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

Chengzhen Yu **Siyuan Ma** **Yang Liu**
Nanyang Technological University

Canran Xiao*
Shenzhen Campus of Sun Yat-sen University

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 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. Embodied reference understanding illustrates the role of context: Touch-Line Transformer combines language with gestural cues to locate referents [Li et al., 2023a]. 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

*Corresponding author: xiaocr3@mail.sysu.edu.cn.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

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., 2024, Dong et al., 2025], while LVLM mitigation is often post hoc and does not curb representation-level over-commitment under weak text constraints [Li et al., 2023b, Leng et al., 2024a]. These findings motivate treating language supervision as *incomplete* and curbing overconfident alignment to weakly supported, view-specific details.

We study how to learn vision–language representations that align to what captions consistently specify and limit overconfident commitment to underspecified details. As illustrated in Fig. 1, we model 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.

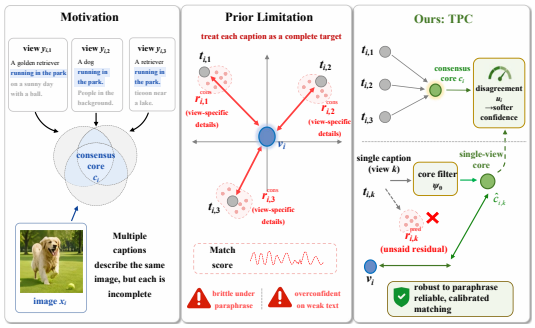

Figure 1: TPC extracts the semantic core, suppresses residual commitment, and produces calibrated, robust vision–language alignment.

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 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. Beyond language supervision, VIBUS combines viewpoint-bottleneck pretraining with uncertainty-spectrum pseudo-label selection for sparse 3D scene parsing [Tian et al., 2022].

Robust alignment under distribution shift and attacks. Robust VLM training has focused on adversarial and distributional *visual* shift; model-guided pose tracking separately addresses visual occlusion [Zhong et al., 2020]. 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., 2024] and weak-to-strong robustness transfer (Adv-W2S) [Dong et al., 2025]. These methods typically assume that text fully specifies the semantics, leaving caption-view ambiguity unmodeled. We 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.

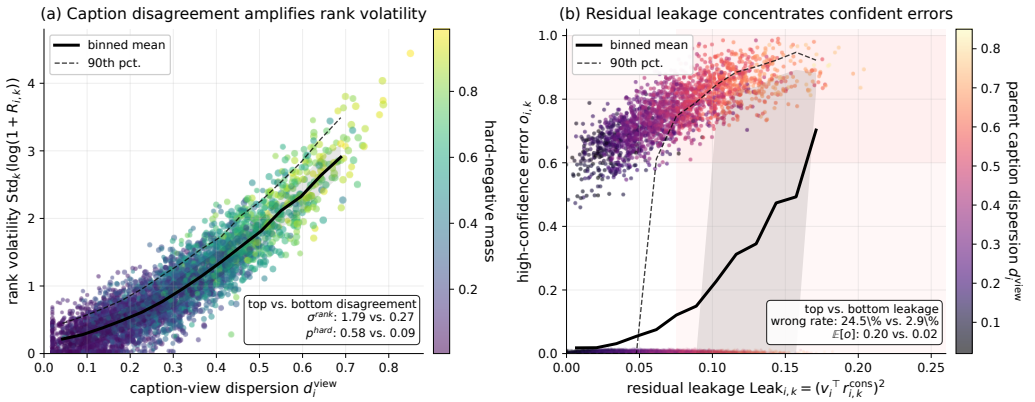

Figure 2: **Frozen VLM failure under partial captions.** Left: larger caption-view dispersion leads to higher rank volatility and stronger hard-negative confusion. Right: residual leakage concentrates confident retrieval errors. All statistics use frozen baseline embeddings.

Related work on continual VL learning preserves compositional structure through reversible primitive–composition alignment and multi-positive training (Compo-ReAlign) [Xiao et al., 2026a], or stabilizes cross-modal geometry under prompt shifts using CCA certificates (Pi-CCA) [Zhang et al., 2026]. CoMem uses a compact concept graph and graph-conditioned feature rehearsal to preserve compositional knowledge during VL adaptation [Zhou et al., 2026]. These methods target retention during sequential adaptation. TPC extracts shared semantics across partial captions of each image and suppresses view-specific residuals. In cyber-threat attribution, Trident combines structural and behavioral evidence through confidence-weighted fusion and replay-based adaptation [Wang et al., 2026].

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., 2023b]. A complementary continual-learning study reports hidden evidence-use forgetting, where answer accuracy is retained while reliance on grounding evidence drifts [Chen et al., 2026]. 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 [Truong et al., 2025]. Their effectiveness is constrained by the underlying shared embedding space. 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

We diagnose a failure mode of frozen vision–language encoders under partial captions. Using a frozen CLIP/OpenCLIP-style dual encoder, we examine whether disagreement among captions of the same image exposes unstable and overconfident matching. The encoders remain frozen throughout the 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 changes retrieval rankings and often leaves the model highly confident in incorrect predictions.

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) suppression of view-specific residuals that dominate similarity and confidence. We now instantiate these requirements in a core–residual alignment framework.

