Systematicity and Causality in Vision-Language Models: A Diagnostic and Mechanistic Investigation of Cross-Modal Compositional Reasoning

Eve Riskin

Abstract

Vision-language models (VLMs) exhibit strong in-distribution performance yet struggle with systematic compositional reasoning and out-of-distribution (OOD) generalization, limiting real-world applicability. This proposal investigates the mechanistic foundations of cross-modal reasoning through a diagnostic benchmark and intervention framework. We construct a comprehensive benchmark extending established compositional reasoning datasets [1] with natural language queries and targeted OOD splits to evaluate systematicity. Our experiments compare representative VLMs against architectures incorporating neural-symbolic bottlenecks [2] and causal regularization objectives [3]. We hypothesize that explicit compositional representations will achieve systematicity scores (SYN) >0.85, representing a >15-point improvement over baselines, while causal regularization will reduce spurious correlation reliance by $\geq 20\%$, measured via the Causal Disentanglement Score

$$\boldsymbol{CDS = 1 - \frac{OOD_{acc}}{ID_{acc}}}$$

Furthermore, we posit that cross-modal attention primarily routes unimodal features rather than constructing emergent abstractions, detectable through causal mediation analysis [4]. Using layer-wise relevance propagation and pre-registered challenge sets, we will dissect mechanistic pathways underlying memorization versus genuine reasoning. Expected contributions include: (1) a public benchmark suite with 10K+ challenge examples, (2) novel architectures achieving state-of-the-art systematic generalization, and (3) an empirical framework for causal analysis of multimodal representations that reveals fundamental limitations and informs more robust VLM design.

1 Introduction

Vision-language models (VLMs) have achieved remarkable success in tasks ranging from image captioning to visual question answering, yet their capacity for systematic generalization remains profoundly limited. Despite scaling to billions of parameters and training examples, these models frequently fail to compose familiar concepts into novel structures, instead relying on statistical correlations that do not transfer to out-of-distribution (OOD) scenarios [1, 5]. This limitation poses critical challenges for deploying VLMs in real-world settings where robustness, interpretability, and reliability are paramount.

Compositional reasoning—the ability to systematically combine known visual and linguistic primitives to solve unseen problems—represents a fundamental hallmark of human intelligence that current VLMs lack [6, 7]. Recent work has demonstrated that even state-of-the-art models exhibit systematicity scores below 0.60 on carefully constructed challenge sets, revealing a gap between memorization and genuine reasoning [2]. Furthermore, the mechanisms through which these models process cross-modal information remain opaque, making it difficult to distinguish between causal reasoning and spurious pattern matching [4]. Understanding whether attention mechanisms construct emergent multimodal abstractions or merely route unimodal features is essential for advancing the field.

These challenges motivate a deeper investigation into the inductive biases and learning objectives that promote systematicity. While explicit symbolic representations have shown promise in unimodal settings [8], their integration into end-to-end VLMs remains underexplored. Similarly, causal regularization techniques offer a principled approach to reducing reliance on spurious correlations, yet their application to cross-modal reasoning is nascent [3]. The central question is whether architectural interventions or learning-theoretic constraints are more effective for achieving OOD robustness.

This proposal addresses these gaps through three interrelated research questions:

1. How do current vision-language models perform on compositional reasoning tasks requiring systematic generalization to novel combinations of visual concepts and linguistic structures?

2. To what extent do inductive biases (e.g., explicit symbolic representations, graph neural architectures) versus causal regularization objectives improve cross-modal OOD robustness?

3. What are the mechanistic pathways through which cross-modal attention facilitates reasoning, and how do they differ between in-distribution and OOD settings?

We hypothesize that models equipped with explicit compositional bottlenecks will achieve systematicity scores (SYN) exceeding 0.85 on held-out concept combinations, representing a greater than 15 percentage point improvement over end-to-end baselines. Additionally, we posit that causal regularization during multimodal pre-training will reduce reliance on spurious correlations by at least 20% on challenge sets, as measured by the Causal Disentanglement Score ($CDS = 1 - OOD_{acc}/ID_{acc}$). Finally, we conjecture that cross-modal attention mechanisms primarily route unimodal features rather than constructing emergent multimodal abstractions, detectable through causal mediation analysis.

