LCD: Semantic–Residual Feature Factorization for Vision-Language Representations

Anonymous ACL submission

Abstract

Contrastive vision-language models align images and text globally, but global compatibility does not determine how visual information should be accounted for when a phrase describes only part of a scene. We study this gap as conditional semantic–residual feature factorization over frozen CLIP image representations and instantiate it with LCD, a lightweight decomposition module trained by selectivity, complementarity, neutrality, and image-consistency constraints. On phrase-region data from Flickr30K Entities, Visual Genome, and VGPhraseCut, LCD gives strong semantic-side selectivity, reaching 0.478 R@1 and 0.577 MRR on Flickr30K. Residual evidence is narrower: LCD trained on VGPhraseCut improves the targeted RAPS metric over a projection proxy on all three test sources. The paper therefore provides controlled evidence for explicit residual semantic modeling, not a general claim of complete complementary information recovery.

1 Introduction

Contrastive vision-language models such as CLIP and SigLIP have established a simple paradigm for image-text representation learning (Radford et al., 2021; Zhai et al., 2023): image and text encoders map paired inputs into a shared space, where matched pairs are encouraged to be closer than mismatched pairs.

This supervision is inherently asymmetric. A real image contains objects, attributes, relations, and background context, while a short caption usually names only part of this content. The resulting image-text information imbalance means that a valid global match does not indicate which visual factors are explained by the text and which remain unaccounted for (Lavoie et al., 2024; Schrodi et al., 2025). Global alignment may therefore favor salient or easily matched cues without providing a structured account of the full visual content.

Motivated by this limitation, we study how to decompose an image representation into target semantics and residual semantics with respect to a text condition. Instead of treating an image-text pair as a single global matching unit, we ask which part of the visual representation is explained by the current text and which part remains as valid but unmentioned visual information. This shifts the focus from global compatibility to condition-relative feature decomposition.

A key challenge is that residual semantics should not be treated as noise. Given a condition such as "a red ball", the unmentioned parts of the image may still contain meaningful objects, attributes, relations, or scene context. These factors are not target semantics under the current condition, but they remain valid same-image semantics. The residual representation should therefore suppress information explained by the current text while preserving complementary visual information not described by it.

To make this problem concrete, we introduce LCD, a preliminary semantic decomposition framework for condition-relative multimodal representations. Given an image and a text condition, LCD separates the visual representation into a target semantic component aligned with the current text and a residual semantic component intended to preserve complementary same-image semantics. Because residual semantics are difficult to evaluate with standard retrieval metrics, we use three diagnostics: LS-MRR for low-salience same-image retrieval, NO-Pref for preference toward non-overlapping complements, and RAPS as their combined residual score.

Our contributions are threefold. First, we formulate condition-relative target-residual semantic decomposition as a feature-accounting problem for multimodal alignment. Second, we introduce LS-MRR, NO-Pref, and RAPS to evaluate whether residual representations preserve meaning-

ful unmentioned same-image semantics. Third, we propose LCD, a lightweight baseline that decomposes frozen vision-language representations into target and residual semantic components. Through controlled phrase-region experiments based on Flickr30K Entities, Visual Genome, and VGPhraseCut (Plummer et al., 2015; Krishna et al., 2017; Wu et al., 2020), we analyze both the promise and the limitations of residual semantic modeling.

2 Related Work

Vision-language alignment and residual accounting. Contrastive vision-language models train image and text encoders so matched pairs are close and mismatched pairs are separated. CLIP made this interface broadly transferable (Radford et al., 2021), and later dual-encoder, locked-image, sigmoid-loss, and fusion-style variants improved scale or training design (Jia et al., 2021; Zhai et al., 2022, 2023; Li et al., 2021, 2022; Kim et al., 2021). These models make global image-text compatibility practical, but compatibility does not say which part of a visual representation is explained by a partial text. Recent analyses of modality gap, object bias, and information imbalance make this limitation more explicit (Schrodi et al., 2025). LCD studies this narrower feature-accounting problem rather than proposing a general alignment objective.

Feature factorization and information bottlenecks. Representation factorization provides the closest conceptual background. Prior multimodal work separates shared and modality-specific factors or studies weakly supervised multimodal representations (Hsu and Glass, 2018; Tsai et al., 2019; Wu and Goodman, 2018); recent contrastive work further argues that task-relevant information can lie in both shared and view-unique factors (Liang et al., 2023). LCD is related in spirit but decomposes the image side relative to an operative semantic condition, not modality-shared and modality-unique information globally. Information-bottleneck work offers a useful contrast because it often compresses non-shared or nuisance information (Federici et al., 2020; Wang et al., 2022; Almudévar et al., 2025). LCD instead treats unmentioned visual semantics as a residual factor to model, while neutrality and complementarity keep it from simply copying the current condition.

