FROM VISUAL FACTS TO SEMANTIC CONCEPTS: STRUCTURING FROZEN VISION–LANGUAGE REPRESENTATIONS WITH EVIDENCE–CONCEPT DECOMPOSITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Frozen vision–language models provide transferable visual features that encode both directly observable cues and higher-level semantic information. We organize one frozen representation into shared, evidence, and concept pathways while keeping the pretrained backbone fixed. Raw visual fact descriptions are mapped to typed evidence tags covering entities, attributes, actions, relations, compositions, and scenes, while abstract concept labels are normalized separately. The model uses specialized residual branches, a concept-conditioned evidence gate, prototype-based predictors, and evidence-aware fusion. On a Flickr30k-based fact–concept benchmark, the full model reaches competitive balanced mAP (19.68 versus 19.29 for a strong shared MLP) while providing separate pathways that can be removed, perturbed, swapped, and probed. A fact bottleneck retains strong fact performance but drops to 7.41 concept mAP. Source removal, gate perturbation, evidence-type analysis, and probing show distinct roles for shared, evidence, and concept information, with the clearest gate effect appearing in the evidence-to-concept route. The structured pathways retain competitive prediction while making evidence and concept use directly measurable.

1 INTRODUCTION

Large-scale vision–language pretraining has made frozen image encoders useful across retrieval, recognition, and multimodal reasoning tasks (Radford et al., 2021; Jia et al., 2021). A common downstream design uses one pretrained visual embedding with a task-specific prediction head. The same embedding may support directly observable cues—such as entities, attributes, actions, and relations—and more abstract interpretations—such as activity type, interaction pattern, or scene-level meaning. A structured interface can make these semantic roles explicit without changing the pretrained encoder.

We ask how a frozen visual feature can support separate evidence and concept representations while still feeding a joint predictor. Compositional vision–language benchmarks continue to expose errors in attribute binding and relational reasoning (Thrush et al., 2022; Zhao et al., 2022; Yuksekgonul et al., 2023a; Ma et al., 2023). Concept-based models also show the value of explicit intermediate variables when their roles are clearly defined and measured (Koh et al., 2020; Yuksekgonul et al., 2023b; Yang et al., 2023).

At the supervision level, we replace a flat list of visual facts with a typed evidence vocabulary. Raw fact descriptions are deterministically converted into evidence tags for entities, attributes, actions, relations, subject–action–object or subject–relation–object compositions, and scene cues. Abstract concept annotations are normalized separately. Section 3 specifies the construction rules, thresholds, and vocabulary sizes used in the experiments.

At the representation level, we keep the CLIP image encoder frozen and learn a shared pathway together with evidence- and concept-oriented residual branches. Concept context modulates the fact-side representation through an evidence gate, and a fusion module combines shared, evidence,

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

and concept information for joint prediction. The gate controls one evidence-selection route, while the concept branch and shared pathway remain direct inputs to the fused predictor.

The evaluation separates prediction from mechanism analysis. Matched frozen-feature baselines compare the proposed decomposition with linear, shared-MLP, prototype, adapter-style, and fact-bottleneck alternatives under the same cached CLIP features and label space. Source removal, gate perturbation, evidence-type analysis, concept swapping, and probing then measure the roles of the learned pathways.

Our contributions are:

- A **structured evidence supervision interface** that maps free-form visual facts into six typed evidence categories while keeping abstract concepts separate;
- An **evidence–concept decomposition** for frozen CLIP features with shared, evidence-oriented, and concept-oriented pathways, concept-conditioned evidence selection, prototype prediction, and fusion; and
- A **matched frozen-feature evaluation with mechanism analysis** that compares predictive trade-offs and measures the functional roles of the shared, evidence, and concept pathways, including the evidence-to-concept gate route.

