TRANSBIND: Explainable Compositional Grounding in LVLMs via Agentic Counterfactual Diagnostics

Anonymous ACL submission

Abstract

Large vision–language models (LVLMs) produce fluent but often unfaithful outputs—correct objects with wrong attributes, reversed relations, or mis-scoped negations. We frame these as *compositional grounding failures* and present TRANSBIND, a framework making such errors testable via *counterfactual minimal pairs* that change one semantic factor while holding others fixed. We introduce two metrics: SEC (selective edit consistency) and AFC (attribution factor coherence), plus a lightweight structured binding module. Experiments on four LVLMs show consistent improvements.

1 Introduction

LVLMs built on vision–language pretraining (Radford et al., 2021; **?**; Li et al., 2023a) and instruction tuning (Liu et al., 2023a; Dai et al., 2023a) achieve strong performance but suffer from hallucinations (Rohrbach et al., 2018; Li et al., 2023b; Dai et al., 2023b). Many errors are *component-level*: the object is correct but the color is wrong; entities are correct but relations reversed; negations are ignored. We view these as *binding failures*—the model fails to bind semantic roles (attributes, relations, operators) to correct visual evidence.

This motivates testing *stability under controlled interventions*: when a minimal pair changes one factor, a grounded model should update only the corresponding output, and explanations should shift to relevant evidence.

Contributions. We propose TRANSBINDwith: (1) DIAGTRANS, an agentic diagnostic suite generating counterfactual minimal pairs targeting attributes, relations, quantities, and negation; (2) SEC/AFC metrics for output consistency and explanation coherence; (3) a structured binding module using object-centric slots (Locatello et al., 2020) and factorized cross-attention.

Factor	Pattern and expected behavior
Attribute	Flip mug color; only color token should change; attribution stays on mug.
Relation	Flip *on*→*under*; truth value flips; referents stable.
Negation	Change *two*→*three* or add *not*; only operator changes.

Table 1: Minimal-pair patterns: selective stability under single-factor interventions.

2 Failure Taxonomy and Minimal Pairs

Based on prior work (Rohrbach et al., 2018; Li et al., 2023b; Dai et al., 2023b; Thrush et al., 2022), we organize grounding failures into: (1) **entity failure**—mentioning unsupported objects; (2) **attribute binding failure**—correct object, wrong attribute; (3) **relational binding failure**—correct entities, wrong relation; (4) **quantifier failure**—wrong counts; (5) **logical operator failure**—ignored negation. Object-level metrics (CHAIR, POPE) capture (1); DIAGTRANStargets (2)–(5) via minimal pairs.

Table 1 illustrates minimal-pair patterns: each intervention changes one factor while expecting selective output change.

3 Related Work

Hallucination and compositionality. CHAIR (Rohrbach et al., 2018) and POPE (Li et al., 2023b) quantify object hallucination. Compositional benchmarks (Winoground (Thrush et al., 2022), VALSE (Parcalabescu et al., 2022), NLVR2 (Suhr et al., 2019), CLEVR (Johnson et al., 2017)) probe systematic grounding but lack counterfactual structure.

Explainability. Attribution methods (Sundararajan et al., 2017; Selvaraju et al., 2016) visualize evidence, but attention may not explain (Jain and Wallace, 2019). ERASER (DeYoung et al., 2020)

Pipeline: Seed → Agent proposes T_f → Instantiate pair → Verify → Evaluate with SEC/AFC

Figure 1: TRANSBINDpipeline overview.

highlights the gap between plausible rationales and faithful evidence. TRANSBINDevaluates explanations via counterfactual tests.

4 Problem Setup

Let $p_\theta(y \mid I, x)$ be an LVLM where I is an image and x is a question/instruction. We view output y as a sequence of semantic components (object mentions, attributes, relations, operators).

Counterfactual minimal pairs. An intervention T_f changes factor f while keeping others fixed:

$$(I, x) \text{ and } T_f(I, x) = (I', x').$$

In practice, I' can be an edited image, a paired image from a dataset, or the same image with a counterfactual prompt.

