GUIDE: Guiding Internal Evidence with Language Instructions

Anonymous ACL submission

Abstract

Large multimodal models can often follow instructions about what to generate, but not necessarily about what evidence to rely on. As a result, models may continue depending on shortcut-associated cues even when instructions suggest otherwise. We introduce GUIDE, a framework for controlling internal evidence usage through language instructions. GUIDE combines grouped parameter-efficient adaptation with instruction-conditioned gating to regulate reliance across task-relevant multimodal evidence pathways during reasoning and generation. We further introduce a pathway-level evaluation framework based on reliance sensitivity, controlled perturbation analysis, pathway modulation, and autoregressive decoding dynamics. Across multimodal reasoning, classification, and generation tasks, GUIDE produces structured and instruction-aligned redistribution of evidence reliance while maintaining competitive task performance. Experiments on GQA, TextVQA, MM-IMDb, CREMA-D, RAVDESS, and Flickr30K show that GUIDE improves robustness under targeted feature perturbations while enabling controllable evidence modulation across diverse multimodal settings. Our findings suggest that multimodal instruction following should extend beyond output control toward explicit regulation of internal evidence utilization.

1 Introduction

Large multimodal models can often follow instructions about what to generate, but not necessarily about what evidence to rely on. For example, a model instructed to *"focus on actions"* may still internally depend on appearance, demographic, or contextual cues while producing the same answer. Similarly, instructions such as *"ignore appearance"* or *"rely on textual evidence"* may alter surface-level outputs while leaving the underlying evidence pathways largely unchanged. This limitation is

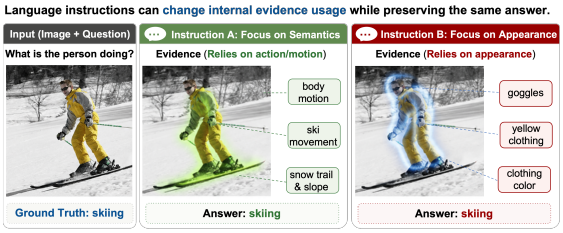

Figure 1: Motivation of GUIDE. Given the same input image and question, different language instructions encourage reliance on different internal evidence pathways while producing the same answer. Semantic-focused instructions emphasize action and interaction cues, whereas appearance-focused instructions increase reliance on visual attribute evidence.

particularly important in multimodal reasoning settings, where multiple evidence sources can independently support the same prediction. In tasks such as GQA and TextVQA, models may rely on object relationships, OCR signals, visual attributes, or background context to produce identical answers. However, shortcut-associated cues are often highly predictive under standard correlation-driven learning, encouraging models to exploit spurious evidence even when instructions suggest otherwise (Geirhos et al., 2020; Yuan et al., 2024; Ye et al., 2024). As a result, controlling outputs alone does not necessarily imply control over how information is internally used. Recent progress in multimodal instruction following has primarily focused on prompt engineering, instruction tuning, and controllable generation objectives. While these approaches improve output controllability, they generally operate at the representation or decoding level and provide limited control over internal evidence usage (Chefer et al., 2021; Nguyen et al., 2023; Turner et al., 2024; Zou et al., 2025).

An important open problem remains: can language instructions systematically regulate the internal evidence pathways used by multimodal models? To address this problem, we introduce **GUIDE**, a framework for controlling internal evidence us-

age with language instructions. GUIDE combines grouped parameter-efficient adaptation with instruction-conditioned gating to dynamically regulate reliance across task-relevant, semantically meaningful feature pathways. Rather than modifying task objectives or input representations, GUIDE operates through internal adaptation pathways to modulate the contribution of multimodal evidence during reasoning and generation. Importantly, GUIDE enables models to produce similar outputs while relying on substantially different internal evidence pathways. For example, semantic-focused instructions increase reliance on action and interaction pathways, whereas appearance-focused instructions increase reliance on visual attribute pathways. Similarly, in TextVQA settings, instructions emphasizing textual evidence increase reliance on OCR-related pathways while suppressing contextual appearance cues.

To study this behavior, we introduce a mechanistic evaluation framework based on pathway modulation, controlled perturbation analysis, reliance sensitivity, and autoregressive decoding dynamics. Across multimodal reasoning, classification, and generation tasks, GUIDE induces structured and instruction-aligned redistribution of feature reliance rather than superficial output variation. We evaluate GUIDE on GQA, TextVQA, MM-IMDb, CREMA-D, RAVDESS, and Flickr30K. Experiments show that GUIDE enables models to produce similar outputs while relying on substantially different evidence pathways, and improves robustness to targeted spurious-feature interventions. Our main contributions are as follows: **1)** We introduce GUIDE, a framework for controlling internal evidence usage with language instructions through grouped adaptation pathways and dynamic instruction-conditioned gating. **2)** We provide pathway-level evidence that language instructions systematically redistribute internal evidence reliance across multimodal pathways. **3)** We demonstrate that instruction-conditioned evidence control generalizes across multimodal reasoning, classification, and generation tasks while preserving task semantics. Together, these findings suggest that evidence-aware instruction following may provide a new direction for controllable and interpretable multimodal reasoning.

2 Related Work

Multimodal Instruction Following. Recent multimodal foundation models (Liu et al., 2024; Dai et al., 2023; Wang et al., 2024; Peng et al., 2023) have substantially improved instruction-guided reasoning and controllable vision–language generation. However, most existing approaches primarily focus on aligning model outputs with user instructions rather than controlling the internal evidence used to produce them. Methods based on prompt engineering, controllable decoding, attention steering, or visual grounding (Chefer et al., 2021; Nguyen et al., 2023; Cheng et al., 2024) can influence generated responses or attention distributions, but provide limited explicit control over feature-level evidence reliance during internal computation. This limitation is particularly important because multimodal models often exploit shortcut-associated cues whenever such correlations are predictive under the training distribution (Geirhos et al., 2020; Yuan et al., 2024; Ye et al., 2024). In multimodal settings, these shortcuts may correspond to appearance attributes, contextual backgrounds, demographic signals, motion patterns, acoustic cues, or OCR artifacts. Such spurious reliance has been repeatedly observed in multimodal benchmarks including MM-IMDb, CREMA-D, and RAVDESS.

Feature Steering and Representation Control. Recent work on feature steering, representation editing, and controllable generation has explored manipulation of latent representations or activations to alter model behavior (Lamb et al., 2025; Zhang et al., 2025; Kong et al., 2024; Turner et al., 2024; Zou et al., 2025). These approaches demonstrate that model behavior can often be influenced through targeted representation-level interventions. However, existing methods primarily focus on text-only settings or activation-level manipulation, primarily through post-hoc activation manipulation rather than instruction-conditioned multimodal evidence routing. They do not explicitly regulate how multimodal evidence is internally selected and combined during reasoning and generation.

Bias Mitigation and Parameter-Efficient Adaptation. Multimodal bias mitigation methods (Agrawal et al., 2018; Zhang et al., 2024; Liu et al., 2025) primarily operate during training to reduce dataset-level spurious correlations. While effective for improving robustness, these approaches generally do not support user-specified evidence control during inference. Parameter-efficient fine-tuning methods such as LoRA (Hu et al., 2022), QLoRA (Dettmers et al., 2023),

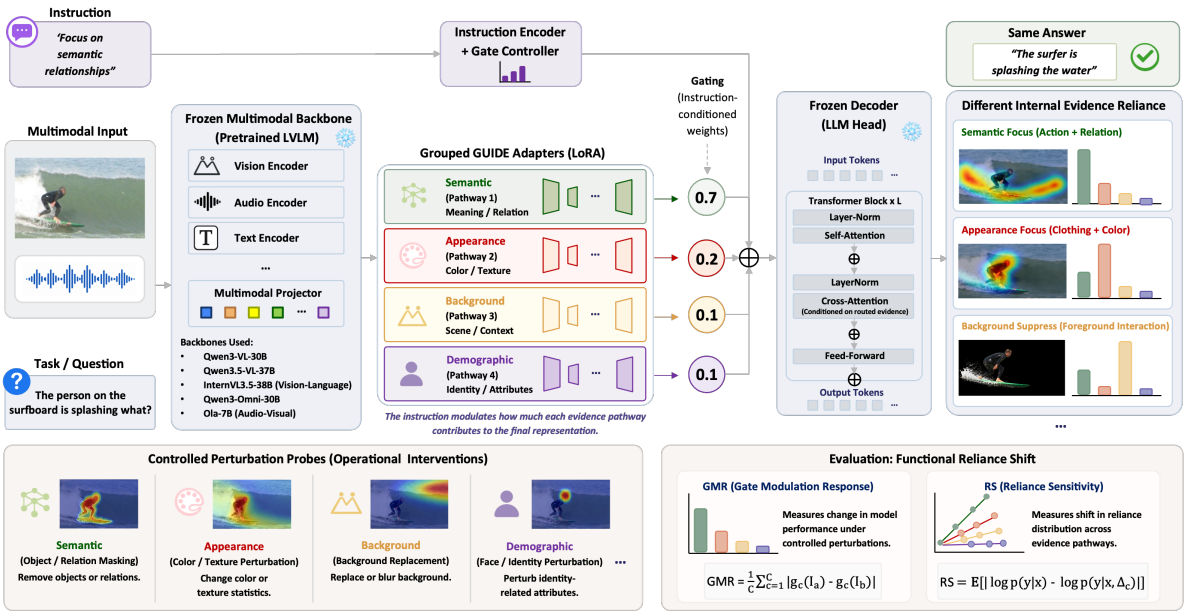

Figure 2: Overview of GUIDE, an instruction-conditioned evidence routing framework for controllable multimodal reasoning. GUIDE dynamically modulates grouped adaptation pathways corresponding to different evidence categories through instruction-conditioned gating, enabling the same prediction to be produced under different internal evidence reliance patterns. Controlled perturbation probes are used to evaluate both pathway modulation and functional reliance shifts across evidence categories.

adapters (Houlsby et al., 2019), FocalLoRA (Shi et al., 2025), and mixture-of-LoRA routing approaches (Huynh et al., 2025; Gao et al., 2025) enable efficient adaptation through lightweight parameter updates, but rely on static adaptation pathways without instruction-conditioned modulation of feature-specific computation.

Our Perspective. Prior work provides limited mechanisms for translating language instructions into explicit control over internal multimodal evidence usage. In contrast, GUIDE treats instructions as interventions on evidence reliance itself, enabling language-conditioned modulation of task-relevant multimodal evidence pathways during reasoning and generation. This enables GUIDE to separate output controllability from evidence controllability, allowing similar predictions to arise from different instructed evidence pathways.