Projection baselines and controlled phrase-region data. Projection residuals remove a specified text direction from a global image embedding, giving an algebraic boundary case related to linear concept erasure (Ravfogel et al., 2020, 2022; Belrose et al., 2023; Shao et al., 2023). They are informative but are not learned residual semantic factors. Our evaluation uses phrase-region resources whose annotations make within-image candidate construction possible: Flickr30K Entities, Visual Genome, and PhraseCut (Plummer et al., 2015; Krishna et al., 2017; Wu et al., 2020). These datasets also bound the evidence, since the results depend on their phrase distributions and filtering rules.

3 Method

3.1 Conditional Feature Factorization

LCD studies conditional feature factorization on the image side of a frozen vision-language model. Let I denote an image and let t denote an operative phrase or short textual condition. A pretrained encoder maps the image to a global embedding $g_I \in \mathbb{R}^d$ and visual tokens $V_I = \{v_i\}_{i=1}^N$, where $v_i \in \mathbb{R}^d$. The paired text encoder maps the condition to $e_t \in \mathbb{R}^d$. Similarity is cosine similarity,

$$\mathrm{sim}(x, y) = \bar{x}^\top \bar{y}, \qquad \bar{x} = x/\|x\|_2. \tag{1}$$

The condition t is used to define a split of the visual representation. It is not the main object being modeled. LCD learns two condition-relative image-side factors,

$$s_{I,t} = f_{\mathrm{sem}}(I, t), \qquad r_{I,t} = f_{\mathrm{res}}(I, t). \tag{2}$$

The factor $s_{I,t}$ should account for the visual semantics specified by t. The factor $r_{I,t}$ should retain complementary semantics that are valid for the same image but not explained by the current condition. This residual factor is not assumed to contain all remaining information in an absolute sense. It is a learned semantic factor whose role is defined by the current condition and by the same-image alternatives available during training.

During training, each image I_b is associated with valid same-image conditions $\mathcal{T}_b^+ = \{t_{b,k}\}_{k=1}^{K_b}$ and cross-image conditions $\mathcal{T}_b^-$. We write $e_{b,k}$, $s_{b,k}$, and $r_{b,k}$ for the text embedding, explained factor, and residual factor for the k-th condition of image I_b. The same-image conditions define possible complementary semantics; cross-image conditions supply negatives.

3.2 Factor Construction

LCD uses OpenAI CLIP ViT-B/16 (Radford et al., 2021) as a frozen backbone. The trainable module begins after the frozen image and text encoders and constructs the two factors in a way that keeps the decomposition tied to the original image representation.

The explained semantic factor is built from condition-guided visual token evidence. A QK-identity value pooler uses the condition embedding as a query and image tokens as keys, while keeping the values in normalized visual-token space. With layer normalization (Ba et al., 2016) $\mathrm{LN}(\cdot)$, trainable projections W_q, W_k, and temperature $\tau = 0.07$,

$$\begin{aligned} q_t &= \mathrm{norm}(W_q \,\mathrm{LN}(e_t)), \\ k_i &= \mathrm{norm}(W_k \,\mathrm{LN}(v_i)), \end{aligned} \tag{3}$$

$$\alpha_i(t) = \frac{\exp(q_t^\top k_i/\tau)}{\sum_{j=1}^{N} \exp(q_t^\top k_j/\tau)}. \tag{4}$$

The pooled visual evidence is

$$h_{I,t} = \mathrm{norm}\left(\sum_{i=1}^{N} \alpha_i(t)\,\mathrm{norm}(v_i)\right). \tag{5}$$

A condition-modulated semantic head then maps the pooled evidence to the explained factor. It first normalizes the visual evidence and uses the condition embedding to produce a FiLM modulation (Perez et al., 2018),

$$x_{I,t} = \mathrm{LN}(h_{I,t}), \tag{6}$$

$$\begin{aligned} \gamma_t &= 1 + \rho \tanh(W_\gamma e_t), \\ \beta_t &= \rho \tanh(W_\beta e_t), \\ \tilde{h}_{I,t} &= \gamma_t \odot x_{I,t} + \beta_t. \end{aligned} \tag{7}$$

The FiLM parameters are initialized so that the modulation starts as the identity transform. The modulated feature is then passed through a bottleneck MLP with a GELU nonlinearity (Hendrycks and Gimpel, 2016):

$$m_{I,t} = W_2 \,\mathrm{Drop}\left(\mathrm{GELU}(W_1 \tilde{h}_{I,t})\right), \tag{8}$$

$$s_{I,t} = \mathrm{norm}\left(\mathrm{LN}(m_{I,t})\right). \tag{9}$$

The main configuration uses $\rho = 0.1$, dropout 0.1, and a $512 \to 128 \to 512$ bottleneck.