Selective stability principle. A grounded model should satisfy: (1) **targeted update**—f-related output changes appropriately; (2) **invariance elsewhere**—unrelated components remain stable; (3) **evidence shift**—attribution shifts to f-relevant regions.

5 DIAGTRANS: Counterfactual Diagnostics

DIAGTRANStargets four factor families: attributes (color/material), relations (spatial), quantification, and negation. Data sources include COCO (Lin et al., 2014), Visual Genome (Krishna et al., 2017), GQA (Hudson and Manning, 2019), and NLVR2 (Suhr et al., 2019).

We support two modes: **paired-image** (mining images differing in one factor) and **paired-prompt** (flipping operators while keeping the image fixed). An LLM agent proposes candidates following templates; verification filters non-minimal pairs using dataset annotations, lightweight detectors (Liu et al., 2023b; Kirillov et al., 2023), and human audit. Each pair carries metadata: target factor, confound checklist, verification evidence, and minimality score.

6 Metrics

We introduce two metrics for counterfactual evaluation over minimal pairs.

6.1 SEC: Selective Edit Consistency

Given outputs y, y', we compute edit alignment via Levenshtein backtrace. Let $\Delta(y, y')$ be edited tokens and S_f be factor-f tokens (extracted via lexicon + pattern rules). We define:

$$\text{SEC} = \frac{|\Delta(y, y') \cap S_f|}{|\Delta(y, y')| + \epsilon},$$

where ϵ avoids division by zero. High SECindicates the model changes only what should change. We canonicalize factor mentions (e.g., "navy"→"blue") before extraction.

6.2 AFC: Attribution Factor Coherence

For attribution method A (e.g., integrated gradients, Grad-CAM), let $A(I, x)$ produce an attribution map. AFCmeasures whether attribution shifts to f-relevant regions R_f:

$$\text{AFC} = \text{sim}(A_f, R_f) - \lambda\, \text{sim}(A_{\neg f}, R_f),$$

where A_f and $A_{\neg f}$ are attributions for factor-related and unrelated tokens. This makes explanation evaluation testable rather than anecdotal.

7 Structured Binding Module

Diagnostics are useful only if they inform improvements. We describe a modular inductive bias for explicit role–filler binding.

7.1 Motivation

Classical compositional views treat meaning as role–filler binding (Smolensky, 1990). In LVLMs, the challenge is binding linguistic roles ("color of the mug") to correct visual referents. Generic cross-attention can represent this implicitly but does not guarantee stable binding under interventions.

7.2 Design

We extract object-centric slots $S = \{s_i\}$ via Slot Attention (Locatello et al., 2020), then apply *factorized cross-attention*: object, attribute, and relation queries attend separately to slots before composition (Algorithm 1). This reduces the chance that frequent language patterns override weak visual signals.

7.3 Training Signals

DIAGTRANScan provide weak supervision via two optional objectives: (1) **selective invariance**—penalize changes in $H_{\neg f}$ under T_f; (2) **factor-aligned change**—encourage controlled change in H_f via contrastive margin.

Algorithm 1 Factorized Binding Fusion

Require: Image features V, text tokens X
Slots $S \leftarrow \text{SlotAttention}(V)$
$Q_{\text{obj}}, Q_{\text{attr}}, Q_{\text{rel}} \leftarrow \text{Proj}(X)$
$H_i \leftarrow \text{CrossAttn}(Q_i, S)$ for $i \in \{\text{obj}, \text{attr}, \text{rel}\}$
return $\text{Compose}(H_{\text{obj}}, H_{\text{attr}}, H_{\text{rel}})$

Family	Pass%	κ	Notes
Attribute	88	.78	color cues
Relation	81	.71	viewpoint
Quantifier	84	.74	counting
Negation	79	.69	scope

Table 2: Human audit (n=600).