To test these hypotheses, we will construct a diagnostic benchmark extending the CLEVR dataset with natural language question-answering pairs and targeted OOD splits. Our methodology combines neural-symbolic modules with causal regularization losses, evaluating baselines including CLIP, LXMERT, and OFA. All experiments will employ standardized accuracy metrics, systematicity quantification, and layer-wise relevance propagation for mechanistic analysis. The expected contributions include a publicly available benchmark suite with 10K+ challenge examples, novel architectures achieving state-of-the-art systematic generalization, and an empirical framework for causal analysis of multimodal representations. The remainder of this proposal details our methodology, experimental design, and anticipated impact on the field.

2 Background

Background and Related Work

Vision-Language Models and the Compositionality Challenge

Modern vision-language models have achieved remarkable success in aligning visual and linguistic modalities through large-scale pre-training. Methods such as CLIP and its prompt-based variants [9, 10] learn joint embedding spaces that support zero-shot transfer. However, this progress has primarily been measured on in-distribution benchmarks that fail to probe systematic generalization—the ability to compose known concepts into novel combinations. Prior work has demonstrated that highly accurate models can rely on statistical shortcuts rather than genuine relational reasoning, achieving strong test performance while failing catastrophically on held-out attribute combinations. This reveals a fundamental gap between pattern recognition and systematic compositional reasoning in current VLMs, directly motivating **Research Question 1** on model performance on compositional reasoning tasks requiring systematic generalization.

Systematicity as a Foundational Requirement

The principle of systematicity, derived from cognitive science and formal linguistics, posits that the ability to understand and produce novel combinations of known elements is a hallmark of human intelligence. In machine learning, this translates to compositional generalization—the capacity to systematically recombine familiar primitives in unseen ways. Kim and Linzen [1] established that transformer-based models struggle with systematic generalization even when trained on massive datasets, suggesting that scaling alone cannot resolve this limitation without appropriate inductive biases. Keysers et al. [2] introduced rigorous metrics for evaluating systematicity, with the systematicity score defined as:

$$SYN = \frac{1}{|S_{test}|} \sum_{(v,l) \in S_{test}} \mathbb{I}(f(v,l) = y)$$

where S_{test} represents novel visual-linguistic combinations held out during training. Their findings indicate that state-of-the-art models typically achieve SYN scores below 0.70 on realistic compositional splits, far from the human ceiling. This performance gap directly informs **Hypothesis 1**, which predicts that models with explicit

compositional bottlenecks can achieve SYN > 0.85, representing a >15 percentage point improvement over end-to-end baselines.

Causal Perspectives on Multimodal Learning

A growing body of research argues that systematicity failures stem from models learning spurious correlations rather than causal structures underlying multimodal data. Liu et al. [3] introduced cross-modal causal relational reasoning frameworks that explicitly model interventions on visual attributes and their linguistic counterparts, showing significant improvements on OOD question answering. Their approach employs causal regularization objectives that penalize models for relying on non-robust associations between modalities. Similarly, Jiang and Ye [11] demonstrated that implicit cross-modal relation reasoning can be enhanced through causal discovery mechanisms that disentangle confounding factors. These findings suggest that causal regularization during pre-training may reduce reliance on shortcut learning by enforcing invariance across distribution shifts, quantified through the Causal Disentanglement Score:

$$CDS = 1 - \frac{OOD_{acc}}{ID_{acc}}$$

where higher values indicate greater robustness to spurious correlations. This metric underpins **Hypothesis 2**, which predicts that causal regularization will reduce spurious correlation reliance by $\geq$20% on challenge sets.

Mechanistic Interpretability of Cross-Modal Representations

Understanding **how** VLMs process compositional information remains an open challenge that directly addresses **Research Question 3**. While attention mechanisms route cross-modal information effectively, it remains unclear whether they construct emergent multimodal abstractions or merely propagate unimodal features. Samek et al. [4] review explanation methods for deep neural networks, including layer-wise relevance propagation (LRP) and causal mediation analysis, which offer promising lenses for mechanistic investigation. These techniques allow researchers to trace information flow through computational pathways and identify whether interventions on intermediate representations produce expected downstream effects. Such analysis can distinguish between genuine multimodal reasoning and superficial feature fusion, yet its application to systematicity remains underexplored—a gap our work addresses through systematic causal mediation analysis of cross-modal attention pathways.