The residual factor is constructed from the global image representation after removing the semantic direction predicted by $s_{I,t}$. Let

$$u_{I,t} = \mathrm{norm}(W_p s_{I,t}) \tag{10}$$

be the projected semantic direction. The projection path is detached for residual-side objectives, so the residual branch does not reshape $s_{I,t}$ through this removal step. The image representation after explained-direction removal is

$$z_{I,t} = \mathrm{norm}\left(g_I - (g_I^\top u_{I,t}) u_{I,t}\right). \tag{11}$$

The residual head then applies condition-dependent FiLM modulation (Perez et al., 2018) and an MLP:

$$\begin{aligned} \tilde{z}_{I,t} &= \mathrm{FiLM}(z_{I,t}, e_t), \\ r_{I,t} &= \mathrm{norm}(f_{\mathrm{res}}(\tilde{z}_{I,t})). \end{aligned} \tag{12}$$

For an image-level consistency check, LCD also forms a recomposed representation

$$\hat{g}_{I,t} = \mathrm{norm}\left(\mathrm{LN}(W_s s_{I,t} + W_r r_{I,t})\right). \tag{13}$$

This construction is intentionally asymmetric: the explained factor is selected from local visual token evidence, while the residual factor begins from the global image representation after the explained direction has been removed.

3.3 Factorization Constraints

LCD is trained by four constraints that define the intended factorization behavior. For each image I_b, let $E_b^+ = \{e_{b,k}\}_{k=1}^{K_b}$ be same-image condition embeddings and $E_b^- = \{e_{b,m}^-\}_{m=1}^{M_b}$ be cross-image condition embeddings. The candidate set is

$$\mathcal{C}_b = E_b^+ \cup E_b^-. \tag{14}$$

Let $\mathrm{CE}(Y, \ell)$ denote cross entropy with target mass distributed uniformly over the positive set Y. For any factor x, write $\ell(x; \mathcal{C}_b) = \{\mathrm{sim}(x, e)/\tau : e \in \mathcal{C}_b\}$.

The selectivity constraint trains $s_{b,k}$ to explain the current condition rather than the whole image:

$$\mathcal{L}_{\mathrm{sem}} = \frac{1}{|\Omega|} \sum_{(b,k)\in\Omega} \mathrm{CE}(\{e_{b,k}\}, \ell(s_{b,k}; \mathcal{C}_b)). \tag{15}$$

The complementarity constraint trains $r_{b,k}$ to prefer other valid same-image semantics. For a condition $t_{b,k}$, define

$$\mathcal{P}_{b,k} = \{e_{b,j} : j \neq k,\ e_{b,j} \in E_b^+\}. \tag{16}$$

Then

$$\mathcal{L}_{\mathrm{res}} = \frac{1}{|\Omega_{\mathrm{res}}|} \sum_{(b,k)\in\Omega_{\mathrm{res}}} \mathrm{CE}(\mathcal{P}_{b,k}, \ell(r_{b,k}; \mathcal{C}_b)). \tag{17}$$

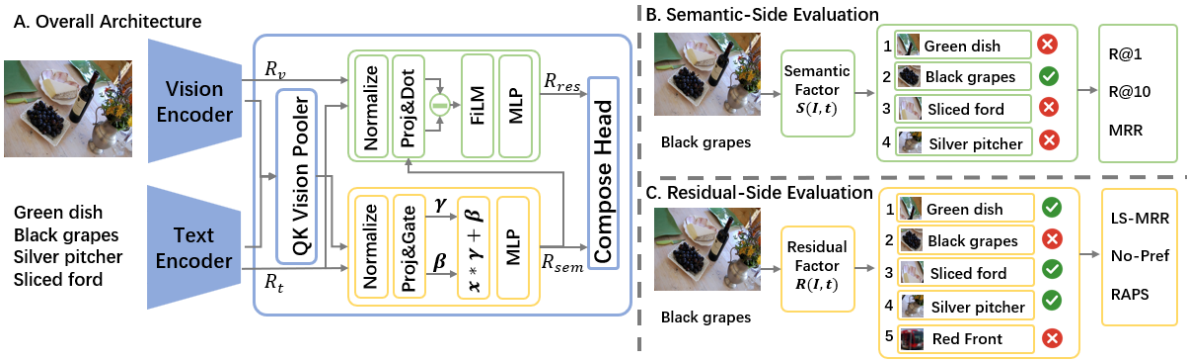

Figure 1: Overview of LCD. Panel A shows the decomposition architecture: frozen image and text encoders feed a QK vision pooler, a condition-modulated semantic branch, a residual branch, and a compose head. Panel B illustrates semantic-side evaluation, where the semantic factor $S(I, t)$ ranks same-image candidate regions for the queried condition. Panel C illustrates residual-side evaluation, where the residual factor $R(I, t)$ is tested on complementary same-image semantics with LS-MRR, No-Pref, and RAPS.