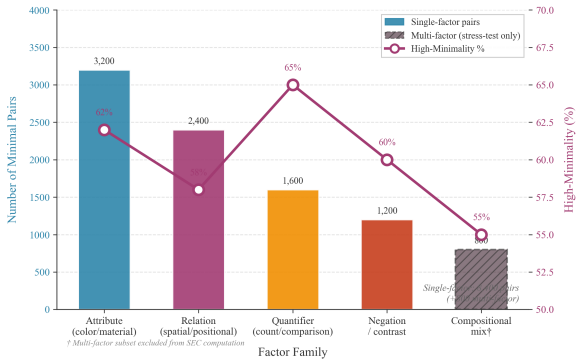

Figure 2: DIAGTRANS-ACL composition by factor family.

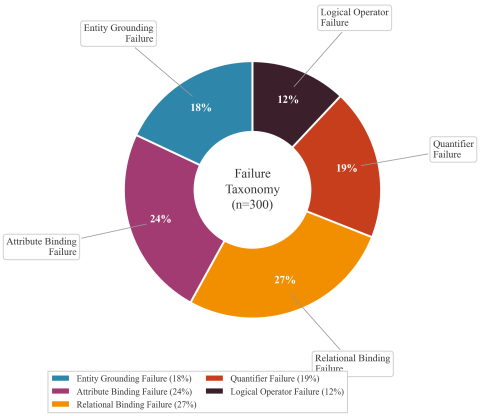

Figure 3: Error distribution on 300 audited pairs.

8 Experiments

8.1 Benchmark Instantiation

We instantiate DIAGTRANS-ACL from COCO (Lin et al., 2014), GQA (Hudson and Manning, 2019), NLVR2 (Suhr et al., 2019), VALSE (Parcalabescu et al., 2022), and Winoground (Thrush et al., 2022). The suite contains 8,400 single-factor pairs. Figure 2 shows the composition. Human audit on 600 pairs (Table 2) achieves 76–88% minimality pass rate with Cohen's κ 0.67–0.78.

8.2 Models

We evaluate four representative open LVLMs:

- **BLIP-2** (Li et al., 2023a): frozen vision encoder + Q-Former.
- **InstructBLIP** (Dai et al., 2023a): instruction-tuned BLIP-2.
- **LLaVA** (Liu et al., 2023a): visual instruction tuning.
- **MiniGPT-4** (Zhu et al., 2023): frozen encoder + LLM alignment.

8.3 Methods Compared

We compare: (1) **Base**—original LVLM; (2) **+TRANSBIND**—add factorized binding fusion; (3) **+TRANSBIND+cf-reg**—additionally apply counterfactual regularization.

8.4 Main Results

Table 3 shows TRANSBINDimproves SECand AFCacross all models and factors, with largest gains on relations and negation.

8.5 Ablations and Analysis

Table 4 shows rankings are stable across minimality levels. Table 5 confirms both slots and factorization contribute. SECcorrelates with POPE (ρ=−.62) and CHAIR (ρ=−.54), suggesting compositional grounding improvements reduce hallucinations. Figure 3 shows binding failures dominate over entity hallucination.

8.6 Implementation Details

We use K=7 slots with 3 refinement iterations. Factor queries are projected via a 2-layer MLP (GELU, hidden 512). Counterfactual regularization fine-tuning uses learning rate 1e-5, batch size 64, and 1 epoch on ∼27k augmented instances.