3 GUIDE

3.1 Problem Formulation

GUIDE addresses the problem of controlling which evidence a multimodal model relies on while preserving task semantics. Given multimodal input x, instruction I, and pretrained model f_θ, conventional instruction following primarily controls the output: $y = f_\theta(x, I)$. However, different evidence sources may independently support the same prediction. For example, a model may answer a question using semantic object relationships, OCR signals, appearance attributes, or contextual background cues. Our goal is therefore not merely to control outputs, but to regulate the internal evidence pathways used to produce them. Specifically, we seek instruction-conditioned evidence modulation such that $f_\theta(x, I_a) \approx f_\theta(x, I_b)$ while relying on substantially different internal evidence pathways under instructions I_a and I_b.

3.2 Taxonomy-Aligned Feature Pathways

GUIDE treats multimodal reasoning as composition over heterogeneous evidence sources. To enable controllable evidence modulation, we partition internal adaptation pathways into task-relevant evidence groups corresponding to different forms of multimodal information. Specifically, GUIDE associates grouped adaptation pathways with semantic interactions, visual appearance, environmental context, motion dynamics, acoustic signals, demographic cues, and OCR/textual grounding information depending on the modality and benchmark. While the overall routing framework is shared across tasks, the specific pathway taxonomy is adapted to the dominant evidence structure of each benchmark. This formulation allows GUIDE to maintain a unified routing mechanism while adapting evidence control to diverse multi-

modal settings. These task-specific evidence pathways enable language instructions to selectively emphasize or suppress different forms of multimodal evidence during reasoning and generation. Each evidence pathway corresponds to a dedicated grouped low-rank adaptation branch attached to shared transformer projections. Pathway assignments are determined using benchmark-specific evidence taxonomies and weak evidence localization heuristics described in Appendix A.

3.3 Grouped LoRA Adaptation

GUIDE implements instruction-conditioned evidence modulation through grouped parameter-efficient adaptation. Given pretrained weight matrix W, standard LoRA introduces a low-rank update of the form $W' = W + BA$, where A and B are trainable low-rank matrices. Instead of using a single shared adaptation pathway, GUIDE decomposes the update into multiple evidence-aligned adaptation groups:

$$W' = W + \sum_{c=1}^{C} g_c(I, h_t)\, B_c A_c,$$

where c indexes feature pathways, $B_c A_c$ denotes the grouped low-rank adaptation associated with pathway c, and $g_c(I, h_t)$ is an instruction-conditioned gate predicted from the instruction representation and current model state. Each grouped adaptation pathway captures a different form of multimodal evidence, while the gating mechanism dynamically regulates their relative contribution during reasoning and generation. Unlike static parameter-efficient adaptation, this formulation allows language instructions to selectively amplify or suppress specific evidence pathways at inference time. Importantly, GUIDE does not require explicit disentanglement of internal representations. Instead, grouped adaptation provides a structured intervention space through which language instructions can modulate feature reliance while preserving the underlying task objective. This grouped parameterization provides a lightweight but persistent mechanism for regulating evidence contributions within grouped internal adaptation pathways.

3.4 Instruction-Conditioned Gating

To regulate evidence usage, GUIDE predicts pathway-specific gates conditioned on both the instruction representation and the current model state. Given instruction embedding e_I and hidden representation h_t, the gating network computes:

$$g(I, h_t) = \mathrm{softmax}\left(W_g[e_I; h_t]\right),$$

where $g(I, h_t) \in \mathbb{R}^C$ denotes normalized activations over C taxonomy-aligned feature pathways. These gates determine the relative contribution of each grouped adaptation pathway during computation. Intuitively, the gating mechanism acts as an instruction-conditioned controller over internal evidence usage, dynamically redistributing reliance across different forms of multimodal evidence. Because gating depends on both the instruction and the evolving model state, pathway activations can change throughout reasoning and generation. This enables GUIDE to modulate evidence usage persistently during autoregressive decoding rather than acting only as a static prompt-level bias. Gates are computed at each decoding step and shared across grouped adaptation modules within the corresponding transformer layer. GUIDE is trained end-to-end using the original task objective while jointly optimizing grouped adaptation pathways and instruction-conditioned gating networks. Additional training details and regularization objectives are provided in Appendix D.

3.5 Reliance Modulation Objective

To evaluate whether language instructions meaningfully alter internal evidence usage, we introduce two complementary measures: Gate Modulation Response (GMR) and Reliance Sensitivity (RS).

Gate Modulation Response (GMR). GMR measures how strongly internal pathway activations respond to instruction changes. Given two instructions I_a and I_b, we compute the average pathway activation difference between their corresponding gate activations:

$$\mathrm{GMR} = \frac{1}{C} \sum_{c=1}^{C} |g_c(I_a) - g_c(I_b)|.$$

For autoregressive tasks, GMR is computed by averaging pathway activation differences across decoding steps and transformer layers, followed by normalization over the number of pathway groups. For classification tasks, activations are aggregated from the final prediction representation. Additional implementation details for GMR computation are provided in Appendix C. Higher GMR indicates stronger instruction-conditioned redistribution of evidence pathways.

Reliance Sensitivity (RS). While GMR measures internal modulation, it does not necessarily imply functional influence on predictions. We therefore introduce RS, which measures how

strongly model predictions depend on a given evidence pathway. For pathway group c, we perturb evidence associated with the corresponding pathway and measure the resulting change in prediction likelihood:

$$\mathrm{RS}_c = \mathbb{E}\left[|\log p(y|x) - \log p(y|x, \Delta_c)|\right],$$

where Δ_c denotes perturbation of pathway c. For each benchmark, pathway perturbations are implemented using taxonomy-aligned controlled interventions targeting the corresponding evidence category. These perturbations include modality-specific masking, corruption, replacement, and feature suppression operations designed to selectively degrade the targeted evidence source while avoiding complete destruction of task-relevant information. Detailed perturbation protocols for all datasets are provided in Table 3. Higher RS indicates greater functional reliance on the corresponding evidence source. Intuitively, GMR measures whether instructions change internal pathway activations, whereas RS measures whether those pathways functionally influence model behavior. Together, GMR and RS distinguish superficial pathway modulation from meaningful changes in multimodal evidence reliance during reasoning and generation.

4 Evaluation Setup

We evaluate whether GUIDE enables controllable modulation of multimodal evidence usage while preserving task semantics. To evaluate both controllability and generalization, we include multimodal reasoning, classification, and generation settings spanning controlled and natural multimodal tasks. All experiments follow a unified training setup with grouped parameter-efficient adaptation and instruction-conditioned gating. Additional implementation details are provided in Appendix L.

4.1 Tasks and Datasets

We evaluate GUIDE across multimodal reasoning, classification, and generation benchmarks selected for their identifiable evidence structure and controllable evidence pathways.

Multimodal Reasoning. We evaluate on GQA (Hudson and Manning, 2019) and TextVQA (Singh et al., 2019). GQA enables analysis of semantic, relational, and contextual reasoning under controlled visual reasoning settings, while TextVQA provides a complementary setting requiring explicit reliance on OCR and textual grounding signals. For GQA, we evaluate on the balanced split to reduce shortcut-associated answer biases.

Multimodal Classification. We evaluate on MM-IMDb (Arevalo et al., 2017), CREMA-D (Cao et al., 2014), and RAVDESS (Livingstone and Russo, 2018). These datasets contain diverse shortcut-associated cues spanning appearance, context, motion, acoustic, and demographic signals, making them suitable for studying controllable evidence reliance under fixed task semantics.

Multimodal Generation. We evaluate on Flickr30K (Young et al., 2014) image captioning. Caption generation provides a more challenging setting in which evidence modulation must persist throughout autoregressive decoding rather than only affecting a single prediction step. We follow the standard benchmark splits for all datasets and perform perturbation-based evaluation only on held-out evaluation partitions.

4.2 Instruction Taxonomy

We construct instruction families corresponding to different evidence preferences, including semantic-focused, appearance-focused, background-suppression, motion-focused, and OCR-focused conditions depending on the benchmark. In addition to fixed templates, we evaluate paraphrased and compositional instruction variants to test generalization beyond narrow lexical formulations. Representative instruction templates are provided in Appendix B.

4.3 Evidence Perturbation Protocol

To measure functional reliance on different evidence pathways, we introduce taxonomy-aligned perturbation protocols targeting semantic, appearance, contextual, OCR/textual, motion, and acoustic evidence categories depending on the benchmark. These interventions include modality-specific masking, corruption, replacement, and suppression operations designed to selectively degrade targeted evidence sources while preserving overall task solvability. For perturbations requiring coarse evidence localization, we use pretrained multimodal backbones to obtain weak evidence-region proposals without additional supervision or task-specific fine-tuning. Additional perturbation details and backbone configurations are provided in Appendix L.1.

4.4 Evaluation Metrics

We evaluate GUIDE using both task-level performance metrics and pathway-level evidence-control

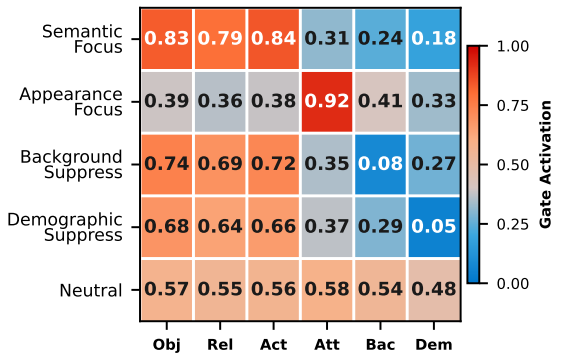

Figure 3: Gate activation heatmap on GQA under different instruction conditions. Obj, Rel, Act, Att, Bac, Dem denote **Obj**ect, **Rel**ation, **Act**ion, **Att**ribute, **Bac**kground, **Dem**ographic respectively

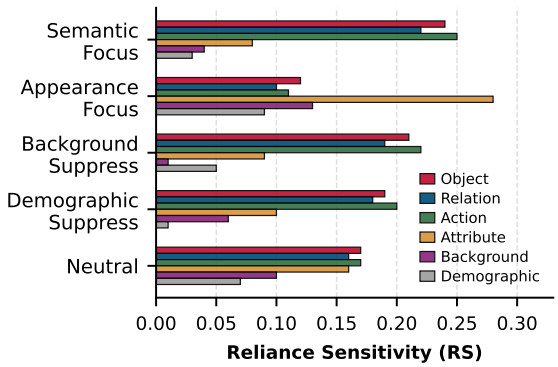

Figure 4: RS under different instruction conditions on GQA. Instructions selectively redistribute functional dependence across evidence pathways.

metrics. To quantify instruction-conditioned evidence modulation, we evaluate: (i) Gate Modulation Response (GMR), which measures redistribution of pathway activations under different instructions, and (ii) Reliance Sensitivity (RS), which measures functional dependence on perturbed evidence pathways. Together, GMR and RS distinguish superficial pathway modulation from meaningful changes in multimodal evidence reliance during reasoning and generation. We additionally report standard task-specific evaluation metrics for each benchmark to measure semantic preservation under instruction-conditioned evidence modulation.