The set Ω_{res} contains valid conditions with at least one same-image complement.

The neutrality constraint prevents the residual factor from simply copying the current condition:

$$\mathcal{L}_{\text{neu}} = \frac{1}{|\Omega|} \sum_{(b,k)\in\Omega} \text{sim}(r_{b,k}, e_{b,k})^2. \tag{18}$$

The image-consistency constraint keeps the two factors tied to the original frozen image representation. With

$$c_{b,k} = \text{norm}(s_{b,k} + r_{b,k}), \tag{19}$$

the constraint is

$$\mathcal{L}_{\text{img}} = \frac{1}{|\Omega|} \sum_{(b,k)\in\Omega} \left(1 - \text{sim}(c_{b,k}, \text{sg}(g_b))\right), \tag{20}$$

where $\text{sg}(\cdot)$ denotes stop-gradient.

The full objective is

$$\mathcal{L}_{\text{LCD}} = \mathcal{L}_{\text{sem}} + \mathcal{L}_{\text{res}} + 0.5\,\mathcal{L}_{\text{neu}} + 0.1\,\mathcal{L}_{\text{img}}. \tag{21}$$

The four terms correspond to selectivity, complementarity, neutrality, and consistency. They define LCD as a feature factorization protocol rather than as a general-purpose condition-driven encoder.

4 Experiments

4.1 Experimental Setup

The experiments evaluate whether LCD separates condition-specified visual semantics from complementary same-image semantics. We use fixed phrase-region indexes from Flickr30K Entities, Visual Genome, and VGPhraseCut, with 1,000 Flickr30K images and 10,371 valid regions, 2,601 Visual Genome images and 111,130 valid regions, and 2,533 VGPhraseCut images and 13,833 valid regions. Semantic-side evaluation ranks valid same-image region records for each query phrase. Residual-side evaluation ranks fixed target-complement candidate sets, where the complement is a same-image region with low spatial and lexical overlap to the target. Dataset construction, split statistics, filtering rules, and candidate generation are detailed in Appendix A.

4.2 Semantic-Side Experiments

The semantic-side experiment evaluates the representation $s(I, t)$. For each query phrase, the model ranks same-image candidate records by cosine similarity. We report recall at rank K, denoted R@K, as the fraction of queries for which the correct candidate appears among the top K ranked candidates. R@1 measures exact top-ranked retrieval, while R@10 measures whether the correct candidate remains among a short list of plausible candidates. We also report mean reciprocal rank,

$$\text{MRR} = \frac{1}{|\mathcal{Q}|} \sum_{q\in\mathcal{Q}} \frac{1}{\text{rank}(q)}, \tag{22}$$

where $\text{rank}(q)$ is the rank assigned to the correct candidate for query q. Higher values indicate that the correct candidate appears earlier in the ranking.

Table 1 shows that LCD substantially improves condition-specified semantic ranking over the frozen encoder baselines. On Flickr30K, the LCD model trained on Flickr30K reaches 0.478 R@1 and 0.577 MRR, compared with 0.029 R@1 and 0.063 MRR for OpenAI CLIP-B/16 and 0.034

R@1 and 0.072 MRR for the strongest frozen open-source baseline in this block. On Visual Genome, the model trained on Visual Genome obtains 0.062 R@1 and 0.126 MRR, compared with 0.006 R@1 and 0.018 MRR for the strongest frozen baseline. VGPhraseCut shows a different generalization pattern: the table highlights the model trained on VGPhraseCut as the matched train-test setting, but the Flickr30K-trained LCD model also transfers strongly to this test source. This cross-dataset behavior suggests that the learned semantic factor is not merely memorizing one evaluation source, although the magnitude of transfer depends on the source distribution.

4.3 Residual-Side Evaluation

The residual-side experiment evaluates whether $r(I, t)$ carries complementary information rather than only removing the exact text direction. The main comparison includes projection proxies, $g - \mathrm{proj}_{e_t}(g)$, which remove the component of each global image embedding along the condition text direction. These proxies are related to linear projection-based concept erasure, but here they are used only as exact text-direction diagnostics (Ravfogel et al., 2020; Belrose et al., 2023). They are strong hand-crafted baselines for exact-direction removal, but they have no learned residual branch and do not adapt through the LCD objectives.

Table 2 reports the residual-side metrics used for the main comparison. Low-salience MRR is mean reciprocal rank on the bottom 30% salience subset, designed to reduce the reward for retrieving visually obvious complementary content already prominent in the global representation. Non-overlap Preference measures whether the residual representation prefers complementary semantics with low overlap to the target condition, rather than simply retaining target-like content. Let M_{low} denote Low-salience MRR and P_{non} denote Non-overlap Preference. The main residual score is Residual Anti-Projection Score:

$$\mathrm{RAPS} = \frac{1}{2}(M_{\mathrm{low}} + P_{\mathrm{non}}). \tag{23}$$

RAPS is used as the main residual-side metric because it focuses on low-salience recovery and non-overlapping complementary preference.