2 Related Work

Frozen VLM adaptation and compositional semantics. CLIP and related models provide transferable cross-modal representations (Radford et al., 2021; Jia et al., 2021; Li et al., 2022; 2023; Liu et al., 2023). Prompt tuning and adapters adapt frozen models with a small number of trainable parameters (Zhou et al., 2022; Jia et al., 2022; Gao et al., 2024). Compositional diagnostics reveal persistent difficulty with attributes, relations, and structured reasoning (Thrush et al., 2022; Zhao et al., 2022; Yuksekgonul et al., 2023a; Ma et al., 2023; Wang et al., 2023), motivating hard-negative, grounding, and region-aware approaches (Yao et al., 2022; Li et al., 2021; Zhang et al., 2021; Dou et al., 2022; Kamath et al., 2021). Our method operates at the downstream representation level: the pretrained CLIP backbone stays fixed, and trainable modules organize its output into evidence-, concept-, and shared pathways. The experiments therefore compare alternative heads that use the same frozen feature cache.

Concept-based and modular representations. Concept bottlenecks and post-hoc concept interfaces connect predictions to interpretable variables (Koh et al., 2020; Yuksekgonul et al., 2023b; Yang et al., 2023). Modular latent models and scene-graph methods organize information into structured components (Higgins et al., 2017; Chen et al., 2016; Xu et al., 2017; Krishna et al., 2017). Our design keeps evidence and concept supervision separate while allowing both to reach the final predictor. The fact-bottleneck baseline in Sec. 5.2 provides a direct comparison with a stricter fact-mediated design.

Representation probing. Linear and nonlinear probes are widely used to measure what information can be recovered from learned representations (Alain & Bengio, 2017; Conneau et al., 2018; Goh et al., 2021). We use probing as a diagnostic of accessibility across the shared, evidence, concept, and fused representations, alongside direct predictive baselines and intervention-based analyses.

3 Problem Setup and Structured Supervision

3.1 Task Definition and Label Spaces

For each image x, the benchmark provides raw visual fact descriptions $\mathcal{T}(x)$ and abstract concept annotations $\mathcal{C}(x)$. These fields describe two different semantic levels: the fact field records directly observable content, while the concept field summarizes higher-level meanings. We therefore keep the two supervision sources separate rather than collapsing them into one vocabulary. Deterministic preprocessing maps the fact field to a retained evidence set $E(x)$ and the concept field to a normalized concept set $C(x)$. The corresponding multi-label targets are

$$y_f = \mathrm{MultiHot}(E(x)), \qquad y_c = \mathrm{MultiHot}(C(x)). \tag{1}$$

The subscript f denotes evidence derived from the raw fact annotations, whereas c denotes abstract concept labels. The annotation file contains 31,731 Flickr30k image records; before image-path filtering, 30,128 records contain at least one retained evidence tag and one retained concept.

3.2 Concept Normalization

Concept strings are lowercased, punctuation is removed, hyphens are split, and whitespace is normalized. A fixed canonical dictionary merges selected lexical variants and near-synonyms, while concepts not covered by the dictionary retain their normalized surface form. We retain concepts occurring at least five times, with a vocabulary cap of 3,000; the resulting concept vocabulary contains 2,487 labels. The same normalization map is used for all experiments.

3.3 Structured Evidence Extraction

Raw fact descriptions are converted into typed visual evidence after text normalization, stop-word removal, lexical normalization, and local phrase matching. For image x, the retained evidence set is

$$\begin{aligned} E(x) = E_{\text{ent}}(x) \cup E_{\text{attr}}(x) \cup E_{\text{act}}(x) \\ \cup E_{\text{rel}}(x) \cup E_{\text{comp}}(x) \cup E_{\text{scene}}(x), \end{aligned} \tag{2}$$

covering entities, attributes, actions, local relations, subject–action/relation–object compositions, and coarse scene cues. Representative tags include `attr blue shirt`, `actobj holding phone`, `rel behind person`, `comp person holding phone`, and `scene outdoor`. The extractor uses small phrase lists, token lexicons, and deterministic local rules rather than an external semantic parser. Duplicate and generic tags are removed, and tags occurring fewer than eight times are discarded. The final evidence vocabulary contains 603 labels. Appendix A gives raw examples and the full construction sequence.

