (endpoint fidelity) of Proposition 1, evaluated at $\bar\theta = \arccos(0.35) \approx 69.5°$, the angle of the mean template–caption cosine. Because $v_\nu(\theta) = \theta^2 \mathrm{Var}_\nu(\Lambda) + O(\theta^4)$ for symmetric laws, $\mathrm{Var}_\nu(\Lambda)$ is the angle-free, small-angle form of endpoint fidelity.

	Endpoints	$\mathrm{Beta}(0.1,0.1)$	$\mathrm{Beta}(0.2,0.2)$	$\mathrm{Beta}(0.5,0.5)$	Uniform
Interior budget $\mathcal{I}(\nu)$	0	0.058	0.101	0.182	0.250
Central mass $C_{0.1}(\nu)$	0%	18.7%	32.7%	59.0%	80.0%
Common-mode gain $r_\nu(\bar\theta)$	0.822	0.851	0.872	0.910	0.940
Endpoint fidelity $v_\nu(\bar\theta)$	0.325	0.273	0.236	0.168	0.114
Variance $\mathrm{Var}_\nu(\Lambda)$	0.250	0.208	0.179	0.125	0.083

Proof of Proposition 1. Write $D = D(\Lambda)$. Then $|\varphi| = (\frac{1}{2} - D)\theta \in [0, \theta/2] \subset [0, \pi/2)$, so $\cos\varphi = \cos((\frac{1}{2} - D)\theta) \in [\cos(\theta/2), 1]$ and $\sin^2\varphi = \sin^2((\frac{1}{2} - D)\theta) \in [0, \sin^2(\theta/2)]$. Since $\sin\varphi$ is odd in $\Lambda - \frac{1}{2}$, its expectation vanishes under a symmetric law; this proves (i). On $[0, \frac{1}{2}]$, $d \mapsto \sin^2((\frac{1}{2} - d)\theta)$ is strictly decreasing, so $v_\nu = \sin^2(\theta/2)$ forces $D = 0$ almost surely, and symmetry then puts mass $\frac{1}{2}$ on each endpoint; likewise $v_\nu = 0$ forces $D = \frac{1}{2}$ almost surely; this proves (ii). For (iii), $d \mapsto \cos((\frac{1}{2} - d)\theta)$ and $d \mapsto d$ are strictly increasing and $d \mapsto \sin^2((\frac{1}{2} - d)\theta)$ is strictly decreasing, so the claims follow from the stochastic dominance established above. □

Proof of Proposition 2. For $\lambda < \frac{1}{2}$, $\mathcal{L}(\lambda) - \mathcal{L}(0) = \int_0^{D(\lambda)} \mathcal{L}'(\xi)\,d\xi$. For $\lambda > \frac{1}{2}$, substituting $\lambda' = 1 - \xi$ in $\mathcal{L}(1) - \mathcal{L}(\lambda) = \int_\lambda^1 \mathcal{L}'(\lambda')\,d\lambda'$ gives $\mathcal{L}(\lambda) - \mathcal{L}(1) = -\int_0^{D(\lambda)} \mathcal{L}'(1-\xi)\,d\xi$. At $\lambda = \frac{1}{2}$ both expressions hold with $D = \frac{1}{2}$, and their average equals $\mathcal{L}(\frac{1}{2}) - \mathcal{L}_{\mathrm{end}}$. Under a symmetric law, the map $\Lambda \mapsto 1-\Lambda$ preserves ν and D, so $\mathbb{E}[\mathbf{1}\{\Lambda < \frac{1}{2}\}\,h(D)] = \mathbb{E}[\mathbf{1}\{\Lambda > \frac{1}{2}\}\,h(D)]$ for every integrable h, and the endpoint values average to $\mathcal{L}_{\mathrm{end}}$. (Equivalently, $\mathbb{E}_\nu \mathcal{L}(\Lambda) = \mathbb{E}_\nu \frac{1}{2}[\mathcal{L}(D) + \mathcal{L}(1-D)]$ and $\frac{1}{2}[\mathcal{L}(D) + \mathcal{L}(1-D)] - \mathcal{L}_{\mathrm{end}} = \int_0^D \rho(\xi)\,d\xi$.) Therefore

$$\mathcal{L}_\nu - \mathcal{L}_{\mathrm{end}} = \mathbb{E}_\nu \int_0^{D(\Lambda)} \rho(\xi)\,d\xi = \int_0^{1/2} \rho(\xi) \Pr_\nu(D(\Lambda) > \xi)\,d\xi,$$