The LCD model trained on VGPhraseCut obtains the highest RAPS in all three test blocks. Relative to the OpenAI CLIP-B/16 projection proxy, RAPS increases from 0.077 to 0.124 on Flickr30K, from 0.071 to 0.131 on Visual Genome, and from 0.126 to 0.146 on VGPhraseCut. The same model gives the highest Low-salience MRR in all three blocks and the highest Non-overlap Preference on Visual Genome and VGPhraseCut; on Flickr30K, OpenCLIP-L/14 DataComp is slightly higher on Non-overlap Preference. The residual evidence is therefore stated through RAPS and its two reported components.

5 Analysis

This section analyzes why phrase-conditioned decomposition is needed. We first check whether the data contains condition-relative information imbalance, then examine global compatibility ambiguity and projection behavior.

5.1 Information Imbalance in Phrase-Conditioned Images

Phrase-conditioned images contain more visual semantics than a single phrase usually names. As Table 3 shows, each image contains 4.298 to 5.831 valid phrases on average, and each target condition has 1.544 to 2.058 valid same-image complements. Median target regions occupy only 0.045 to 0.077 of the image, and median complement regions occupy 0.021 to 0.085. Thus the operative phrase typically describes a localized part of the image rather than the whole visual representation.

Annotated complements are also broadly available: every retained target has at least one low-overlap same-image complement, and target and best-complement salience are close in median. This does not mean the annotations exhaust all unmentioned visual information; it only establishes comparable same-image semantics for feature-accounting analysis.

5.2 Compatibility Does Not Resolve Feature Accounting

Frozen CLIP global image-text scores are informative about whether a phrase is compatible with an image, but this compatibility relation is not the same as condition-relative feature accounting. A global score can say that an image and a phrase match, yet it does not decide which visual semantics have already been explained by the current phrase and which same-image semantics should remain as residual complements.

Table 4 summarizes the Section 5.2 evidence. In 563 pooled target-complement cases, the target phrase and best same-image complement have

	Flickr30K			VG			VGP		
	R@1 ↑	R@10 ↑	MRR ↑	R@1 ↑	R@10 ↑	MRR ↑	R@1 ↑	R@10 ↑	MRR ↑
OpenAI CLIP-B/16	0.029	0.126	0.063	0.005	0.030	0.014	0.013	0.057	0.029
OpenAI CLIP-L/14	0.029	0.132	0.066	0.005	0.031	0.015	0.011	0.054	0.028
OpenCLIP-B/16 DataComp	0.031	0.129	0.064	0.005	0.032	0.015	0.014	0.062	0.031
OpenCLIP-L/14 DataComp	0.034	0.149	0.072	0.006	0.035	0.016	0.015	0.064	0.033
FLAIR	0.032	0.140	0.069	0.006	0.038	0.018	0.015	0.068	0.034
DreamLIP	0.030	0.134	0.066	0.006	0.033	0.016	0.013	0.063	0.031
LCD CLIP B16 Trained on Flickr30K	**0.478**	**0.769**	**0.577**	0.049	0.195	0.099	0.279	0.559	0.373
LCD CLIP B16 Trained on VG	0.162	0.441	0.256	**0.062**	**0.252**	**0.126**	0.193	0.510	0.298
LCD CLIP B16 Trained on VGP	0.150	0.432	0.243	0.023	0.116	0.057	**0.239**	**0.547**	**0.340**

Table 1: Semantic-side results. Each dataset block reports R@1, R@10, and MRR; higher is better for all metrics. Bold marks the LCD model trained on the matching dataset for that block.

	Flickr30K			VG			VGP		
	LS-MRR ↑	NO-Pref. ↑	RAPS ↑	LS-MRR ↑	NO-Pref. ↑	RAPS ↑	LS-MRR ↑	NO-Pref. ↑	RAPS ↑
OpenAI CLIP-B/16 proj.	0.126	0.028	0.077	0.059	0.083	0.071	0.219	0.032	0.126
OpenAI CLIP-L/14 proj.	0.086	0.027	0.056	0.048	0.082	0.065	0.167	0.031	0.099
OpenCLIP-B/16 DataComp proj.	0.064	0.033	0.049	0.035	0.103	0.069	0.123	0.041	0.082
OpenCLIP-L/14 DataComp proj.	0.048	**0.034**	0.041	0.029	0.106	0.067	0.094	0.040	0.067
FLAIR proj.	0.017	0.010	0.014	0.016	0.066	0.041	0.036	0.018	0.027
DreamLIP proj.	0.051	0.024	0.038	0.051	0.087	0.069	0.104	0.030	0.067
LCD trained on Flickr30K	0.200	0.021	0.111	0.086	0.116	0.101	0.214	0.040	0.127
LCD trained on VG	0.133	0.010	0.072	0.094	0.107	0.101	0.168	0.040	0.104
LCD trained on VGP	**0.215**	0.033	**0.124**	**0.125**	**0.136**	**0.131**	**0.237**	**0.055**	**0.146**

