Text as Partial Constraint: Core–Residual Alignment for Robust Vision–Language Learning

Chengzhen Yu[1] Canran Xiao[2,*] Siyuan Ma[1] Yang Liu[1]

[1]Nanyang Technological University (NTU)
[2]Shenzhen Campus of Sun Yat-sen University
[*]Corresponding author: xiaocr3@mail.sysu.edu.cn

Abstract

Vision–language alignment powers open-vocabulary recognition, retrieval, and LVLM grounding, yet natural captions are often underspecified, making similarity brittle and overly confident under paraphrase and omitted details. We aim to learn representations whose matching is stable across caption views and whose confidence reflects how strongly text constrains an image. We propose TEXT AS PARTIAL CONSTRAINT (TPC), a core–residual alignment framework that treats multi-view captions as incomplete supervision: it distills a consensus semantic core as the alignment target, learns a single-view core predictor for standard inference with one query, and explicitly discourages vision–language similarity from depending on the orthogonal "unsaid" residual. An uncertainty-aware contrastive objective further softens alignment when caption views disagree, reducing overconfident updates under weak language constraints. Across zero-shot recognition and adversarial robustness, TPC achieves 81.42/64.05 Top-1 clean/robust accuracy on ImageNet and 76.19/52.03 on an Avg-14 transfer suite, while improving LVLM transfer with 85.16 POPE F1 and 59.57 OKVQA accuracy under an LLaVA-1.5-7B stack. These results suggest that modeling text as a partial constraint is a practical and principled route to more reliable vision–language representations under underspecified language supervision.

1 Introduction

Vision–language alignment underpins open-vocabulary recognition, cross-modal retrieval, and increasingly serves as the perceptual backbone of large vision–language models (LVLMs) used for interactive reasoning and decision support (Radford et al., 2021; Jia et al., 2021; Liu et al., 2023; Awadalla et al., 2023). Its practical appeal comes from learning with abundant natural captions, yet the same supervision is inherently incomplete: users and annotators routinely omit details, rely on shared context, and phrase descriptions in diverse ways. As these models move from curated benchmarks to real queries and safety-critical settings, failures driven by underspecified language become costly—mis-ranked retrieval, unstable predictions under paraphrase, and overconfident downstream generations.

Prior work has advanced vision–language pretraining by scaling contrastive dual encoders and improving training objectives (Radford et al., 2021; Jia et al., 2021; Zhai et al., 2023; Cherti et al., 2023; Sun et al., 2023). Yet most methods still treat each caption/prompt as a *complete* target, despite natural supervision being *partial*: different textual views share a stable core but diverge in omitted or ambiguous details. This mismatch surfaces in compositional and negation stress tests, where strong VLMs remain sensitive to small but meaning-relevant textual changes (Parcalabescu et al., 2022; Thrush et al., 2022; Hsieh et al., 2023; Alhamoud et al., 2025), and in real under-specified user queries where adding missing constraints yields large gains (Choi et al., 2026). In parallel, robustness training mainly targets *visual* shift (Mao et al., 2024; Wang et al., 2024a; Schlarmann et al., 2024b;a), while LVLM mitigation is often post hoc and does not curb representation-level over-commitment under weak text constraints (Li et al., 2023; Leng et al., 2024a). Thus, a key gap is the lack of a principled treatment of language supervision as *incomplete* that prevents overconfident alignment to weakly supported, view-specific details.

This paper asks: *Can we learn vision–language representations that align to what captions consistently specify, while avoiding over-confident commitment to what they leave underspecified?* As illustrated in Fig. 1, we answer this question by reframing natural captions as partial constraints: multiple caption views reveal a shared consensus core, while their view-specific deviations expose an "unsaid" residual space. Building on this observation, TPC aligns images to the consensus core, learns a single-caption core filter for standard inference, suppresses residual over-commitment, and calibrates confidence according to caption-view disagreement.

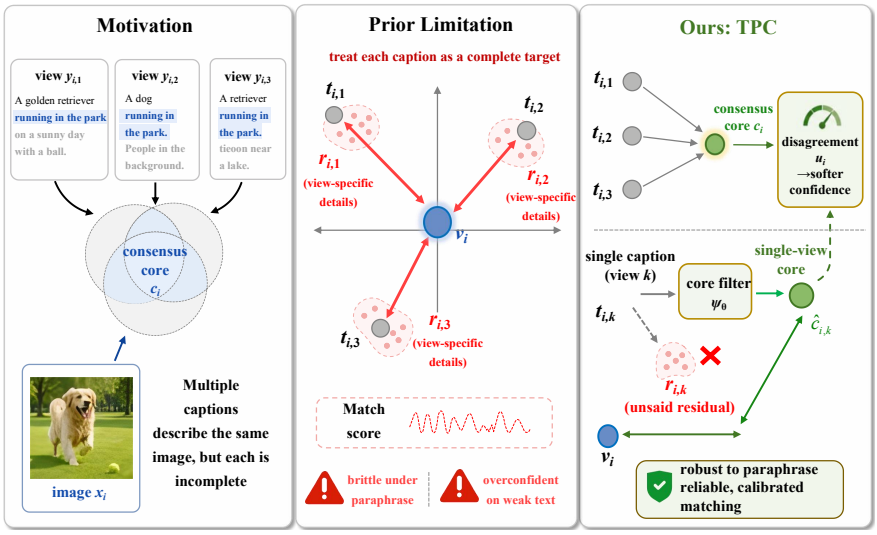

Figure 1: Unlike prior methods that over-align to each caption as a complete target, TPC extracts the core, suppresses residual commitment, and produces calibrated, robust vision–language alignment.

Our contributions are as follows: ***(i)*** We formalize caption-view variation as a form of language underspecification and identify over-commitment to what is "unsaid" as a central mechanism behind brittleness and downstream hallucination. ***(ii)*** We introduce a training principle that uses multi-view language to learn alignment that prioritizes view-invariant semantics and produces confidence that tracks how strongly text constrains the instance, while remaining deployable with a single query text. ***(iii)*** Across zero-shot recognition, adversarial robustness, and LVLM transfer, the resulting representations improve both accuracy and reliability, yielding stronger robustness and reduced hallucination under underspecified language supervision.

2 Related Work

Contrastive VLP under noisy and underspecified text. CLIP and ALIGN showed that contrastive dual encoders trained on large-scale image–text pairs enable strong zero-shot transfer (Radford et al., 2021; Jia et al., 2021), and later work mostly improved results through scaling and refined objectives/recipes (Cherti et al., 2023; Zhai et al., 2023; Sun et al., 2023). Yet natural captions are partial: annotators and paraphrases share core semantics but differ in omitted details, making similarity brittle under lexical/semantic edits and negation (Hsieh et al., 2023; Dumpala et al., 2024; Alhamoud et al., 2025) and under-specified real queries (Choi et al., 2026). Because training often treats each caption as complete (or merely adds positives), models can over-align to view-specific residuals. TPC instead treats multi-caption supervision as a partial constraint, aligning to a consensus core while suppressing the "unsaid" residual so similarity reflects what is consistently supported.

Robust alignment under distribution shift and attacks. Robust VLM training has focused on adversarial and distributional *visual* shift. Related methods include TeCoA (Mao et al., 2024), PMG-AFT (Wang et al., 2024a), and TGA-ZSR (Yu et al., 2024), as well as plug-in robustness via unsupervised adversarial fine-tuning (RobustCLIP/FARE) (Schlarmann et al., 2024b) and weak-to-strong robustness transfer (Adv-W2S) (Schlarmann et al., 2024a). While effective, they typically assume the text fully specifies the semantics and thus do not address caption-view ambiguity. We instead treat caption variation as the shift source, model views via an instance-wise ambiguity set, and improve worst-case similarity by strengthening core alignment while reducing sensitivity to unsaid residuals, with disagreement-aware temperature scaling to curb overconfident updates. Adjacent continual vision–language and multimodal adaptation work studies how to preserve compositional structure, prompt-invariant certificates, or concept-graph memory across task streams (Xiao et al., 2026; Zhang et al., 2026; Zhou et al., 2026). These methods address temporal adaptation, while our focus is alignment under underspecified language supervision; the two directions are complementary.

Grounding and hallucination in LVLMs. LVLMs such as LLaVA and OpenFlamingo largely inherit a frozen CLIP-like vision encoder with a lightweight connector (Liu et al., 2023; Awadalla et al., 2023), and benchmarks like POPE highlight persistent object hallucination (Li et al., 2023). Mitigations span decoding-time calibration (VCD) (Leng et al., 2024a), connector-side fine-grained alignment (Jiang et al., 2025), and post-hoc concept shaping (VL-SAE) (Shen et al., 2025), among other LVLM alignment objectives

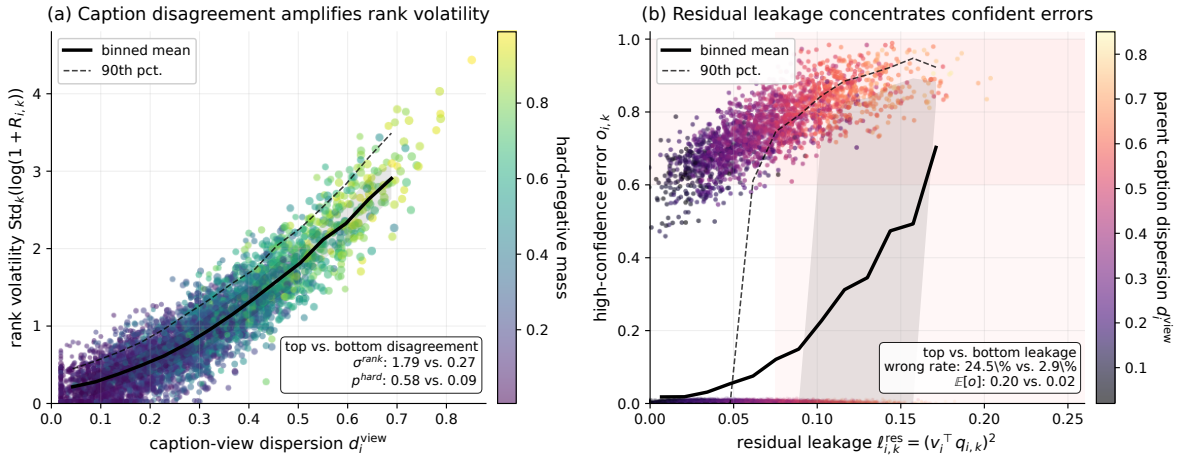

Figure 2: **Frozen VLM failure under partial captions.** Left: larger caption-view dispersion leads to higher rank volatility and stronger hard-negative confusion. Right: residual leakage concentrates confident retrieval errors. All statistics are computed with frozen embeddings; no proposed module or training loss is used.

(Truong et al., 2025). Yet these fixes are constrained by the shared embedding space they build upon. TPC improves this foundation by strengthening the alignment core and reducing residual over-commitment, serving as a drop-in vision encoder that complements decoding- and connector-level approaches.

3 Preliminary Study

Before introducing our method, we first diagnose a failure mode of frozen vision–language encoders under partial captions. Using a frozen CLIP/OpenCLIP-style dual encoder, we evaluate images with multiple captions and ask: *when different captions describe the same image, does caption disagreement expose unstable and overconfident matching?* No model is trained or modified in this study. We compute four diagnostic quantities: caption-view dispersion, rank volatility across captions, residual leakage into view-specific textual components, and high-confidence retrieval error. Full definitions and the evaluation protocol are provided in Appendix B.1.

Finding 1: caption disagreement predicts unstable matching. Fig. 2(a) shows that images with larger caption-view dispersion exhibit substantially higher rank volatility across captions. The same image can be retrieved reliably with one valid caption but poorly with another, even though both captions refer to the same visual instance. The hard-negative posterior mass also increases with disagreement, suggesting that underspecified captions make semantically nearby negatives increasingly competitive.

Finding 2: view-specific residuals drive confident errors. Fig. 2(b) shows that high-confidence retrieval errors concentrate in examples with large residual leakage. This indicates that frozen VLM embeddings can over-commit to caption-specific components that are not consistently supported across views. Such over-commitment is harmful not only because it changes the ranking, but also because the model often remains highly confident when it is wrong.

Implication. These observations suggest that robust alignment under natural captions requires three ingredients: (i) a stable semantic target shared across caption views, (ii) a deployable mechanism that can recover this target from a single caption at inference time, and (iii) an explicit way to prevent view-specific residuals from dominating similarity and confidence. We now instantiate these requirements in a core–residual alignment framework.

4 Method

Motivated by the observation that multi-view captions share a stable core but differ in "unsaid" details that induce brittleness, we formulate TEXT AS PARTIAL CONSTRAINT as a minimal core–residual alignment framework, as shown in Fig.3. Given K views, we align to a consensus core, learn a single-view core filter for one-text inference, and treat the orthogonal component as an unsaid residual that is discouraged from affecting similarity. An uncertainty-aware contrastive objective with a non-commitment regularizer jointly optimizes these parts to improve robustness under underspecified language supervision.