Limitations

First, minimality is hard to guarantee in the wild: real images correlate attributes and relations, and editing images at scale is non-trivial. We mitigate this via a high-minimality subset and metadata-driven analysis. Second, SECdepends on factor token identification, which can be imperfect for

Model	SEC ↑				AFC ↑	
	Attr.	Rel.	Quant.	Neg.	IG	GradCAM
BLIP-2	.612	.438	.524	.387	.281	.253
+ TRANSBIND	.658	.527	.571	.476	.342	.308
+ TRANSBIND+ cf-reg	**.687**	**.564**	**.598**	**.518**	**.371**	**.339**
InstructBLIP	.641	.472	.553	.421	.304	.271
+ TRANSBIND	.683	.551	.592	.503	.358	.327
+ TRANSBIND+ cf-reg	**.709**	**.589**	**.621**	**.541**	**.392**	**.358**
LLaVA	.623	.451	.508	.406	.268	.241
+ TRANSBIND	.671	.538	.561	.489	.331	.297
+ TRANSBIND+ cf-reg	**.698**	**.576**	**.593**	**.527**	**.364**	**.331**
MiniGPT-4	.598	.426	.497	.378	.259	.232
+ TRANSBIND	.647	.516	.548	.467	.318	.286
+ TRANSBIND+ cf-reg	**.678**	**.557**	**.581**	**.509**	**.352**	**.321**

Table 3: Main results on DIAGTRANS-ACL (8,400 pairs).

Method	SEC (full)	SEC (high-min.)
Base	.508	.561
+ TRANSBIND	.559	.608
+ TRANSBIND+ cf-reg	**.591**	**.642**

Table 4: Robustness: full vs. high-minimality subset.

Variant	SEC	AFC
Base	.522	.288
+ slots only	.563	.324
+ factorization only	.571	.331
+ TRANSBIND(both)	**.592**	**.358**

Table 5: Ablation on InstructBLIP.

paraphrased outputs; we use canonicalization but edge cases remain. Third, AFCrequires region proxies (annotations or detectors), which may inject their own biases. These limitations motivate transparent reporting and ablation of verification strength.

Ethical Considerations

More faithful LVLMs can improve safety in assistive and decision-support settings, but diagnostic tools can also be misused to create adversarial prompts. We recommend releasing diagnostics with clear intended use and avoiding sensitive images or personal data. When human verification is used, annotators should be fairly compensated and protected from harmful content exposure.

9 Conclusion

We presented TRANSBIND, a framework for making compositional grounding in LVLMs measurable and explainable through agentic counterfactual diagnostics. By linking minimal interventions to output consistency (SEC) and explanation coherence (AFC), TRANSBINDprovides an evaluation loop that can guide architectural improvements. Experiments show consistent gains across four LVLM families.

Acknowledgments

We thank the authors of the public datasets and benchmarks used in this work.

References

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023a. Instructblip: Towards general-purpose vision-language models with instruction tuning. *CoRR*, abs/2305.06500.

Wenliang Dai, Zihan Liu, Ziwei Ji, Dan Su, and Pascale Fung. 2023b. Plausible may not be faithful: Probing object hallucination in vision-language pre-training. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2136–2148, Dubrovnik, Croatia. Association for Computational Linguistics.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4443–4458. Association for Computational Linguistics.

Drew A. Hudson and Christopher D. Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20,*

2019, pages 6700–6709. Computer Vision Foundation / IEEE.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 3543–3556. Association for Computational Linguistics.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. 2017. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1988–1997. IEEE Computer Society.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. 2023. Segment anything. *CoRR*, abs/2304.02643.

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. *Int. J. Comput. Vis.*, 123(1):32–73.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023a. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. *CoRR*, abs/2301.12597.

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

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *CoRR*, abs/2304.08485.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. 2023b. Grounding DINO: marrying DINO with grounded pre-training for open-set object detection. *CoRR*, abs/2303.05499.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. 2020. Object-centric learning with slot attention. *CoRR*, abs/2006.15055.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2022. 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, Dublin, Ireland. Association for Computational Linguistics.

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. *CoRR*, abs/2103.00020.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, Brussels, Belgium. Association for Computational Linguistics.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2016. Grad-cam: Visual explanations from deep networks via gradient-based localization. *CoRR*, abs/1610.02391.

Paul Smolensky. 1990. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artif. Intell.*, 46(1-2):159–216.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 5228–5238. IEEE.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *CoRR*, abs/2304.10592.