Table 2: Residual-side results. Each dataset block reports LS-MRR (Low-salience MRR), NO-Pref. (Non-overlap Preference), and RAPS; higher is better for all metrics. RAPS is the average of LS-MRR and NO-Pref. Projection rows use the named frozen model as the global representation. Bold marks the block-wise best value for each metric; the VGP-trained LCD row is best on RAPS in all three test blocks.

the same rounded mean global similarity, 0.228 versus 0.228. The mean margin is effectively zero, with a bootstrap 95% confidence interval of $[-0.004, 0.003]$; top-1 compatibility is split between target and complement, and 35.2% of cases fall inside the ± 0.02 ambiguous region. These numbers do not show that CLIP fails; they show that global compatibility is under-specified for deciding the current condition boundary.

Figure 2 shows the same pattern at distribution level: target and complement scores overlap heavily, margins concentrate near zero, and many cases lie near the equal-score diagonal. Several phrases from the same image can be globally compatible at once, while only one phrase is the operative condition for the decomposition.

Salience buckets explain the ambiguity. High-salience complements often dominate global compatibility, low-salience complements are usually dominated by the target, and the medium bucket is most ambiguous. This reversal is useful for interpreting the decomposition problem: a global image embedding follows whichever phrase is most compatible with the whole image, not necessarily the phrase selected as the current condition. Thus residual analysis cannot be reduced to taking the next strongest global match; it requires a condition-relative accounting rule. We therefore use the pooled result as the main evidence and the buckets as interpretation.

5.3 Direction Removal Is Not Complement Acquisition

Projection-based methods provide a useful boundary baseline because they can remove linearly accessible directions from representations (Ravfogel et al., 2020, 2022; Belrose et al., 2023). Here the projection proxy directly removes the current text direction from the global image representation and asks whether the remaining vector becomes a representation of same-image complementary semantics.

Table 5 summarizes the Section 5.3 evidence. Across 3,912 projection rows, the proxy has mean target-direction removal of 0.168, but its mean acquisition toward the best same-image complement is -0.133. Its complement-target margin remains positive at 0.029, so the proxy is informative rather than failed; exact removal and complement acquisition are simply different operations.

Figure 3 makes this distinction visible. Points farther right have stronger target-direction suppression, while only points above zero on the vertical

Dataset	Images	Target conditions	Avg phrases/image	Avg valid complements	Median target area	Median complement area
Flickr30K	402	1,215	4.935	1.646	0.077	0.085
VG	400	1,295	5.831	1.544	0.045	0.021
VGPhraseCut	402	972	4.298	2.058	0.069	0.085

Table 3: Dataset-level motivation statistics for phrase-conditioned images. Area columns report median region area ratios. Complements are valid annotated same-image phrases after the filtering rules used for this analysis.

Group	n	Target score	Best comp. score	Margin	95% CI	Target top-1	Ambig.
Pooled	563	0.228	0.228	-0.000	[-0.004, 0.003]	51.2%	35.2%
High-salience comp.	188	0.232	0.265	-0.033	[-0.038, -0.027]	18.6%	27.1%
Medium-salience comp.	187	0.229	0.226	0.003	[-0.001, 0.008]	51.3%	46.5%
Low-salience comp.	188	0.222	0.193	0.029	[0.025, 0.033]	83.5%	31.9%

Table 4: Compatibility-accounting gap diagnostics. Scores are global CLIP image-text similarities for the target phrase and the best same-image complement. Ambig. is the fraction of cases whose target-complement margin lies within ± 0.02.

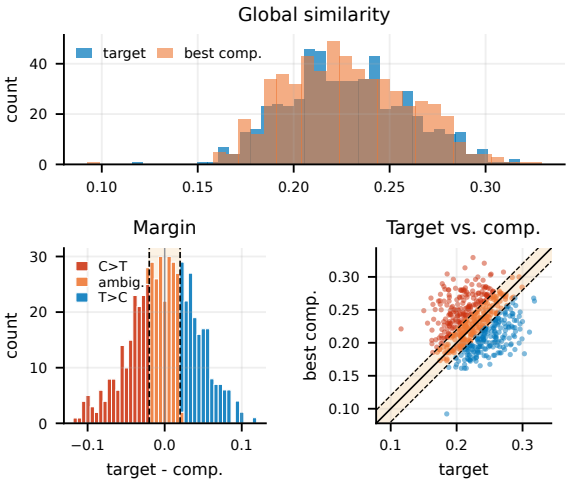

Figure 2: Global compatibility distributions for target phrases and best same-image complements. The top panel compares target and complement similarities; the lower panels show target-minus-complement margins and the ± 0.02 ambiguity band.