Setup. Different textual views of the same image agree on the main semantics but diverge on omitted or

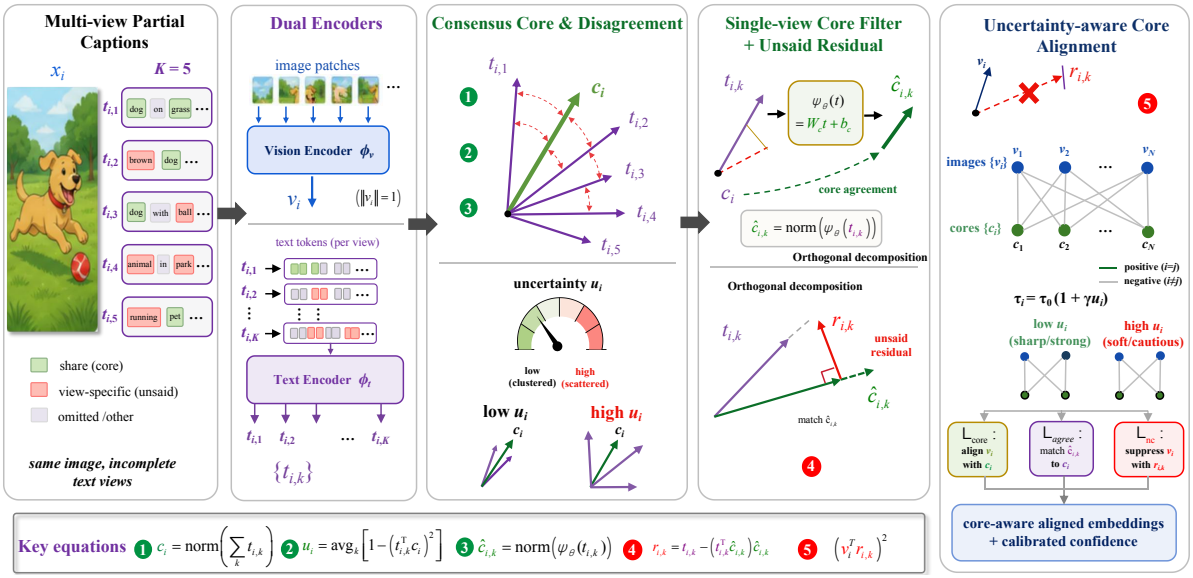

Figure 3: **Overview of TPC.** TPC treats captions as partial constraints: it distills a multi-view consensus core, learns a single-view core predictor, decomposes each caption into core and unsaid residual components, and performs uncertainty-aware alignment that suppresses residual over-commitment.

ambiguous details. A robust aligner should match vision to the shared semantics and avoid committing to view-specific details that are not consistently supported.

In each minibatch, we sample B instances $\{(x_i, \{y_{i,k}\}_{k=1}^{K})\}_{i=1}^{B}$. Let ϕ_v and ϕ_t be a vision encoder and a text encoder that map inputs to $\mathbb{R}^d$. We use $\mathrm{norm}(z) \triangleq z/(\|z\|_2 + \varepsilon)$ with $\varepsilon > 0$.

$$v_i \triangleq \mathrm{norm}(\phi_v(x_i)) \in \mathbb{R}^d, \qquad t_{i,k} \triangleq \mathrm{norm}(\phi_t(y_{i,k})) \in \mathbb{R}^d. \tag{1}$$

Here v_i is the normalized image embedding, $t_{i,k}$ is the normalized embedding of the k-th text view for image i, and d is the embedding dimension.

4.1 Multi-View Consensus Core and Disagreement

The intersection of semantics across views provides the most reliable supervision. Disagreement across views signals underspecification and omissions, which should reduce confidence rather than sharpen alignment.

We define the consensus core as the normalized mean direction of multi-view text embeddings:

$$c_i \triangleq \mathrm{norm}\Big(\sum_{k=1}^{K} t_{i,k}\Big) \in \mathbb{R}^d. \tag{2}$$

The vector c_i summarizes the shared semantics among $\{t_{i,k}\}$ and serves as the text-side target for alignment.

We quantify disagreement-based uncertainty by measuring the average squared cosine deviation from the consensus:

$$u_i \triangleq \frac{1}{K}\sum_{k=1}^{K}\Big(1 - (t_{i,k}^{\top} c_i)^2\Big) \in [0,1]. \tag{3}$$

Since $\|t_{i,k}\|_2 = \|c_i\|_2 = 1$, each $(t_{i,k}^{\top} c_i)^2 \in [0,1]$ and thus u_i is nonnegative and bounded. Large u_i indicates that views scatter around c_i and the language constraint is weak.

4.2 Learning a Single-View Core Filter

Benchmarks and applications provide a single query text at inference time. We therefore learn a deterministic map that extracts the core from a single text embedding, using the multi-view consensus c_i as supervision.

We introduce a lightweight core filter $\psi_\theta : \mathbb{R}^d \to \mathbb{R}^d$ implemented as a linear layer. For each view embedding $t_{i,k}$, we predict a single-view core direction

$$\hat{c}_{i,k} \triangleq \mathrm{norm}\big(\psi_\theta(t_{i,k})\big) = \mathrm{norm}\big(W_c t_{i,k} + b_c\big) \in \mathbb{R}^d, \tag{4}$$

where $W_c \in \mathbb{R}^{d\times d}$ and $b_c \in \mathbb{R}^d$ are trainable parameters. The agreement objective (defined below) enforces $\hat{c}_{i,k}$ to match the consensus c_i, making the mapping usable with a single text at test time.

4.3 Unsaid Residual and Non-Commitment

Even when the core is correct, the remaining components of language contain view-specific details and ambiguity. To prevent over-alignment, we explicitly discourage the image embedding from aligning with these unsaid components.

Given the predicted core direction $\hat{c}_{i,k}$, we define the *unsaid residual* as the component of $t_{i,k}$ orthogonal to $\hat{c}_{i,k}$:

$$r_{i,k} \triangleq t_{i,k} - (t_{i,k}^\top \hat{c}_{i,k})\, \hat{c}_{i,k} \in \mathbb{R}^d. \tag{5}$$

Because $\|\hat{c}_{i,k}\|_2 = 1$, Eq. (5) removes the projection of $t_{i,k}$ onto $\hat{c}_{i,k}$, yielding $\hat{c}_{i,k}^\top r_{i,k} = 0$.

We enforce *non-commitment* by penalizing correlation between image embeddings and residuals:

$$\mathcal{L}_{\mathrm{nc}} \triangleq \frac{1}{BK} \sum_{i=1}^{B} \sum_{k=1}^{K} \big(v_i^\top r_{i,k}\big)^2. \tag{6}$$

Here $v_i^\top r_{i,k}$ is a scalar measuring image–residual alignment; squaring yields a smooth penalty that drives v_i away from unsaid directions.

4.4 Uncertainty-Aware Core Alignment

When language is underspecified (large u_i), alignment should be less brittle and similarity should be less overconfident. We implement this by scaling the contrastive sharpness using the disagreement-based uncertainty.

We set an instance-dependent temperature $\tau_i \triangleq \tau_0(1 + \gamma u_i)$ with $\tau_0 > 0$ and $\gamma \geq 0$. We then align images to consensus cores using a symmetric InfoNCE loss:

$$\begin{aligned} \mathcal{L}_{\mathrm{core}} \triangleq & -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp\big(v_i^\top c_i/\tau_i\big)}{\sum_{j=1}^{B}\exp\big(v_i^\top c_j/\tau_i\big)} \\ & -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp\big(c_i^\top v_i/\tau_i\big)}{\sum_{j=1}^{B}\exp\big(c_i^\top v_j/\tau_i\big)}. \end{aligned} \tag{7}$$

In Eq. (7), negatives are other batch elements. A larger u_i increases τ_i and smooths the softmax distribution, reducing overconfident gradients when text constraints are weak.

4.5 Training Objective

We require the single-view core filter to reproduce the multi-view consensus and require vision to align to the consensus while ignoring residuals.

We train the core filter with a squared agreement loss

$$\mathcal{L}_{\mathrm{agree}} \triangleq \frac{1}{BK} \sum_{i=1}^{B} \sum_{k=1}^{K} \|\hat{c}_{i,k} - c_i\|_2^2. \tag{8}$$

Eq. (8) uses the consensus c_i from Eq. (2) as a target, making $\hat{c}_{i,k}$ a single-view estimator of the shared semantics.

The full objective is

$$\mathcal{L} \triangleq \mathcal{L}_{\mathrm{core}} + \lambda_{\mathrm{agree}}\mathcal{L}_{\mathrm{agree}} + \lambda_{\mathrm{nc}}\mathcal{L}_{\mathrm{nc}}, \tag{9}$$

with fixed weights $\lambda_{\mathrm{agree}} \geq 0$ and $\lambda_{\mathrm{nc}} \geq 0$. We optimize parameters of ϕ_v, ϕ_t, and ψ_θ by minimizing Eq. (9).

4.6 Inference and Calibrated Confidence

Standard retrieval uses a single text query and requires a single embedding. We output a core embedding for ranking and a calibrated confidence for downstream decision-making.

Given a query text y, we compute $t = \mathrm{norm}(\phi_t(y))$ and $\hat{c} = \mathrm{norm}(\psi_\theta(t))$. We rank candidates with the core similarity $s(x, y) = v^\top \hat{c}$. We compute a scalar confidence temperature using the residual energy $\hat{u}(y) = 1 - (t^\top \hat{c})^2$ and $\tau(y) = \tau_0(1 + \gamma \hat{u}(y))$, then convert similarities into calibrated probabilities for downstream selection via $p(j \mid y) = \mathrm{softmax}\big(s(x_j, y)/\tau(y)\big)$. This calibration changes the sharpness of the retrieval distribution while preserving the ranking induced by $s(x, y)$.

5 Theory

We provide two theoretical justifications for TPC: caption-view variation induces a worst-case residual penalty, and multi-view consensus denoises incomplete textual supervision.

5.1 Robustness to Caption-View Shift as Distributional Robust Optimization

Let c_i be the consensus core of instance i and let $\mathcal{U}_i = \mathrm{span}\{t_{i,k} - (t_{i,k}^\top c_i)c_i : k \in [K]\} \subseteq c_i^\perp$ denote the subspace of caption-specific deviations. With residual radius $\rho_i = \max_k \|t_{i,k} - (t_{i,k}^\top c_i)c_i\|_2$, we model admissible caption views by the spherical-cap ambiguity set

$$\mathcal{A}_i(\rho_i) \triangleq \Big\{ t \in \mathbb{R}^d : \|t\|_2 = 1,\ t = \sqrt{1 - \|r\|_2^2}\, c_i + r, \quad r \in \mathcal{U}_i,\ \|r\|_2 \le \rho_i \Big\}. \tag{10}$$

This set contains all observed caption views under the mild hemisphere condition and its radius is controlled by the disagreement statistic u_i; see Lemma A.1.

Theorem 5.1 (Tight invariance bound under caption-view shift)**.** *Fix $c \in \mathbb{R}^d$ with $\|c\|_2 = 1$, a subspace $\mathcal{U} \subseteq c^\perp$, and $\rho \in [0, 1)$. Let $\mathcal{A}(c, \mathcal{U}, \rho)$ be defined as in Eq. (10) by replacing $(c_i, \mathcal{U}_i, \rho_i)$ with $(c, \mathcal{U}, \rho)$. For any unit vector v, define*

$$a \triangleq v^\top c, \qquad b \triangleq \|\mathbf{P}_{\mathcal{U}} v\|_2. \tag{11}$$

Then

$$\inf_{t \in \mathcal{A}(c, \mathcal{U}, \rho)} v^\top t = a\sqrt{1 - \rho^2} - \rho b. \tag{12}$$

Consequently, for any monotone non-increasing loss ℓ,

$$\sup_{t \in \mathcal{A}(c, \mathcal{U}, \rho)} \ell(v^\top t) = \ell\left(a\sqrt{1 - \rho^2} - \rho b\right). \tag{13}$$

Theorem 5.1 shows that robustness improves by increasing core alignment $a = v^\top c$ and decreasing residual sensitivity $b = \|\mathbf{P}_{\mathcal{U}} v\|_2$. This directly motivates aligning image embeddings to consensus cores while suppressing their correlation with unsaid residual directions. Since Lemma A.1 gives $\rho_i^2 \le K u_i$, the same analysis also motivates using caption-view disagreement to modulate confidence. The full proof is in Appendix A.1.

5.2 Generalization via Multi-View Denoising

We further view multiple captions as noisy observations of a latent semantic core. For each instance i, assume

$$t_{i,k} = \mu_i + \varepsilon_{i,k}, \qquad \mathbb{E}[\varepsilon_{i,k} \mid \mu_i] = 0, \qquad \|\varepsilon_{i,k}\|_2 \le \sigma < 1, \tag{14}$$

where $\mu_i \in \mathbb{S}^{d-1}$ is the latent core. Let $c_i = \mathrm{norm}(\frac{1}{K}\sum_k t_{i,k})$ be the normalized consensus. For a predictor class $\mathcal{F}$ and an L-Lipschitz loss, define

$$R_\mu(f) = \mathbb{E}\big[\ell(f(x)^\top \mu)\big], \qquad R_c(f) = \mathbb{E}\big[\ell(f(x)^\top c)\big].$$

Theorem 5.2 (PAC bound with K-view denoising). *Let* $\widehat{f} \in \arg\min_{f\in\mathcal{F}} \widehat{R}_c(f)$, *where* $\widehat{R}_c(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i)^\top c_i)$. *Then, with probability at least* $1-\delta$,

$$R_\mu(\widehat{f}) \le \inf_{f\in\mathcal{F}} R_\mu(f) + 4L\,\mathfrak{R}_n(\mathcal{F}) + 2\sqrt{\frac{\log(4/\delta)}{2n}} + 2L\,\varepsilon_{K,n}(\delta), \tag{15}$$

where

$$\varepsilon_{K,n}(\delta) \triangleq \frac{2\eta_{K,n}(\delta)}{1-\eta_{K,n}(\delta)}, \qquad \eta_{K,n}(\delta) \triangleq \frac{\sigma}{\sqrt{K}} + \sigma\sqrt{\frac{2\log(2n/\delta)}{K}}. \tag{16}$$

In particular, when $\eta_{K,n}(\delta) \le \frac{1}{2}$, $\varepsilon_{K,n}(\delta) = O(\sigma\sqrt{\log(n/\delta)/K})$.

Corollary 5.3 (Sample complexity in K)**.** *Under the conditions of Theorem 5.2, to make the denoising contribution at most* ϵ, *it suffices that*

$$K = \Omega\left(\frac{\sigma^2\log(n/\delta)}{\epsilon^2}\right). \tag{17}$$

Theorem 5.2 and Corollary 5.3 show that the consensus target becomes statistically cleaner as the number of views increases. This supports using multi-view consensus for training while learning a single-view core estimator for deployment. The full proof is in Appendix A.2.