4 Method

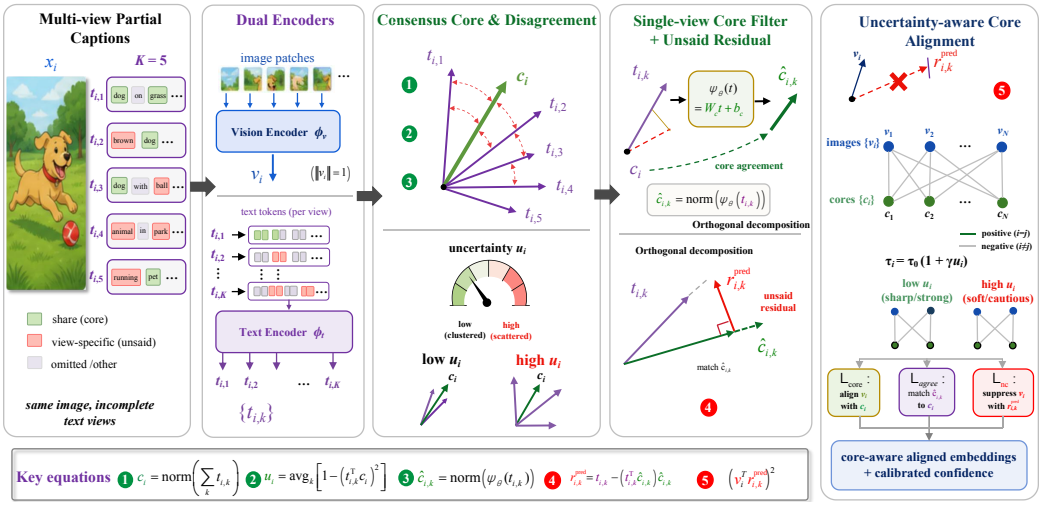

Figure 3: **Overview of TPC.** TPC treats captions as partial constraints: it distills a multi-view consensus core, learns a single-view core predictor, decomposes each caption into core and unsaid residual components, and performs uncertainty-aware alignment that suppresses residual over-commitment.

TEXT AS PARTIAL CONSTRAINT models the shared semantics and view-specific "unsaid" details of multi-view captions through core–residual alignment (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 ambiguous details. The alignment objective is to match vision to shared semantics and limit commitment to inconsistently supported, view-specific details.

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 $\mathrm{norm}(z) \triangleq z/\|z\|_2$ for $z \neq 0$.

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

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, motivating lower-confidence 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 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 $\hat{c}_{i,k}$:

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

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

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

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

Here $v_i^{\top} r_{i,k}^{\mathrm{pred}}$ 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

We soften alignment under underspecified language (large u_i) by scaling contrastive sharpness with 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

Training aligns the single-view core filter and image embeddings to the multi-view consensus while suppressing residual alignment.

We train the core filter with a squared agreement loss

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

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

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

with fixed weights $\lambda_{\text{agree}} \geq 0$ and $\lambda_{\text{nc}} \geq 0$. 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 $t = \text{norm}(\phi_t(y))$ and $\hat{c} = \text{norm}(\psi_\theta(t))$. 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 $p(j \mid y) = \text{softmax}\big(s(x_j, y)/\tau(y)\big)$. 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. Define $r_{i,k}^{\text{cons}} = t_{i,k} - (t_{i,k}^\top c_i)c_i$ and the consensus-orthogonal subspace $\mathcal{U}_i = \text{span}\{r_{i,k}^{\text{cons}} : k \in [K]\} \subseteq c_i^\perp$. With residual radius $\rho_i = \max_k \|r_{i,k}^{\text{cons}}\|_2$, we model admissible caption views by the spherical-cap ambiguity set

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

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 nonzero 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 with $v^\top c \geq 0$, 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,

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

If $\mathcal{U} = \{0\}$, then $\mathcal{A}(c, \mathcal{U}, \rho) = \{c\}$ for any $\rho \in [0, 1)$, so the worst-case similarity is a and the worst-case loss is $\ell(a)$.

Under its nonnegative-alignment condition, 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 $1-\delta$,*

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

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 a target tolerance $\epsilon_{\mathrm{tol}} > 0$, it suffices that*

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

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 [Lin et al., 2014] and Flickr30K [Young et al., 2014], which provide multiple captions per image and are standard for alignment evaluation. (ii) *Zero-shot recognition*: ImageNet-1K [Deng et al., 2009, Russakovsky et al., 2015] and a broad transfer suite covering natural, fine-grained, texture, remote sensing, medical, and sketch/rendition shifts. We evaluate zero-shot recognition on 14 datasets, including ImageNet-1K and 13 transfer benchmarks. The Avg-14 metric denotes the unweighted macro-average across all 14 datasets, including ImageNet, following the evaluation protocol of Adv-W2S. (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, including evaluation on GQA [Hudson and Manning, 2019], ScienceQA [Lu et al., 2022], VizWiz [Gurari et al., 2018], and OK-VQA [Marino et al., 2019].

Method	ImageNet (Clean↑)	ImageNet (Robust↑)	Avg-14 (Clean↑)	Avg-14 (Robust↑)
Standard CLIP [Radford et al., 2021]	74.90	0.00	<u>73.08</u>	0.01
TeCoA [Mao et al., 2024]	80.00	<u>61.74</u>	61.56	43.26
PMG-AFT [Wang et al., 2024b]	77.84	60.02	64.46	45.74
FARE [Schlarmann et al., 2024]	72.96	43.56	65.50	42.97
TGA-ZSR [Yu et al., 2024]	<u>80.26</u>	61.46	62.11	45.19
Adv-W2S [Dong et al., 2025]	75.84	58.30	68.75	<u>48.38</u>
TPC (Ours)	**81.42**	**64.05**	**76.19**	**52.03**

Table 1: Zero-shot recognition and adversarial robustness. Top-1 clean and robust accuracy (%) on ImageNet and the macro-average over 14 datasets (ImageNet plus 13 transfer benchmarks) under AutoAttack ($\epsilon_{\text{adv}} = 2/255$). Best and second-best results are shown in bold and underlined, respectively.

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. Best and second-best results are shown in **bold** and <u>underlined</u>, respectively.

	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	<u>82.29</u>	<u>84.05</u>	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	<u>63.0</u>	<u>68.7</u>	<u>52.3</u>	<u>58.3</u>
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., 2024]	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 [Dong et al., 2025]	<u>85.60</u>	<u>84.90</u>	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**

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 AutoAttack [Croce and Hein, 2020]. 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., 2024], Robust SuperAlignment (Adv-W2S) [Dong et al., 2025]); (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. Implementation details are in §B.4.

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 the role of core isolation and residual suppression in reducing brittle over-commitment.

Table 3: **Single-factor ablations of TPC.** We report zero-shot clean/robust accuracy (%) and LVLM transfer (POPE Avg F1, VQA Avg). Best and second-best results are shown in **bold** and underlined, respectively.

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)

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.

6.3 Ablation and Analysis

Single-factor ablations. Table 3 shows that each component contributes to performance. The removal of the multi-view consensus core consistently degrades robustness and LVLM transfer, while removing core agreement causes the largest overall performance regression. Uncertainty scaling mainly affects robust accuracy (ImageNet robust (↓1.17)). 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.