5 Results

5.1 Controlled Evidence Routing

We first evaluate whether instruction-conditioned gating produces structured redistribution of internal evidence pathways. Figure 3 shows gate activation patterns on GQA under different instruction conditions, while Figure 13–Figure 17 present corresponding results on additional multimodal benchmarks. Across all datasets, GUIDE produces consistent and instruction-aligned changes in pathway activations. Semantic-focused instructions increase activation of semantic and relational pathways, whereas appearance-focused instructions sharply increase reliance on appearance-related pathways. Similarly, suppression-oriented instructions selectively reduce activation of the targeted pathway while preserving activation in task-relevant semantic pathways. Neutral instructions produce comparatively balanced activation patterns, suggesting that GUIDE deviates from default evidence usage only when explicitly instructed. Importantly, these redistribution patterns remain consistent across multimodal reasoning, classification, and generation settings despite substantial differences in modality and task structure. Overall, GUIDE induces structured and instruction-consistent redistribution of internal evidence routing rather than superficial prompt-conditioned output variation.

5.2 Functional Evidence Redistribution

We next evaluate whether instruction-conditioned pathway modulation produces corresponding changes in functional prediction reliance. Figure 4 reports Reliance Sensitivity (RS) across evidence pathways on GQA under different instruction conditions, while Figure 19–Figure 22 present corresponding results on additional benchmarks. Across all tasks, RS patterns consistently reflect the intended evidence preference specified by the instruction. Semantic-focused instructions increase prediction sensitivity to semantic, relational, and interaction pathways, whereas appearance-focused instructions produce dominant reliance on appearance-related pathways. Suppression-oriented instructions substantially reduce reliance on the targeted pathway while preserving sensitivity to task-relevant semantic pathways. For example, background suppression consistently reduces functional dependence on contextual pathways without collapsing overall prediction sensitivity. These results show that instruction-conditioned gating alters not only internal activations, but also the evidence pathways that functionally influence model predictions during reasoning and generation.

5.3 Counterfactual Evidence Control

We next evaluate whether instruction-conditioned evidence routing produces corresponding changes in robustness under controlled counterfactual per-

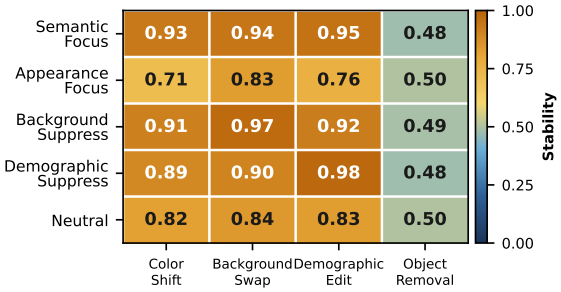

Figure 5: Counterfactual stability under different instruction conditions and perturbation types on GQA. Semantic-focused and background-suppression instructions remain robust under appearance and contextual perturbations, whereas object removal substantially reduces stability across all settings.

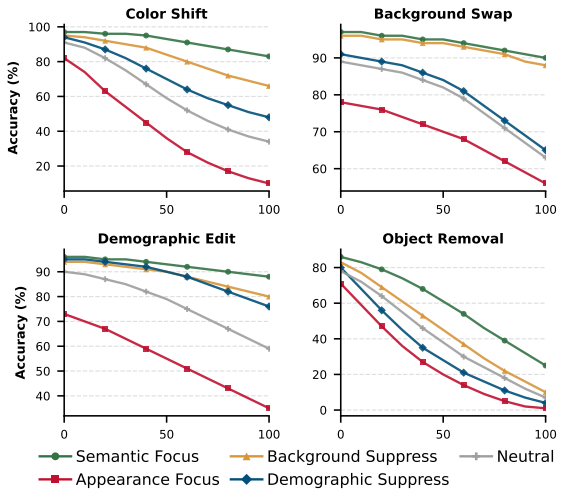

Figure 6: Prediction robustness under increasing perturbation strength on GQA. Appearance-focused instructions degrade most rapidly under color shifts, whereas semantic-focused and background-suppression instructions remain comparatively robust.

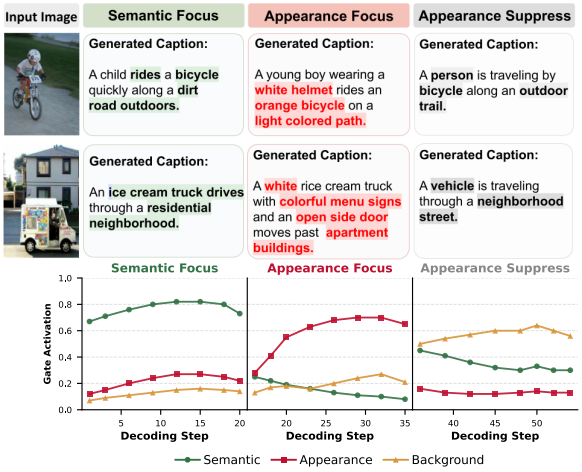

Figure 7: Instruction-conditioned decoding-time evidence routing on Flickr30K. Different instructions induce persistent changes in pathway activations throughout autoregressive generation, resulting in captions emphasizing either semantic/appearance attributes.

turbations. Figure 8 provides qualitative examples showing that different instructions induce distinct evidence reliance patterns while preserving the same final prediction. Under semantic-focused instructions, the model exhibits increased reliance on action-relevant and semantically meaningful regions, whereas appearance-focused instructions redirect attention toward figure-centric visual attributes. To quantify this behavior, Figure 5 reports counterfactual stability under four perturbation types. Semantic-focused and background-suppression instructions remain highly stable under color and background perturbations, whereas appearance-focused instructions exhibit substantially lower robustness under appearance-related shifts. In contrast, object removal consistently reduces stability across all instruction conditions, indicating that object-level evidence remains a task-critical dependency that cannot be fully suppressed. Figure 6 analyzes robustness under progressively increasing perturbation strength. Appearance-focused instructions exhibit the steepest degradation under color shifts, while semantic-focused and background-suppression instructions retain substantially higher accuracy throughout the perturbation range. Additional perturbation stability analyses across MM-IMDb and Flickr30K are provided in Appendix K. All robustness curves are averaged across the evaluation set over multiple perturbation strengths. Together, these results demonstrate that GUIDE modulates not only internal pathway activations, but also functional evidence on which predictions depend during multimodal reasoning.

5.4 Decoding-Time Evidence Routing

Figure 7 illustrates pathway-level gate activations across autoregressive decoding steps on Flickr30K under different instruction conditions. Distinct instructions produce consistently different routing dynamics throughout generation rather than only at initialization. Under semantic-focused instructions, semantic pathways remain dominant across decoding steps and resulting in captions that selectively prioritize actions and scene interactions. In contrast, appearance-focused instructions increase activation of appearance-related pathways, leading the model to generate captions centered on visual attributes such as clothing, color, and object appearance. Appearance-suppression instructions selectively reduce appearance-related gate activation

while preserving semantic and contextual pathways, resulting in captions that retain scene-level semantics while avoiding detailed appearance descriptions. Importantly, these instruction-conditioned routing patterns persist throughout autoregressive decoding, suggesting that GUIDE dynamically regulates evidence usage during generation rather than merely biasing the initial decoding state. The plotted gate trajectories correspond to representative decoding sequences and exhibit similar trends across additional evaluation examples. Additional analyses of compositional and conflict-aware routing are provided in Appendix F.

5.5 Same Answer, Different Evidence

We evaluate whether GUIDE can preserve task behavior while substantially altering internal evidence reliance. Table 1 reports output consistency together with changes in pathway modulation (ΔGMR) and functional reliance (ΔRS) across conflicting instruction conditions. Across all benchmarks, GUIDE maintains high output consistency despite large redistribution of internal evidence pathways. For example, GQA preserves the same answer in 93.8% of cases while producing substantial shifts in both pathway modulation and functional reliance. These results provide direct quantitative evidence that GUIDE can alter how multimodal models internally use evidence without necessarily changing the final prediction. Additional analyses are provided in Appendix E and G.

Table 1: Same-answer behavior under semantic–appearance evidence-routing shifts. GUIDE preserves high output consistency despite substantial changes in pathway modulation (ΔGMR) and functional reliance (ΔRS). Flickr30K reports caption similarity (0–1), while other datasets report same-prediction rate (%).

Dataset	Output Consistency	ΔGMR	ΔRS
GQA	93.8	0.61	0.19
TextVQA	91.8	0.51	0.16
MM-IMDb	92.1	0.58	0.17
CREMA-D	91.7	0.55	0.18
RAVDESS	91.1	0.53	0.17
Flickr30K	0.84	0.63	0.20

5.6 Ablation Study

We evaluate the contribution of grouped routing and dynamic instruction-conditioned gating. Table 2 reports representative ablation results on GQA, CREMA-D, and Flickr30K. Removing taxonomy-aligned grouped pathways substantially reduces both Gate Modulation Response (GMR) and Reliance Sensitivity (RS), indicating that structured pathway decomposition is necessary for controllable evidence redistribution. Replacing dynamic instruction-conditioned gates with static routing also consistently decreases perturbation robustness, suggesting that instruction-conditioned modulation is critical for stable evidence control. Prompt-only control exhibits the weakest modulation behavior, producing only minor output variation without meaningful shifts in pathway modulation or functional evidence reliance. Overall, the results show that both grouped adaptation pathways and dynamic instruction-conditioned gating are necessary for effective multimodal evidence control. Full ablation results across all benchmarks are provided in Appendix H.