6 Experiments

6.1 Experimental Setup

Datasets. We evaluate TEXT AS PARTIAL CONSTRAINT on three complementary suites. (i) *Image–text retrieval*: MS-COCO and Flickr30K, which provide multiple captions per image and are standard for alignment evaluation. (ii) *Zero-shot recognition*: ImageNet-1K and a broad transfer suite covering natural, fine-grained, texture, remote sensing, medical, and sketch/rendition shifts. (iii) *Robustness to underspecified language*: datasets with multi-caption annotations (COCO/Flickr30K) where we can evaluate sensitivity to caption choice. We additionally report LVLM-side results by swapping in our vision encoder for fixed LVLM frameworks.

Evaluation metrics. For retrieval, we report Recall@{1,5,10} for both image-to-text and text-to-image. For zero-shot recognition, we report Top-1 accuracy and, when applicable, robust accuracy under standardized adversarial evaluation. To assess overconfidence under weak language constraints, we report calibration metrics (ECE and NLL) on both classification probabilities and retrieval softmax scores (Appendix §B.3).

Compared methods. We compare TEXT AS PARTIAL CONSTRAINT to: (i) *contrastive dual-encoder baselines* and strong recent training recipes (CLIP (Radford et al., 2021), OpenCLIP (Cherti et al., 2023), SigLIP (Zhai et al., 2023), EVA-CLIP (Sun et al., 2023)); (ii) *robust/anti-misalignment alignment* methods targeting zero-shot robustness or robustness transfer (TeCoA (Mao et al., 2024), RobustCLIP/FARE (Schlarmann et al., 2024b), Robust SuperAlignment (Adv-W2S) (Schlarmann et al., 2024a)); (iii) *post-hoc representation enhancement* at the concept level (VL-SAE (Shen et al., 2025)); and (iv) **LVLM-side alignment** under a fixed LVLM stack (LLaVA (Liu et al., 2023), OpenFlamingo (Awadalla et al., 2023)), including patch-aligned training (Jiang et al., 2025) and Directed-Tokens (Truong et al., 2025). We match backbone/data/steps when feasible; methods requiring adversarial generation or extra supervision are additionally reported in a separate, compute-matched block. Refer to §B.4 for Implementation details.

6.2 Main Results

Zero-shot generalization & robustness. In Table 1, TPC is best on both clean and robust accuracy, reaching 81.42/64.05 on ImageNet and 76.19/52.03 on Avg-14 (clean/robust), outperforming the strongest robust baseline Adv-W2S by +5.58/+5.75 on ImageNet and +7.44/+3.65 on Avg-14. This supports that isolating a shared core and suppressing residual alignment reduces brittle over-commitment.

LVLM transfer. In Table 2, using TPC as a drop-in vision encoder achieves the best POPE F1 across all settings (Avg 85.16), exceeding Adv-W2S by +1.36. It also improves VQA accuracy on all benchmarks (e.g., OKVQA 59.57), surpassing Patch-Aligned Training, indicating that core–residual alignment complements projector- or decoding-level fixes.

Method	ImageNet (Clean↑)	ImageNet (Robust↑)	Avg-14 (Clean↑)	Avg-14 (Robust↑)
Standard CLIP (Radford et al., 2021)	74.90	0.00	**73.08**	0.01
TeCoA (Mao et al., 2024)	**80.00**	**61.74**	61.56	43.26
PMG-AFT (Wang et al., 2024b)	77.84	60.02	64.46	**45.74**
FARE (Schlarmann et al., 2024b)	72.96	43.56	65.50	42.97
TGA-ZSR (Yu et al., 2024)	**80.26**	**61.46**	62.11	45.19
Adv-W2S (Schlarmann et al., 2024a)	75.84	58.30	**68.75**	**48.38**
TPC (Ours)	**81.42**	**64.05**	**76.19**	**52.03**

Table 1: **Zero-shot recognition and adversarial robustness.** Top-1 clean / robust accuracy (%) on ImageNet and the average over 14 datasets under AutoAttack (ϵ=2/255).

Table 2: **Transfer to LVLMs: hallucination and VQA.** Left: POPE hallucination F1 (%) under three sampling schemes and their average (higher is better). Right: VQA accuracy (%) under an LLaVA-1.5-7B stack.

	POPE Hallucination (F1↑)				**VQA Accuracy↑**			
Method	Random	Popular	Adversarial	Avg	GQA	SciQA	VizWiz	OKVQA
Decoding / concept-level enhancement (frozen LVLM)								
LLaVA1.5 (Regular) (Liu et al., 2023)	80.87	79.27	77.16	79.10	**62.0**	**66.8**	**50.0**	**53.4**
VCD (Leng et al., 2024b)	84.04	82.31	80.13	82.16	60.8	65.2	48.7	51.9
VL-SAE (Shen et al., 2025)	**85.50**	**84.37**	**82.29**	**84.05**	61.5	66.1	49.8	53.0
Projector / fine-grained alignment (LVLM-side training)								
Patch-Aligned Training (Jiang et al., 2025)	81.2	80.5	78.1	79.93	**63.0**	**68.7**	**52.3**	**58.3**
Robust vision encoder replacement								
TeCoA (Mao et al., 2024)	79.80	79.10	75.20	78.00	60.2	64.5	47.3	51.8
PMG-AFT (Wang et al., 2024b)	81.70	80.90	76.30	79.60	61.0	65.8	48.2	52.6
FARE (Schlarmann et al., 2024b)	82.20	81.50	78.60	80.80	61.8	66.5	49.5	54.1
TGA-ZSR (Yu et al., 2024)	80.40	79.80	76.00	78.70	60.5	64.9	47.8	52.0
Adv-W2S (Schlarmann et al., 2024a)	**85.60**	**84.90**	**81.00**	**83.80**	62.5	67.9	51.4	56.8
Ours (core–residual alignment)								
TPC (Ours)	**86.93**	**85.81**	**82.74**	**85.16**	**63.38**	**69.12**	**52.94**	**59.57**

6.3 Ablation and Analysis

Single-factor ablations. Table 3 shows that each component matters. Removing the multi-view consensus core most harms robustness and LVLM transfer (Avg-14 robust (↓2.30), POPE (↓1.74)), indicating the shared intersection is a better target than any single caption. Uncertainty scaling mainly affects robust accuracy (ImageNet robust (↓1.17)), while dropping core agreement yields the largest overall regression (ImageNet robust (↓3.13)), underscoring the need for a reliable single-view core at test time. Removing non-commitment also consistently degrades robustness and hallucination metrics (Avg-14 robust (↓2.37), POPE (↓2.13)), supporting residual suppression.

Hyperparameter sensitivity. Fig. 4 shows TPC is stable across a wide range of settings. Robustness increases with the number of views and saturates around $K \approx 4$–5; γ exhibits a mild sweet spot (too small under-calibrates, too large over-smooths). λ_{nc} primarily controls robustness/hallucination and remains stable for $\lambda_{\text{nc}} \in [0.05, 0.2]$, while λ_{agree} is flat around 1–2. Varying τ_0 within the standard CLIP range changes little, suggesting calibration is dominated by the relative modulation from u.

Core–Residual Mechanistic Evidence. We test whether the non-commitment regularizer reduces alignment to view-specific residual directions. On multi-caption data, we measure residual leakage $\ell = (v^\top r)^2$ before and after fine-tuning for standard CLIP and TPC, using the caption consensus to define the residual space. Fig. 5 shows that standard CLIP fine-tuning increases leakage from ≈ 0.049 to ≈ 0.071, whereas TPC reduces it to ≈ 0.040 and tightens the tail. This supports that λ_{nc} suppresses over-commitment to unsaid, view-specific components rather than merely improving aggregate accuracy.

Table 3: **Single-factor ablations of TPC.** We report zero-shot clean/robust accuracy (%) and LVLM transfer (POPE Avg F1, VQA Avg).

Method	ImageNet (Clean↑)	ImageNet (Robust↑)	Avg-14 (Clean↑)	Avg-14 (Robust↑)	POPE (Avg F1↑)	VQA (Avg↑)
TPC (Full)	**81.42**	**64.05**	**76.19**	**52.03**	**85.16**	**61.25**
w/o multi-view consensus core ($c_i \leftarrow t_{i,k}$)	80.61 (↓0.81)	62.47 (↓1.58)	75.18 (↓1.01)	49.73 (↓2.30)	83.42 (↓1.74)	60.21 (↓1.04)
w/o uncertainty scaling (γ=0)	81.07 (↓0.35)	62.88 (↓1.17)	75.73 (↓0.46)	50.21 (↓1.82)	84.11 (↓1.05)	60.82 (↓0.43)
w/o core agreement (λ_{agree}=0)	80.03 (↓1.39)	60.92 (↓3.13)	74.62 (↓1.57)	48.31 (↓3.72)	82.96 (↓2.20)	59.74 (↓1.51)
w/o non-commitment (λ_{nc}=0)	81.16 (↓0.26)	62.34 (↓1.71)	75.50 (↓0.69)	49.66 (↓2.37)	83.03 (↓2.13)	60.05 (↓1.20)

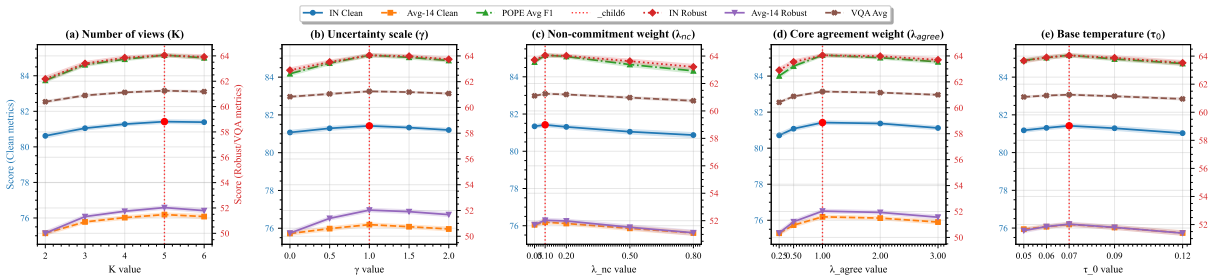

Figure 4: **Hyperparameter sensitivity of TPC.** Mean±std over 3 seeds.

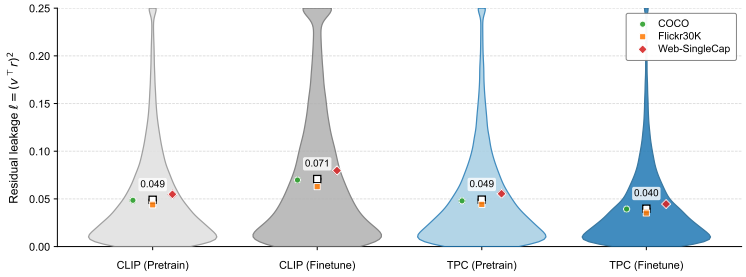

Figure 5: **Residual leakage.** Distribution of $\ell = (v^\top r)^2$ before/after fine-tuning; lower indicates less residual commitment. Protocol in App. B.6.

Uncertainty Calibration. We evaluate whether the inferred uncertainty $\hat{u}$ ranks queries by reliability under weak text constraints. We sort queries from low to high $\hat{u}$ and compute risk–coverage curves by retaining increasingly uncertain queries. Fig. 6 shows that TPC achieves lower risk than CLIP at the same coverage, with the largest gains in the selective low-coverage regime. This indicates that $\hat{u}$ is a meaningful abstention signal for underspecified text.

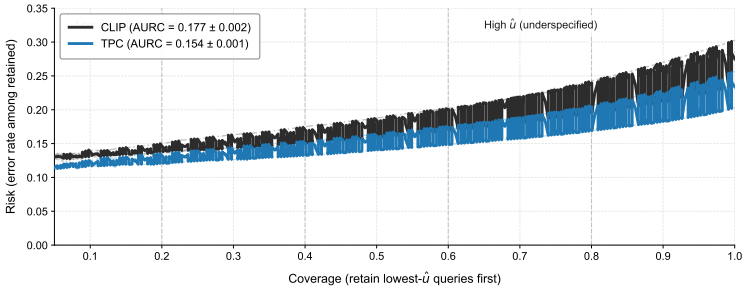

Figure 6: **Risk–coverage from $\hat{u}$.** Protocol in App. B.6.

7 Conclusion

We address brittleness and overconfidence in vision–language alignment under underspecified captions, and propose TEXT AS PARTIAL CONSTRAINT (TPC) to align images to view-invariant semantics while suppressing view-specific residuals with single-text deployability. Our results show that modeling language as an incomplete constraint and calibrating by view disagreement yields more reliable robustness and LVLM transfer. Future work will scale view acquisition (generation/retrieval) and extend

partial-constraint alignment to multimodal prompting and selective, safety-aware deployment.

References

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*, pages 29612–29622, 2025.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2818–2829, 2023.

Dasol Choi, Guijin Son, Hanwool Lee, Minhyuk Kim, Hyunwoo Ko, Teabin Lim, Ahn Eungyeol, Jungwhan Kim, Seunghyeok Hong, and Youngsook Song. What users leave unsaid: Under-specified queries limit vision-language models. *arXiv preprint arXiv:2601.06165*, 2026.

Sri Harsha Dumpala, Aman Jaiswal, Chandramouli Shama Sastry, Evangelos Milios, Sageev Oore, and Hassan Sajjad. Sugarcrepe++ dataset: Vision-language model sensitivity to semantic and lexical alterations. *Advances in Neural Information Processing Systems*, 37:17972–18018, 2024.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *Advances in neural information processing systems*, 36:31096–31116, 2023.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

Jiachen Jiang, Jinxin Zhou, Bo Peng, Xia Ning, and Zhihui Zhu. Analyzing fine-grained alignment and enhancing vision understanding in multimodal language models. *arXiv preprint arXiv:2505.17316*, 2025.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882, 2024a.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882, 2024b.

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*, pages 292–305, 2023.

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

Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, and Carl Vondrick. Understanding zero-shot adversarial robustness for large-scale models. In *The Eleventh International Conference on Learning Representations*, 2024.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8253–8280, 2022.

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*, pages 8748–8763. PmLR, 2021.

Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. Robust clip: unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 43685–43704, 2024a.

Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. Robust clip: unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 43685–43704, 2024b.

Shufan Shen, Junshu Sun, Qingming Huang, and Shuhui Wang. Vl-sae: Interpreting and enhancing vision-language alignment with a unified concept set. *arXiv preprint arXiv:2510.21323*, 2025.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

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*, pages 5238–5248, 2022.

Thanh-Dat Truong, Huu-Thien Tran, Tran Thai Son, Bhiksha Raj, and Khoa Luu. Directed-tokens: A robust multi-modality alignment approach to large language-vision models. *arXiv preprint arXiv:2508.14264*, 2025.

Sibo Wang, Jie Zhang, Zheng Yuan, and Shiguang Shan. Pre-trained model guided fine-tuning for zero-shot adversarial robustness. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24502–24511, 2024a.

Sibo Wang, Jie Zhang, Zheng Yuan, and Shiguang Shan. Pre-trained model guided fine-tuning for zero-shot adversarial robustness. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24502–24511, 2024b.

Canran Xiao, Tianxiang Xu, Siyuan Ma, Yiyang Jiang, Haoyu Gao, and Yuhan Wu. Reversible primitive–composition alignment for continual vision–language learning. In *The Fourteenth International Conference on Learning Representations*, 2026.

Lu Yu, Haiyang Zhang, and Changsheng Xu. Text-guided attention is all you need for zero-shot robustness in vision-language models. *Advances in Neural Information Processing Systems*, 37:96424–96448, 2024.

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*, pages 11975–11986, 2023.

Jiayu Zhang, Chuangxin Zhao, Canran Xiao, Ruibo Duan, Wenyi Mo, Haoyu Gao, and Wenshuo Wang. Pi-cca: Prompt-invariant cca certificates for replay-free continual multimodal learning. In *The Fourteenth International Conference on Learning Representations*, 2026.

Heng Zhou, Jing Tang, Jusheng Zhang, Yanshu Li, Canran Xiao, Liwei Hou, Zong Ke, and Jiawei Yao. Comem: Compositional concept-graph memory for vision–language adaptation. In *The Fourteenth International Conference on Learning Representations*, 2026.

A Proofs

A.1 Full Proofs for §5.1

Lemma A.1 (Containment of observed views and radius bounds). *Fix an instance i with normalized text views $\{t_{i,k}\}_{k=1}^K \subset \mathbb{R}^d$ and consensus core $c_i = \mathrm{norm}(\sum_{k=1}^K t_{i,k})$. Define residuals $r_{i,k} \triangleq t_{i,k} - (t_{i,k}^\top c_i)c_i$ and radii $\delta_{i,k} \triangleq \|r_{i,k}\|_2$. Let $\mathcal{U}_i \triangleq \mathrm{span}\{r_{i,k} : k \in [K]\} \subseteq c_i^\perp$ and $\rho_i \triangleq \max_{k\in[K]} \delta_{i,k}$. Assume $t_{i,k}^\top c_i \geq 0$ for all $k \in [K]$. Then:*

1. *(Containment) For every $k \in [K]$, $t_{i,k} \in \mathcal{A}_i(\rho_i)$ where $\mathcal{A}_i(\rho_i)$ is defined in Eq. (10).*
2. *(Disagreement controls radius) With $u_i = \frac{1}{K}\sum_{k=1}^K \left(1 - (t_{i,k}^\top c_i)^2\right)$, the maximal radius satisfies*

$$u_i \;\leq\; \rho_i^2 \;\leq\; K u_i. \tag{18}$$

Proof. **Step 1: orthogonal decomposition.** Since $\|t_{i,k}\|_2 = \|c_i\|_2 = 1$, define $r_{i,k} \triangleq t_{i,k} - (t_{i,k}^\top c_i)c_i$. Then $c_i^\top r_{i,k} = c_i^\top t_{i,k} - (t_{i,k}^\top c_i)\|c_i\|_2^2 = 0$, hence $r_{i,k} \in c_i^\perp$ and in particular $r_{i,k} \in \mathcal{U}_i$.

Step 2: identify the coefficient on c_i. Using Pythagoras on the orthogonal decomposition,

$$\|t_{i,k}\|_2^2 = \|(t_{i,k}^\top c_i)c_i\|_2^2 + \|r_{i,k}\|_2^2 = (t_{i,k}^\top c_i)^2 + \delta_{i,k}^2.$$

Since $\|t_{i,k}\|_2 = 1$, we obtain

$$\delta_{i,k}^2 = 1 - (t_{i,k}^\top c_i)^2. \tag{19}$$

Under the assumption $t_{i,k}^\top c_i \geq 0$, we have $t_{i,k}^\top c_i = \sqrt{1 - \delta_{i,k}^2}$.

Step 3: prove containment. Rewrite each view as

$$t_{i,k} = (t_{i,k}^\top c_i)c_i + r_{i,k} = \sqrt{1 - \delta_{i,k}^2}\, c_i + r_{i,k}.$$

Since $r_{i,k} \in \mathcal{U}_i$ and $\|r_{i,k}\|_2 = \delta_{i,k} \leq \rho_i$, this matches Eq. (10), hence $t_{i,k} \in \mathcal{A}_i(\rho_i)$.

Step 4: relate ρ_i and u_i. By Eq. (19),

$$u_i = \frac{1}{K}\sum_{k=1}^K \left(1 - (t_{i,k}^\top c_i)^2\right) = \frac{1}{K}\sum_{k=1}^K \delta_{i,k}^2.$$

The left inequality in Eq. (18) follows since $\max_k \delta_{i,k}^2 \geq \frac{1}{K}\sum_k \delta_{i,k}^2$:

$$\rho_i^2 = \max_k \delta_{i,k}^2 \;\geq\; \frac{1}{K}\sum_{k=1}^K \delta_{i,k}^2 = u_i.$$

For the right inequality, use $\max_k \delta_{i,k}^2 \leq \sum_{k=1}^K \delta_{i,k}^2$:

$$\rho_i^2 = \max_k \delta_{i,k}^2 \;\leq\; \sum_{k=1}^K \delta_{i,k}^2 = K u_i.$$

This proves Eq. (18). □

Theorem A.2 (Tight invariance bound under caption-view shift). *Fix $c \in \mathbb{R}^d$ with $\|c\|_2 = 1$, a subspace $\mathcal{U} \subseteq c^\perp$, and $\rho \in [0, 1)$. Let*

$$\mathcal{A}(c, \mathcal{U}, \rho) = \left\{ t \in \mathbb{R}^d :\; \|t\|_2 = 1,\; t = \sqrt{1 - \|r\|_2^2}\, c + r,\; r \in \mathcal{U},\; \|r\|_2 \leq \rho \right\}.$$

For any $v \in \mathbb{R}^d$ with $\|v\|_2 = 1$ and $a = v^\top c \geq 0$, define $b = \|\mathbf{P}_{\mathcal{U}} v\|_2$. Then

$$\inf_{t \in \mathcal{A}(c,\mathcal{U},\rho)} v^\top t = a\sqrt{1 - \rho^2} - \rho b.$$

Moreover, for any monotone non-increasing $\ell : \mathbb{R} \to \mathbb{R}$,

$$\sup_{t \in \mathcal{A}(c,\mathcal{U},\rho)} \ell(v^\top t) = \ell\Big(a\sqrt{1 - \rho^2} - \rho b\Big).$$

Proof. **Step 1: reduce to the residual variable.** Take any $t \in \mathcal{A}(c, \mathcal{U}, \rho)$. By definition, there exists $r \in \mathcal{U}$ with $\|r\|_2 \le \rho$ such that $t = \sqrt{1 - \|r\|_2^2}\, c + r$. Then

$$v^\top t = v^\top\big(\sqrt{1 - \|r\|_2^2}\, c + r\big) = (v^\top c)\sqrt{1 - \|r\|_2^2} + v^\top r. \tag{20}$$

Since $r \in \mathcal{U}$ and $\mathbf{P}_\mathcal{U}$ is the orthogonal projector, $v^\top r = (\mathbf{P}_\mathcal{U} v)^\top r$. Denote $u \triangleq \mathbf{P}_\mathcal{U} v$ so that $\|u\|_2 = b$. Eq. (20) becomes

$$v^\top t = a\sqrt{1 - \|r\|_2^2} + u^\top r, \qquad \text{with} \quad r \in \mathcal{U},\ \|r\|_2 \le \rho. \tag{21}$$

Step 2: minimize over the direction of r for fixed radius. Fix any radius $s \in [0, \rho]$ and consider all $r \in \mathcal{U}$ with $\|r\|_2 = s$. By Cauchy–Schwarz,

$$u^\top r \ge -\|u\|_2 \|r\|_2 = -bs, \tag{22}$$

with equality achieved by choosing $r = -s\, u / \|u\|_2$ when $b > 0$ (and any r when $b = 0$). Plugging Eq. (22) into Eq. (21) yields, for fixed s,

$$\inf_{\substack{r \in \mathcal{U} \\ \|r\|_2 = s}} v^\top t = a\sqrt{1 - s^2} - bs. \tag{23}$$

Step 3: minimize over the radius $s \in [0, \rho]$. Define the scalar function $f(s) \triangleq a\sqrt{1 - s^2} - bs$ on $[0, \rho]$. Because $a \ge 0$ and $b \ge 0$, its derivative satisfies, for all $s \in (0, \rho)$,

$$f'(s) = -\frac{as}{\sqrt{1 - s^2}} - b \;<\; 0. \tag{24}$$

Thus f is strictly decreasing on $[0, \rho]$, and the minimum is attained at $s = \rho$:

$$\min_{s \in [0, \rho]} f(s) = f(\rho) = a\sqrt{1 - \rho^2} - b\rho. \tag{25}$$

Combining Eq. (23) and Eq. (25) proves

$$\inf_{t \in \mathcal{A}(c, \mathcal{U}, \rho)} v^\top t = a\sqrt{1 - \rho^2} - \rho b.$$

Tightness follows by selecting $r^\star = -\rho\, u / \|u\|_2$ when $b > 0$ (and $r^\star = 0$ when $b = 0$), which satisfies $\|r^\star\|_2 = \rho$ and attains equality in Eq. (22).

Step 4: convert worst-case similarity to worst-case loss. Since ℓ is monotone non-increasing, the maximizer of $\ell(v^\top t)$ over $t \in \mathcal{A}(c, \mathcal{U}, \rho)$ is achieved at the minimizer of $v^\top t$:

$$\sup_{t \in \mathcal{A}(c, \mathcal{U}, \rho)} \ell(v^\top t) = \ell\Big(\inf_{t \in \mathcal{A}(c, \mathcal{U}, \rho)} v^\top t\Big) = \ell\Big(a\sqrt{1 - \rho^2} - \rho b\Big).$$

□

A.2 Full Proofs for §5.2

A.2.1 Auxiliary Lemmas

Lemma A.3 (Second-moment bound for the averaged view noise). *Let $\{\varepsilon_k\}_{k=1}^K$ be independent random vectors in $\mathbb{R}^d$ such that $\mathbb{E}[\varepsilon_k] = 0$ and $\|\varepsilon_k\|_2 \le \sigma$ almost surely. Then*

$$\mathbb{E}\Big\|\frac{1}{K}\sum_{k=1}^K \varepsilon_k\Big\|_2^2 \;\le\; \frac{\sigma^2}{K}, \qquad \mathbb{E}\Big\|\frac{1}{K}\sum_{k=1}^K \varepsilon_k\Big\|_2 \;\le\; \frac{\sigma}{\sqrt{K}}. \tag{26}$$

Proof. Let $\bar{\varepsilon} = \frac{1}{K}\sum_{k=1}^K \varepsilon_k$. By expanding the squared norm and using independence and zero mean,

$$\begin{aligned}
\mathbb{E}\|\bar{\varepsilon}\|_2^2 &= \mathbb{E}\Big\langle \frac{1}{K}\sum_{k=1}^K \varepsilon_k, \frac{1}{K}\sum_{\ell=1}^K \varepsilon_\ell \Big\rangle = \frac{1}{K^2}\sum_{k=1}^K \mathbb{E}\|\varepsilon_k\|_2^2 + \frac{1}{K^2}\sum_{k \ne \ell} \mathbb{E}\langle \varepsilon_k, \varepsilon_\ell \rangle \\
&= \frac{1}{K^2}\sum_{k=1}^K \mathbb{E}\|\varepsilon_k\|_2^2 + \frac{1}{K^2}\sum_{k \ne \ell} \Big\langle \mathbb{E}[\varepsilon_k], \mathbb{E}[\varepsilon_\ell] \Big\rangle \le \frac{1}{K^2}\sum_{k=1}^K \sigma^2 = \frac{\sigma^2}{K}.
\end{aligned}$$

The second inequality follows from Jensen: $\mathbb{E}\|\bar{\varepsilon}\|_2 \le \sqrt{\mathbb{E}\|\bar{\varepsilon}\|_2^2} \le \sigma/\sqrt{K}$. □

Lemma A.4 (McDiarmid concentration for the averaged noise norm). *Let $\{\varepsilon_k\}_{k=1}^K$ be independent random vectors with $\|\varepsilon_k\|_2 \le \sigma$ almost surely. Define $g(\varepsilon_1, \dots, \varepsilon_K) \triangleq \left\|\frac{1}{K}\sum_{k=1}^K \varepsilon_k\right\|_2$. Then for any $\delta \in (0,1)$,*

$$\mathbb{P}\left(g - \mathbb{E}[g] \ge \sigma\sqrt{\frac{2\log(1/\delta)}{K}}\right) \le \delta. \tag{27}$$