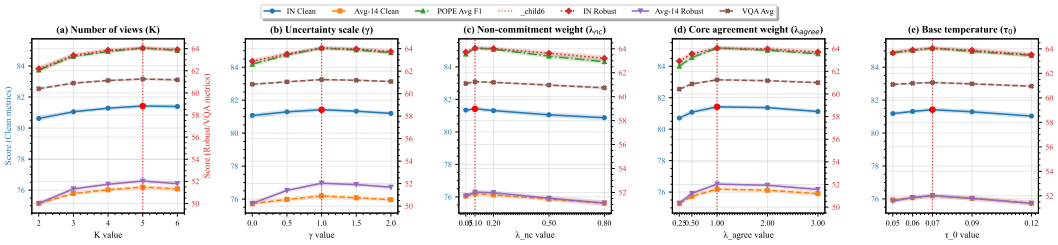

Figure 4: **Hyperparameter sensitivity of TPC.** Mean±std over 3 seeds.

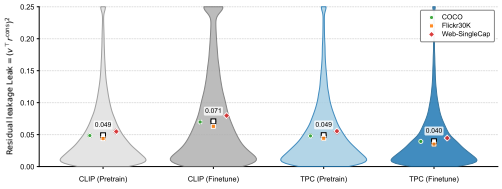

Figure 5: **Residual leakage.** Distribution of $\text{Leak} = (v^\top r^{\text{cons}})^2$ before/after fine-tuning; lower indicates less residual commitment. Protocol in App. B.6.

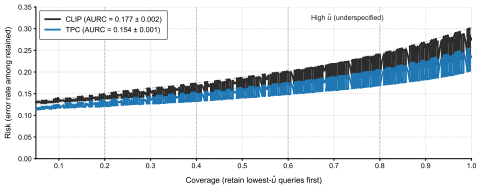

Figure 6: **Risk–coverage from $\hat{u}$.** Protocol in App. B.6.

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 $\text{Leak} = (v^\top r^{\text{cons}})^2$ 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 the role of λ_{nc} in suppressing over-commitment to unsaid, view-specific components.

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.

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 and retaining single-text inference. Our results show that modeling language as an incomplete constraint and calibrating by view disagreement improves 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.

Acknowledgments and Disclosure of Funding

Funding. This work was supported by the China Postdoctoral Science Foundation (2026M791721) and the Postdoctoral Fellowship Program of CPSF (GZC20260899).

Competing interests. The authors declare no competing interests.

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.

Qianyu Chen, Canran Xiao, and Runxuan Tang. Hidden forgetting in continual multimodal learning: When accuracy survives but grounding fails. *arXiv preprint arXiv:2607.02020*, 2026. URL `https://arxiv.org/abs/2607.02020`.

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: Underspecified queries limit vision-language models. *arXiv preprint arXiv:2601.06165*, 2026.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2206–2216. PMLR, 2020. URL `https://proceedings.mlr.press/v119/croce20b.html`.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848. URL `https://doi.org/10.1109/CVPR.2009.5206848`.

Junhao Dong, Cong Zhang, Xinghua Qu, Zejun Ma, Piotr Koniusz, and Yew-Soon Ong. Robust SuperAlignment: Weak-to-strong robustness generalization for vision-language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/1a87980b9853e84dfb295855b425c262-Abstract-Conference.html.

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.

Enguang Fan, Anfeng Peng, Matthew Caesar, Jae Kim, Josh Eckhardt, Greg Kimberly, and Denis Osipychev. Towards effective swarm-based GPS spoofing detection in disadvantaged platforms. In *2023 IEEE Military Communications Conference (MILCOM)*, pages 722–728. IEEE, 2023. doi: 10.1109/MILCOM58377.2023.10356314.

Enguang Fan, Yifan Chen, Zihan Shan, Klara Nahrstedt, Matthew Caesar, and Jae Kim. Communication-aware multi-agent reinforcement learning for decentralized cooperative UAV deployment, 2026. URL https://arxiv.org/abs/2603.16141.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. VizWiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3608–3617, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Gurari_VizWiz_Grand_Challenge_CVPR_2018_paper.html.

Jiani He, Dingyan Shang, Yihua Xu, Shiqi Huang, Yan Lyu, Jize Li, and Shangjing Tang. Semantic signal-assisted inspection and recovery allocation in reverse logistics, 2026. URL https://arxiv.org/abs/2609.02116.

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.

Saisai Hu. Research on security enhancement methods for adversarial robust large language model intelligent agents for medical decision-making tasks. In *2026 7th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT)*, pages 599–603. IEEE, 2026. URL https://arxiv.org/abs/2605.08257.

Shiqi Huang, Jiani He, Dingyan Shang, Yihua Xu, Jize Li, Yan Lyu, and Lakshmi Muraleedharan Nair. DACRI: Decision-aware causal intervention ranking for critical supply chains, 2026. URL https://arxiv.org/abs/2608.11154.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6700–6709, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Hudson_GQA_A_New_Dataset_for_Real-World_Visual_Reasoning_and_Compositional_CVPR_2019_paper.html.

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.

Jize Li, Jiani He, Dishu Yang, Dingyan Shang, Jingjing Liu, and Shiqi Huang. Accuracy-preserving stability regularization for large-scale retail demand forecasting, 2026. URL https://arxiv.org/abs/2607.13331.

Yang Li, Xiaoxue Chen, Hao Zhao, Jiangtao Gong, Guyue Zhou, Federico Rossano, and Yixin Zhu. Understanding embodied reference with Touch-Line Transformer. In *The Eleventh International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=ugA1HX69sf.

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, 2023b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, 2014. doi: 10.1007/978-3-319-10602-1_48. URL https://link.springer.com/chapter/10.1007/978-3-319-10602-1_48.

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

Jingjing Liu, Dishu Yang, Hongbo Qin, and Hansong Zhang. Passing tests, missing semantics: Evidence of interpretive lock-in in plan-based DataFrame code generation, 2026. URL https://arxiv.org/abs/2610.15797.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/11332b6b6cf4485b84afadb1352d3a9a-Abstract-Conference.html.

Mingjun Ma, Saisai Hu, Tiantian Zhu, Yitian Lin, Jiachen Sun, Zhengqiu Weng, Dehao Li, and Yajie Zhang. MuSK: Multi-scale knowledge learning for provenance-graph anomaly detection. *Computer Networks*, 289:112728, 2026. doi: 10.1016/j.comnet.2026.112728. URL https://doi.org/10.1016/j.comnet.2026.112728.