4 Method

4.1 Overview

A frozen vision–language embedding can simultaneously encode directly observable cues and higher-level semantics, but a single downstream representation does not make the roles of these two levels explicit. Our goal is therefore not simply to add capacity to a frozen CLIP feature. Instead, we organize the downstream computation so that evidence-oriented and concept-oriented information can be separated, interact in a controlled way, and still contribute jointly to the final prediction.

Given an image x, the frozen CLIP image encoder produces $h = \phi(x) \in \mathbb{R}^{d_{\text{CLIP}}}$. A shared pathway first extracts z_s, which preserves information useful to both tasks. Two specialized residual branches then produce a fact-oriented precursor z_f and a concept representation z_c. The distinction is intentional: z_f provides candidate evidence features, whereas z_c represents higher-level semantic context. Concept information is subsequently used to modulate the fact pathway, producing a refined evidence representation z_e rather than replacing the fact representation with a hard concept bottleneck.

Finally, the shared representation z_s, refined evidence representation z_e, and concept representation z_c are fused into z_u for joint evidence and concept prediction. This organization exposes four representations—shared, evidence, concept, and fused—that can be removed, swapped, perturbed, or probed. The resulting interventions are central to our evaluation because they test whether the decomposition has functional consequences beyond adding parameters.

The architecture deliberately keeps parallel routes instead of imposing a strict evidence bottleneck. A hard bottleneck would make concept prediction depend entirely on the quality and coverage of the extracted evidence vocabulary, even though the frozen CLIP feature may contain useful semantic information that is not represented by the retained evidence tags. Retaining z_s and z_c gives the predictor access to this information, while the gated $z_f \to z_e$ route provides an explicit place to study how concept context changes evidence selection. In this sense, the decomposition is designed as an analyzable interface around a frozen representation rather than as a claim that all abstract semantics must be mediated by the extracted evidence labels. Figure 1 summarizes the information flow.

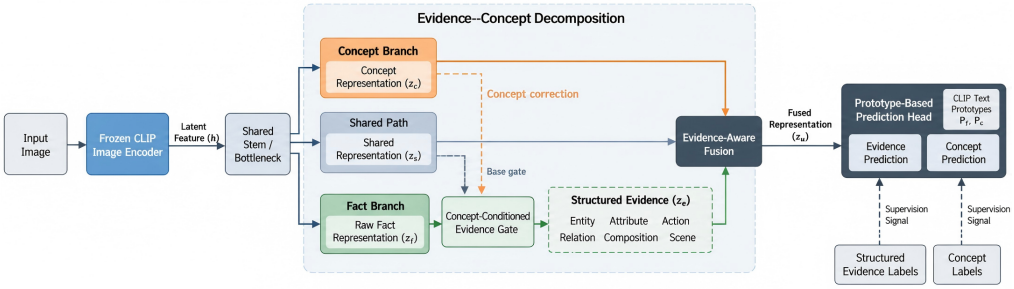

Figure 1: Overview of the evidence–concept decomposition. A frozen CLIP feature feeds a shared pathway and two specialized residual branches. Concept context modulates the fact-oriented branch to form a refined evidence representation, after which shared, evidence, and concept information are fused for prediction.

4.2 Shared and Specialized Residual Pathways

Let $d_b = 768$ denote the branch dimension and let $\mathrm{norm}(v) = v/\|v\|_2$. A trainable shared transformation F_s maps the frozen CLIP feature to

$$z_s = \mathrm{norm}(F_s(h)) \,. \tag{3}$$

The fact and concept branches then apply independent gated residual updates. For $t \in \{f, c\}$,

$$\Delta_t = B_t(z_s) \odot q_t(z_s), \tag{4}$$

$$z_t = \mathrm{norm}(a_t z_s + \Delta_t) \,, \tag{5}$$

where B_t is a branch-specific residual transformation, q_t is a channel gate, and a_t is a learned residual coefficient. The two branches share the same input but have independent parameters, yielding z_f and z_c. This design allows specialization while retaining a direct shared component inherited from the same frozen feature. Exact layer widths, residual blocks, and initialization details are listed in Appendix B.