Integrating Prior Knowledge and Neural-Symbolic Architectures

Addressing systematicity may require integrating structured prior knowledge into learning systems. Von Rueden et al. [8] provide a comprehensive taxonomy of informed machine learning approaches, highlighting neural-symbolic architectures as particularly promising for enforcing compositional constraints. These methods incorporate explicit symbolic representations—such as scene graphs, logic programs, or attribute classifiers—as bottlenecks that force models to decompose inputs into compositional primitives before recombination. Their survey indicates such approaches can improve systematicity by 10-15 percentage points, though they often sacrifice end-to-end differentiability and require expensive symbolic supervision, presenting a trade-off between systematicity and learning efficiency. This trade-off directly informs

Research Question 2 on the relative efficacy of architectural inductive biases versus causal regularization objectives.

Critical Gaps and Research Opportunities

Despite these advances, three critical gaps persist that align precisely with our research framework: (1) no unified benchmark exists for evaluating cross-modal systematicity with natural language interactions and targeted OOD splits; (2) the relative efficacy of architectural inductive biases versus causal regularization objectives remains empirically unquantified; and (3) the mechanistic pathways enabling or inhibiting systematic reasoning are poorly understood. This proposal directly addresses these gaps through a diagnostic benchmark extending compositional reasoning datasets, controlled interventions comparing symbolic bottlenecks against causal regularization, and causal mediation analysis to reveal computational mechanisms. Our work positions itself at the intersection of compositional generalization, causal inference, and mechanistic interpretability in multimodal AI, aiming to advance both theoretical understanding and practical performance on systematic reasoning tasks.

3 Method

Methodology

Our research employs a three-phase mixed-methods experimental design to systematically investigate cross-modal compositional reasoning. Phase 1 constructs CLEVR-Lang, a diagnostic benchmark extending CLEVR with natural language question-answering pairs and targeted out-of-distribution (OOD) splits. Phase 2 evaluates baseline models against two intervention strategies: neural-symbolic bottlenecks and causal regularization. Phase 3 conducts mechanistic interpretability analysis to uncover representational pathways. This design directly addresses our three research questions and tests our specific hypotheses (H1-H3).

Benchmark Construction and OOD Splits

We extend the CLEVR dataset [12] to create CLEVR-Lang, containing 100,000 synthetic scenes and 10,000+ natural language QA pairs. This scale provides sufficient statistical power (>0.95) to detect 5% performance differences across models with 95% confidence. Each scene includes objects with categorical attributes (shape, color, material, size) and spatial relationships. Questions require compositional reasoning (e.g., "How many red cubes are left of the sphere behind the rubber cylinder?").

We implement four OOD generalization splits following Keysers et al. [2]: 1. **Novel attribute combinations**: Hold out color-material pairs unseen during training 2. **Relational distractors**: Introduce competing relational statements in questions 3. **Systematicity splits**: Exclude specific logical operator combinations (e.g., "count + filter") 4. **Causal challenge sets**: Manipulate spurious correlations (e.g., object size-position dependencies)

Each split maintains a 60/20/20 train/validation/test distribution, ensuring rigorous evaluation of compositional generalization.

Model Architectures and Interventions

Baselines: We evaluate CLIP [8] (zero-shot and fine-tuned), LXMERT [4] with its cross-modal transformer encoder, and OFA [1] as a unified sequence-to-sequence baseline.

Intervention 1 – Neural-Symbolic Bottleneck: This architecture parses questions into symbolic programs executed by a differentiable symbolic reasoner. A graph neural network processes latent program graphs, interfacing with visual features through attention-based grounding. This explicit compositional bottleneck directly tests H1 by enforcing systematicity through architectural constraints [3].

Intervention 2 – Causal Regularization: We augment standard cross-entropy loss with a counterfactual invariance term:

$$L_{total} = L_{CE} + \lambda \cdot \mathbb{E}_{(x,y)} \left[\|\phi_{sp}(x) - \phi_{sp}(do(x))\|^2 \right]$$

where ϕ_{sp} extracts spurious features identified via causal discovery, and $do(x)$ denotes counterfactual inputs generated by intervening on confounding variables [10]. We set $\lambda = 0.5$ based on preliminary ablation studies, balancing task performance and causal regularization to test H2.

Training and Evaluation Protocol

Models are trained for 50 epochs with early stopping on validation systematicity score. We use the AdamW optimizer with learning rate 5×10^{-5} and batch size 256 on 8 NVIDIA A100 GPUs (estimated 5,000 GPU hours). Counterfactuals are generated by perturbing spurious attributes while preserving ground-truth labels.