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.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3195–3204, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Marino_OK-VQA_A_Visual_Question_Answering_Benchmark_Requiring_External_Knowledge_CVPR_2019_paper.html.

Meta. Llama-3.1-8B-Instruct model card. Hugging Face, 2024. URL https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct. Official model card.

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.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y. URL https://link.springer.com/article/10.1007/s11263-015-0816-y.

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, 2024.

Wanru Shao. Interpretable ensemble learning for network traffic anomaly detection: A SHAP-based explainable AI framework for embedded systems security. In *2026 14th International Symposium on Digital Forensics and Security (ISDFS)*, pages 1–4. IEEE, 2026a. doi: 10.1109/ISDFS69419.2026.11459069. URL https://doi.org/10.1109/ISDFS69419.2026.11459069.

Wanru Shao. Policy-driven vulnerability risk quantification framework for large-scale cloud infrastructure data security. In *Proceedings of the 2026 5th International Conference on Cyber Security, Artificial Intelligence and Digital Economy*, pages 40–45, New York, NY, USA, 2026b. Association for Computing Machinery. doi: 10.1145/3819836.3819842. URL https://doi.org/10.1145/3819836.3819842.

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.

Beiwen Tian, Liyi Luo, Hao Zhao, and Guyue Zhou. VIBUS: Data-efficient 3D scene parsing with VIewpoint Bottleneck and Uncertainty-Spectrum modeling. *ISPRS Journal of Photogrammetry and Remote Sensing*, 194:302–318, 2022. doi: 10.1016/j.isprsjprs.2022.10.013.

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.

Mengxiao Wang and Nitesh Saxena. Trading agents or market crashers? dissecting robustness and security failures in academic financial LLM trading schemes, 2026. URL https://arxiv.org/abs/2609.19705.

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.

Zijun Wang, Saisai Hu, Tiantian Zhu, Wenrui Cheng, Zhengqiu Weng, Haiting Chen, Xiangyang Zheng, and Suyu Zhang. Trident: Dual-stream APT attribution over heterogeneous threat knowledge graphs. *Computers & Security*, 171:105094, 2026. doi: 10.1016/j.cose.2026.105094. URL https://doi.org/10.1016/j.cose.2026.105094.

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*, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/8f0d446441a938d9de420a8ab8d7fd36-Abstract-Conference.html.

Linqi Xiao, Venkateswarlu Gudepu, and Andrea Fumagalli. CORA: A conversational AI-agent for intelligent automation of OpenROADM multi-vendor optical transport networks. In *Applications of Machine Learning 2026*, volume 14218 of *Proceedings of SPIE*, page 142180Y. SPIE, 2026b. doi: 10.1117/12.3113861.

Linqi Xiao, Aparaajitha Gomathinayakam Latha, Venkateswarlu Gudepu, and Andrea Fumagalli. A YANG-grounded LLM agent supporting multi-vendor OpenROADM optical transport networks. In *Proceedings of the ACM SIGCOMM 2026 Conference*, pages 2109–2115. ACM, 2026c. doi: 10.1145/3789240.3822606.

Dishu Yang, Jingjing Liu, and Jize Li. Balancing privacy, utility, and safety in LLM alignment through preference optimization, 2026a. URL https://arxiv.org/abs/2608.30141.

Dishu Yang, Qi Su, Hongbo Qin, and Hansong Zhang. Measurement-first auditing of agentic leaderboards: Contamination susceptibility, matched-control re-evaluation, and scorer validation, 2026b. URL https://arxiv.org/abs/2610.05830.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. doi: 10.1162/tacl_a_00166. URL https://aclanthology.org/Q14-1006/.

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. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/e4165c96702bac5f4962b70f3cf2f136-Abstract-Conference.html.

Zihao Zheng, Baichuan Li, Junyi Yao, and Jiayu Long. Reliable financial named entity recognition under domain shift: Confidence estimation and selective prediction, 2026. URL https://arxiv.org/abs/2608.19558.

Leisheng Zhong, Yu Zhang, Hao Zhao, An Chang, Wenhao Xiang, Shunli Zhang, and Li Zhang. Seeing through the occluders: Robust monocular 6-DOF object pose tracking via model-guided video object segmentation. *IEEE Robotics and Automation Letters*, 5(4):5159–5166, 2020. doi: 10.1109/LRA.2020.3003866.

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. URL https://openreview.net/forum?id=xp7wDU9JBW.

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 $r_{i,k}^{\mathrm{cons}} \triangleq t_{i,k} - (t_{i,k}^\top c_i)c_i$ and radii $\delta_{i,k} \triangleq \|r_{i,k}^{\mathrm{cons}}\|_2$. Let $\mathcal{U}_i \triangleq \mathrm{span}\{r_{i,k}^{\mathrm{cons}} : k \in [K]\} \subseteq c_i^\perp$ 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 $u_i = \frac{1}{K}\sum_{k=1}^K \big(1 - (t_{i,k}^\top c_i)^2\big)$, 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 $r_{i,k}^{\mathrm{cons}} \triangleq t_{i,k} - (t_{i,k}^\top c_i)c_i$. Then $c_i^\top r_{i,k}^{\mathrm{cons}} = c_i^\top t_{i,k} - (t_{i,k}^\top c_i)\|c_i\|_2^2 = 0$, hence $r_{i,k}^{\mathrm{cons}} \in c_i^\perp$ and in particular $r_{i,k}^{\mathrm{cons}} \in \mathcal{U}_i$.

