GUIDE: Guiding Internal Evidence with Language Instructions

Soyeon Caren Han[1*†], **Hyunsuk Chung**[1*], **Jinwoo Kim**[1‡],
Seungyeon Ji[2,3‡], **Kyungreem Han**[2,4†]

[1]School of Computing and Information Systems, The University of Melbourne, Australia
[2]Brain Science Institute, Korea Institute of Science and Technology, Korea
[3]Department of Computer Science and Engineering, Korea University, Korea
[4]Division of Bio-Medical Science & Technology,
University of Science and Technology KIST School, Korea

{caren.han, hyunsuk.chung.1, jinwoo.kim.1}@unimelb.edu.au
{syji, khan}@kist.re.kr

Abstract

Large multimodal models follow instructions about what to generate, but not necessarily about what evidence to rely on. Hence, models may continue to depend 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 modulate multimodal evidence pathways during reasoning and generation. We further introduce a pathway-level evaluation framework that characterizes instruction-conditioned evidence modulation through reliance sensitivity, controlled perturbation analysis, pathway modulation, and autoregressive decoding dynamics. Across multimodal reasoning, classification, and generation, GUIDE induces structured and instruction-aligned redistribution of evidence reliance while largely preserving task behavior. Experiments on GQA, TextVQA, MM-IMDb, CREMA-D, RAVDESS, and Flickr30K show that GUIDE improves robustness under targeted evidence perturbations and enables controllable modulation across diverse multimodal settings. This suggests that multimodal instruction following can extend beyond output control toward regulating how different evidence sources contribute to model predictions.

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 depend on appearance, demographic, or contextual cues while producing the same answer. Similarly, instructions such as *"ignore appearance"* or *"rely on textual evidence"* may change model outputs without systematically changing the evidence sources that contribute to the prediction. This limitation is particularly important in multimodal settings, where multiple evidence sources can independently support the same task outcome. In tasks such as GQA and TextVQA, models can rely on object relationships, OCR signals, visual attributes, or background context to produce identical answers. Moreover, shortcut-associated cues are often highly predictive under 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 contributes to model predictions. Recent progress in multimodal instruction following has primarily focused on prompt engineering, instruction tuning, and controllable generation. While these approaches improve output controllability, they provide limited mechanisms for explicitly regulating instruction-conditioned evidence reliance (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 evidence pathways used by multimodal models? To address this problem, we introduce **GUIDE**, a framework for controlling multimodal evidence usage with language instructions. GUIDE combines grouped parameter-efficient adaptation with instruction-conditioned gating to dynamically modulate task-relevant and taxonomy-aligned evidence

[*]Equal contribution.

[†]Corresponding authors: Soyeon Caren Han caren.han@unimelb.edu.au, Kyungreem Han khan@kist.re.kr

[‡]These authors contributed equally as second authors.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 52363–52384
October 24-29, 2026 ©2026 Association for Computational Linguistics

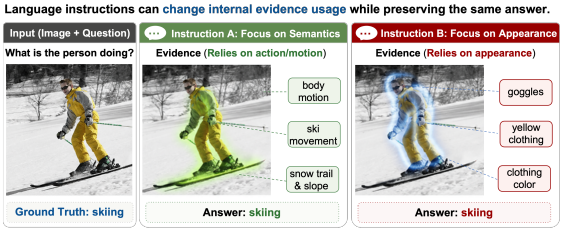

Figure 1: Motivation of GUIDE. Given the same input image and question, different language instructions encourage reliance on different internal evidence pathways while producing the same answer. Semantic-focused instructions emphasize action and interaction cues, whereas appearance-focused instructions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

pathways. Rather than modifying task objectives or input representations, GUIDE operates through grouped internal adaptation pathways to regulate the contribution of multimodal evidence during reasoning and generation. Importantly, GUIDE enables models to produce similar outputs while exhibiting substantially different pathway activation and functional reliance patterns. For example, semantic-focused instructions increase reliance on action and interaction pathways, whereas appearance-focused instructions increase reliance on visual attribute pathways. Similar to this, in TextVQA, instructions that emphasize textual evidence increase reliance on OCR-related pathways while reducing dependence on contextual appearance cues.

To study this behavior, we introduce a pathway-level 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 evidence reliance while largely preserving task behavior. We evaluate GUIDE on GQA, TextVQA, MM-IMDb, CREMA-D, RAVDESS, and Flickr30K. Experiments further show improved robustness under targeted evidence perturbations. Our main contributions are as follows: **1)** We introduce GUIDE, a framework for instruction-conditioned control of multimodal evidence usage through grouped adaptation pathways and dynamic gating. **2)** We provide pathway-level and perturbation-based evidence that language instructions systematically redistribute model reliance across task-relevant multimodal evidence sources. **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 multimodal instruction following can extend beyond output control toward regulating how different evidence sources contribute to model predictions.

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.