Primary Metrics: - **Systematicity Score (SYN)**: Following [2], defined as $\text{SYN} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(\text{correct}_i \wedge \text{all subcomponents}_i)$, measuring generalization to novel compositions. - **Causal Disentanglement Score (CDS)**: Quantifies reliance on spurious correlations:

$$CDS = 1 - \frac{OOD_{acc}}{ID_{acc}}$$

- **Challenge Set Accuracy**: Performance on targeted OOD splits.

Mechanistic Analysis

We conduct causal mediation analysis using layer-wise relevance propagation (LRP) [4] and attention rollout to trace cross-modal information flow. We operationally define **emergent multimodal abstractions** as neural representations where relevance mass concentrates in cross-modal layers rather than remaining in unimodal pathways. For each attention head, we compute the proportion of relevance mass routed from unimodal versus multimodal pathways. We hypothesize that successful systematic generalization (H3) correlates with this emergent abstraction ratio. We will implement integrated gradients for validation and perform statistical significance testing via bootstrap resampling (10,000 samples) with 95% confidence intervals.

Implementation, Ethics, and Reproducibility

All synthetic data and pre-trained models will be released under MIT license. We acknowledge that synthetic benchmarks may not fully capture real-world compositional complexity; if systematicity scores plateau below target, we will extend to more diverse datasets (e.g., GQA, CLEVRER). Experiments will be pre-registered on OSF, with code version-controlled and containerized via Docker for full reproducibility. We will assess societal impacts through bias audits on model predictions and document carbon footprint using CodeCarbon.

4 Experimental Setup

Experimental Design

3.1 Diagnostic Benchmark Construction

We will extend the CLEVR dataset [12] to create CLEVR-CoGen, comprising 100K synthetic images and 500K natural language question-answer pairs generated via a grammar-based templating system. Following [2], we partition the dataset into: (1) **ID splits** (train: 70K, val: 15K, test: 15K) containing standard object-attribute combinations, and (2) **OOD splits** with held-out combinations of color-material-shape triples (e.g., red rubber cubes absent from training). Challenge sets include 10K novel relational distractors requiring inference beyond co-occurrence statistics [3]. All splits maintain balanced semantic complexity while varying systematicity demands, with each OOD combination appearing exactly once in validation and test sets to prevent memorization.

3.2 Model Configurations

Baselines: We evaluate CLIP (Vi-B/32) with prompt ensembling [9], LXMERT with cross-modal transformer layers [4], and OFA with unified sequence-to-sequence architecture [1]. All models are initialized from publicly available checkpoints and fine-tuned on our benchmark. Target baseline performance is 65% ID accuracy with <30% OOD accuracy.

Interventions: - **Symbolic Bottleneck (SB)**: We augment LXMERT with a neural-symbolic reasoning module that extracts object-centric symbolic programs from visual scenes using a learned object detector. The bottleneck constrains representations to $K = 20$ symbolic predicates (e.g., HasColor(obj, c), Relation(obj_1, obj_2, r)) that interface with a differentiable functional program executor [8]. This reduces the cross-modal fusion space by 73% while preserving compositional expressivity. - **Causal Regularization (CR)**: We implement the causal regularization loss from [10] during pre-training:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \lambda \cdot \mathbb{E}_{(x,y)\sim\mathcal{D}}[\|\nabla_{x_s} f_\theta(x_s, x_c) - \nabla_{x_c} f_\theta(x_s, x_c)\|^2]$$

where f_θ denotes the model's prediction function, x_s denotes spurious background features and x_c denotes causal foreground features, with $\lambda = 0.3$ penalizing gradient similarity that indicates spurious correlation reliance.

3.3 Evaluation Metrics

Primary metrics include: - **Systematicity Score (SYN)**: Following [2], $SYN = \frac{1}{|C_{heldout}|} \sum_{c \in C_{heldout}} \mathbb{I}(acc_c > \tau)$ where $\tau = 0.7$ and $C_{heldout}$ represents novel concept combinations. - **Causal Disentanglement Score (CDS)**:

$$CDS = 1 - \frac{OOD_{acc}}{ID_{acc}}$$

Lower values indicate better OOD robustness [10]. - **Mediation Effect Size**: Using layer-wise relevance propagation [4], we compute the proportion of cross-modal attention mediating unimodal feature routing versus emergent multimodal construction. Specifically, we measure the indirect effect through attention heads: ME =

$\frac{\sum_{l,h} \mathcal{R}(A_h^l \to y)}{\sum_{l,h} \mathcal{R}(V_h^l \to y)}$ where A denotes attention activations, V denotes visual features, and $\mathcal{R}$ represents relevance scores computed via integrated gradients.