Proof. We verify bounded differences. Let $(\varepsilon_1, \dots, \varepsilon_K)$ and $(\varepsilon_1, \dots, \varepsilon_k', \dots, \varepsilon_K)$ differ only at coordinate k. Then by the reverse triangle inequality,

$$\begin{aligned}\left|g(\varepsilon_1, \dots, \varepsilon_K) - g(\varepsilon_1, \dots, \varepsilon_k', \dots, \varepsilon_K)\right| &\le \left\|\frac{1}{K}\sum_{j=1}^K \varepsilon_j - \frac{1}{K}\left(\varepsilon_k' + \sum_{j\ne k}\varepsilon_j\right)\right\|_2 \\ &= \frac{1}{K}\|\varepsilon_k - \varepsilon_k'\|_2 \le \frac{1}{K}\left(\|\varepsilon_k\|_2 + \|\varepsilon_k'\|_2\right) \le \frac{2\sigma}{K}.\end{aligned}$$

Thus the bounded difference constants are $c_k = 2\sigma/K$ for all k. McDiarmid's inequality yields, for any $t > 0$,

$$\mathbb{P}(g - \mathbb{E}g \ge t) \le \exp\left(-\frac{2t^2}{\sum_{k=1}^K c_k^2}\right) = \exp\left(-\frac{2t^2}{K\cdot(4\sigma^2/K^2)}\right) = \exp\left(-\frac{Kt^2}{2\sigma^2}\right).$$

Setting $t = \sigma\sqrt{2\log(1/\delta)/K}$ gives Eq. (27). □

Lemma A.5 (Normalization perturbation inequality). *Let $\mu \in \mathbb{R}^d$ satisfy $\|\mu\|_2 = 1$ and let $e \in \mathbb{R}^d$ satisfy $\|e\|_2 < 1$. Define $c = \mathrm{norm}(\mu + e) = (\mu + e)/\|\mu + e\|_2$. Then*

$$\|c - \mu\|_2 \le \frac{2\|e\|_2}{1 - \|e\|_2}. \tag{28}$$

Proof. Let $u = \mu + e$ and note that $\|u\|_2 \ge \|\mu\|_2 - \|e\|_2 = 1 - \|e\|_2$ by the triangle inequality. Then

$$\begin{aligned}\|c - \mu\|_2 &= \left\|\frac{u}{\|u\|_2} - \mu\right\|_2 = \frac{1}{\|u\|_2}\|u - \|u\|_2\mu\|_2 \\ &= \frac{1}{\|u\|_2}\|\mu + e - \|u\|_2\mu\|_2 = \frac{1}{\|u\|_2}\|e + (1 - \|u\|_2)\mu\|_2 \\ &\le \frac{1}{\|u\|_2}\left(\|e\|_2 + |1 - \|u\|_2| \cdot \|\mu\|_2\right) = \frac{1}{\|u\|_2}\left(\|e\|_2 + |1 - \|u\|_2|\right).\end{aligned}$$

Using the reverse triangle inequality, $|1 - \|u\|_2| = |\|\mu\|_2 - \|u\|_2| \le \|\mu - u\|_2 = \|e\|_2$. Therefore,

$$\|c - \mu\|_2 \le \frac{1}{\|u\|_2}(2\|e\|_2) \le \frac{2\|e\|_2}{1 - \|e\|_2},$$

which proves Eq. (28). □

Lemma A.6 (Lipschitz transfer from core error to loss error). *Let $\ell : [-1,1] \to \mathbb{R}$ be L-Lipschitz. For any $v \in \mathbb{S}^{d-1}$ and any $a, b \in \mathbb{S}^{d-1}$,*

$$\left|\ell(v^\top a) - \ell(v^\top b)\right| \le L\,\|a - b\|_2. \tag{29}$$

Proof. Since $\|v\|_2 = 1$, Cauchy–Schwarz gives $|v^\top a - v^\top b| = |v^\top(a - b)| \le \|a - b\|_2$. By Lipschitzness of ℓ, $|\ell(v^\top a) - \ell(v^\top b)| \le L|v^\top a - v^\top b| \le L\|a - b\|_2$. □

Lemma A.7 (Rademacher contraction for inner-product losses). *Let $\ell : [-1,1] \to \mathbb{R}$ be L-Lipschitz and let $\mathcal{F} \subseteq \{f : \mathcal{X} \to \mathbb{S}^{d-1}\}$. For any fixed sample $\{(x_i, c_i)\}_{i=1}^n$ with $\|c_i\|_2 = 1$,*

$$\mathfrak{R}_n(\ell \circ \mathcal{F}) \triangleq \mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i\,\ell\big(f(x_i)^\top c_i\big)\right] \le L\,\mathfrak{R}_n(\mathcal{F}), \tag{30}$$

where $\mathfrak{R}_n(\mathcal{F}) = \mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i)^\top c_i\right]$.

Proof. This is a direct application of the Ledoux–Talagrand contraction principle. For completeness, let $\phi_i(s) = \ell(s)$. Each ϕ_i is L-Lipschitz on $[-1,1]$ and $\phi_i(0)$ is constant, hence

$$\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i\,\phi_i(f(x_i)^\top c_i)\right] \le L\,\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i\, f(x_i)^\top c_i\right],$$

which is Eq. (30). □

A.2.2 Proof of Theorem 5.2

Proof. We split the proof into four steps corresponding to the proof sketch.

Step 1: uniform bound on the averaged noise. Fix $i \in [n]$ and write $\bar{\varepsilon}_i = \frac{1}{K}\sum_{k=1}^{K} \varepsilon_{i,k}$. By Lemma A.4, for any $\delta_i \in (0,1)$,

$$\mathbb{P}\left(\|\bar{\varepsilon}_i\|_2 \geq \mathbb{E}\|\bar{\varepsilon}_i\|_2 + \sigma\sqrt{\frac{2\log(1/\delta_i)}{K}} \right) \leq \delta_i. \tag{31}$$

By Lemma A.3, $\mathbb{E}\|\bar{\varepsilon}_i\|_2 \leq \sigma/\sqrt{K}$. Substituting into Eq. (31) gives

$$\mathbb{P}\left(\|\bar{\varepsilon}_i\|_2 \geq \frac{\sigma}{\sqrt{K}} + \sigma\sqrt{\frac{2\log(1/\delta_i)}{K}} \right) \leq \delta_i. \tag{32}$$

Set $\delta_i = \delta/(2n)$ and apply a union bound over $i \in [n]$:

$$\begin{aligned} \mathbb{P}\left(\max_{i\in[n]} \|\bar{\varepsilon}_i\|_2 \geq \frac{\sigma}{\sqrt{K}} + \sigma\sqrt{\frac{2\log(2n/\delta)}{K}} \right) &\leq \sum_{i=1}^{n} \mathbb{P}\left(\|\bar{\varepsilon}_i\|_2 \geq \frac{\sigma}{\sqrt{K}} + \sigma\sqrt{\frac{2\log(2n/\delta)}{K}} \right) \\ &\leq \sum_{i=1}^{n} \frac{\delta}{2n} = \frac{\delta}{2}. \end{aligned} \tag{33}$$

Define

$$\eta_{K,n}(\delta) \triangleq \frac{\sigma}{\sqrt{K}} + \sigma\sqrt{\frac{2\log(2n/\delta)}{K}}. \tag{34}$$

Then Eq. (33) states that, with probability at least $1 - \delta/2$,

$$\max_{i\in[n]} \|\bar{\varepsilon}_i\|_2 \leq \eta_{K,n}(\delta). \tag{35}$$

Step 2: uniform denoising bound on consensus cores. Recall $\bar{t}_i = \mu_i + \bar{\varepsilon}_i$ and $c_i = \mathrm{norm}(\bar{t}_i)$. On the event in Eq. (35), Lemma A.5 with $e = \bar{\varepsilon}_i$ implies, for every $i \in [n]$,

$$\|c_i - \mu_i\|_2 \leq \frac{2\|\bar{\varepsilon}_i\|_2}{1 - \|\bar{\varepsilon}_i\|_2} \leq \frac{2\eta_{K,n}(\delta)}{1 - \eta_{K,n}(\delta)} \triangleq \varepsilon_{K,n}(\delta). \tag{36}$$

Thus, with probability at least $1 - \delta/2$,

$$\max_{i\in[n]} \|c_i - \mu_i\|_2 \leq \varepsilon_{K,n}(\delta). \tag{37}$$

Step 3: transfer from observed risk to true risk. Fix any $f \in \mathcal{F}$. For each $i \in [n]$, Lemma A.6 with $v = f(x_i)$, $a = c_i$, and $b = \mu_i$ yields

$$\left|\ell(f(x_i)^\top c_i) - \ell(f(x_i)^\top \mu_i)\right| \leq L\,\|c_i - \mu_i\|_2. \tag{38}$$

On the event in Eq. (37), Eq. (38) implies

$$\ell(f(x_i)^\top \mu_i) \leq \ell(f(x_i)^\top c_i) + L\,\varepsilon_{K,n}(\delta), \qquad \forall i \in [n]. \tag{39}$$

Averaging over i gives the empirical relation

$$\widehat{R}_\mu(f) \triangleq \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i)^\top \mu_i) \leq \widehat{R}_c(f) + L\,\varepsilon_{K,n}(\delta). \tag{40}$$

Taking expectation over fresh draws (and using the same Lipschitz argument pointwise) yields the population relation

$$R_\mu(f) \leq R_c(f) + L\,\varepsilon_{K,n}(\delta). \tag{41}$$

Step 4: PAC generalization on the observed risk and ERM decomposition. Apply a standard Rademacher generalization bound to the class $\ell \circ \mathcal{F}$ on the sample $\{(x_i, c_i)\}_{i=1}^{n}$: with probability at least $1 - \delta/2$,

$$\sup_{f\in\mathcal{F}} \left(R_c(f) - \widehat{R}_c(f)\right) \leq 2\,\mathfrak{R}_n(\ell \circ \mathcal{F}) + \sqrt{\frac{\log(4/\delta)}{2n}}. \tag{42}$$

By Lemma A.7, $\mathfrak{R}_n(\ell \circ \mathcal{F}) \le L\,\mathfrak{R}_n(\mathcal{F})$, hence

$$\sup_{f\in\mathcal{F}} \big(R_c(f) - \widehat{R}_c(f)\big) \le 2L\,\mathfrak{R}_n(\mathcal{F}) + \sqrt{\frac{\log(4/\delta)}{2n}}. \tag{43}$$

On the intersection of events (37) and (43) (which holds with probability at least $1-\delta$ by a union bound), we bound the excess true risk of $\widehat{f}$.

Let $f^\star \in \arg\min_{f\in\mathcal{F}} R_\mu(f)$. Starting from Eq. (41) and using Eq. (43) twice:

$$\begin{aligned}
R_\mu(\widehat{f}) &\le R_c(\widehat{f}) + L\,\varepsilon_{K,n}(\delta) && (44)\\
&\le \widehat{R}_c(\widehat{f}) + \Big(2L\,\mathfrak{R}_n(\mathcal{F}) + \sqrt{\tfrac{\log(4/\delta)}{2n}}\Big) + L\,\varepsilon_{K,n}(\delta) && (45)\\
&\le \widehat{R}_c(f^\star) + \Big(2L\,\mathfrak{R}_n(\mathcal{F}) + \sqrt{\tfrac{\log(4/\delta)}{2n}}\Big) + L\,\varepsilon_{K,n}(\delta) && (46)\\
&\le R_c(f^\star) + 2\Big(2L\,\mathfrak{R}_n(\mathcal{F}) + \sqrt{\tfrac{\log(4/\delta)}{2n}}\Big) + L\,\varepsilon_{K,n}(\delta) && (47)\\
&\le R_\mu(f^\star) + 2\Big(2L\,\mathfrak{R}_n(\mathcal{F}) + \sqrt{\tfrac{\log(4/\delta)}{2n}}\Big) + 2L\,\varepsilon_{K,n}(\delta), && (48)
\end{aligned}$$

where (46) uses ERM optimality of $\widehat{f}$ on $\widehat{R}_c$, and (48) uses Eq. (41) for $f^\star$. Rearranging yields Eq. (15).

Finally, if $\eta_{K,n}(\delta) \le \frac{1}{2}$, then $\varepsilon_{K,n}(\delta) = \frac{2\eta}{1-\eta} \le 4\eta$, giving the stated $O(\sigma\sqrt{\log(n/\delta)/K})$ rate. □

A.2.3 Proof of Corollary 5.3

Proof. From Theorem 5.2, it suffices to enforce $2L\,\varepsilon_{K,n}(\delta) \le \epsilon$. When $\eta_{K,n}(\delta) \le \frac{1}{2}$, $\varepsilon_{K,n}(\delta) \le 4\eta_{K,n}(\delta)$, so it suffices that $8L\,\eta_{K,n}(\delta) \le \epsilon$. By the definition of $\eta_{K,n}(\delta)$ in Eq. (34), this holds when

$$\frac{\sigma}{\sqrt{K}} + \sigma\sqrt{\frac{2\log(2n/\delta)}{K}} \le \frac{\epsilon}{8L},$$

which is implied by $K = \Omega(\sigma^2 \log(n/\delta)/\epsilon^2)$ (absorbing constants and L into $\Omega(\cdot)$). □

B Additional Experimental Details

B.1 Preliminary Study Protocol

This subsection provides the full protocol for the preliminary study in Sec. 3. The study is purely diagnostic.

Data and encoder. We use multi-caption image–text datasets such as MS-COCO and Flickr30K, where each image x_i is paired with K captions $\{y_{i,k}\}_{k=1}^{K}$. Unless stated otherwise, we use all available human captions and set $K = 5$. A frozen dual encoder maps images and captions to normalized embeddings:

$$v_i = \mathrm{norm}(\phi_v(x_i)) \in \mathbb{R}^d, \qquad t_{i,k} = \mathrm{norm}(\phi_t(y_{i,k})) \in \mathbb{R}^d. \tag{49}$$

All image–text similarities are cosine similarities $v^\top t$.

Caption-view dispersion. We quantify how much the captions of the same image disagree in the frozen text embedding space:

$$d_i^{\mathrm{view}} = \frac{2}{K(K-1)} \sum_{1\le k<\ell\le K} \big(1 - t_{i,k}^\top t_{i,\ell}\big). \tag{50}$$

A larger d_i^{view} means that different valid captions share less overlap and leave more details unspecified.