4.3 Concept-Conditioned Evidence Selection

The fact branch contains candidate evidence, but not every channel needs to be equally useful for every semantic context. We therefore let the concept representation provide a correction to a base evidence gate computed from the fact and shared paths:

$$b_{\mathrm{base}} = G_{\mathrm{base}}([z_f; z_s]), \qquad d_c = G_\Delta(z_c), \tag{6}$$

$$m = \sigma(b_{\mathrm{base}} + \alpha d_c + b_0) \,. \tag{7}$$

Here $m \in (0,1)^{d_b}$ is the concept-conditioned channel mask and α controls the contribution of the concept correction. The refined evidence representation is

$$z_e = \mathrm{norm}(G_e(z_f \odot (0.10 + 0.90m)) + 0.15 z_f) \,, \tag{8}$$

where G_e is the evidence-refinement module. The offset and residual term keep information available even when individual gate values are small. Thus the concept route modulates evidence selection, but does not become the only path through which evidence or concepts can reach the final predictor.

4.4 Evidence-Aware Fusion and Prototype Prediction

We combine the three semantic sources only after the specialized representations have been formed. Let $z_{\mathrm{cat}} = [z_e; z_c; z_s]$. A second channel-wise gate produces a fusion residual,

$$q_u = G_{\mathrm{gate}}(z_{\mathrm{cat}}), \tag{9}$$

$$\Delta_u = G_u(z_{\mathrm{cat}} \odot (0.05 + 0.90 q_u)) \,, \tag{10}$$

$$z_u = \mathrm{norm}(\beta_s z_s + \beta_e z_e + \beta_c z_c + 0.25 \Delta_u) \,. \tag{11}$$

The direct coefficients $\beta_s, \beta_e, \beta_c$ are learned, so the fused representation can preserve source-specific information while also using the learned interaction term Δ_u.

Evidence and concept outputs are predicted with separate CLIP text-prototype heads. For task $t \in \{f, c\}$ and retained label $\ell_{t,k}$,

$$\begin{aligned} p_{t,k} &= \mathrm{norm}(\phi_{\mathrm{text}}(\texttt{``a photo of } \ell_{t,k}\texttt{''})), \\ H_t(v)_k &= \tau^{-1} v^\top p_{t,k} + b_{t,k}, \end{aligned} \tag{12}$$

where the text prototypes are frozen, τ is learnable, and $b_{t,k}$ is a class bias. The final logits retain direct branch readouts in addition to the fused prediction:

$$\ell_f^u = H_f(z_u) + r_f H_f(z_e), \tag{13}$$

$$\ell_c^u = H_c(z_u) + r_c H_c(z_c) + 0.15 H_c(z_e). \tag{14}$$

The term $H_c(z_e)$ is also evaluated independently as the evidence-to-concept diagnostic used later in the paper.

4.5 Training Objectives

Base training uses asymmetric multi-label loss for fused evidence and concept prediction together with branch-level supervision. Denote the corresponding mean losses by $\mathcal{L}_f^u$, $\mathcal{L}_c^u$, $\mathcal{L}_f^e$, $\mathcal{L}_c^c$, and $\mathcal{L}_c^{e\to c}$. The base objective is

$$\mathcal{L}_{\mathrm{base}} = \mathcal{L}_f^u + 1.2\mathcal{L}_c^u + 0.35\mathcal{L}_c^{e\to c} + 0.3\big(\mathcal{L}_f^e + \mathcal{L}_c^c\big). \tag{15}$$

After base training, a short refinement stage reinitializes the concept-conditioned gate and updates gate-related modules. It adds anchor, concept-swap, group, and selectivity regularizers to the task loss,

$$\mathcal{L}_{\mathrm{refine}} = \mathcal{L}_{\mathrm{task}} + 0.05\mathcal{L}_{\mathrm{anchor}} + 0.05\mathcal{L}_{\mathrm{swap}} + 0.02\mathcal{L}_{\mathrm{group}} + 0.02\mathcal{L}_{\mathrm{sel}}. \tag{16}$$

The exact asymmetric loss, four regularizers, and the two-stage optimization protocol are given in Appendix B.4. Keeping these implementation-level definitions in the appendix avoids interrupting the main representation flow while preserving full reproducibility.