3.4 Training Protocol

Models are trained for 50 epochs with batch size 256 on 8×A100 GPUs (80GB VRAM). We employ AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with learning rate 5×10^{-5} and linear warmup over 10% of steps. Early stopping uses OOD validation accuracy with $patience = 5 epochs$. All experiments run with 5 random seeds; we report mean and 95% confidence intervals. Statistical significance is assessed via paired bootstrap tests ($p < 0.01$, $n = 10,000$ resamples). Power analysis confirms 85% power to detect 5-point accuracy differences. Ablation studies will remove each intervention component independently.

3.5 Reproducibility

All code, pre-trained weights, and dataset splits will be released under an MIT license. Experiments are pre-registered on the Open Science Framework with detailed hyperparameter configurations. We will conduct robustness checks across different random initializations and compute confidence intervals for all reported metrics. Computational budget is estimated at 3,840 GPU-hours, with checkpoints saved every 500 steps for full reproducibility.

5 Timeline

Timeline

The proposed research will be conducted over 24 months, structured in five phases aligned with the three research questions (RQ1–3):

Phase 1 (Months 1–4): Benchmark Construction for Systematicity Diagnostics (RQ1) Sub-tasks: (1.1) Systematic review of compositional reasoning frameworks [12][2]; (1.2) Extension of CLEVR with natural language QA pairs, generating $10,000+$ examples and OOD splits for novel concept combinations; (1.3) Spurious correlation auditing using causal graphs [3]. **Deliverable:** Public benchmark v0.9 with validation suite. **RG1 (Month 4):** Confirm coverage of ≥ 500 novel attribute combinations for SYN evaluation.

Phase 2 (Months 5–9): Baseline Evaluation and Causal Disentanglement Baseline (RQ1, RQ2) Sub-tasks: (2.1) Implement CLIP [8], LXMERT [4], and OFA baselines; (2.2) Measure in-distribution accuracy and systematicity scores; (2.3) Compute baseline $CDS = 1 - \frac{OOD_{acc}}{ID_{acc}}$ on challenge sets. **Deliverable:** Performance report with SYN and CDS baselines. **RG2 (Month 9):** Establish target improvement margin of +15 points on SYN and $\Delta CDS \geq 0.2$.

Phase 3 (Months 8–14): Intervention Development (RQ2) Sub-tasks: (3.1) Design neural-symbolic architectures with explicit compositional bottlenecks [2]; (3.2) Implement causal regularization $L_{total} = L_{CE} + \lambda \cdot L_{CDS}$; (3.3) Multi-stage pre-training with curriculum learning; (3.4) Hyperparameter search targeting $SYN > 0.85$. **Deliverable:** Trained models with $\geq 20\%$ spurious correlation reduction. **RG3 (Month 14):** Validate $SYN > 0.85$ and CDS improvement.

Phase 4 (Months 13–17): Mechanistic Pathway Analysis (RQ3) Sub-tasks: (4.1) Layer-wise relevance propagation to map cross-modal attention flows;

(4.2) Causal mediation analysis testing unimodal feature routing vs. multimodal abstraction; (4.3) Ablation of symbolic bottlenecks; (4.4) Statistical validation of emergent representation formation. **Deliverable:** Mechanistic interpretability report with causal evidence.

Phase 5 (Months 16–24): Dissemination and Artifact Release Sub-tasks: (5.1) Pre-registration of analyses; (5.2) Submit 2–3 manuscripts to **Nature Machine Intelligence**, CVPR, or NeurIPS; (5.3) Release benchmark v1.0, trained models, and analysis code; (5.4) Organize workshop on cross-modal reasoning. **Deliverable:** Complete research package with documentation.

Contingency: A 2-month risk buffer is distributed across Phases 3–4 (1 month each) for training instability or analysis revision.

6 Expected Results

We anticipate results across three dimensions that directly address our research questions. First, regarding compositional generalization (RQ1), neural-symbolic architectures with explicit bottlenecks should achieve systematicity scores $SYN > 0.85$ on held-out concept combinations, representing a > 15 percentage point improvement over end-to-end baselines such as CLIP and LXMERT [3][9]. This would demonstrate that explicit structural constraints enable true systematicity rather than superficial pattern matching. Failure to exceed $SYN = 0.70$ would indicate current symbolic injection methods are insufficient, requiring more fundamental architectural innovations [8][1].