Caption-induced rank volatility. For each caption $y_{i,k}$, we retrieve images from a fixed gallery $\mathcal{G}$ and record the rank of the ground-truth image:

$$R_{i,k} = 1 + \sum_{x_j\in\mathcal{G},\, j\ne i} \mathbf{1}\Big[v_j^\top t_{i,k} > v_i^\top t_{i,k}\Big]. \tag{51}$$

We then define the rank volatility across captions as

$$\sigma_i^{\mathrm{rank}} = \mathrm{Std}_{k\in[K]}\big(\log(1 + R_{i,k})\big). \tag{52}$$

The logarithm reduces the domination of extreme ranks while preserving caption-induced instability.

Hard-negative posterior mass. To measure whether caption disagreement makes negatives more competitive, we compute a retrieval posterior over the same gallery:

$$p(j \mid y_{i,k}) = \frac{\exp(v_j^\top t_{i,k}/\tau_0)}{\sum_{m:x_m \in \mathcal{G}} \exp(v_m^\top t_{i,k}/\tau_0)}, \tag{53}$$

where τ_0 is the frozen encoder's evaluation temperature. The hard-negative mass for image i is

$$p_i^{\text{hard}} = \frac{1}{K} \sum_{k=1}^{K} \max_{j \neq i} p(j \mid y_{i,k}). \tag{54}$$

A larger value indicates that at least one incorrect image receives high posterior probability under the caption query.

Diagnostic residual leakage. We next examine whether the image embedding aligns with view-specific textual components. We first define a diagnostic multi-caption centroid

$$\tilde{t}_i = \text{norm}\left(\sum_{k=1}^{K} t_{i,k} \right), \tag{55}$$

which is used only for analysis. For each caption, we remove the component parallel to this centroid:

$$q_{i,k} = t_{i,k} - (t_{i,k}^\top \tilde{t}_i)\tilde{t}_i. \tag{56}$$

The residual leakage score is

$$\ell_{i,k}^{\text{res}} = \left(v_i^\top q_{i,k}\right)^2. \tag{57}$$

Large $\ell_{i,k}^{\text{res}}$ means that the image embedding is correlated with the caption component that is not shared by the multi-caption centroid.

High-confidence retrieval error. For each caption query, we define the top prediction and its confidence as

$$\hat{j}_{i,k} = \arg\max_{j:x_j \in \mathcal{G}} p(j \mid y_{i,k}), \qquad \pi_{i,k} = \max_{j:x_j \in \mathcal{G}} p(j \mid y_{i,k}). \tag{58}$$

The high-confidence error score is

$$o_{i,k} = \pi_{i,k} \cdot \mathbf{1}\left[\hat{j}_{i,k} \neq i\right]. \tag{59}$$

This score is large only when the frozen model retrieves a wrong image with high confidence.

Visualization. Fig. 2(a) plots each image as one point, with x-axis d_i^{view} and y-axis σ_i^{rank}. Point color encodes p_i^{hard}, so the plot jointly shows caption disagreement, rank instability, and hard-negative confusion. Fig. 2(b) plots each caption view as one point, with x-axis $\ell_{i,k}^{\text{res}}$ and y-axis $o_{i,k}$. Point color encodes the parent image's caption-view dispersion d_i^{view}. For both panels, we overlay binned means and upper quantiles to show the global trend without hiding the dense point distribution.

Interpretation. The preliminary study is designed to reveal a failure mechanism rather than to evaluate TPC. It supports the following empirical pattern: caption disagreement is associated with unstable retrieval ranks, and residual leakage is associated with confident errors. These two observations motivate the method design in Sec. 4: align to a view-shared semantic component, learn a single-caption estimator for deployment, suppress view-specific residual alignment, and calibrate confidence when captions provide weak constraints.

B.2 Datasets and Protocols

Training data and multi-view construction. TEXT AS PARTIAL CONSTRAINT requires K textual views per image to instantiate a *partial* language constraint and expose an "unsaid" subspace. We therefore use a *two-tier* view construction strategy:

(A) Human multi-caption datasets. For datasets that naturally provide multiple captions, we use **all available human captions** and set K=5 (MS-COCO, Flickr30K). During training, we encode each image *once* and encode its K captions independently, so multi-view supervision increases text-side compute but does not multiply vision-side cost. To match the step budget of single-caption baselines, we subsample $K' \in \{3, 4, 5\}$ captions per step (uniformly without replacement) and cycle through captions across epochs; the consensus c_i and uncertainty u_i are always computed on the sampled set.

(B) Single-caption web-style corpora. When only one caption is available, we synthesize view diversity that is faithful but intentionally underspecified. Given the original caption y, we generate $K-1$ additional views consisting of: (i) a *core-only* view that retains only the main entities/actions (drops attributes such as colors, counts, and fine descriptors), (ii) a *paraphrase* view with similar semantics but different surface form, and (iii) a *back-translation* view (EN→DE→EN) to inject lexical variation. Unless stated otherwise, we set K=4 for this regime (original + three synthesized views). We filter synthesized views to avoid semantic drift by enforcing: (1) length in $[5, 30]$ tokens, (2) no new named entities compared to y, and (3) cosine similarity to the original caption above a threshold under a *frozen* text encoder (we use ≥ 0.25 as default). If a view fails, we resample it; if repeated failures occur, we fall back to another core-only deletion sample to preserve correctness. This construction produces controlled omissions, which is precisely the regime where our residual suppression and uncertainty-aware sharpness are expected to help.

Image–text retrieval benchmarks. We evaluate retrieval on MS-COCO and Flickr30K using standard splits. At test time, each query caption y is mapped to the single-view core $\hat{c} = \mathrm{norm}(\psi_\theta(\mathrm{norm}(\phi_t(y))))$ and each candidate image is mapped to $v = \mathrm{norm}(\phi_v(x))$; ranking is performed by the core similarity $s(x, y) = v^\top \hat{c}$. We report image-to-text and text-to-image Recall@{1,5,10}. To stress robustness to underspecification, we additionally report: (i) **Mean-over-captions** performance (average Recall@K over all captions of an image), (ii) **Worst-caption** performance (minimum over captions), and (iii) **Caption sensitivity** (rank standard deviation across captions), see Appendix §B.3. These explicitly measure whether the model "over-commits" to view-specific details.

Zero-shot classification and distribution shift. We follow a CLIP-style zero-shot protocol with prompt ensembling. For ImageNet-1K (val), we use a fixed set of templates (e.g., the standard CLIP template set) and average the normalized text embeddings across templates per class. For transfer, we use dataset-specific template sets when available and otherwise use a small shared template pool (e.g., 8 generic templates). Crucially, our method uses the *core* text embedding for classification, i.e., class scores are $v^\top \hat{c}_{\text{class}}$. We report Top-1 accuracy on clean images and, when enabled, robust accuracy under AutoAttack (§B.3). We also include fine-grained alignment diagnostics that probe attribute/relation sensitivity, which closely matches our "unsaid" motivation: Winoground (compositional correctness), ARO-style attribute/object/relation tests, and caption-contrast sets (e.g., hard negatives formed by swapping attributes/relations). These are reported as accuracy (higher is better).

LVLM downstream and hallucination benchmarks. To test whether improved alignment transfers to generation settings, we plug our vision encoder into a fixed LVLM stack. Concretely, we keep the LLM weights, tokenizer, and instruction-tuning data fixed, and only change the vision encoder. Because changing the vision encoder can alter feature statistics/dimensions, we retrain (or re-fit) only the vision–language projector for a small number of steps while freezing the LLM (and optionally freezing the vision encoder after a short alignment warmup). We evaluate: (i) VQAv2 and TextVQA (accuracy) and (ii) POPE (accuracy/precision/recall/F1) for object hallucination. For open-ended caption hallucination, we report CHAIR. We present LVLM results separately from dual-encoder retrieval/classification to avoid conflating architectural and training differences.

B.3 Evaluation Metrics

Retrieval and caption robustness. We report Recall@{1,5,10} for both directions. For caption robustness, given an image with captions $\{y_k\}_{k=1}^K$ we compute: (i) mean Recall@K over captions, (ii) worst-caption Recall@K, and (iii) caption-sensitivity as the standard deviation of the (1-indexed) rank of the ground-truth match across captions. We additionally report a **disagreement-to-error** analysis: we bucket queries by predicted residual energy $\hat{u}(y) = 1 - (t^\top \hat{c})^2$ and report Recall@K and calibration within each bucket, which directly tests whether our uncertainty signal is meaningful.

Calibration for retrieval and classification. For classification, we compute ECE with M=15 equal-width bins over top-1 confidence and report NLL. For retrieval, we convert similarities into a calibrated distribution using our per-query temperature $\tau(y)$. Because a full softmax over the entire gallery can be expensive, we compute calibration on a **standardized candidate set**: for each query we include the ground-truth match, the top-N retrieved candidates (default N=100), and a fixed number of uniformly sampled negatives (default 300), then apply softmax on this set. We report ECE/NLL on top-1 correctness, Brier score, and risk–coverage curves (selective retrieval) obtained by abstaining when max softmax probability is below a threshold.

Adversarial robustness . When comparing to adversarially trained or adversarially fine-tuned baselines, we evaluate robust Top-1 accuracy on ImageNet-1K under AutoAttack with an ℓ_∞ threat model and a fixed perturbation radius. We use a single, fixed radius across all methods (default ϵ=4/255) and the same attack settings (steps, restarts) without per-method tuning.

B.4 Implementation Details

Backbone, resolution, and tokenization. We use an OpenCLIP-style dual encoder with a ViT vision backbone and a Transformer text encoder. Unless stated otherwise, images are processed at 224×224 with standard CLIP augmentations (random resized crop and horizontal flip for training; center crop for evaluation). Texts are truncated/padded to a fixed maximum length (77 tokens in the CLIP tokenizer) and encoded with the backbone tokenizer. We report results for two model scales: ViT-B/16 for fast ablations and ViT-L/14 for main results.

Core filter initialization and stability. The single-view core filter $\psi_\theta(t) = W_c t + b_c$ is initialized to be near-identity ($W_c \leftarrow I$ and $b_c \leftarrow 0$), so $\hat{c}$ starts close to the original text direction and training focuses on *removing* inconsistent components rather than rotating the space arbitrarily. We clamp the instance temperature to avoid extremes, i.e., $\tau_i = \tau_0(1 + \gamma u_i)$ is clipped to $[\tau_0, 3\tau_0]$ by default, which stabilizes early training when disagreement estimates are noisy.

Optimization and schedule. We use AdamW with cosine learning-rate decay and linear warmup. A typical configuration is: global batch size 4096 (via gradient accumulation if needed), warmup 5% of total steps, and total 100k–200k steps depending on data scale. We use mixed precision (bf16) and gradient clipping (max norm 1.0). We set separate learning rates for (i) encoders and (ii) the core filter: $\mathrm{LR}_{\mathrm{enc}} = 5 \times 10^{-5}$ and $\mathrm{LR}_{\psi} = 5 \times 10^{-4}$ by default, with weight decay 0.1. Unless otherwise stated, we tune only $\mathrm{LR}_{\mathrm{enc}}$ and keep LR_{ψ} fixed.

Method hyperparameters. We use K=5 on COCO/Flickr30K and K=4 on synthesized-view corpora. Unless otherwise stated, we set τ_0=0.07, γ=1.0, λ_{agree}=1.0, and λ_{nc}=0.1. We tune $\gamma \in \{0, 0.5, 1, 2\}$, $\lambda_{\mathrm{agree}} \in \{0.5, 1, 2\}$, and $\lambda_{\mathrm{nc}} \in \{0.05, 0.1, 0.2, 0.5\}$ on a held-out validation split. For ablations, we also report: (i) removing uncertainty scaling (γ=0), (ii) removing residual suppression (λ_{nc}=0), and (iii) using consensus-only training without the single-view filter.

Compute and reproducibility. Training is run with distributed data parallel on A100-class GPUs. We match global batch size and total optimization steps across methods whenever feasible. For multi-view training, the image forward pass is shared across views to keep compute comparable. We run each main setting with three random seeds and report mean±std.

B.5 Compared Methods

CLIP-style contrastive alignment. We use the standard CLIP objective as the primary baseline (Radford et al., 2021), i.e., symmetric InfoNCE over image–text pairs with cosine similarity. We also include OpenCLIP (Cherti et al., 2023) as an open, large-scale reproduction with widely used checkpoints and evaluation protocols.

Stronger contemporary training recipes. We compare to SigLIP (Zhai et al., 2023), which replaces softmax-normalized contrastive loss with a pairwise sigmoid objective, and EVA-CLIP (Sun et al., 2023), which introduces improved scaling/training techniques for CLIP at scale. When these baselines are used as *off-the-shelf* checkpoints trained on different corpora, we report them as "external checkpoints" under the same evaluation suite; when data access permits, we additionally retrain with our data/budget and report the matched-data results.

Robust/anti-misalignment training and adaptation. We include TeCoA (Mao et al., 2024), which adversarially fine-tunes CLIP-style models (min–max optimization) to improve zero-shot adversarial robustness under distribution shifts. We include RobustCLIP/FARE (Schlarmann et al., 2024b), which performs *unsupervised* adversarial fine-tuning of the *vision embedding* while explicitly preserving the original CLIP features, enabling plug-and-play robustness gains in downstream LVLMs. We include Robust SuperAlignment (Adv-W2S) (Schlarmann et al., 2024a), which extends weak-to-strong generalization to *robustness transfer* by incorporating adversarial examples into the alignment objective. These methods incur additional adversarial-example generation and/or teacher guidance; we therefore (i) report them in a separate robustness block, (ii) keep the backbone and evaluation protocol identical, and (iii) match training steps/compute as closely as possible.

Post-hoc concept-level enhancement. We compare against VL-SAE (Shen et al., 2025), which learns a unified concept set via sparse autoencoding of vision–language representations and can be used to strengthen alignment at the concept level. We apply VL-SAE as a post-hoc module on top of the same frozen encoders (no encoder retraining), isolating the effect of the enhancement mechanism.