Table 2: Representative ablation results across multimodal reasoning (GQA), classification (CREMA-D), and generation (Flickr30K) tasks. Removing grouped pathways or replacing dynamic instruction-conditioned gates substantially reduces pathway modulation (GMR), functional reliance redistribution (RS), and perturbation robustness (Robust.). Full benchmark results are provided in Appendix H.

Dataset	Variant	GMR↑	RS↑	Robust.↑
GQA	Full	**0.61**	**0.19**	**0.93**
	No grouped	0.22	0.07	0.81
	Static gates	0.38	0.10	0.86
	Prompt-only	0.18	0.06	0.79
CREMA-D	Full	**0.55**	**0.18**	**0.91**
	No grouped	0.19	0.06	0.79
	Static gates	0.33	0.09	0.84
	Prompt-only	0.16	0.05	0.77
Flickr30K	Full	**0.63**	**0.20**	**0.84**
	No grouped	0.23	0.07	0.73
	Static gates	0.39	0.11	0.78
	Prompt-only	0.19	0.06	0.71

6 Conclusion

We introduced GUIDE, a framework for controlling internal multimodal evidence usage through natural language instructions. By combining grouped adaptation pathways with instruction-conditioned gating, GUIDE enables multimodal models to redistribute reliance across different forms of evidence during reasoning and generation while preserving overall task behavior. Our findings suggest that multimodal instruction following should regulate not only what models generate, but also how they internally use information.

Limitations

GUIDE does not explicitly disentangle semantic factors or guarantee perfectly isolated evidence pathways. Instead, the proposed grouped routing mechanism provides an operational intervention space for modulating functional evidence reliance under natural language instructions. As a result, pathway activations should not be interpreted as uniquely corresponding to fully disentangled semantic concepts. In addition, the evidence perturbation protocols used in this work are approximate operational interventions rather than precise semantic manipulations. Although we employ controlled perturbations and weak localization strategies, some perturbations may partially affect multiple evidence categories simultaneously. Future work may explore more principled causal intervention and localization methods for multimodal evidence control. Finally, GUIDE is evaluated primarily on controlled multimodal reasoning, classification, and generation benchmarks. Additional evaluation on larger-scale open-ended multimodal agents and real-world deployment settings remains an important direction for future work.

Potential Risks

Techniques for controllable evidence routing may potentially be misused to selectively amplify or suppress specific evidence sources in ways that obscure model reasoning behavior. For example, evidence suppression mechanisms could unintentionally hide undesirable biases while preserving similar output behavior. In addition, imperfect evidence localization and perturbation protocols may introduce unintended artifacts or incomplete suppression of sensitive cues. We therefore emphasize that GUIDE should not be interpreted as providing guaranteed fairness, bias removal, or fully interpretable reasoning. Instead, the framework is intended as a research tool for studying controllable multimodal evidence reliance under operational intervention settings.

Ethics Statement

Human evaluation was conducted using voluntary annotators recruited for research purposes. Annotators were informed about the evaluation procedure and provided consent prior to participation. The annotation tasks involved evaluating multimodal model outputs and did not involve sensitive personal data, medical information, or high-risk decision making. All data used in human evaluation were obtained from publicly available research benchmarks under their original licenses. No personally identifiable information was collected or stored during the annotation process.

References

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4971–4980.

John Arevalo, Thamar Solorio, Manuel Montes y Gómez, and Fabio A. González. 2017. Gated multimodal units for information fusion. *Preprint*, arXiv:1702.01992.

Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. 2014. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390.

Hila Chefer, Shir Gur, and Lior Wolf. 2021. Transformer interpretability beyond attention visualization. *Preprint*, arXiv:2012.09838.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. 2024. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37:135062–135093.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

Chongyang Gao, Kezhen Chen, Jinmeng Rao, Ruibo Liu, Baochen Sun, Yawen Zhang, Daiyi Peng, Xiaoyuan Guo, and VS Subrahmanian. 2025. Mola: Moe lora with layer-wise expert allocation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, page 5097–5112.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019.

Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Tuan-Luc Huynh, Thuy Vu, Weiqing Wang, Trung Le, Dragan Gasevic, Yuan-Fang Li, and Thanh-Toan Do. 2025. Mixlora-dsi: Dynamically expandable mixture-of-lora experts for rehearsal-free generative retrieval over dynamic corpora. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 380–396.

Lingkai Kong, Haorui Wang, Wenhao Mu, Yuanqi Du, Yuchen Zhuang, Yifei Zhou, Yue Song, Rongzhi Zhang, Kai Wang, and Chao Zhang. 2024. Aligning large language models with representation editing: A control perspective. In *Advances in Neural Information Processing Systems*, volume 37, pages 37356–37384. Curran Associates, Inc.

Tom A. Lamb, Adam Davies, Alasdair Paren, Philip Torr, and Francesco Pinto. 2025. Focus on this, not that! steering LLMs with adaptive feature specification. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Yishu Liu, Jiawei Zhu, Congcong Wen, Guangming Lu, Hui Lin, and Bingzhi Chen. 2025. Towards robust visual question answering via prompt-driven geometric harmonization. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*. AAAI Press.

Steven R Livingstone and Frank A Russo. 2018. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391.

Thao Nguyen, Yuheng Li, Utkarsh Ojha, and Yong Jae Lee. 2023. Visual instruction inversion: Image editing via visual prompting. *Preprint*, arXiv:2307.14331.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.

Zitong Shi, Guancheng Wan, Haixin Wang, Ruoyan Li, Zijie Huang, Wanjia Zhao, Yijia Xiao, Xiao Luo, Carl Yang, Yizhou Sun, and Wei Wang. 2025. Don't forget the enjoin: FocalloRA for instruction hierarchical alignment in large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2024. Steering language models with activation engineering. *Preprint*, arXiv:2308.10248.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wenqian Ye, Guangtao Zheng, Yunsheng Ma, Xu Cao, Bolin Lai, James Matthew Rehg, and Aidong Zhang. 2024. MM-spubench: Towards better understanding of spurious biases in multimodal LLMs. In *Workshop on Responsibly Building the Next Generation of Multimodal Foundational Models*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*.

Yu Yuan, Lili Zhao, Kai Zhang, Guangting Zheng, and Qi Liu. 2024. Do llms overcome shortcut learning? an evaluation of shortcut challenges in large language models. *Preprint*, arXiv:2410.13343.

Hanyu Zhang, Xiting Wang, Chengao Li, Xiang Ao, and Qing He. 2025. Controlling large language models through concept activation vectors. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(24):25851–25859.

Xiaohui Zhang, Jaehong Yoon, Mohit Bansal, and Huaxiu Yao. 2024. Multimodal representation learning by alternating unimodal adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 27456–27466.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2025. Representation engineering: A top-down approach to ai transparency. *Preprint*, arXiv:2310.01405.

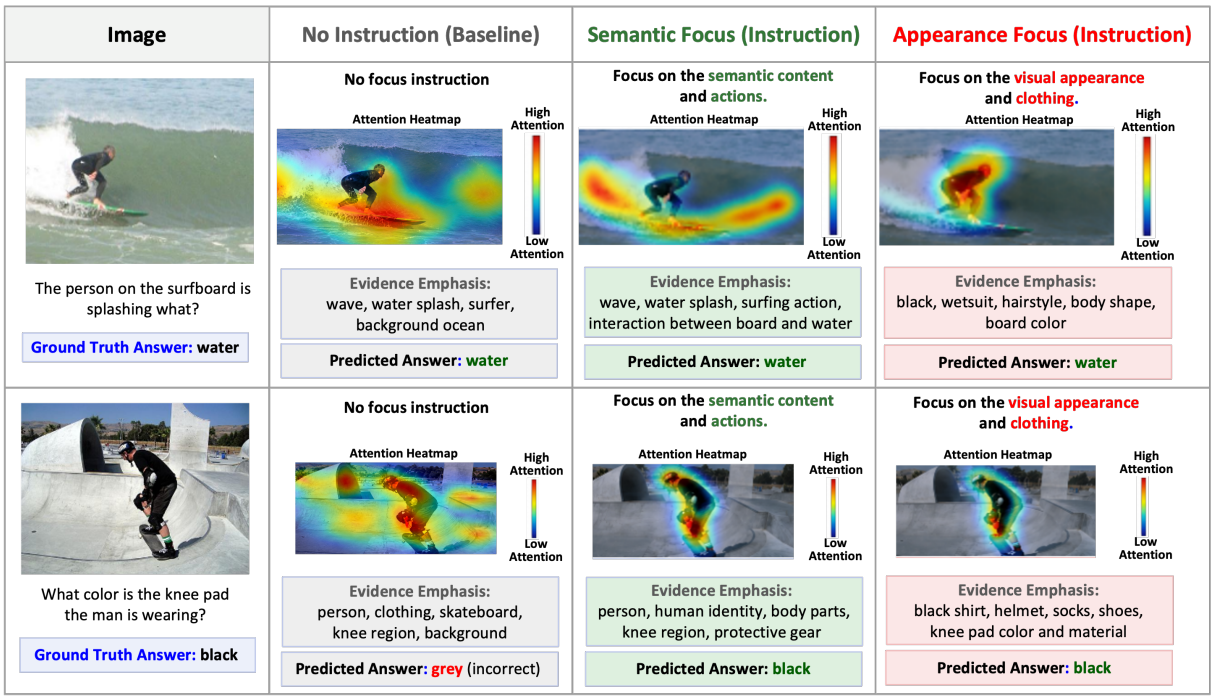

Figure 8: Qualitative examples of instruction-conditioned evidence routing on GQA. Semantic-focused instructions emphasize action-relevant and interaction-based evidence, whereas appearance-focused instructions prioritize visual attribute cues while preserving the same final prediction.