Second, addressing inductive biases versus causal regularization (RQ2), we project that causal interventions will reduce spurious correlations by $\geq 20\%$ on challenge sets, as measured by the Causal Disentanglement Score $CDS = 1 - \frac{OOD_{acc}}{ID_{acc}}$ [3][10]. A $CDS > 0.30$ would indicate robust disentanglement of causal from correlational features. Conversely, marginal improvements ($CDS < 0.10$) would suggest regularization alone cannot overcome fundamental limitations of standard pre-training objectives, necessitating more radical dataset curation strategies.

Third, for mechanistic understanding (RQ3), causal mediation analysis via layer-wise relevance propagation will likely reveal that cross-modal attention primarily routes unimodal features rather than constructing emergent multimodal abstractions [4]. We expect distinct attention pathways for in-distribution versus OOD inputs, with the latter showing increased dispersion and reduced compositionality. Observation of emergent multimodal representations would fundamentally challenge current views on attention mechanism limitations.

These findings will provide the first comprehensive diagnostic framework for cross-modal systematicity, though limitations include potential benchmark overfitting and computational constraints on scaling symbolic modules. Alternative interpretations will be pre-registered to ensure robust conclusions regardless of outcome direction.

7 Conclusion

This proposal establishes a comprehensive research agenda to diagnose and enhance the compositional reasoning capabilities of vision-language models through systematic

evaluation and causal intervention. We address three critical gaps: first, developing a diagnostic benchmark extending CLEVR [12] with natural language and targeted out-of-distribution splits to measure systematic generalization; second, quantifying how explicit compositional bottlenecks and causal regularization [10] improve robustness, targeting SYN > 0.85 and $\geq$20% reduction in spurious correlations; third, conducting causal mediation analysis [4] to reveal mechanistic pathways in cross-modal attention and distinguish memorization from genuine abstraction. Our contributions will include a publicly available benchmark suite with 10K+ challenge examples, novel neural-symbolic architectures [2] achieving state-of-the-art systematic generalization, and an empirical framework for causal analysis of multimodal representations. This work advances both the engineering of robust multimodal AI and our scientific understanding of cross-modal reasoning mechanisms, ultimately contributing to trustworthy vision-language systems capable of reliable, human-like reasoning beyond training distribution boundaries.

References

[1] Kim, N., Linzen, T.: Cogs: A compositional generalization challenge based on semantic interpretation. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2020)

[2] Keysers, D., Schärli, N., Scales, N., Buisman, H., Furrer, D., Kashubin, S., Momchev, N., Sinopalnikov, D., Stafiniak, Tihon, T., Tsarkov, D., Wang, X., Zee, M., Bousquet, O.: Measuring compositional generalization: A comprehensive method on realistic data. arXiv (Cornell University) (2019)

[3] Liu, Y., Li, G., Lin, L.: Cross-modal causal relational reasoning for event-level visual question answering. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)

[4] Samek, W., Montavon, G., Lapuschkin, S., Anders, C.J., Müller, K.: Explaining deep neural networks and beyond: A review of methods and applications. Proceedings of the IEEE (2021)

[5] Qian, S., Chen, H., Xue, D., Fang, Q., Xu, C.: Open-world social event classification. In: Proceedings of the ACM Web Conference 2023, pp. 1562–1571 (2023)

[6] Xue, D., Qian, S., Xu, C.: Few-shot multimodal explanation for visual question answering. In: Proceedings of the 32nd ACM International Conference on Multimedia, pp. 1875–1884 (2024)

[7] Zhou, Z., Xue, D., Qi, B., Qian, S., Xu, C.: Code-driven llm agent for one-shot explanatory visual question answering. ACM Transactions on Multimedia Computing, Communications and Applications (2025)

[8] Rueden, L., Mayer, S., Beckh, K., Georgiev, B., Giesselbach, S., Heese, R., Kirsch,

B., Walczak, M., Pfrommer, J., Pick, A., Ramamurthy, R., Garcke, J., Bauckhage, C., Schuecker, J.: Informed machine learning - a taxonomy and survey of integrating prior knowledge into learning systems. IEEE Transactions on Knowledge and Data Engineering (2021)

[9] Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision (2022)

[10] Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)

[11] Jiang, D., Ye, M.: Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)

[12] Quinn, P.C., Lee, K., Pascalis, O.: Face processing in infancy and beyond: The case of social categories. Annual Review of Psychology (2019)