where the exchange of expectation and integral is justified by Fubini's theorem because $\mathcal{L}$ is smooth, so $\mathcal{L}'$ is bounded on $[0,1]$. This is Eq. (9). For the derivative, write $\ell_i^{\mathrm{I}\to\mathrm{T}} = \log\sum_j e^{a_{ij}} - a_{ii}$ and $\ell_j^{\mathrm{T}\to\mathrm{I}} = \log\sum_i e^{a_{ij}} - a_{jj}$, so that $\mathcal{L} = \frac{1}{2B}\sum_i \ell_i^{\mathrm{I}\to\mathrm{T}} + \frac{1}{2B}\sum_j \ell_j^{\mathrm{T}\to\mathrm{I}}$. Then $\frac{d}{d\lambda}\ell_i^{\mathrm{I}\to\mathrm{T}} = \sum_j (p_{ij} - \delta_{ij})a'_{ij}$ and $\frac{d}{d\lambda}\ell_j^{\mathrm{T}\to\mathrm{I}} = \sum_i (q_{ij} - \delta_{ij})a'_{ij}$ with $a'_{ij} = u_i^\top \partial_\lambda z_j/\tau$, hence

$$\mathcal{L}'(\lambda) = \frac{1}{2B\tau} \sum_{i,j} \left(P_{ij}^\lambda - 2\delta_{ij}\right) u_i^\top \partial_\lambda z_j(\lambda). \tag{17}$$

Substituting $\partial_\lambda z_j(0) = \theta_j w_j^{s\to t}$ and $\partial_\lambda z_j(1) = -\theta_j w_j^{t\to s}$ into $\rho(0) = \frac{1}{2}[\mathcal{L}'(0) - \mathcal{L}'(1)]$ gives the tangent form

$$\rho(0) = \frac{1}{4B\tau} \sum_{i,j} \theta_j \Big[\big(P_{ij}^0 - 2\delta_{ij}\big)\, u_i^\top w_j^{s\to t} + \big(P_{ij}^1 - 2\delta_{ij}\big)\, u_i^\top w_j^{t\to s} \Big]. \tag{18}$$

By Eq. (15), $u_i^\top w_j^{s\to t} = \sin(\frac{\theta_j}{2})c_{ij} + \cos(\frac{\theta_j}{2})d_{ij}$ and $u_i^\top w_j^{t\to s} = \sin(\frac{\theta_j}{2})c_{ij} - \cos(\frac{\theta_j}{2})d_{ij}$. Collecting the c_{ij} and d_{ij} terms gives Eq. (10); in the differential part the two $-2\delta_{ij}$ terms cancel. For the expansion, Taylor's theorem gives $\rho(\xi) = \rho(0) + \rho'(0)\xi + \frac{1}{2}\rho''(\eta_\xi)\xi^2$ for some $\eta_\xi \in [0, \xi]$. Inserting this into Eq. (9) and using the layer-cake identity $\int_0^{1/2} \xi^k S_\nu(\xi)\, d\xi = \mathbb{E}_\nu[D^{k+1}]/(k+1)$ yields the stated expansion, with $|R_\nu| \le \frac{1}{2} \sup|\rho''| \cdot \mathbb{E}_\nu[D^3]/3$. The second derivative needed for $\rho'(0) = \frac{1}{2}[\mathcal{L}''(0) + \mathcal{L}''(1)]$ follows by differentiating Eq. (17) once more with $\partial_\lambda p_{ij} = p_{ij}(a'_{ij} - \sum_k p_{ik}a'_{ik})$, $\partial_\lambda q_{ij} = q_{ij}(a'_{ij} - \sum_k q_{kj}a'_{kj})$, and $a''_{ij} = -\theta_j^2 a_{ij}$ from Eq. (14). □

Scope of the expansion. Each derivative of ρ brings a factor of order $\theta_j |u_i^\top w_j|/\tau$ through the logits, so at CLIP's temperature ($\tau \approx 0.01$) the ratio $\rho'(0)/\rho(0)$ can be large, and the leading term $\mathcal{I}(\nu)\rho(0)$ need not dominate the remainder. The ratio $\mathbb{E}_\nu[D^2]/\mathcal{I}(\nu)$ is 0.29, 0.29, 0.31, and 0.33 for $\beta = 0.1, 0.2, 0.5, 1$, so this caveat applies to every law in our sweep, not only to U-shaped ones. The interpretation in Section 5.2 is therefore a description of the leading term; only the identity (9) is exact.