Step 2: identify the coefficient on c_i. Using Pythagoras on the orthogonal decomposition,
$$\|t_{i,k}\|_2^2 = \|(t_{i,k}^\top c_i)c_i\|_2^2 + \|r_{i,k}^{\mathrm{cons}}\|_2^2 = (t_{i,k}^\top c_i)^2 + \delta_{i,k}^2.$$
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
$$t_{i,k} = (t_{i,k}^\top c_i)c_i + r_{i,k}^{\mathrm{cons}} = \sqrt{1-\delta_{i,k}^2}\, c_i + r_{i,k}^{\mathrm{cons}}.$$
Since $r_{i,k}^{\mathrm{cons}} \in \mathcal{U}_i$ and $\|r_{i,k}^{\mathrm{cons}}\|_2 = \delta_{i,k} \leq \rho_i$, 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),
$$u_i = \frac{1}{K}\sum_{k=1}^K \big(1 - (t_{i,k}^\top c_i)^2\big) = \frac{1}{K}\sum_{k=1}^K \delta_{i,k}^2.$$
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 nonzero subspace $\mathcal{U} \subseteq c^\perp$, and $\rho \in [0,1)$. Let*
$$\mathcal{A}(c,\mathcal{U},\rho) = \Big\{ t \in \mathbb{R}^d :\; \|t\|_2 = 1,\; t = \sqrt{1-\|r\|_2^2}\, c + r,\; r \in \mathcal{U},\; \|r\|_2 \leq \rho \Big\}.$$
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. For $\rho = 0$, the ambiguity set is $\{c\}$ and both conclusions hold immediately. Assume $\rho > 0$ below. **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,

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

Equality holds for $r = -s\, u/\|u\|_2$ when $b > 0$. When $b = 0$, choose $r = sw$ for any unit $w \in \mathcal{U}$, which exists because $\mathcal{U} \neq \{0\}$. 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 $s \in (0, \rho)$,

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

Thus f is non-increasing on $[0, \rho]$, and a 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 $b > 0$, or $r^\star = \rho w$ for any unit $w \in \mathcal{U}$ when $b = 0$. In both cases, $\|r^\star\|_2 = \rho$ and equality holds 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 $v^\top t$:

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

□

If $\mathcal{U} = \{0\}$, the only admissible residual is $r = 0$. Thus $\mathcal{A}(c, \mathcal{U}, \rho) = \{c\}$, and for any unit v the worst-case similarity and loss are $a = v^\top c$ and $\ell(a)$, respectively.

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,

$$\mathbb{E}\|\bar{\varepsilon}\|_2^2 = \mathbb{E}\Big\langle \frac{1}{K}\sum_{k=1}^{K}\varepsilon_k, \frac{1}{K}\sum_{\ell=1}^{K}\varepsilon_\ell \Big\rangle = \frac{1}{K^2}\sum_{k=1}^{K}\mathbb{E}\|\varepsilon_k\|_2^2 + \frac{1}{K^2}\sum_{k\neq\ell}\mathbb{E}\langle\varepsilon_k,\varepsilon_\ell\rangle$$
$$= \frac{1}{K^2}\sum_{k=1}^{K}\mathbb{E}\|\varepsilon_k\|_2^2 + \frac{1}{K^2}\sum_{k\neq\ell}\Big\langle \mathbb{E}[\varepsilon_k], \mathbb{E}[\varepsilon_\ell]\Big\rangle \leq \frac{1}{K^2}\sum_{k=1}^{K}\sigma^2 = \frac{\sigma^2}{K}.$$

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

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

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

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

$$\big|g(\varepsilon_1,\ldots,\varepsilon_K) - g(\varepsilon_1,\ldots,\varepsilon_k',\ldots,\varepsilon_K)\big| \leq \Big\|\frac{1}{K}\sum_{j=1}^{K}\varepsilon_j - \frac{1}{K}\Big(\varepsilon_k' + \sum_{j\neq k}\varepsilon_j\Big)\Big\|_2$$
$$= \frac{1}{K}\|\varepsilon_k - \varepsilon_k'\|_2 \leq \frac{1}{K}\big(\|\varepsilon_k\|_2 + \|\varepsilon_k'\|_2\big) \leq \frac{2\sigma}{K}.$$

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

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

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*

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

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

$$\|c-\mu\|_2 = \Big\|\frac{u}{\|u\|_2} - \mu\Big\|_2 = \frac{1}{\|u\|_2}\|u - \|u\|_2\mu\|_2$$
$$= \frac{1}{\|u\|_2}\|\mu + e - \|u\|_2\mu\|_2 = \frac{1}{\|u\|_2}\|e + (1-\|u\|_2)\mu\|_2$$
$$\leq \frac{1}{\|u\|_2}\Big(\|e\|_2 + |1-\|u\|_2|\cdot\|\mu\|_2\Big) = \frac{1}{\|u\|_2}\big(\|e\|_2 + |1-\|u\|_2|\big).$$

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

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

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 $a, b \in \mathbb{S}^{d-1}$,*

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

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 $\|c_i\|_2 = 1$,*

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

where $\xi_i \in \{-1, +1\}$ are independent Rademacher signs and $\mathfrak{R}_n(\mathcal{F}) = \mathbb{E}_\xi \big[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \xi_i f(x_i)^\top c_i \big]$.

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

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

which is Eq. (30). □

A.2.2 Proof of Theorem 5.2

Proof. The proof has four steps.

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 $\delta_i \in (0,1)$,

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

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

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

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

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

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 $1 - \delta/2$,

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

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 $i \in [n]$,

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

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

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

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

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

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

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

Averaging over i gives the empirical relation

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

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

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

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 $1 - \delta/2$,

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

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

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

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:

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

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 $2L\,\varepsilon_{K,n}(\delta) \le \epsilon_{\text{tol}}$. When $\eta_{K,n}(\delta) \le \frac{1}{2}$, $\varepsilon_{K,n}(\delta) \le 4\eta_{K,n}(\delta)$, so it suffices that $8L\,\eta_{K,n}(\delta) \le \epsilon_{\text{tol}}$. By the definition of $\eta_{K,n}(\delta)$ in Eq. (34), this holds when

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

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

B Additional Experimental Details

B.1 Preliminary Study Protocol

We detail the diagnostic study in Sec. 3.

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

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

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

$$c_i = \mathrm{norm}\left(\sum_{k=1}^{K} t_{i,k}\right), \tag{55}$$

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

$$r_{i,k}^{\mathrm{cons}} = t_{i,k} - (t_{i,k}^\top c_i) c_i. \tag{56}$$

The residual leakage score is

$$\mathrm{Leak}_{i,k} = \big(v_i^\top r_{i,k}^{\mathrm{cons}}\big)^2. \tag{57}$$

Large $\mathrm{Leak}_{i,k}$ 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 $\text{Leak}_{i,k}$ 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 diagnostic study links caption disagreement to unstable retrieval ranks and residual leakage to 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 its K captions independently, increasing text-side compute while sharing the image forward pass. 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. These controlled omissions test the setting targeted by residual suppression and uncertainty-aware sharpness.