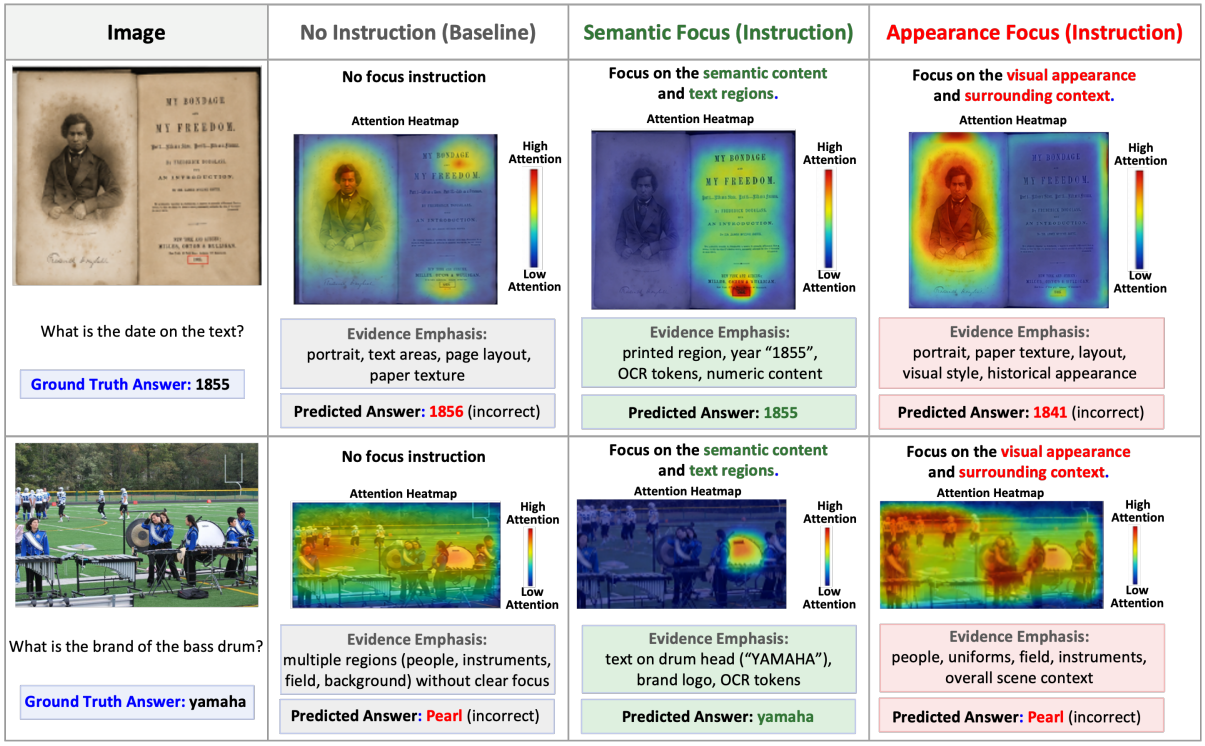

Figure 9: Qualitative examples of instruction-conditioned evidence routing on TextVQA. Semantic-focused instructions increase reliance on OCR and text-relevant regions, whereas appearance-focused instructions emphasize visual appearance and contextual cues, often leading to different prediction behavior under the same input.

A Dataset-Specific Evidence Taxonomy

GUIDE uses dataset-specific evidence taxonomies to define controllable multimodal evidence pathways aligned with the dominant evidence structure of each benchmark. Rather than assuming perfectly disentangled semantic factors, the taxonomy provides coarse functional partitions corresponding to task-relevant multimodal cues that can independently influence model predictions. The taxonomy is constructed based on three principles: (1) identifying the primary evidence sources required for successful task completion, (2) separating shortcut-associated or potentially confounding evidence categories, and (3) defining perturbation categories that can be independently manipulated through controlled interventions.

Table 3 summarizes the perturbation protocols associated with each evidence category. These perturbations are designed to measure relative changes in functional reliance under targeted evidence suppression rather than simulate naturally occurring corruptions. For visual reasoning benchmarks such as GQA and Flickr30K, the taxonomy separates semantic interaction cues, appearance attributes, environmental context, and person-related appearance signals. Semantic pathways primarily correspond to object interactions, action understanding, and relational reasoning, while appearance pathways capture visual attribute evidence such as clothing color, texture, and style cues. Background pathways capture environmental and scene-level contextual information that may independently support prediction. Person-related appearance pathways isolate facial and identity-associated appearance signals that may correlate spuriously with task outputs. For TextVQA, the taxonomy additionally includes an OCR/Text pathway corresponding to explicit textual grounding signals. This separation is necessary because textual evidence often provides a distinct predictive source independent of object appearance or scene context. For MM-IMDb, the taxonomy partitions semantic narrative information, interaction-related character relationships, poster appearance cues, and contextual visual style information. This formulation enables controlled analysis of whether models rely on narrative semantics or shortcut-associated visual correlations contained in movie posters. For audiovisual emotion benchmarks such as CREMA-D and RAVDESS, the taxonomy separates facial-expression signals, acoustic emotion cues, motion dynamics, appearance information, and contextual background signals. This separation enables analysis of whether language instructions modulate reliance across different modalities while preserving emotion classification behavior. Importantly, these evidence categories are not treated as perfectly independent causal mechanisms. Instead, they provide structured functional partitions that enable controlled pathway modulation and perturbation-based reliance analysis across diverse multimodal settings.

B Instruction Taxonomy and Pathway Conditioning

GUIDE uses dataset-specific instruction templates to provide semantically consistent pathway-conditioning signals across multimodal benchmarks. The templates are designed to encourage emphasis or suppression of task-relevant evidence categories rather than optimize benchmark-specific prompt engineering behavior. Table 4 shows representative instruction templates used throughout training and evaluation. Multiple paraphrased variants were used for each pathway category to reduce sensitivity to fixed prompt wording. Instruction templates were used during both training and evaluation, while paraphrased variants were additionally introduced during evaluation to assess robustness beyond fixed lexical formulations.

C GMR Computation Details

GMR measures the magnitude of instruction-conditioned redistribution across taxonomy-aligned evidence pathways. For each input sample, we compute gate activations under two semantically contrasting instructions applied to the same multimodal input. Typical instruction pairs include semantic-focused versus appearance-focused prompts, OCR-focused versus contextual prompts, or motion-focused versus appearance-focused prompts depending on the benchmark.

For autoregressive generation tasks, gate activations are recorded at each decoding step and aggregated across generated tokens and transformer layers. Specifically, pathway activation differences are first averaged over decoding steps within each layer and then aggregated across all gated transformer layers. The resulting value is normalized by the number of pathway groups to ensure comparability across benchmarks with different taxonomy sizes. For classification tasks, gate activations are extracted from the final prediction representation and aggregated across gated transformer layers. Be-

Table 3: Dataset-specific perturbation protocols used for evaluating functional reliance on different evidence pathways. Perturbations are designed as controlled interventions targeting task-relevant evidence categories. Perturbations are applied only during evaluation and are not used for training GUIDE.

Dataset	Evidence Category	Perturbation Protocol
GQA	Semantic	Object masking, relation disruption, interaction-region masking, action-region occlusion
	Appearance	Color jitter, texture perturbation, attribute-preserving recoloring
	Background	Contextual background replacement, environmental region masking, scene blurring
	Demographic	Facial anonymization, skin-tone perturbation, hairstyle masking
TextVQA	Semantic	Object-region masking and interaction-region perturbation
	OCR/Text	OCR-region masking, character corruption, localized text blurring
	Appearance	Color perturbation and texture modification
	Background	Contextual scene replacement and environmental masking
MM-IMDb	Semantic	Plot-summary truncation, semantic phrase removal, narrative corruption
	Interaction	Character-relation sentence removal and interaction phrase masking
	Appearance	Poster recoloring, texture modification, visual style perturbation
	Background	Poster background replacement and contextual style masking
	Demographic	Face-region perturbation and identity-related appearance masking
CREMA-D	Facial Expression	Eye/mouth-region masking and expression suppression
	Vocal Emotion	Pitch perturbation, spectral masking, temporal audio corruption
	Motion	Temporal frame shuffle, motion blur, frame corruption
	Appearance	Clothing/color perturbation and illumination modification
	Background	Background scene masking and environmental audio suppression
RAVDESS	Facial Emotion	Facial-region masking and expression suppression
	Acoustic Tone	Pitch shifting, spectral masking, temporal audio perturbation
	Motion	Motion blur and temporal frame corruption
	Appearance	Clothing/style perturbation and illumination modification
	Context	Background suppression and contextual scene masking
Flickr30K	Semantic	Foreground object masking and semantic object removal
	Interaction	Contact-region masking and interaction-region occlusion
	Appearance	Clothing recoloring, texture perturbation, color jitter
	Background	Scene-context replacement while preserving foreground entities
	Demographic	Facial anonymization and identity-cue suppression

cause classification models do not involve autoregressive token generation, no temporal decoding aggregation is required. To reduce noise from individual prompts or samples, all reported GMR values are averaged across the evaluation set. Higher GMR indicates stronger instruction-conditioned redistribution of internal evidence reliance across taxonomy-aligned pathways.

D Training Objective

GUIDE is trained end-to-end using the original task objective while learning instruction-conditioned evidence modulation through grouped adaptation pathways and dynamic gating. Given input x, instruction I, and target output y, the model is optimized using the standard task loss:

$$\mathcal{L}_{\text{task}} = -\log p(y \mid x, I).$$

During training, only grouped low-rank adaptation parameters and gating networks are updated, while the pretrained backbone model remains frozen. This preserves the original multimodal capabilities of the pretrained model while enabling lightweight instruction-conditioned modulation of evidence usage.

The gating network is trained jointly with the task objective without explicit supervision over pathway assignments. Instead, GUIDE learns instruction-conditioned redistribution implicitly through task optimization across taxonomy-aligned grouped adaptation pathways. This formulation allows the model to discover instruction-sensitive evidence modulation strategies without requiring manually annotated evidence labels or disentangled internal representations.

To encourage stable utilization of multiple evidence pathways, we additionally apply entropy regularization over gate activations:

$$\mathcal{L}_{\text{gate}} = -\lambda \sum_{c=1}^{C} g_c \log g_c,$$

where g_c denotes the normalized activation of pathway group c and λ controls the strength of regularization. The final training objective is:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \mathcal{L}_{\text{gate}}.$$

Table 4: Representative instruction templates used for dataset-specific pathway conditioning.

Dataset	Pathway	Instruction Type	Example Instruction
GQA	Semantic	Focus	"Focus on actions and object interactions."
	Appearance	Focus	"Focus on visual appearance and clothing attributes."
	Appearance	Suppress	"Ignore clothing and appearance-related cues."
	Background	Suppress	"Ignore environmental context and focus on foreground interactions."
TextVQA	OCR/Text	Focus	"Rely primarily on textual evidence in the image."
	OCR/Text	Suppress	"Ignore textual content and rely on visual context."
	Appearance	Suppress	"Ignore visual appearance and focus on readable text."
MM-IMDb	Semantic	Focus	"Focus on narrative and plot-related information."
	Interaction	Focus	"Focus on character relationships and interactions."
	Appearance	Suppress	"Ignore poster style and rely on semantic content."
CREMA-D	Vocal Emotion	Focus	"Focus on vocal tone and acoustic emotion cues."
	Facial Expression	Suppress	"Ignore facial appearance and rely on speech emotion."
	Motion	Focus	"Focus on motion dynamics and temporal behavior."
RAVDESS	Acoustic Tone	Focus	"Focus on vocal tone and acoustic emotion cues."
	Motion	Focus	"Focus on temporal motion and behavioral dynamics."
	Appearance	Suppress	"Ignore clothing and appearance-related cues."
Flickr30K	Interaction	Focus	"Focus on human-object interactions."
	Appearance	Suppress	"Ignore clothing and appearance-related attributes."
	Background	Suppress	"Ignore background context and focus on foreground entities."