Remark 1 (The two scalars do not rank all laws)**.** *For $\theta \le \pi/2$, the map $d \mapsto \sin^2((\frac{1}{2} - d)\theta)$ is convex on $[0, \frac{1}{2}]$, so among symmetric laws with a given interior budget $\mathcal{I}$, endpoint fidelity is maximised by the mixture $(1 - 2\mathcal{I}) \cdot \frac{1}{2}(\delta_0 + \delta_1) + 2\mathcal{I} \cdot \delta_{1/2}$. At $\mathcal{I} = 0.101$ and $\bar{\theta}$, this mixture has $v = 0.259$, compared with 0.236 for $\mathrm{Beta}(0.2, 0.2)$. Endpoint fidelity and interior budget therefore do not single out the Beta family. We use them only to order laws along the Beta path, on which they move in opposite directions (Proposition 1(iii)). Which interior law is best for downstream metrics is an empirical question that the analysis does not settle. Moreover, the 73% fidelity ratio of $\mathrm{Beta}(0.2, 0.2)$ to the endpoint law depends only mildly on angle: it is 0.714 in the small-angle limit, 0.725 at $\bar{\theta}$, and 0.738 at $100°$.*

Averaging targets versus averaging losses. The loss is nonlinear in its logits, although each logit is linear in its target. The following identity characterises the full symmetric objective, including all negative logits.

Proposition 3 (Mean logits and prediction variation)**.** *Let $A(\lambda) \in \mathbb{R}^{B\times B}$ have entries $a_{ij}(\lambda) = u_i^\top z_j(\lambda)/\tau$, let $\bar{A} = \mathbb{E}_\nu[A(\Lambda)]$, and let $F(A)$ be the symmetric InfoNCE loss as a function of its logit matrix. Index its $2B$ queries (all rows and all columns) by q, with logit vectors $a_q(\lambda)$, and let $p_q(\lambda) = \mathrm{softmax}(a_q(\lambda))$ and $\bar{p}_q = \mathrm{softmax}(\mathbb{E}_\nu[a_q(\Lambda)])$. Then*

$$\mathbb{E}_\nu[F(A(\Lambda))] = F(\bar{A}) + \frac{1}{2B} \sum_{q=1}^{2B} \mathbb{E}_\nu\big[\mathrm{KL}\left(\bar{p}_q \,\|\, p_q(\Lambda)\right)\big]. \tag{19}$$

For a symmetric law, the mean target is $\mathbb{E}_\nu[z_j(\Lambda)] = r_\nu(\theta_j)\, e_j^+$.

Proof. Let $\psi(a) = \log \sum_k e^{a_k}$ and $\bar{a} = \mathbb{E}[a]$. By direct substitution, $\mathrm{KL}(\mathrm{softmax}(\bar{a}) \,\|\, \mathrm{softmax}(a)) = \psi(a) - \psi(\bar{a}) - \mathrm{softmax}(\bar{a})^\top (a - \bar{a})$, and the last term has expectation zero. The positive logit enters the cross-entropy $\psi(a) - a_y$ linearly and commutes with expectation. Hence the expected cross-entropy of each query equals its value at $\bar{a}$ plus the expected KL divergence; summing over queries proves Eq. (19). The mean target follows by taking expectations in Eq. (7), where symmetry removes the sine term. □

The KL term is nonnegative under every law, so training on the averaged target $r_\nu(\theta_j)e_j^+$ is not equivalent to arc smoothing. Because $r_\nu(\theta_j)$ varies across examples, the mean-logit term is also not a single global temperature change.

Batch-shared versus per-example draws. Let $\delta_q = a_q(\Lambda) - \bar{a}_q$ and $H_q = \mathrm{diag}(\bar{p}_q) - \bar{p}_q\bar{p}_q^\top$. Under per-example draws, column j uses its own coefficient $\Lambda_j \sim \nu$, independently across j; under the shared draw, all columns use one Λ. The two schemes have the same mean logits $\bar{A}$ and therefore differ only in the prediction-variation term of Eq. (19).

Proposition 4 (Batch-shared versus per-example draws)**.** *To second order in δ_q, $\mathbb{E}_\nu \mathrm{KL}(\bar{p}_q \| p_q(\Lambda)) = \frac{1}{2}\mathbb{E}_\nu[\delta_q^\top H_q \delta_q]$, with*

$$\mathbb{E}_\nu[\delta_q^\top H_q \delta_q] = \sum_k \bar{p}_{qk}(1-\bar{p}_{qk})\,\mathrm{Var}(\delta_{qk}) - \sum_{k\neq l} \bar{p}_{qk}\bar{p}_{ql}\,\mathrm{Cov}(\delta_{qk},\delta_{ql}). \tag{20}$$

Column queries have the same joint law under both schemes. For a row query i, per-example draws make every cross-covariance vanish, whereas the shared draw gives

$$\mathrm{Cov}(\delta_{ik},\delta_{il}) = \tau^{-2}\big[\,\mathrm{Cov}(\cos\varphi_k,\cos\varphi_l)\,c_{ik}c_{il} + \mathbb{E}[\sin\varphi_k \sin\varphi_l]\,d_{ik}d_{il}\big], \tag{21}$$

in which both bracketed coefficients are nonnegative.