Image–text retrieval benchmarks. We evaluate retrieval on MS-COCO [Lin et al., 2014] and Flickr30K [Young et al., 2014] using standard splits. At test time, each query caption y is mapped to the single-view core $\hat{c} = \text{norm}(\psi_\theta(\text{norm}(\phi_t(y))))$ and each candidate image is mapped to $v = \text{norm}(\phi_v(x))$; 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 measure over-commitment 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). 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 probe attribute/relation sensitivity through fine-grained alignment diagnostics: 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 report LVLM results separately from dual-encoder retrieval/classification to distinguish their architectures and training settings.

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 [Croce and Hein, 2020] with an ℓ_∞ threat model and a fixed perturbation radius. We use a single, fixed radius across all methods (default $\epsilon_{\text{adv}}{=}4/255$) 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 $224{\times}224$ 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. 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: $\text{LR}_{\text{enc}} = 5 \times 10^{-5}$ and $\text{LR}_\psi = 5 \times 10^{-4}$ by default, with weight decay 0.1. Unless otherwise stated, we tune only LR_{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 $\tau_0{=}0.07$, $\gamma{=}1.0$, $\lambda_{\text{agree}}{=}1.0$, and $\lambda_{\text{nc}}{=}0.1$. We tune $\gamma \in \{0, 0.5, 1, 2\}$, $\lambda_{\text{agree}} \in \{0.5, 1, 2\}$, and $\lambda_{\text{nc}} \in \{0.05, 0.1, 0.2, 0.5\}$ on a held-out validation split. For ablations, we also report: (i) removing uncertainty scaling ($\gamma{=}0$), (ii) removing residual suppression ($\lambda_{\text{nc}}{=}0$), 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$\pm$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., 2024], 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) [Dong et al., 2025], 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]. We report LVLM results separately from dual-encoder retrieval results, keeping the LLM, connector architecture, and SFT data identical unless the baseline 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

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

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

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

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

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

The residual leakage score is then

$$\mathrm{Leak}_{i,k} = (v_i^{\top} r_{i,k}^{\mathrm{cons}})^2. \tag{63}$$

A larger $\mathrm{Leak}_{i,k}$ indicates stronger correlation between the image representation and a view-specific textual component orthogonal to the shared semantic core. We compute $\mathrm{Leak}_{i,k}$ before and after fine-tuning for standard CLIP and TPC. Across methods, the diagnostic residual space uses the multi-caption consensus c_i as its reference, independent of 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 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

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

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. The core filter 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.

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

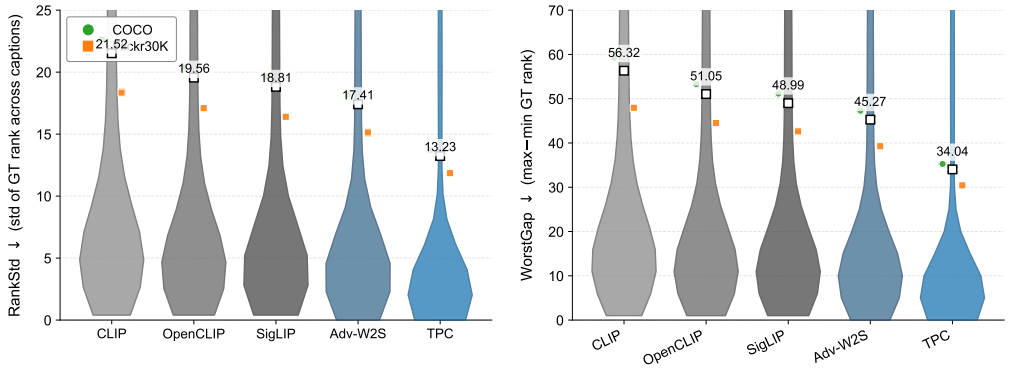

Figure 7: **Caption-choice sensitivity.** Violin plots show per-image distributions of RankStd and WorstGap on COCO and Flickr30K. Dots and error bars indicate mean and 95% CI; faint points show dataset-wise means. Lower is better: less variation in retrieval rank across captions of the same image.

C Additional Experimental Results

C.1 Consensus Core Stability and Caption-Choice Sensitivity

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.

C.2 Single-Caption Core Filter: ψ_θ and Multi-View Consensus

We evaluate how closely the single-view core filter ψ_θ approximates 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 $c_i = \mathrm{norm}(\sum_k t_{i,k})$ with $t_{i,k} = \mathrm{norm}(\phi_t(y_{i,k}))$. For each caption view, we predict a single-caption core $\hat{c}_{i,k} = \mathrm{norm}(\psi_\theta(t_{i,k}))$ and measure

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

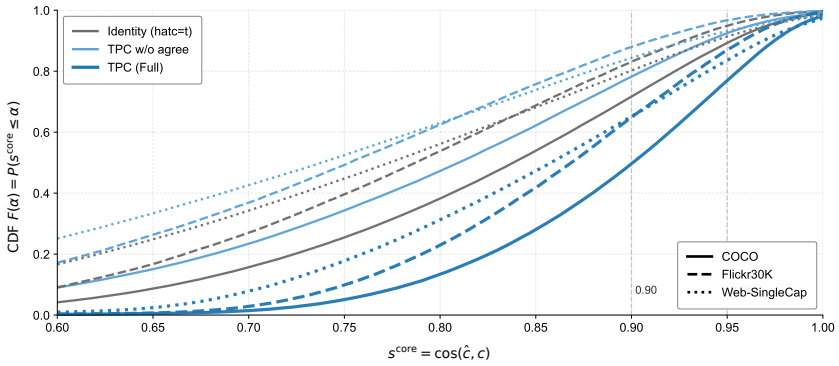

Figure 8: **Consensus-core prediction by ψ_θ.** Empirical CDFs of $s^{\text{core}} = \cos(\hat{c}, c)$ on COCO, Flickr30K, and a single-caption web-style set. Colors denote methods; line styles denote datasets. A right-shifted curve indicates that single-caption cores $\hat{c}$ better match the multi-view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