Subset	n	Rem.	Acq.	Mgn.	95% CI
Pooled	3,912	.168	-.133	.029	[.027,.031]
High sal.	361	.173	-.136	.050	[.044,.055]
Med. sal.	359	.166	-.134	.025	[.020,.030]
Low sal.	365	.166	-.133	.004	[-.002,.010]

Table 5: Direction removal and complement acquisition for the projection proxy. Rem. is target-direction removal; Acq. is movement toward the best same-image complement relative to the original global representation; Mgn. is the complement-target margin.

axis gain complement similarity. The plotted finite projection rows remain below zero, including high-removal samples. This "direction removed, no complement gain" pattern motivates learned semantic–residual factorization without claiming that LCD uniformly dominates the projection proxy.

6 Limitations

The study is limited by the scale and coverage of its evaluation data. The benchmark contains 135,334 valid region records, but the reported image sources are still modest: 1,000 Flickr30K images, 2,601 Visual Genome images, and 2,533 VGPhraseCut images. These datasets provide controlled phrase-region annotations, but they do not cover the full diversity of open-domain scenes, expressions, and long-tail relations.

The residual definition is also rule-specified. Area filters, IoU thresholds, lexical overlap thresholds, phrase-type checks, salience buckets, and hard-negative construction determine which phrases count as targets, complements, and distractors. These choices make the benchmark reproducible, but they also mean that different complement definitions could change the measured residual behavior, especially for nested phrases, shared head nouns, and short generic descriptions.

The metrics capture only part of the intended decomposition behavior. R@1, R@10, and MRR measure same-image phrase ranking, while RAPS focuses on low-salience recovery and non-overlapping complementary preference. These metrics are appropriate for the controlled questions in this paper, but they do not establish downstream usefulness or an exhaustive measure of residual quality. Broader datasets, less rule-dependent complement construction, and downstream evaluations are needed before treating semantic–residual factorization as a general-purpose vision-language repre-

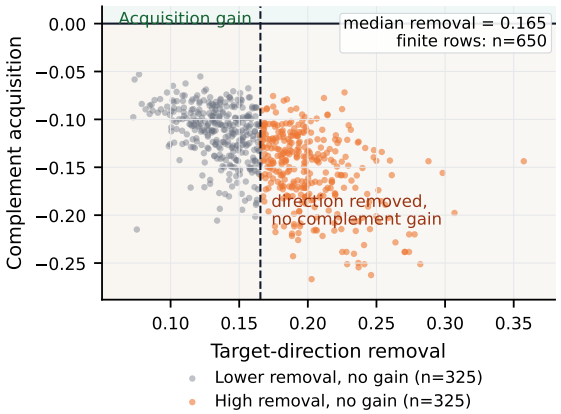

Figure 3: Removal versus complement acquisition for the projection proxy. The horizontal line marks zero acquisition gain, and the dashed vertical line marks median target-direction removal. Orange points have high removal but no complement-acquisition gain; gray points have lower removal and no gain.

sentation improvement.

7 Conclusion

This paper introduced LCD as a controlled study of semantic–residual feature factorization in frozen vision-language representations. Rather than treating the operative text as the main contribution, LCD uses it to specify which visual semantics should be explained away, then learns an explained semantic factor and a complementary residual factor. The results suggest that frozen vision-language representations can be partially reorganized into condition-specified semantic and residual factors. The evidence is strongest for semantic-side selectivity and more limited for residual semantics, where LCD improves targeted low-salience and non-overlap metrics over projection-based residuals. This positions LCD as a controlled first step toward residual semantic modeling, with broader claims left to larger datasets, less rule-specified complements, and downstream evaluations.

References

Antonio Almudévar, José Miguel Hernández-Lobato, Sameer Khurana, Ricard Marxer, and Alfonso Ortega. 2025. Aligning multimodal representations through an information bottleneck. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 1250–1270. PMLR.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. 2023. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems*.

Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. 2020. Learning robust representations via multi-view information bottleneck. In *International Conference on Learning Representations*.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*.

Wei-Ning Hsu and James Glass. 2018. Disentangling by partitioning: A representation learning framework for multimodal sensory data. *arXiv preprint arXiv:1805.11264*.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. ViLT: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5583–5594. PMLR.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73.