5 Experiments

5.1 Experimental Setup

Data and evaluation protocol. We use the Flickr30k image collection together with the paired fact–concept annotation file and the label construction from Sec. 3. All trainable comparison methods use the same cached CLIP features, data partitions, evidence vocabulary, and concept vocabulary. The matched-head comparison reports three training seeds (42, 43, and 44), while zero-shot CLIP is evaluated once. F1 thresholds are selected on validation data and then applied unchanged to the test set.

Backbone and optimization. The visual backbone is CLIP ViT-L/14 and remains frozen throughout training. The branch dimension is 768, the stem hidden dimension is 1,024, and dropout is 0.22. Base training uses AdamW with learning rate 10^{-4} and weight decay 10^{-4}. Gate refinement uses learning rate 5×10^{-4}, with an optional lower-rate fusion stage beginning at refinement epoch 3. Full architecture and optimization settings are provided in Appendix B.

Metrics. For matched test comparisons we report evidence/fact mAP, concept mAP, balanced mAP (the mean of the two), and validation-thresholded micro-F1. Trained methods report mean and sample standard deviation across three seeds. A paired image bootstrap with 2,000 resamples measures test-sample variation for the seed-42 models; the resulting intervals are reported in Table 2. Mechanism analyses use evidence and concept micro-F1 together with evidence-to-concept micro-F1 and $\mathrm{FusionScore} = \frac{1}{2}(\mathrm{F1}_f^{\mathrm{micro}} + \mathrm{F1}_c^{\mathrm{micro}})$.

Table 1: Held-out test performance (%). Trained methods report mean ± sample standard deviation over seeds 42, 43, and 44.

Method	Fact mAP	Concept mAP	Bal. mAP	Fact F_1	Concept F_1
CLIP zero-shot	7.73	9.82	8.77	5.67	5.06
Linear	7.67 ± 0.01	4.09 ± 0.03	5.88 ± 0.01	47.39 ± 0.08	29.59 ± 0.05
Shared MLP	17.40 ± 0.22	21.19 ± 0.22	19.29 ± 0.13	54.02 ± 0.23	38.86 ± 0.19
Flat prototype	16.91 ± 0.44	16.16 ± 0.15	16.53 ± 0.22	53.43 ± 0.30	34.21 ± 0.18
CLIP-Adapter-ML	10.72 ± 0.31	11.00 ± 0.24	10.86 ± 0.28	40.48 ± 0.45	10.96 ± 0.71
Fact bottleneck	$\mathbf{17.82 \pm 0.32}$	7.41 ± 0.14	12.61 ± 0.23	$\mathbf{54.90 \pm 0.07}$	31.36 ± 0.05
SED (ours)	17.71 ± 0.28	$\mathbf{21.42 \pm 0.27}$	$\mathbf{19.68 \pm 0.23}$	54.54 ± 0.22	$\mathbf{39.10 \pm 0.30}$

Table 2: Paired image-bootstrap differences in test micro-F1 (ours minus baseline) for seed 42. Brackets give 95% intervals from 2,000 resamples.

Baseline	Fact difference [95% CI]	Concept difference [95% CI]
Linear	+6.64 [+5.92, +7.36]	+8.02 [+7.30, +8.75]
Shared MLP	+0.05 [-0.47, +0.61]	-1.08 [-1.70, -0.46]
Flat prototype	+0.79 [+0.26, +1.34]	+3.46 [+2.85, +4.07]
CLIP-Adapter-ML	+13.77 [+12.97, +14.61]	+27.10 [+26.20, +27.96]
Fact bottleneck	-0.91 [-1.43, -0.36]	+6.18 [+5.48, +6.85]

5.2 Predictive Comparison with Matched Frozen-Feature Heads

We compare alternative heads under the same frozen CLIP feature cache and label space: a linear head, a shared nonlinear MLP, a flat prototype predictor, a CLIP-Adapter-style multi-label head (Gao et al., 2024), a fact-bottleneck variant, zero-shot CLIP, and the full SED model. These baselines vary the amount and organization of downstream capacity while holding the pretrained visual representation fixed.