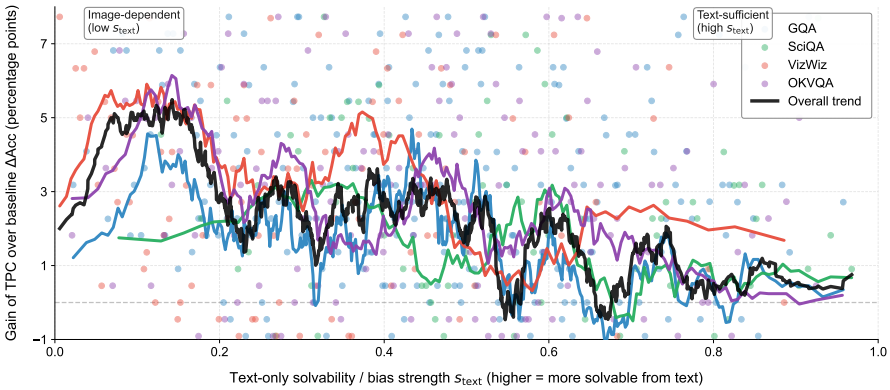

Figure 9: **Gain vs. text-only solvability.** Each point is an equal-count bin (hundreds of bins per dataset); colors indicate datasets. The trend line is a smoothed running mean over bins (same color per dataset) and an overall trend (black). Higher gain at lower solvability indicates TPC is most beneficial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

We plot the empirical CDF $F(\alpha) = \mathbb{P}(s^{\text{core}} \le \alpha)$ 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 s^{core}), 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 eliminates much of the gain, indicating that ψ_θ learns a deployable single-caption estimator of shared semantics.

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

We test whether suppressing residual commitment is most beneficial when text is underspecified (weak constraint), with diminishing gains when text is already sufficient (strong bias).

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

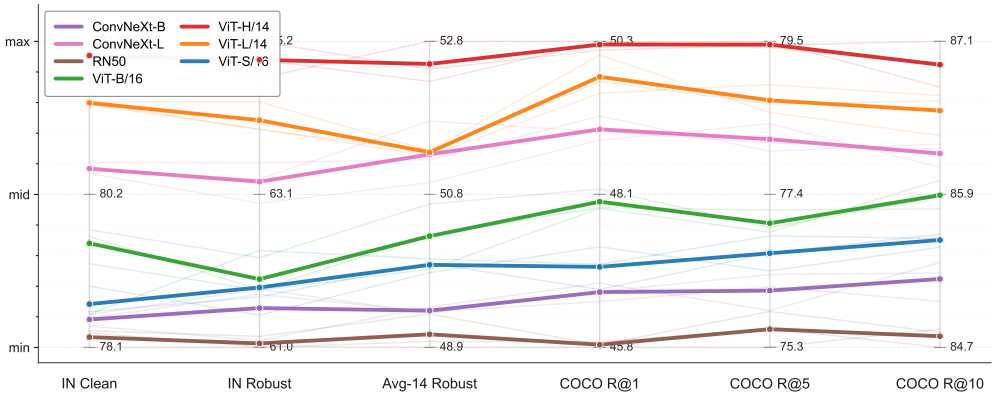

Figure 10: **Cross-backbone robustness and retrieval.** Parallel coordinates over multiple metrics (each axis is min–max normalized for readability). Each backbone is shown with three seed runs (faint) and the mean (bold). Consistent trends across architectures indicate that TPC generalizes across backbone families.

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 ($\epsilon_{\text{adv}}=2/255$), 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).

Fig. 10 shows that TPC yields stable performance across both ViT and ConvNeXt families: larger backbones improve clean accuracy and COCO recall, while robust accuracy remains consistently high. 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. We test the hypothesis that residual suppression and uncertainty-aware smoothing reduce the performance and calibration degradation associated with over-commitment to view-specific text.

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 text-only span extractor driven by an off-the-shelf instruction-tuned LLM. Concretely, we use **Llama-3.1-8B-Instruct** [Meta, 2024] 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. The extractor returns only verbatim spans from the original caption, preserving its wording and content. 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

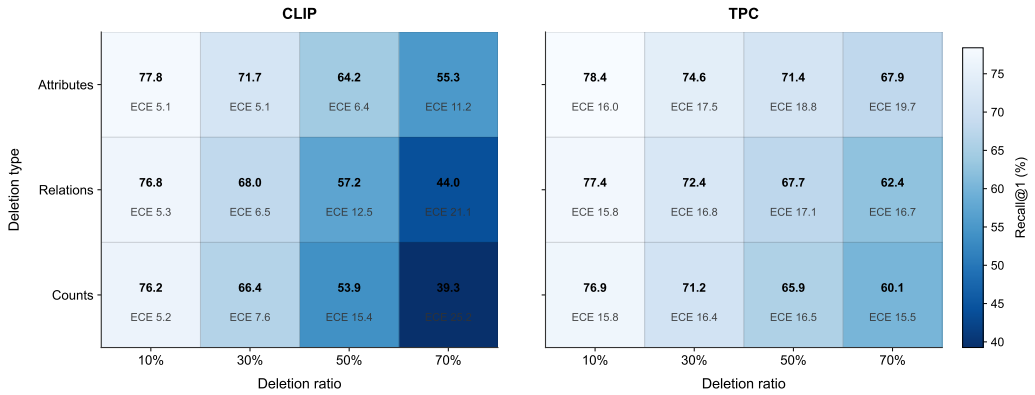

Figure 11: **Controlled caption deletion stress test.** Heatmaps show Recall@1 (%) under increasing deletion ratios (x-axis) and deletion types (y-axis). Each cell is annotated with *Recall@1 / ECE* (ECE shown as a small number). TPC maintains higher Recall@1 throughout. Its ECE is higher at 10–50% deletion. At 70% deletion, its ECE is lower only for relations and counts; attribute deletion retains higher ECE at every ratio.

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 retains higher Recall@1 and exhibits smaller performance drops as deletions increase. Calibration shows a trade-off: TPC starts with higher ECE at 10% deletion and remains higher at 30–50% across all three deletion types. For relations and counts, its ECE changes less as deletion severity increases, becoming lower than CLIP's at 70% deletion (16.7 vs. 21.1 and 15.5 vs. 25.2, respectively). For attributes, TPC's ECE remains higher at every ratio. These results support improved retrieval robustness and more favorable calibration degradation trends in some severe-deletion settings; lower ECE is confined to those settings.