LVLM-side alignment baselines. To test transfer to generative multimodal models, we follow the plug-and-play LVLM protocol: we fix the LVLM framework (e.g., LLaVA (Liu et al., 2023) or OpenFlamingo (Awadalla et al., 2023)) and only swap the vision encoder (ours vs. baseline encoders). We further compare against LVLM alignment improvements that modify the connector/training, including

patch-aligned training (Jiang et al., 2025) and Directed-Tokens (Truong et al., 2025). LVLM-side methods are not directly comparable to dual-encoder retrieval training; we therefore report LVLM results separately and keep the LLM, connector architecture, and SFT data identical unless the baseline explicitly changes them.

B.6 Main Experiment Details

Residual Leakage Diagnostic. This diagnostic evaluates whether training changes the degree to which image embeddings align with caption-specific components that are not shared across views. For each image x_i with K captions $\{y_{i,k}\}_{k=1}^{K}$, we compute normalized embeddings

$$v_i = \mathrm{norm}(\phi_v(x_i)), \qquad t_{i,k} = \mathrm{norm}(\phi_t(y_{i,k})). \tag{60}$$

We define a caption-consensus core using all available caption views:

$$c_i = \mathrm{norm}\left(\sum_{k=1}^{K} t_{i,k}\right). \tag{61}$$

For each caption view, we remove the component parallel to the consensus core and obtain the diagnostic residual

$$r_{i,k} = t_{i,k} - (t_{i,k}^{\top} c_i) c_i. \tag{62}$$

The residual leakage score is then

$$\ell_{i,k} = (v_i^{\top} r_{i,k})^2. \tag{63}$$

A larger $\ell_{i,k}$ means that the image representation is more strongly correlated with a view-specific textual component rather than the shared semantic core. We compute $\ell_{i,k}$ before and after fine-tuning for standard CLIP and TPC. For a fair comparison, the residual space is always defined by the multi-caption consensus c_i rather than by the learned single-view core predictor ψ_θ. Fig. 5 visualizes the distribution over all image–caption pairs, with mean and 95% confidence intervals estimated across seeds and dataset-wise means shown as small markers.

Selective Risk–Coverage from $\hat{u}$. This diagnostic evaluates whether the uncertainty proxy $\hat{u}$ identifies queries whose text constraints are weak. Given a query caption y, we compute

$$t = \mathrm{norm}(\phi_t(y)), \qquad \hat{c} = \mathrm{norm}(\psi_\theta(t)), \qquad \hat{u}(y) = 1 - (t^{\top}\hat{c})^2. \tag{64}$$

Intuitively, $\hat{u}(y)$ is large when the original text embedding contains substantial energy outside the predicted core direction, indicating that the query may contain view-specific or underspecified information. Let $\mathcal{Q}$ denote the evaluation query set and let $e_m(q) \in \{0, 1\}$ be the top-1 error of method m on query q. For threshold θ, we retain the low-uncertainty subset

$$\mathcal{Q}_\theta = \{q \in \mathcal{Q} \mid \hat{u}(q) \le \theta\}. \tag{65}$$

We then compute coverage and method-specific risk as

$$\mathrm{coverage}(\theta) = \frac{|\mathcal{Q}_\theta|}{|\mathcal{Q}|}, \qquad \mathrm{risk}_m(\theta) = \frac{1}{|\mathcal{Q}_\theta|} \sum_{q \in \mathcal{Q}_\theta} e_m(q). \tag{66}$$

Sweeping θ from small to large yields a risk–coverage curve. Lower risk at the same coverage indicates that the retained low-$\hat{u}$ queries are indeed more reliable. In Fig. 6, all methods are evaluated on the same retained query subsets induced by TPC's $\hat{u}$, so differences reflect prediction reliability under the same abstention policy. Solid curves report the overall average, faint curves report per-dataset trends, and shaded bands indicate ± 1 standard deviation over three seeds.

B.7 Compute Resources

All experiments are run with distributed data parallel training on NVIDIA A100-class GPUs. Unless otherwise stated, training uses bf16 mixed precision, gradient accumulation to maintain the stated global batch size, and shared image forward passes across caption views. Each worker has 80GB GPU memory, 64 CPU cores, approximately 512GB host memory, and local NVMe storage for dataset caching. The main software stack uses PyTorch, CUDA, NCCL, OpenCLIP-style data loading, and deterministic evaluation scripts.

Table 4 reports the approximate compute required to reproduce the reported experiments. GPU-hours are computed as

$$\text{GPU-hours} = \#\text{GPUs} \times \text{wall-clock hours}.$$

Table 4: **Approximate compute resources.** Runtime is measured per run unless otherwise stated.

Experiment	Backbone / Model	GPUs	Peak Mem. / GPU	Steps / Eval Size	Wall Time	GPU-hours
Main TPC training	ViT-L/14 dual encoder	8×A100-80GB	67GB	160K steps	31h	248
Fast ablations	ViT-B/16 dual encoder	4×A100-80GB	43GB	100K steps	11h	44
Hyperparameter sensitivity	ViT-B/16 dual encoder	4×A100-80GB	44GB	80K–100K steps	9–12h	36–48
Backbone generalization	ViT / ConvNeXt / RN variants	4–8×A100-80GB	38–72GB	80K–160K steps	10–39h	40–312
Retrained CLIP-style baselines	matched backbone	4–8×A100-80GB	40–66GB	100K–160K steps	12–30h	48–240
AutoAttack robustness evaluation	ViT-L/14	4×A100-80GB	36GB	ImageNet val	8.5h/model	34/model
COCO/Flickr retrieval evaluation	ViT-L/14	1×A100-80GB	24GB	full test split	1.2h/model	1.2/model
LVLM projector fitting	LLaVA-1.5-7B stack	8×A100-80GB	58GB	12K steps	5.5h	44
POPE/VQA evaluation	LLaVA-1.5-7B stack	4×A100-80GB	48GB	benchmark eval	3–6h/model	12–24/model
Controlled-deletion span extraction	Llama-3.1-8B-Instruct	1×A100-80GB	31GB	cached captions	4.5h	4.5
Controlled-deletion retrieval eval	ViT-L/14	1×A100-80GB	24GB	deletion grid	2.8h/model	2.8/model

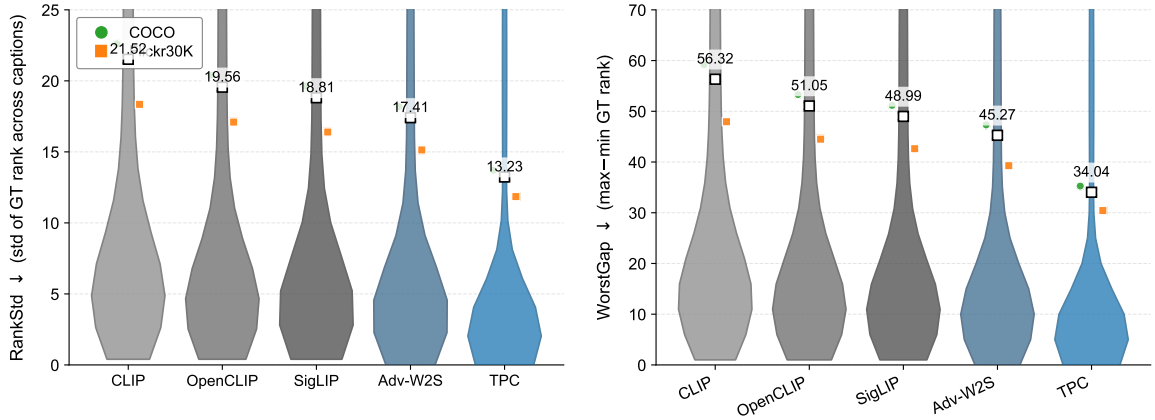

Figure 7: **Caption-choice sensitivity.** Violin plots show per-image distributions of RankStd and WorstGap on COCO and Flickr30K. Dots and error bars indicate mean and 95% CI; faint points show dataset-wise means. Lower is better: less variation in retrieval rank across captions of the same image.

The total compute for experiments reported in the paper is approximately 2.8K–3.2K A100 GPU-hours, including main training, ablations, sensitivity runs, robustness evaluation, LVLM projector fitting, and controlled-deletion evaluation. Including preliminary pilot runs and unsuccessful early variants, the full research process used approximately 3.6K A100 GPU-hours.

The most expensive components are main ViT-L/14 training, backbone generalization, and robustness evaluation. In contrast, the proposed core filter itself adds only $O(d^2)$ parameters and negligible inference overhead relative to the dual encoder. Multi-view training increases text-side computation, but the image encoder is evaluated once per image, so the dominant vision-side cost remains comparable to CLIP-style training under the same batch size and optimization steps.

C Additional Experimental Results

C.1 Consensus Core Stability: Does Caption-Choice Sensitivity Decrease?

We test the central "text as partial constraint" hypothesis: if the model aligns vision to a shared consensus core, retrieval should be less sensitive to which caption view is used as the query.

For each image x_i with captions $\{y_{i,k}\}_{k=1}^{K}$, we run text-to-image retrieval using each caption as the query and record the rank of the ground-truth image in the gallery, denoted $r_{i,k}$ (1 is best). We then compute two caption-sensitivity metrics:

$$\mathrm{RankStd}(i) = \mathrm{Std}\big(\{r_{i,k}\}_{k=1}^{K}\big), \qquad \mathrm{WorstGap}(i) = \max_k r_{i,k} - \min_k r_{i,k}. \tag{67}$$

Lower values indicate that the model is more stable to caption choice. We report distributions over images and compare CLIP-style training vs. TPC.

Fig. 7 shows that TPC substantially reduces caption-choice sensitivity: both RankStd and WorstGap distributions shift downward and exhibit shorter tails than CLIP, indicating fewer brittle failures driven by view-specific details. This supports our design choice of aligning to a consensus core while suppressing residual commitment.

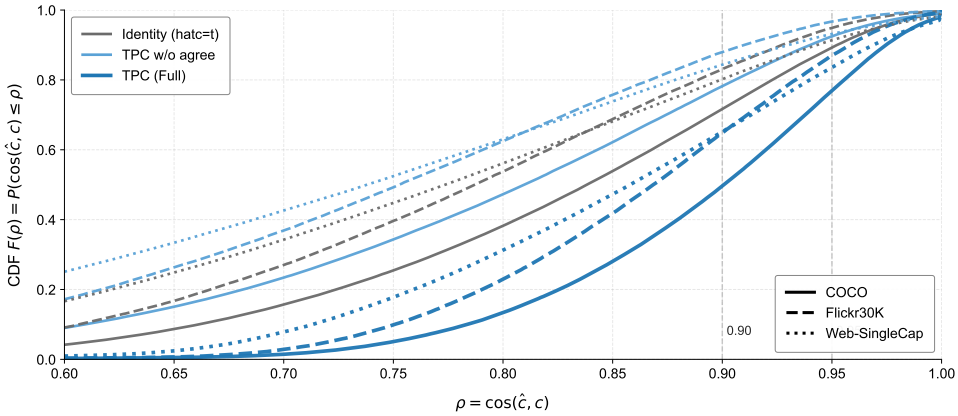

Figure 8: **Does ψ_θ predict the consensus core?** Empirical CDFs of $\rho = \cos(\hat{c}, c)$ on COCO, Flickr30K, and a single-caption web-style set. Colors denote methods; line styles denote datasets. A right-shifted curve indicates that single-caption cores $\hat{c}$ better match the multi-view consensus c.

C.2 Single-Caption Core Filter: Does ψ_θ Recover the Multi-View Consensus?

We verify that the single-view core filter ψ_θ is not a shortcut, but actually learns to approximate the multi-view intersection semantics captured by the consensus core c.

For each image x_i with captions $\{y_{i,k}\}_{k=1}^{K}$, we compute the consensus core $c_i = \text{norm}(\sum_k t_{i,k})$ with $t_{i,k} = \text{norm}(\phi_t(y_{i,k}))$. For each caption view, we predict a single-caption core $\hat{c}_{i,k} = \text{norm}(\psi_\theta(t_{i,k}))$ and measure

$$\rho_{i,k} \;=\; \cos(\hat{c}_{i,k}, c_i) \;=\; \hat{c}_{i,k}^{\top} c_i \in [-1, 1]. \tag{68}$$

We plot the empirical CDF $F(\alpha) = \mathbb{P}(\rho \le \alpha)$ across caption views. As baselines, we include an **Identity** "filter" ($\hat{c} = t$) and an **w/o-agree** variant that removes the agreement loss (λ_{agree}=0). A better core filter yields a CDF that is shifted right (more mass at high ρ), consistently across datasets.

Fig. 8 shows that TPC yields a clear right-shift in the $\cos(\hat{c}, c)$ distribution across datasets, while the Identity baseline exhibits a heavier low-similarity tail. Removing agreement supervision collapses much of the gain, indicating that ψ_θ learns a deployable single-caption estimator of shared semantics rather than exploiting dataset artifacts.

C.3 When Helpful: Gain vs. Text-Only Solvability

when is TPC *most useful?* If text is underspecified (weak constraint), a robust aligner should benefit more from suppressing residual commitment; if text is already sufficient (strong bias), gains should diminish.

For each VQA query $q = (x, y)$, we compute a *text-only solvability* score by running the same LVLM with the image masked (or replaced by a null image) and taking the maximum answer confidence: $s_{\text{text}}(q) = \max_a p(a \mid y, \varnothing) \in [0, 1]$. We then evaluate the full LVLM with (i) a baseline vision encoder and (ii) the TPC vision encoder, and compute the *accuracy gain* within bins of similar s_{text}. Concretely, we sort examples by s_{text} and form equal-count bins; each plotted point is a bin with $x = \mathbb{E}[s_{\text{text}}]$ and $y = \Delta\text{Acc} = \text{Acc}_{\text{TPC}} - \text{Acc}_{\text{Base}}$ (in percentage points).

Fig. 9 shows a clear pattern: gains are largest when s_{text} is low (image-dependent queries), and shrink toward zero as s_{text} increases (text already sufficient or strongly biased). This supports our hypothesis that TPC is particularly helpful under underspecified language supervision, where suppressing residual alignment prevents brittle over-commitment.