This objective encourages GUIDE to maintain task performance while learning instruction-conditioned redistribution of internal evidence reliance across taxonomy-aligned pathways.

E Extended Conflict Routing Results

To further evaluate whether GUIDE supports compositional evidence control beyond fixed instruction pairs, we analyze additional conflicting evidence-routing conditions across multimodal reasoning, classification, and generation tasks. Table 5 reports output consistency together with changes in pathway modulation (ΔGMR) and functional reliance (ΔRS) under multiple instruction conflicts.

Across datasets, GUIDE consistently preserves high output consistency despite substantial redistribution of internal evidence pathways. Different conflict combinations induce distinct magnitudes of pathway modulation depending on the targeted evidence categories. Conflicts involving semantic versus appearance preferences generally produce the largest routing shifts, whereas conflicts involving neutral or partially aligned evidence conditions yield comparatively smaller modulation.

Importantly, GUIDE maintains relatively stable task behavior even under highly conflicting evidence preferences across visual, textual, acoustic, and multimodal generation settings. These results further support that GUIDE enables continuous and compositional evidence control rather than relying on isolated prompt templates or single-pathway switching behavior.

F Instruction Conflict and Compositional Evidence Routing

We next evaluate whether GUIDE can resolve conflicting evidence preferences within a single instruction. To this end, we construct compositional instructions such as *"focus on semantic relationships while ignoring appearance"* and *"focus on appearance while suppressing semantic interactions."* Figure 10 shows semantic–appearance trade-off behavior on GQA, MM-IMDb, and Flickr30K, where λ controls the relative weighting between semantic-focused and appearance-focused evidence preferences. As λ shifts from appearance-oriented to semantic-oriented reasoning, GUIDE produces smooth redistribution between semantic and appearance pathways rather than abrupt switching behavior. Balanced-conflict conditions yield intermediate pathway activations instead of collapsing to a single dominant evidence source, indicating compositional redistribution of multimodal evidence. Despite substantial changes in internal routing, GUIDE maintains relatively stable task behavior, suggesting continuous and conflict-aware evidence control beyond fixed instruction templates.

G Layer-wise Evidence Routing

We next analyze how instruction-conditioned evidence routing evolves across transformer depth

Table 5: Same-answer behavior under conflicting evidence-routing instructions across multimodal tasks. GUIDE preserves output consistency (Flickr30K reports caption similarity (0–1), and shows high output consistency despite substantial changes in pathway modulation (ΔGMR) and functional reliance (ΔRS).

Dataset	Conflict Condition	Output Consistency	ΔGMR	ΔRS
GQA	Semantic ↔ Appearance	93.8	0.61	0.19
	Semantic ↔ Background Suppress	95.1	0.47	0.14
	Appearance ↔ Background Suppress	91.4	0.54	0.17
	Semantic ↔ Demographic Suppress	94.6	0.42	0.12
	Neutral ↔ Semantic	89.7	0.31	0.09
TextVQA	Semantic ↔ OCR/Text	92.6	0.57	0.18
	Semantic ↔ Appearance	91.8	0.51	0.16
	OCR/Text ↔ Background Suppress	93.4	0.49	0.15
	Appearance ↔ Background Suppress	90.8	0.46	0.14
	Neutral ↔ OCR/Text	88.9	0.30	0.09
MM-IMDb	Semantic ↔ Appearance	92.1	0.58	0.17
	Semantic ↔ Background Suppress	94.3	0.43	0.13
	Appearance ↔ Demographic Suppress	90.6	0.49	0.15
	Semantic ↔ Demographic Suppress	93.8	0.40	0.11
	Neutral ↔ Semantic	88.4	0.28	0.08
CREMA-D	Semantic ↔ Appearance	91.7	0.55	0.18
	Semantic ↔ Motion	89.9	0.46	0.16
	Motion ↔ Appearance	88.3	0.51	0.17
	Semantic ↔ Background Suppress	93.6	0.41	0.12
	Neutral ↔ Semantic	87.8	0.26	0.07
RAVDESS	Semantic ↔ Appearance	91.1	0.53	0.17
	Semantic ↔ Motion	89.4	0.44	0.15
	Motion ↔ Appearance	87.9	0.49	0.16
	Semantic ↔ Background Suppress	92.8	0.39	0.11
	Neutral ↔ Semantic	87.1	0.24	0.07
Flickr30K	Semantic ↔ Appearance	0.84	0.63	0.20
	Semantic ↔ Appearance Suppress	0.88	0.51	0.15
	Appearance ↔ Background Suppress	0.81	0.56	0.18
	Semantic ↔ Background Suppress	0.90	0.44	0.12
	Neutral ↔ Semantic	0.79	0.29	0.08

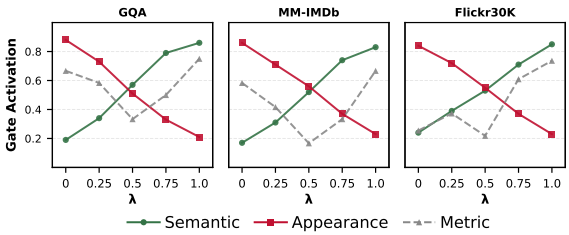

Figure 10: Instruction-conditioned semantic–appearance trade-off curves across GQA, MM-IMDb, and Flickr30K. As instruction emphasis shifts toward semantic reasoning, GUIDE smoothly redistributes pathway activations while maintaining relatively stable task performance. Dashed gray curves denote task consistency metrics (GQA, MM-IMDb) and CIDEr scores (Flickr30K).

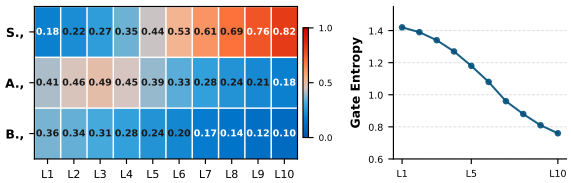

Figure 11: Layer-wise pathway modulation and gate entropy on GQA. y-labels on heatmap denote **S**emantic, **A**ppearance, and **B**ackground. Semantic pathway activations increase in deeper layers, whereas gate entropy progressively decreases, indicating increasingly selective evidence routing during generation.

during autoregressive generation. Figure 11 shows layer-wise pathway modulation on Flickr30K. Early decoding layers exhibit stronger appearance-related routing, whereas later layers increasingly emphasize semantic pathways. Intermediate layers show more balanced routing behavior, suggesting progressive integration of multimodal evidence before stronger semantic pathway dominance in deeper layers. In parallel, gate entropy steadily decreases across layers, indicating that pathway activations become increasingly specialized during generation. These results suggest that GUIDE does not induce isolated routing changes at a single layer,

but instead produces progressively structured evidence routing across transformer layers throughout the decoding process.

G.1 Generalization to Paraphrased and Compositional Instructions

To evaluate whether GUIDE generalizes beyond fixed instruction templates, we project Reliance Sensitivity (RS) vectors from template-based, paraphrased, and compositional instruction variants into a 2D space using t-SNE. As shown in Figure 12, semantic-intent instructions and appearance-intent instructions form consistently separated clusters across all benchmarks. Importantly, this separation remains stable across template, paraphrased, and compositional instruction variants, suggesting that GUIDE captures instruction-conditioned evidence preferences beyond narrow lexical formulations. These results further support that GUIDE generalizes instruction-conditioned evidence routing to semantically equivalent and compositionally novel instructions.

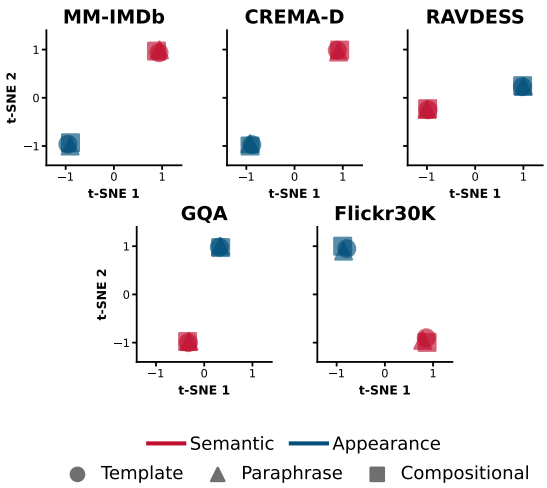

Figure 12: t-sne of semantic-intent (red) and appearance-intent instruction (blue) embeddings on each benchmark across various instruction types.

H Full Ablation Results

Table 6 reports full ablation results across all multimodal reasoning, classification, and generation benchmarks. Consistent trends are observed across datasets: removing taxonomy-aligned grouped pathways substantially reduces pathway modulation and functional reliance redistribution, while replacing dynamic instruction-conditioned gates with static routing decreases robustness under taxonomy-aligned perturbations. Prompt-only control consistently exhibits the weakest evidence modulation behavior, suggesting that GUIDE cannot be explained solely through conventional prompt steering or superficial decoding effects.

Table 6: Ablation study on grouped routing and dynamic gating. Removing taxonomy-aligned grouped pathways or replacing dynamic instruction-conditioned gates substantially reduces pathway modulation (GMR), functional reliance redistribution (RS), and perturbation robustness (Robust.) across all benchmarks. Robust. denotes counterfactual prediction stability under taxonomy-aligned perturbations.