Proof of Proposition 4. From the proof above, $\mathrm{KL}(\bar{p}_q\|p_q) = \psi(\bar{a}_q+\delta_q) - \psi(\bar{a}_q) - \bar{p}_q^\top\delta_q$. Since $\nabla\psi(\bar{a}_q) = \bar{p}_q$ and $\nabla^2\psi(\bar{a}_q) = H_q$, Taylor's theorem gives $\mathrm{KL}(\bar{p}_q\|p_q) = \frac{1}{2}\delta_q^\top H_q\delta_q + O(\|\delta_q\|^3)$. Expanding, $\mathbb{E}[\delta_q^\top H_q\delta_q] = \sum_k \bar{p}_{qk}\,\mathrm{Var}(\delta_{qk}) - \sum_{k,l}\bar{p}_{qk}\bar{p}_{ql}\,\mathrm{Cov}(\delta_{qk},\delta_{ql})$, which rearranges to Eq. (20). Every entry of column j depends only on the coefficient assigned to pair j, whose law is ν under both schemes, so column queries are unaffected. For row i, entry k is $a_{ik} = [\cos\varphi_k\, c_{ik} + \sin\varphi_k\, d_{ik}]/\tau$. Independent draws make the entries independent, while each marginal variance is unchanged. Under the shared draw, $\delta_{ik} = [(\cos\varphi_k - r_\nu(\theta_k))c_{ik} + \sin\varphi_k\, d_{ik}]/\tau$, and the cross terms $\mathbb{E}[(\cos\varphi_k - r_\nu(\theta_k))\sin\varphi_l]$ vanish because the first factor is even and the second odd in $\Lambda - \frac{1}{2}$; this gives Eq. (21). Both $\cos\varphi_k$ and $\cos\varphi_l$ are increasing functions of D, so their covariance is nonnegative by Chebyshev's association inequality, and $\sin\varphi_k\sin\varphi_l \geq 0$ pointwise because both factors have the sign of $\Lambda - \frac{1}{2}$. □

When an image's common-mode scores against the other pairs share a sign (random cross-pair cosines are positive; Table 6) and dominate the differential terms, the shared draw's positive cross-covariances reduce the second-order prediction-variation penalty. This comparison assumes small logit variation ($|\delta_{qk}| \ll 1$, i.e. $\theta_k|d_{ik}| + \theta_k^2|c_{ik}| \ll \tau$), which need not hold at CLIP's temperature, so we offer it only as a heuristic motivation. Per-example draws, by contrast, add the full marginal variance of every logit without this cancellation. The shared draw also makes the training objective exactly the law average $\mathcal{L}_\nu$ of the one-parameter family analysed in Section 5.

Numerical checks. We verified the identities numerically on a random batch ($B = 5$, $d = 8$, $\tau = 0.2$) of unit vectors with softmax InfoNCE. The profile identity (9) (by quadrature) and Propositions 3 and 4 (the latter's covariance formula) agree to relative error below 10^{-6}; the two forms of $\rho(0)$, Eqs. (10) and (18), agree to machine precision, and both match a finite difference to 10^{-5}. All entries of Table 7 and the numbers in Appendix A were computed by quadrature.

E Broader Societal Impacts

This work aims to improve the fine-grained spatial perception of CLIP-style vision–language models while preserving their zero-shot transfer. Potential positive impacts include more reliable visual reasoning in downstream multimodal systems, especially for counting, spatial relations, object layout, and local grounding, which may benefit assistive visual interfaces, image retrieval, educational tools, robotics, and document or scene understanding.

At the same time, better spatial perception may increase the effectiveness of systems used for surveillance, monitoring, or other privacy-sensitive applications. The method relies on captions from an external multimodal captioner, which may inherit or amplify biases, stereotypes, omissions, or hallucinations; if such captions systematically misdescribe people, objects, or environments, the fine-tuned model may encode those errors. Improved spatial reasoning also does not guarantee reliability in high-stakes domains such as medical, legal, security, or safety-critical decision-making.

We therefore view this method as a representation-learning improvement rather than a complete safety solution. Deployment should include task-specific validation, bias and robustness evaluation

across demographic and geographic groups, privacy safeguards for image data, and human oversight in high-stakes settings.