SED and the shared MLP are close on the main task metrics. SED is higher by 0.39 points in balanced mAP, 0.52 points in fact F1, and 0.24 points in concept F1. The task-level gap is therefore small; the main difference is that SED exposes separate representations that can be intervened on and probed. The fact bottleneck illustrates a different trade-off: it attains the highest fact mAP/F1 but its concept mAP drops to 7.41, compared with 21.42 for SED.

The seed-42 paired bootstrap gives a complementary view of test-sample variation. Table 2 reports paired differences in micro-F1 for the same images. The intervals are clearly positive relative to the linear, flat-prototype, and adapter-style baselines. Against the shared MLP, the fact difference is essentially zero while the concept difference is negative for seed 42; this differs from the three-seed mean in Table 1, where SED has a slightly higher concept F1. We therefore treat the two models as close in predictive performance rather than using the bootstrap to claim a uniform advantage. The fact-bottleneck comparison shows the opposite trade-off: it is stronger on fact prediction for this seed but substantially weaker on concepts.

5.3 Functional Roles of Shared, Evidence, and Concept Pathways

We next intervene on the refined seed-42 checkpoint at inference time. "No evidence," "No concept," and "No shared" remove one semantic source from the integration pathway. "Fixed gate" replaces the learned evidence gate by 0.5, while "Shuffled gate" assigns gate vectors from different samples.

Removing evidence lowers FusionScore from 0.4240 to 0.3681 and reduces evidence-to-concept micro-F1 from 0.0992 to 0.0157. Removing concept information slightly raises evidence micro-F1 but nearly eliminates concept prediction, while removing the shared path mainly harms evidence prediction. The different failure patterns indicate that the three sources do not act as interchangeable copies of the same representation.

Table 3: Key source-removal and gate-perturbation results. Full macro-F1 and gate statistics are reported in Appendix C.1.

Setting	Evidence Micro-F1	Concept Micro-F1	Evidence→Concept Micro-F1	FusionScore
Full	0.5043	0.3437	0.0992	0.4240
No evidence	0.4509	0.2853	0.0157	0.3681
No concept	0.5189	0.0125	0.0850	0.2657
No shared	0.3347	0.3473	0.0850	0.3410
Fixed gate	0.5103	0.3322	0.0621	0.4213
Shuffled gate	0.5093	0.3323	0.0564	0.4208

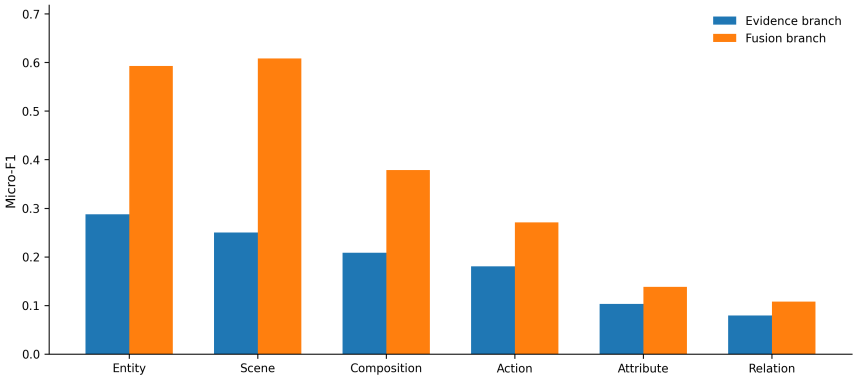

Figure 2: Micro-F1 across the six evidence types. Fusion improves all categories, while attribute and relation remain the most difficult.

Evidence types. The 603 evidence labels are grouped into six semantic types. Figure 2 compares readout from the evidence and fusion representations. Fusion improves micro-F1 in all six categories. The largest absolute gains occur for scene and entity labels, while composition improves from 0.2087 to 0.3788. Attribute and relation remain the most difficult categories. Full macro-F1 values, label counts, and positive counts are reported in Appendix C.2.