Dataset	Variant	GMR↑	RS↑	Robust.↑
GQA	Full	**0.61**	**0.19**	**0.93**
	No grouped	0.22	0.07	0.81
	Static gates	0.38	0.10	0.86
	Prompt-only	0.18	0.06	0.79
TextVQA	Full	**0.59**	**0.18**	**0.92**
	No grouped	0.21	0.06	0.80
	Static gates	0.36	0.09	0.85
	Prompt-only	0.17	0.05	0.78
MM-IMDb	Full	**0.58**	**0.17**	**0.92**
	No grouped	0.20	0.06	0.80
	Static gates	0.35	0.09	0.85
	Prompt-only	0.17	0.05	0.78
CREMA-D	Full	**0.55**	**0.18**	**0.91**
	No grouped	0.19	0.06	0.79
	Static gates	0.33	0.09	0.84
	Prompt-only	0.16	0.05	0.77
RAVDESS	Full	**0.53**	**0.17**	**0.91**
	No grouped	0.18	0.06	0.78
	Static gates	0.31	0.08	0.83
	Prompt-only	0.15	0.05	0.76
Flickr30K	Full	**0.63**	**0.20**	**0.84**
	No grouped	0.23	0.07	0.73
	Static gates	0.39	0.11	0.78
	Prompt-only	0.19	0.06	0.71

I Additional Evidence Routing Visualizations

Figure 13–Figure 17 provide additional gate activation visualizations across multimodal classification and generation benchmarks. Consistent with the main GQA analysis, instruction-conditioned gating produces structured redistribution of pathway activations aligned with the intended evidence preference. Semantic-focused instructions increase reliance on semantic and interaction-related pathways, whereas appearance-focused instructions amplify appearance-related activations. Suppression-oriented instructions consistently reduce activation of the targeted evidence category while preserving task-relevant semantic routing. These trends remain consistent across multimodal reasoning, classification, and generation settings despite substantial differences in modality and task structure.

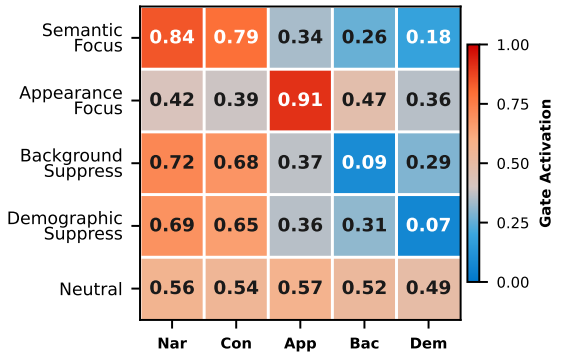

Figure 13: Gate activation heatmap on MM-IMDb under different instruction conditions. X-labels denote **Nar**rative, **Con**ceptual, **App**earance, **Bac**kground, **Dem**ographic respectively.

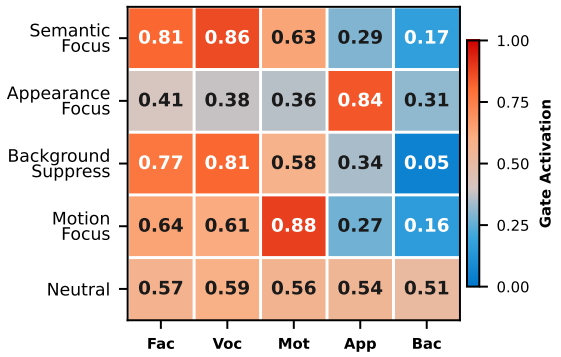

Figure 14: Gate Activation Heatmap on CREMA-D under different instruction conditions. X-labels denote **Fac**ial Expression, **Voc**al Emotion, **Mot**ion, **App**earance, **Bac**kground respectively.

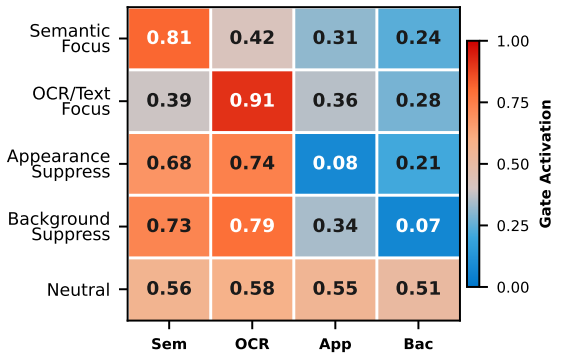

Figure 15: Gate Activation Heatmap on TextVQA under different instruction conditions. X-labels denote **Sem**antic, **OCR**/Text, **App**earance, **Bac**kground respectively

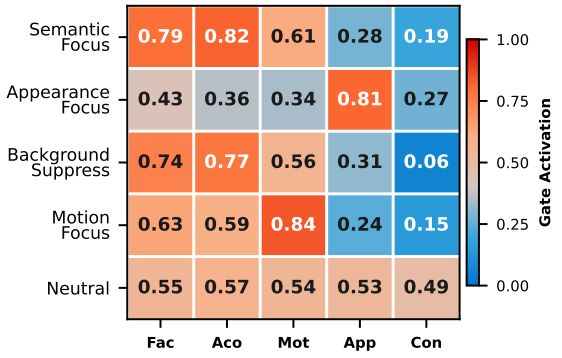

Figure 16: Gate Activation Heatmap on RAVDESS under different instruction conditions. X-labels denote **Fac**ial Emotion, **Aco**ustic Tone, **Mot**ion, **App**earance, **Con**text respectively.

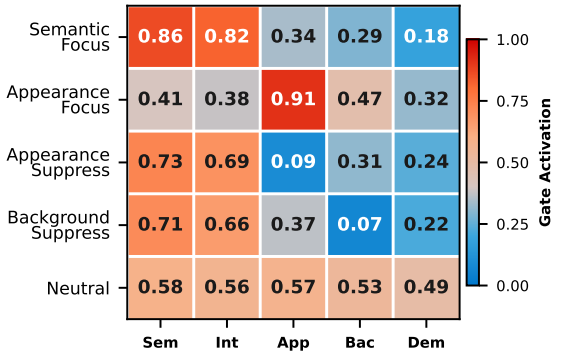

Figure 17: Gate Activation Heatmap on Flickr30K under different instruction conditions. X-labels denote **Sem**antic, **Int**eraction, **App**earance, **Bac**kground, **Dem**ographic respectively

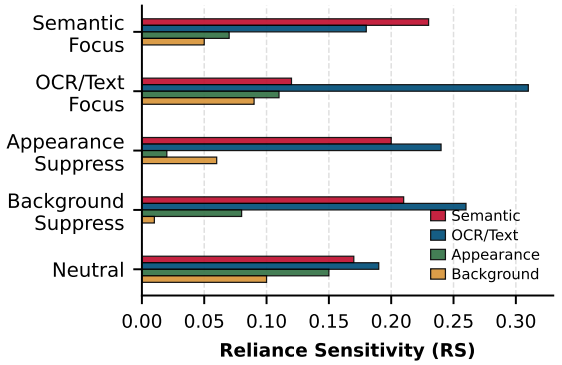

Figure 18: Reliance Sensitivity (RS) across evidence pathways under different instruction conditions on TextVQA. Higher RS indicates greater functional dependence on the corresponding evidence pathway.

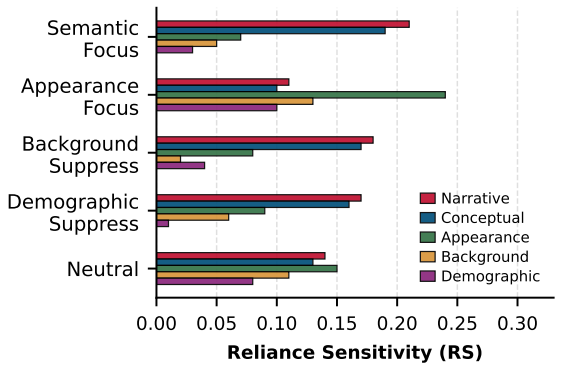

Figure 19: Reliance Sensitivity (RS) across evidence pathways under different instruction conditions on MM-IMDb. Higher RS indicates greater functional dependence on the corresponding evidence pathway.

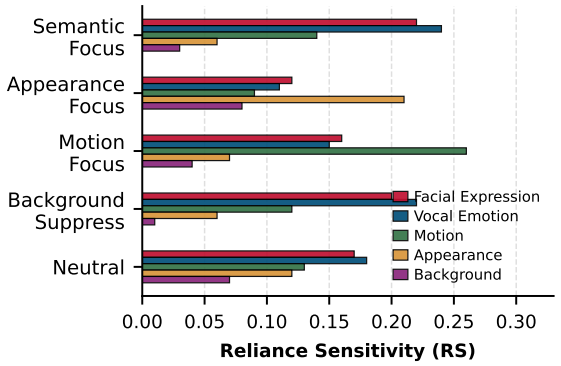

Figure 20: Reliance Sensitivity (RS) across evidence pathways under different instruction conditions on CREMA-D. Higher RS indicates greater functional dependence on the corresponding evidence pathway.

J Additional Functional Reliance Results

Figure 19–Figure 22 report additional Reliance Sensitivity (RS) analyses across multimodal classification and generation benchmarks. Consistent with the main GQA results, instruction-conditioned gating systematically redistributes functional dependence across different evidence pathways. Semantic-focused instructions increase sensitivity to semantic and interaction-related pathways, whereas appearance-focused instructions selectively amplify reliance on appearance-associated evidence. Suppression-oriented instructions consistently reduce dependence on the targeted evidence category while preserving sensitivity to task-relevant semantic pathways. These results further suggest that GUIDE modulates not only internal activations, but also the evidence pathways that functionally influence model predictions across diverse multimodal settings.

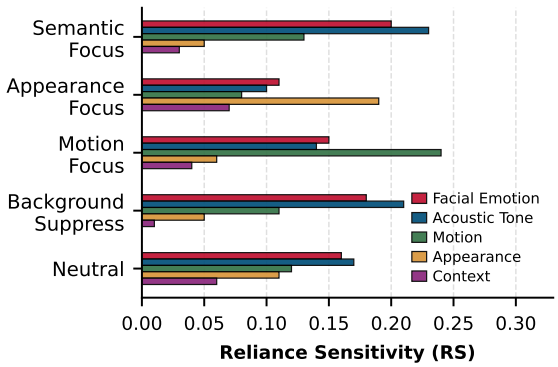

Figure 21: Reliance Sensitivity (RS) across evidence pathways under different instruction conditions on RAVDESS. Higher RS indicates greater functional dependence on the corresponding evidence pathway.