Samuel Lavoie, Polina Kirichenko, Mark Ibrahim, Mahmoud Assran, Andrew Gordon Wilson, Aaron Courville, and Nicolas Ballas. 2024. Modeling caption diversity in contrastive vision-language pretraining. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven

C. H. Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems*.

Paul Pu Liang, Zihao Deng, Martin Q. Ma, James Zou, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2023. Factorized contrastive learning: Going beyond multi-view redundancy. In *Advances in Neural Information Processing Systems*.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. 2018. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2641–2649.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256. Association for Computational Linguistics.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan Cotterell. 2022. Linear adversarial concept erasure. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 18400–18421. PMLR.

Simon Schrodi, David T. Hoffmann, Max Argus, Volker Fischer, and Thomas Brox. 2025. Two effects, one trigger: On the modality gap, object bias, and information imbalance in contrastive vision-language models. In *International Conference on Learning Representations*.

Shun Shao, Yftah Ziser, and Shay B. Cohen. 2023. Erasure of unaligned attributes from neural representations. *Transactions of the Association for Computational Linguistics*, 11:488–510.

Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Learning factorized multimodal representations. In *International Conference on Learning Representations*.

Haoqing Wang, Xun Guo, Zhi-Hong Deng, and Yan Lu. 2022. Rethinking minimal sufficient representation in contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16020–16029.

Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. 2020. PhraseCut: Language-based image segmentation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10216–10225.

Mike Wu and Noah Goodman. 2018. Multimodal generative models for scalable weakly-supervised learning. In *Advances in Neural Information Processing Systems*, volume 31, pages 5580–5590.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. 2022. LiT: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

A Dataset Construction Details

This appendix records the deterministic construction of the training and test data used in the experiments. The training data uses image-level text pools only. Bounding boxes, crops, masks, polygons, and IoU values are not exported as training supervision. The test data uses region annotations to build fixed evaluation records, crops, and candidate sets.

A.1 Training Splits

Training examples are built from Flickr30K Entities, Visual Genome, and VGPhraseCut-derived processed records with seed 42. Each source record is converted into an image-level pool of valid text conditions. Text is lowercased, stripped, whitespace-normalized, and filtered when empty, reserved, numeric, punctuation-only, or shorter than two valid characters. For each target text, the training index stores same-image positives and cross-image negatives. Same-image positive pools are available for 99.94% of training targets, with mean pool size 36.91; cross-image negative pools contain 256 negatives per target with no recorded same-image violations.

The resulting training index contains 4,601,499 LCD training samples and 238,271 validation samples. Flickr30K and VGP use explicit validation

Source	Train Img.	Train Text	Val Img.	Val Text
Flickr30K	28,984	274,975	1,013	9,728
VG	100,257	4,029,135	5,219	209,871
VGP	71,566	300,120	2,869	18,752

Table A.1: Training and validation split sizes after constructing image-level text pools.

splits. Visual Genome uses the processed training split that excludes VGPhraseCut test images, then reserves 5% for validation. This prevents the reported VGPhraseCut test images from appearing in Visual Genome training.

A.2 Test Records

The test benchmark is built from held-out region-phrase records. Each raw annotation is converted into an image id, image path, image size, phrase text, and box when available. Boxes are converted to x_1, y_1, x_2, y_2, clamped to image boundaries, and removed when the resulting area is non-positive. The crop bank contains 135,334 valid regions and has no missing crop files in the recorded benchmark report.

Source	Test Img.	Raw Rec.	Valid Reg.
Flickr30K	1,000	10,375	10,371
VG	2,601	111,130	111,130
VGP	2,533	13,833	13,833

Table A.2: Held-out test records used for semantic-side and residual-side evaluation.

For semantic-side evaluation, each query phrase is paired with candidate region records from the same image. The correct phrase-region record is ranked among these same-image candidates by cosine similarity. This construction makes the task more local than global image-text retrieval because all candidates share the same image context.

A.3 Residual Candidate Sets

Residual-side evaluation starts from target-complement pairs. A target must be a valid non-scene region with phrase and box quality passing the benchmark filters, area ratio between 0.001 and 0.5, and an existing crop. A complement is a same-image region that is not a duplicate of the target and has low overlap with it. The construction uses IoU below 0.2 and text Jaccard similarity below 0.6, then ranks candidates by a deterministic complement score.

Each residual candidate set contains the target, one selected complement, same-image hard negatives, and cross-image hard negatives. Same-image hard negatives are ranked using spatial overlap, lexical overlap, and phrase-type compatibility. Cross-image hard negatives are drawn from other images using phrase similarity, phrase type, and box-size compatibility. The sampled pair pools contain 43,846 pairs for Flickr30K, 165,972 for Visual Genome, and 36,880 for VGPhraseCut; the high-shared subsets contain 4,678, 54,244, and 3,020 pairs, respectively.