5.4 Effect of Concept-Conditioned Gating

The fixed- and shuffled-gate rows in Table 3 isolate the evidence gate without removing an entire semantic source. Both interventions leave the final FusionScore close to the full model (0.4213 and 0.4208 versus 0.4240), but their larger effect appears on the evidence-to-concept route: micro-F1 falls from 0.0992 to 0.0621 and 0.0564. The learned gate therefore has its clearest measured effect on how evidence participates in concept prediction rather than on the final fused score alone.

We further replace z_c with the concept representation of another sample while keeping the image-side representation fixed. Table 4 summarizes the resulting changes. Concept substitution changes the gate substantially more than the downstream evidence and fusion representations: the mean gate distance is 0.1704, compared with 0.0048 for z_e and 0.0124 for z_u. The final top-1 concept changes for 96.19% of the swaps. This pattern is consistent with the residual design: concept context can strongly alter channel selection while the ungated and shared routes preserve most of the image-side representation. Figure 3 shows one representative case; the full statistics, including top-concept score changes, and additional examples are reported in Appendices C.3 and D.1.

5.5 Representation Accessibility

We freeze z_s, z_e, z_c, and z_u and train either a linear probe or a two-layer MLP probe for evidence and concept prediction. The linear probe measures direct linear accessibility, while the MLP probe

Table 4: Concept-swap diagnostics with the image-side representation fixed.

Metric	Value
Gate distance	0.1704 ± 0.0616
Evidence distance	0.0048 ± 0.0019
Fusion distance	0.0124 ± 0.0020
Top-1 concept change	96.19%

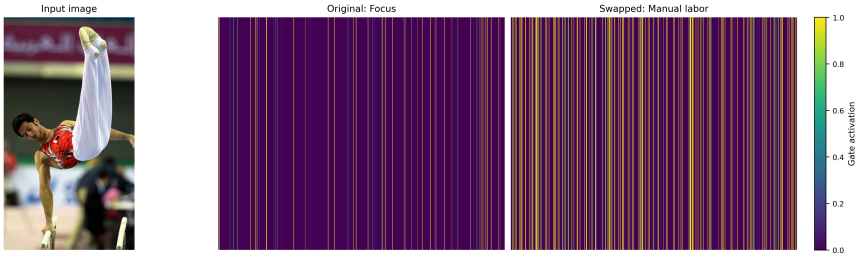

Figure 3: Representative concept-swap example. The image-side representation is fixed while z_c is replaced (*Focus→Manual labor*). The concept-conditioned gate changes under the substituted semantic context.

measures recoverability with a more flexible readout. We report the two probe families separately because they use different capacities and metrics.

Evidence is more linearly accessible from z_e than from z_s (0.4530 versus 0.4107), whereas the shared representation gives slightly higher evidence mAP with the nonlinear probe. For concept prediction, z_c has low linear F1 but reaches 0.1465 mAP with the MLP probe, the highest value among the three source representations. The fused representation z_u is highest on all four probe columns. These differences support the view that the learned pathways expose different readout structure rather than forming a strict information bottleneck.

5.6 Summary of Findings

The predictive comparisons and mechanism analyses answer complementary questions. At the task level, the decomposition remains close to the strongest shared-head baseline rather than obtaining its main benefit from a large accuracy increase. At the representation level, however, the learned sources respond differently to removal, the concept-conditioned gate is most visible in the evidence-to-concept route, and concept substitution changes the gate while the image-side representation is held fixed. Evidence-type results and probing provide an additional view of these differences across label groups and readout capacities. Together, these results support the narrower claim that the proposed structure makes evidence and concept use measurable while retaining competitive prediction; they do not require the source representations to be perfectly disentangled.

6 Scope and Limitations

Our analysis concerns structured access to information in a frozen visual representation rather than complete semantic disentanglement. The evidence vocabulary is deterministically derived from the available fact descriptions, so its coverage and granularity depend on this annotation interface. Evidence-to-concept prediction should therefore be interpreted as predictive association under the constructed label space, not as evidence that abstract concepts are causally mediated by the extracted evidence. Likewise, removal, swapping, and probing reveal functional differences among z_s, z_e, and z_c, but do not imply that these representations contain mutually exclusive information. All pathways originate from the same frozen CLIP feature, and the probing results themselves indicate partial overlap. Finally, the present experiments use CLIP ViT-L/14 on a Flickr30k-based fact–concept benchmark; testing other frozen vision–language backbones and annotation schemes remains an important direction for future work.