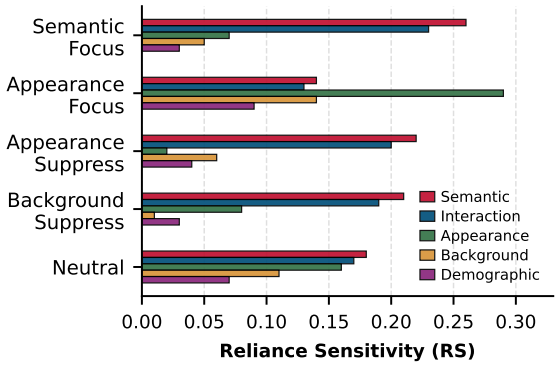

Figure 22: Reliance Sensitivity (RS) across evidence pathways under different instruction conditions on Flickr30K. Higher RS indicates greater functional dependence on the corresponding evidence pathway.

K Additional Perturbation Stability Analyses

Figure 23 reports additional perturbation stability analyses across multimodal reasoning, classification, and generation benchmarks. Consistent with the main GQA results, semantic-focused and suppression-oriented instructions remain comparatively robust under appearance and contextual perturbations, whereas appearance-focused instructions exhibit substantially larger performance degradation under appearance-related shifts. Across all datasets, object-level evidence removal produces the largest degradation, indicating that core semantic evidence remains fundamentally task-critical despite instruction-conditioned redistribution of auxiliary evidence pathways.

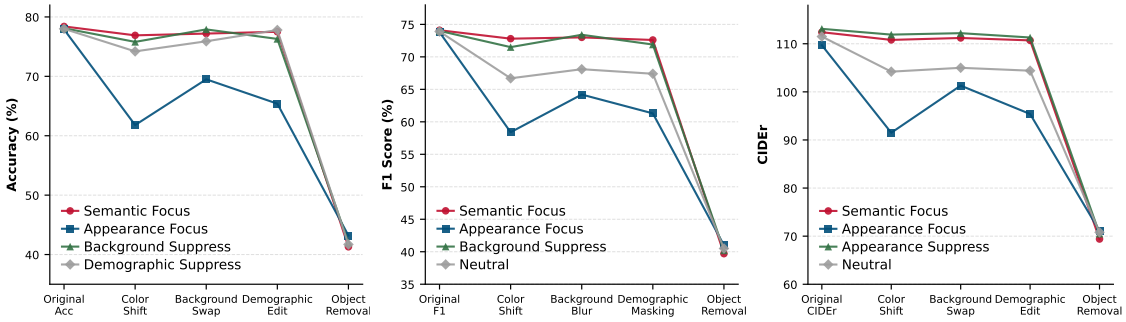

Figure 23: Perturbation Stability Analysis on GQA, MM-IMDb, and Flickr30K.

Table 7: Dataset Statistics.

Dataset	Task	Modality	Train	Val	Test	Evidence Pathway
GQA	Reasoning	Vision–Language	943,000	132,062	12,578	Semantic, Appearance, Background, Demographic
TextVQA	Reasoning	Vision–Language	34,602	5,000	5,734	Semantic, OCR/Text, Appearance, Background
MM-IMDb	Classification	Vision–Language	2,142	1,071	1,072	Semantic, Appearance, Background, Demographic
CREMA-D	Classification	Audio–Visual	5,152	736	1,472	Semantic, Motion, Appearance, Background
RAVDESS	Classification	Audio–Visual	1,470	490	491	Semantic, Motion, Appearance, Background
Flickr30K	Captioning	Vision–Language	28,998	1,024	992	Semantic, Appearance, Background

L Implementation Details

Table 7 summarizes the datasets used throughout our experiments, including task type, modality, dataset splits, and the high-level evidence pathway categories associated with each benchmark.

Unlike conventional parameter-efficient adaptation methods that rely on a single shared adaptation space, GUIDE organizes internal routing into semantically meaningful evidence pathways aligned with the dominant multimodal structure of each task. These pathways define the controllable evidence categories that instructions can selectively emphasize or suppress during reasoning and generation.

We use a shared high-level taxonomy across tasks consisting of semantic, appearance, background, motion, OCR/text, and demographic evidence categories when applicable. While the overall routing framework remains consistent, the fine-grained interpretation of each pathway depends on the underlying modality and benchmark. For example, semantic pathways in GQA primarily correspond to object interactions and relational reasoning, whereas semantic pathways in CREMA-D and RAVDESS correspond more strongly to emotional and acoustic cues. Similarly, OCR/text pathways are primarily relevant in TextVQA, while motion pathways are particularly important in audio–visual emotion recognition benchmarks.

This benchmark-specific pathway organization enables GUIDE to support controllable evidence modulation across heterogeneous multimodal settings while maintaining a unified routing mechanism.

L.1 Pretrained Multimodal Backbones

We evaluate GUIDE across multiple pretrained multimodal foundation models spanning both vision–language and audio–visual settings.

For vision–language benchmarks (GQA, Flickr30K, TextVQA, and MM-IMDb), we instantiate GUIDE on top of Qwen3-VL-30B, Qwen3.5-VL-37B, and InternVL3.5-38B. For audio–visual benchmarks (CREMA-D and RAVDESS), we use Qwen3-Omni-30B and Ola-7B as multimodal backbones.

These models serve as the underlying multimodal reasoning architectures on top of which grouped adaptation and instruction-conditioned routing are applied. We additionally observe that the overall evidence-routing behavior and reliance modulation patterns remain qualitatively consistent across different pretrained multimodal backbones.

Table 8: Training hyperparameters used for GUIDE across multimodal benchmarks. The number of pathway groups is determined by the dataset-specific evidence taxonomy.

Dataset	Backbone	Groups	Rank	LR	Batch	Epochs	λ
GQA	Qwen3-VL-30B	6	8	2×10^{-4}	32	5	0.01
TextVQA	Qwen3-VL-30B	5	8	2×10^{-4}	32	5	0.01
MM-IMDb	InternVL3.5-38B	5	8	1×10^{-4}	64	10	0.01
CREMA-D	Qwen3-Omni-30B	5	8	1×10^{-4}	32	10	0.01
RAVDESS	Ola-7B	5	8	1×10^{-4}	32	10	0.01
Flickr30K	Qwen3.5-VL-37B	5	8	2×10^{-4}	32	5	0.01

M Computational Experiments

All experiments are conducted using grouped parameter-efficient adaptation on top of frozen pretrained multimodal backbones. Training and evaluation are performed using standard GPU-based distributed training environments. Because GUIDE modifies only lightweight grouped LoRA parameters and gating modules while keeping backbone parameters frozen, the computational overhead remains substantially smaller than full model finetuning. Detailed hyperparameters, backbone configurations, and implementation details are provided in the appendix.

M.1 Training Details

Pretrained multimodal backbone parameters remain frozen during training, and only grouped LoRA parameters together with the instruction-conditioned gating network are updated. We use AdamW for optimization with learning rates selected from $\{1 \times 10^{-4}, 2 \times 10^{-4}\}$ based on validation performance. The LoRA rank is fixed to 8 across all experiments, and gate activations are normalized using softmax normalization. Models are trained using mixed precision with batch sizes between 32 and 64 depending on the benchmark and backbone size. For compositional routing analyses, the semantic–appearance tradeoff coefficient λ is varied linearly between 0 and 1 during inference. All hyperparameters are selected on validation splits and kept fixed across instruction variants within each benchmark.

M.2 License for Artifacts

All experiments are conducted using publicly available datasets and pretrained multimodal foundation models subject to their respective licenses and terms of use. We release only the code necessary to reproduce the GUIDE routing framework, training procedures, and evaluation protocols. Users are responsible for ensuring compliance with the licenses and usage restrictions associated with the underlying datasets and pretrained backbones.

Table 9: Human Evaluation Results. Each field is rated on a Likert scale of 1–4.

Condition	Alignment	Preservation
Semantic Focus	3.76	3.63
Appearance Focus	3.82	3.67
Background Suppress	3.59	3.47
Grand Total	**3.72**	**3.59**

N Human Evaluation

N.1 Evaluation Setup

We conducted our human evaluation process with eleven annotators, including one undergraduate student, one graduated assistant, four Master's students, three PhD students, and two Postdoctoral researchers. These annotators were recruited based on their expertise in the relevant research to ensure high-quality annotations. The human evaluation was conducted on a voluntary basis, and no monetary compensation was provided. Before participation, we provided clear instructions explaining the purpose of the evaluation and how the collected data would be used. The annotators were informed that their feedback would contribute to the assessment and refinement of our system's performance, and they participated with full awareness and consent.

N.2 Evaluation Results

We conducted a human evaluation study on Flickr30K to assess whether instruction-conditioned evidence control remains semantically coherent during generation. Annotators were given an image, instruction, and generated caption, and asked to rate: (i) instruction alignment and (ii) semantic preservation on a 4-point scale. The evaluation was conducted on 50 randomly sampled examples across different instruction conditions.

Table 9 reports the average human ratings. All instruction conditions achieve high alignment and preservation scores, indicating that GUIDE produces captions that both follow the intended evidence preference and preserve overall scene semantics. Importantly, semantic preservation remains consistently high despite substantial differences in the routing of instruction-conditioned evidence.

O Additional Qualitative Examples

We provide additional qualitative examples illustrating how GUIDE alters internal evidence reliance under different instruction conditions while preserving the underlying task setting. Figure 8 and Figure 9 show representative examples from GQA and TextVQA under baseline, semantic-focused, and appearance-focused instructions.

Without explicit instructions, the model exhibits comparatively broad and mixed evidence usage across both semantic and appearance-related regions. Under semantic-focused instructions, GUIDE increases reliance on semantically meaningful and task-relevant evidence regions. For GQA, the model emphasizes actions, object interactions, and relational cues relevant to answering the question. For TextVQA, semantic-focused routing increases reliance on OCR-relevant text regions and localized textual evidence required for accurate prediction.

In contrast, appearance-focused instructions redirect evidence routing toward visual appearance, contextual layout, stylistic patterns, clothing attributes, and other superficial visual cues that are less directly relevant to the underlying reasoning objective. These routing changes frequently alter prediction behavior despite identical inputs and task semantics.

Importantly, the observed differences are not limited to superficial attention visualization. The corresponding evidence emphasis patterns and prediction outputs demonstrate that instruction-conditioned routing produces functionally meaningful changes in multimodal evidence reliance. In several cases, baseline routing produces diffuse or unstable evidence usage, whereas semantic-focused routing improves evidence localization and prediction accuracy.

Overall, these qualitative examples further support the quantitative findings presented in the main paper, demonstrating that GUIDE systematically modulates how multimodal models internally use evidence during reasoning and generation.