C.4 Backbone/Scale Generalization

We test whether TPC is robust to architectural choices by swapping the vision backbone while keeping the training recipe as consistent as possible. We train TPC with identical objectives and a matched schedule across ViT-S/16, ViT-B/16, ViT-L/14, ViT-H/14, ConvNeXt-B, ConvNeXt-L, and RN50. We evaluate (i) ImageNet Top-1 clean and AutoAttack robust accuracy (ϵ=2/255), and (ii) MS-COCO text→image Recall@{1,5,10}. We visualize results with a parallel-coordinates plot where each polyline corresponds to one backbone (faint lines: individual seeds; bold line: mean over 3 seeds).

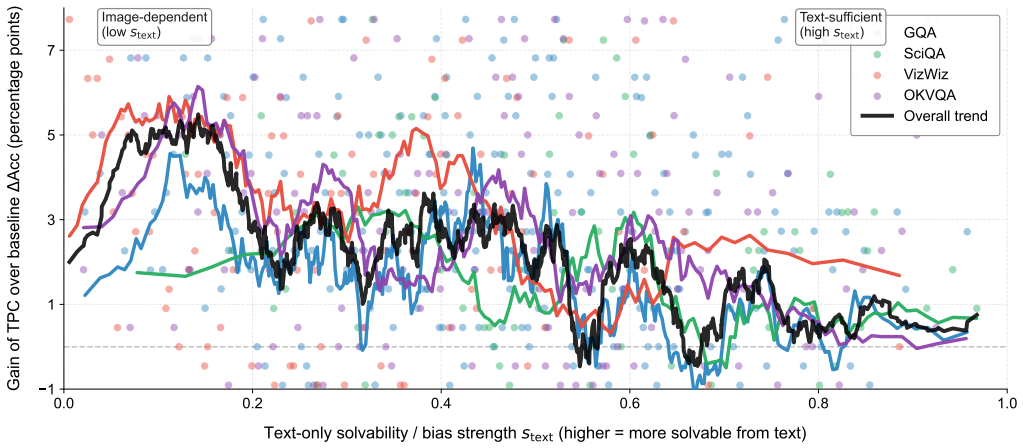

Figure 9: **Gain vs. text-only solvability.** Each point is an equal-count bin (hundreds of bins per dataset); colors indicate datasets. The trend line is a smoothed running mean over bins (same color per dataset) and an overall trend (black). Higher gain at lower solvability indicates TPC is most beneficial when text is a weak/underspecified constraint.

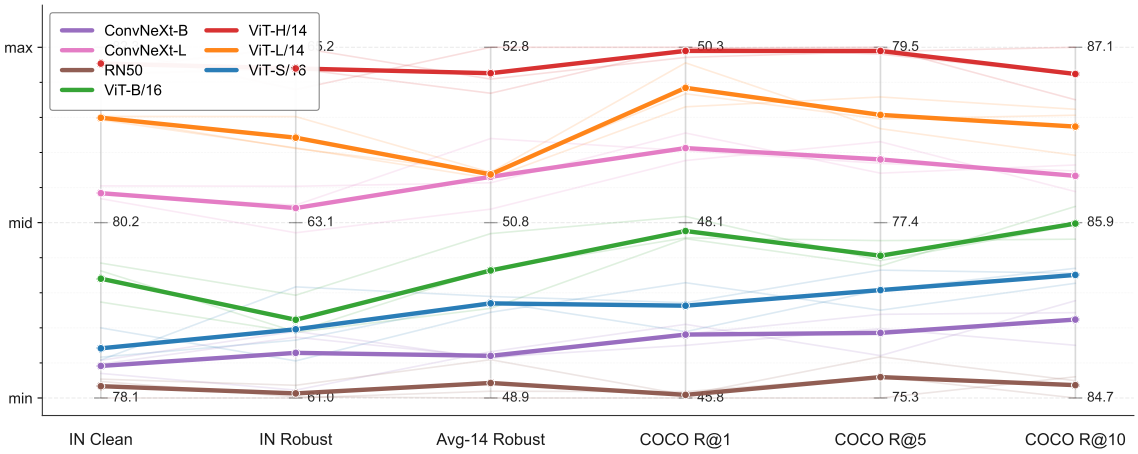

Figure 10: **Cross-backbone robustness and retrieval.** Parallel coordinates over multiple metrics (each axis is min–max normalized for readability). Each backbone is shown with three seed runs (faint) and the mean (bold). Consistent trends across architectures indicate TPC does not rely on a particular backbone family.

Fig. 10 shows that TPC yields stable performance across both ViT and ConvNeXt families: larger backbones improve clean accuracy and COCO recall as expected, while robust accuracy remains consistently high without collapse. This suggests the core–residual decomposition and uncertainty modulation transfer across architectural inductive biases.

C.5 Systematic Missing-Detail Stress Test via Controlled Caption Deletion

We perform a direct *text-as-partial-constraint* stress test by systematically removing critical details from captions. If a method over-commits to view-specific textual fragments, its performance and calibration should degrade sharply as deletions increase. In contrast, TPC should degrade more gracefully due to residual suppression and uncertainty-aware smoothing.

Given a caption y, we apply a controlled deletion operator that removes a fixed fraction of tokens belonging to one of three semantic categories: **Attributes** (color/size/material/adjectives), **Relations** (spatial/interaction predicates), and **Counts** (numerals/quantifiers). We test deletion ratios in $\{10\%, 30\%, 50\%, 70\%\}$ and evaluate text→image retrieval on COCO/Flickr30K. For each condition, we report **Recall@1** (performance; higher is better) and **ECE** (calibration; lower is better) computed from retrieval confidence (top-1 softmax probability over a standardized candidate set).

Data construction: LLM-assisted controlled deletion. The controlled deletion operator is built with a

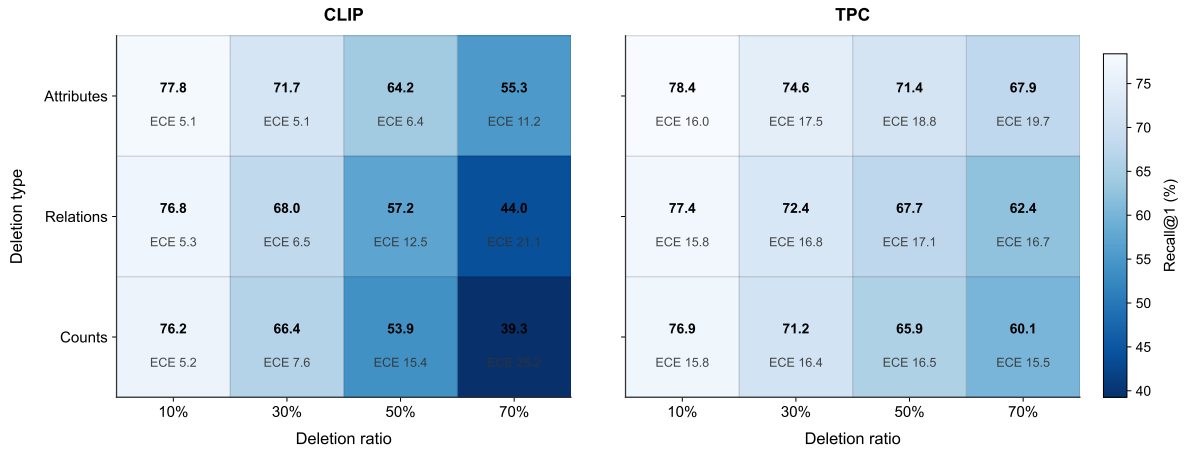

Figure 11: **Controlled caption deletion stress test.** Heatmaps show Recall@1 (%) under increasing deletion ratios (x-axis) and deletion types (y-axis). Each cell is annotated with *Recall@1 / ECE* (ECE shown as a small number). TPC maintains higher accuracy and lower ECE as deletions grow, indicating robustness to systematically weakened language constraints.

text-only span extractor driven by an off-the-shelf instruction-tuned LLM. Concretely, we use **Llama-3.1-8B-Instruct** to tag verbatim spans in the caption that correspond to *Attributes*, *Relations*, and *Counts*, and we then perform *programmatic* deletions to meet the target ratio. We run the extractor with deterministic decoding (temperature 0, top-p=1) and cache all extracted spans for reproducibility. To enforce *strict controllability*, the LLM is not allowed to paraphrase or introduce new content: it only returns spans that appear exactly in the original caption. Given the returned spans for a target type, we uniformly sample spans (with a fixed random seed) until the removed token count reaches the specified deletion ratio (10/30/50/70%) of all tokens covered by that type; if the caption contains too few tokens of a type, we delete all available ones. Finally, we apply a minimal deterministic cleanup (collapse whitespace, remove dangling commas/prepositions) without any rewriting.

Prompt template for LLM span extraction (used to build deletion sets)

```
System: You are a precise annotation tool. You must not rewrite text. You only copy exact substrings
that already appear in the caption.
User: Given the following image caption, extract verbatim spans that belong to three semantic
categories:
(1) ATTRIBUTES: colors, sizes, materials, adjectival modifiers, descriptive properties (e.g., "red",
"wooden", "small", "striped"). (2) RELATIONS: spatial or interaction relations/predicates connecting
entities (e.g., "next to", "on top of", "holding", "behind"). (3) COUNTS: numerals and quantifiers
(e.g., "two", "3", "several", "a pair of").
Rules: - Output MUST be a single valid JSON object with keys: "attributes", "relations", "counts". -
Each value is a list of strings, and each string MUST be an exact substring copied from the caption
(verbatim, case-preserving). - Spans should be minimal and non-overlapping. Do NOT invent spans
that are not present. - If a category has no span, output an empty list for that key. - Do NOT
output anything other than the JSON.
Caption: <<<CAPTION_TEXT>>>
```

Fig. 11 shows that TPC is consistently more resilient to missing details: (1) performance drops are milder under severe deletions, especially for *relations* and *counts* where underspecification is most damaging; (2) calibration degrades more slowly, suggesting $\hat{u}$-modulated sharpness prevents overconfident retrieval when the text constraint is weakened. Overall, the stress test supports our thesis that language supervision is partial and that explicitly suppressing residual commitment improves robustness.

D Responsible Research and Reproducibility

D.1 Scope and Limitations

TPC is designed as a representation-level alignment method for settings where language supervision is naturally partial or can be instantiated through multiple faithful caption views. Its main scope is therefore robust image–text matching, calibrated retrieval, zero-shot transfer, and plug-in transfer to LVLMs through a stronger vision encoder. The method is not intended to replace task-specific safety

Table 5: **Seed variability for main TPC results.** Mean±standard deviation over three random seeds.

Setting	ImageNet Clean	ImageNet Robust	Avg-14 Clean	Avg-14 Robust	POPE Avg F1	VQA Avg
TPC (ViT-L/14)	81.42±0.12	64.05±0.27	76.19±0.18	52.03±0.31	85.16±0.22	61.25±0.19

filters, factuality checkers, or policy-level controls in deployed LVLM systems; rather, it provides a more reliable alignment backbone that can complement these downstream safeguards.

The method uses multi-view captions during training to estimate a consensus semantic core. When human multi-caption annotations are available, we use the provided views directly; when only one caption is available, we construct faithful auxiliary views through controlled deletion, paraphrasing, and back-translation, followed by deterministic filtering. This setting matches many existing vision–language resources and can be extended with retrieval- or generation-based view construction. In our experiments, performance is stable across $K \in \{3, 4, 5\}$ and saturates around four to five views, suggesting that TPC does not require a large number of captions per image.

Computationally, TPC adds text-side view encoding and a lightweight linear core filter, while sharing the image forward pass across views. As a result, the additional overhead is moderate compared with full adversarial training or LVLM instruction tuning. In downstream use, inference remains single-text and uses only the predicted core embedding, so the deployment cost is essentially the same order as a standard CLIP-style dual encoder.

D.2 Statistical Reporting Across Seeds

Unless otherwise stated, trainable TPC variants are run with three random seeds. The seeds affect model initialization of the core filter, data shuffling, caption-view subsampling, and minibatch ordering. We report mean values in the main result tables for readability and use mean±standard deviation in sensitivity and diagnostic figures. For metrics whose distributions are asymmetric, such as residual leakage and rank-based quantities, we additionally use bootstrap confidence intervals over examples or caption views.

Table 5 reports seed-level variability for the principal TPC metrics. The standard deviations are small relative to the margins over the strongest baselines, indicating that the main conclusions are not driven by a single favorable run.

For external checkpoints that are evaluated without retraining, we report deterministic evaluation numbers under the same preprocessing and evaluation scripts. For retrained baselines and ablations, we use the same three-seed protocol whenever the corresponding method is computationally comparable.

D.3 Broader Impact and Responsible Use

TPC aims to improve the reliability of vision–language representations under underspecified language supervision. Potential positive impacts include more stable image–text retrieval, better calibrated confidence under ambiguous captions, and reduced object hallucination when the resulting vision encoder is used inside LVLM systems. These properties are useful for applications where users provide incomplete or paraphrased descriptions and where overconfident visual grounding can lead to unreliable downstream responses.

The same improvements may also increase the capability of downstream multimodal systems. If such systems are deployed without proper safeguards, stronger visual grounding could still be misused in settings such as surveillance, automated profiling, or misleading content generation. Moreover, calibrated confidence should not be interpreted as a guarantee of factual correctness or safety. We therefore view TPC as a representation-level reliability component rather than a complete deployment solution.

Responsible use should combine TPC with application-level safety filters, dataset and demographic audits, human oversight in high-stakes settings, and clear communication of model uncertainty. We do not release a new user-facing model or scraped dataset at submission time.

D.4 LLM Usage Disclosure

The core TPC method does not use an LLM to generate training labels, optimize model parameters, or produce predictions. The only methodological use of an LLM is in the controlled caption-deletion stress test in Appendix C.5. Specifically, we use Llama-3.1-8B-Instruct as a deterministic span extractor to identify verbatim caption spans corresponding to attributes, relations, and counts. The LLM is constrained

to return exact substrings from the original caption and is not allowed to paraphrase, introduce new content, or modify the image–text label.