D Responsible Research and Reproducibility

D.1 Scope and Limitations

TPC addresses representation-level alignment under naturally partial language supervision or multiple faithful caption views. Its scope covers robust image–text matching, calibrated retrieval, zero-shot transfer, and plug-in LVLM transfer through a stronger vision encoder. This alignment backbone can complement task-specific safety filters, factuality checkers, and policy-level controls in deployed LVLM systems. Complementary approaches include evidence constraints, consistency checks, and

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

confidence reweighting for medical decision-making agents [Hu, 2026], optical-network agents [Xiao et al., 2026c,b], and preference optimization balancing privacy, utility, and safety [Yang et al., 2026a].

Training estimates a consensus core from human multi-caption annotations or, for single-caption data, faithful auxiliary views constructed through controlled deletion, paraphrasing, and back-translation with deterministic filtering. This setting accommodates many existing vision–language resources and allows retrieval- or generation-based extensions. Performance is stable across $K \in \{3, 4, 5\}$ and saturates around four to five views in our experiments, suggesting that a small caption set per image is sufficient for TPC.

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

D.2 Statistical Reporting Across Seeds

Unless otherwise stated, trainable TPC variants use three random seeds for core-filter initialization, data shuffling, caption-view subsampling, and minibatch ordering. Main tables report means; sensitivity and diagnostic figures report mean±standard deviation. For asymmetric metrics, including residual leakage and rank-based quantities, we additionally report 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, supporting consistent conclusions across seeds.

External checkpoints use deterministic evaluation with the same preprocessing and scripts, without retraining. Retrained baselines and ablations follow the same three-seed protocol whenever computationally comparable.

D.3 Broader Impact and Responsible Use

Potential benefits of TPC include more stable image–text retrieval, better calibrated confidence under ambiguous captions, and reduced object hallucination in downstream LVLMs. These properties support applications with incomplete or paraphrased user descriptions, where overconfident visual grounding can produce unreliable responses.

Stronger visual grounding may also enable misuse in surveillance, automated profiling, or misleading content generation when downstream systems lack proper safeguards. Calibrated confidence alone cannot ensure factual correctness or safety; passing generated tests can also mask semantic errors [Liu et al., 2026], while financial named entity recognition illustrates limits to confidence-based abstention under severe domain shift [Zheng et al., 2026]. TPC addresses representation-level reliability; related application-level objectives include cooperative decision-making under limited observations [Fan et al., 2026], semantic-signal-based inspection allocation [He et al., 2026], causal intervention ranking [Huang et al., 2026], and forecast stability [Li et al., 2026].

Responsible use calls for application-level safety filters, dataset and demographic audits, evaluation-pipeline auditing [Yang et al., 2026b], human oversight in high-stakes settings, and clear communication of model uncertainty. Complementary deployment-security concerns include policy-driven vulnerability risk assessment for cloud infrastructure [Shao, 2026b], interpretable network-traffic anomaly detection [Shao, 2026a], provenance-graph anomaly detection [Ma et al., 2026], swarm-based GPS spoofing detection [Fan et al., 2023], and security failures in financial LLM trading agents [Wang and Saxena, 2026]. No new user-facing model or scraped dataset accompanies this paper.

D.4 LLM Usage Disclosure

LLM use is confined to the controlled caption-deletion stress test in Appendix C.5. Llama-3.1-8B-Instruct serves as a deterministic span extractor for attributes, relations, and counts. Extraction is restricted to verbatim substrings of the original caption, preserving its content and image–text label.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state that the paper proposes TPC, a core–residual alignment framework for underspecified vision–language supervision, and these claims are supported by the method, theory, main results, ablations, and additional analyses in Sections 4–6 and Appendix C.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Appendix D.1 discusses the intended scope of TPC, including its focus on representation-level alignment, its multi-view supervision setting, and the role of deployment-time safeguards for downstream LVLM use.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Section 5 states the assumptions for the distributional-robustness and multi-view denoising analyses, and Appendix A provides full proofs for all lemmas, theorems, and corollaries.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper specifies datasets, training/evaluation protocols, metrics, baselines, model scales, hyperparameters, seed protocol, and compute resources in Sections 6.1, 6.2, and Appendices B–B.7.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The study uses public benchmark datasets and publicly available model families where applicable. The implementation is not publicly released.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Appendix B.2 describes dataset construction and splits, Appendix B.3 defines the metrics, and Appendix B.4 reports the optimizer, schedules, batch size, view construction, and hyperparameter choices.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Main sensitivity and analysis figures report mean and standard deviation over three seeds, and Appendix D.2 further reports seed-level variability for the principal metrics supporting the main claims.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes].

   Justification: Appendix B.7 reports the GPU type, memory, number of workers, approximate runtime, peak memory, and GPU-hour estimates for main training, ablations, robustness evaluation, LVLM transfer, and controlled-deletion experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes].

   Justification: The work uses public research benchmarks and existing model families under their intended research settings, does not collect private user data, and does not involve human-subject experiments or personally identifiable information.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes].

    Justification: Appendix D.3 discusses positive impacts from more calibrated and robust vision–language alignment, as well as potential misuse or over-reliance risks when improved vision encoders are integrated into downstream LVLM systems.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A].

    Justification: No new high-risk pretrained model, image generator, scraped dataset, or user-facing system accompanies this paper. Responsible-use considerations are discussed in Appendix D.3.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes].

Justification: Existing datasets, codebases, and model families are cited in the paper.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A].

Justification: No new standalone dataset, benchmark, model checkpoint, or software package accompanies this paper. Generated caption views and controlled-deletion spans are used as internal experimental artifacts and are documented through deterministic protocols in Appendices B.2 and C.5.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A].

Justification: The paper does not conduct crowdsourcing experiments and does not involve new human-subject data collection.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A].

    Justification: The paper does not involve human-subject research, new participant recruitment, user studies, or private user data collection.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes].

    Justification: Appendix C.5 and Appendix D.4 disclose that Llama-3.1-8B-Instruct is used only to extract verbatim spans for the controlled caption-deletion stress test; it is not used to train TPC, generate labels for the main benchmarks, or produce model predictions.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.