Table 5: Probing different learned representations. Linear probes report micro-F1; MLP probes report mAP.

Representation	Visual Evidence		Abstract Concept	
	Linear F1	MLP mAP	Linear F1	MLP mAP
Shared (z_s)	0.4107	0.4491	0.1702	0.1231
Evidence (z_e)	0.4530	0.4323	0.1564	0.0758
Concept (z_c)	0.1372	0.4038	0.0141	0.1465
Fusion (z_u)	0.4871	0.5060	0.2024	0.1627

7 Conclusion

We presented an evidence–concept interface for organizing frozen vision–language representations. The method separates a shared representation into evidence- and concept-oriented pathways, uses concept context to modulate evidence selection, and reunifies the resulting sources through gated fusion. Under matched frozen-feature comparisons, its predictive performance is close to a strong shared MLP, while the structured representation exposes semantic sources that can be directly removed, swapped, perturbed, and probed.

The mechanism analyses clarify what this added structure contributes. Source removal produces different failure patterns, concept-conditioned gate perturbations most strongly affect the evidence-to-concept route, evidence-type analysis shows consistent gains after fusion, and representation probes reveal different accessibility patterns across z_s, z_e, z_c, and z_u. The fact-bottleneck comparison further indicates that forcing concept prediction through facts is substantially more restrictive than retaining parallel evidence, concept, and shared routes. These results motivate structured downstream organization as a practical interface for studying how a frozen visual backbone supports observable evidence and higher-level semantic concepts.

Reproducibility Statement

The main paper states the task definition, core model equations, evaluation protocol, and principal comparisons. Appendix A gives full label-construction details; Appendices B–B.4 specify the architecture, prediction heads, and training losses; Appendix C contains the extended quantitative diagnostics; and Appendix D contains qualitative gate analyses. A complete release includes the normalization maps, retained vocabularies, dataset partitions, frozen-feature cache specification, baseline configurations, base checkpoint used to initialize gate refinement, and the training and evaluation scripts.

AI Use Statement

Generative AI tools were used to polish manuscript wording. The authors verified the final text, figures, experimental results, and technical claims.

References

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations (ICLR)*, 2017.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, 2016.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 2126–2136, 2018.

Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Yuanzhi Lu, Fayao Wu, et al. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18166–18176, 2022.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. CLIP-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision (IJCV)*, 132(2):581–595, 2024.

Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. Multimodal neurons in artificial neural networks. *Distill*, 6(3):e30, 2021.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. Beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations (ICLR)*, 2017.

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 (ICML)*, pp. 4904–4916. PMLR, 2021.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision (ECCV)*, pp. 709–727. Springer, 2022.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. MDETR-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1780–1790, 2021.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning (ICML)*, pp. 5338–5348. PMLR, 2020.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision (IJCV)*, 123(1):32–73, 2017.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 9694–9705, 2021.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning (ICML)*, pp. 12888–12900. PMLR, 2022.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 34892–34916, 2023.

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

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 (ICML)*, pp. 8748–8763. PMLR, 2021.

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 (CVPR)*, pp. 5238–5248, 2022.

Tan Wang, Kevin Lin, Linjie Li, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Equivariant similarity for vision-language foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11998–12008, 2023.

Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5410–5419, 2017.

Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19187–19197, 2023.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. FILIP: Fine-grained interactive language-image pre-training. In *International Conference on Learning Representations (ICLR)*, 2022.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations (ICLR)*, 2023a.

Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. In *International Conference on Learning Representations (ICLR)*, 2023b.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. VinVL: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5579–5588, 2021.

Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaoshuai Lu, and Jian Yin. VL-CheckList: Evaluating pre-trained vision-language models with objects, attributes and relations. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4611–4619, 2022.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 130(9):2337–2348, 2022.