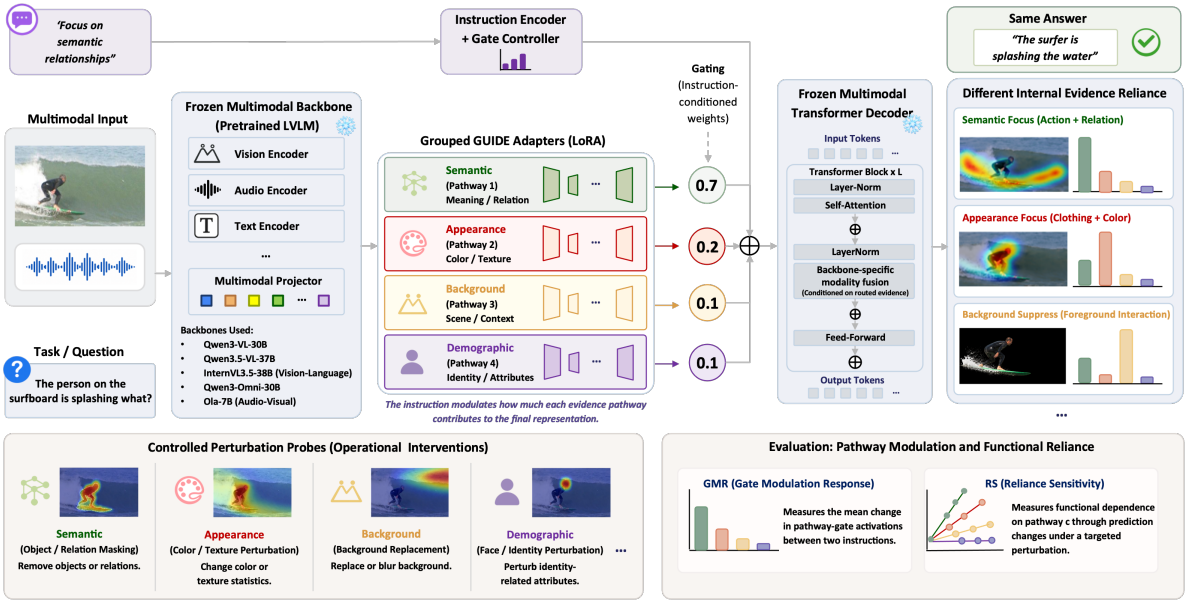

Figure 2: Overview of GUIDE, an instruction-conditioned evidence routing framework for controllable multimodal reasoning. GUIDE dynamically modulates grouped adaptation pathways corresponding to different evidence categories through instruction-conditioned gating, enabling the same prediction to be produced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

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), 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 object relationships, OCR signals, appearance attributes, or contextual background cues. Our goal is therefore not merely to control outputs, but to modulate the evidence pathways contributing to them. Specifically, we seek instruction-conditioned evidence modulation such that $f_\theta(x, I_a) \approx f_\theta(x, I_b)$ while exhibiting substantially different pathway activation and functional reliance patterns 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 organize internal adaptation pathways into task-relevant groups associated with different forms of multimodal evidence. These include semantic interactions, visual appearance, environmental context,

motion dynamics, acoustic signals, demographic cues, and OCR/textual grounding information depending on the benchmark. While the routing mechanism is shared across tasks, the pathway taxonomy is adapted to the dominant evidence structure of each dataset. High-level evidence families may be instantiated as multiple fine-grained operational routing groups; for example, GQA decomposes semantic evidence into object, relation, and action groups, while appearance evidence includes an attribute group. Each operational pathway corresponds to a dedicated grouped low-rank adaptation branch attached to shared transformer projections. Pathway assignments follow 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 $W' = W + BA$, where A and B are trainable low-rank matrices. Instead of using a single shared adaptation pathway, GUIDE decomposes the update into multiple taxonomy-aligned adaptation groups:

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

where c indexes operational routing groups, $B_c A_c$ denotes the corresponding low-rank adaptation, and $g_c(I, h_t)$ is an instruction-conditioned gate predicted from the instruction representation and current model state. The gating mechanism dynamically modulates the contribution of each grouped adaptation pathway during reasoning and generation. Unlike static parameter-efficient adaptation, this formulation allows language instructions to selectively amplify or suppress specific pathways at inference time. GUIDE does not require explicit disentanglement of internal representations; instead, grouped adaptation provides a structured intervention space for modulating evidence reliance while preserving the underlying task objective.

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:

$$g(I, h_t) = \sigma(W_g[e_I; h_t])\,,$$

where $g(I, h_t) \in [0, 1]^C$ contains independently activated gates over C taxonomy-aligned routing groups. Unlike mutually exclusive routing, these gates allow multiple evidence pathways to be active simultaneously while enabling instructions to selectively increase or suppress individual pathway contributions. Because gating depends on both the instruction and the evolving model state, pathway activations can change throughout reasoning and generation. For autoregressive tasks, gates are computed at each decoding step and shared across grouped adaptation modules within the corresponding transformer layer. GUIDE is trained 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 alter evidence usage, we use two complementary measures: Gate Modulation Response (GMR) and Reliance Sensitivity (RS).

Gate Modulation Response (GMR). GMR measures how strongly pathway activations respond to instruction changes. Given two instructions I_a and I_b, we compute:

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

For autoregressive tasks, pathway differences are averaged across decoding steps and gated transformer layers. For classification tasks, activations are aggregated from the final prediction representation. Additional implementation details are provided in Appendix C. Higher GMR indicates stronger instruction-conditioned pathway modulation.

Reliance Sensitivity (RS). While GMR measures pathway modulation, it does not necessarily imply functional influence on predictions. We therefore use RS to measure how strongly model predictions depend on evidence associated with a given pathway. For pathway group c,

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

where Δ_c denotes a controlled perturbation targeting evidence associated with pathway c. Perturba-

tions include modality-specific masking, corruption, replacement, and suppression operations designed to selectively degrade the targeted evidence source while preserving overall task solvability. Detailed protocols are provided in Table 3.

When comparing two instruction conditions, we summarize the change in functional reliance as

$$\Delta\mathrm{RS}(I_a, I_b) = \frac{1}{C}\sum_{c=1}^{C} \left|\mathrm{RS}_c(I_a) - \mathrm{RS}_c(I_b)\right|.$$

Higher RS_c indicates greater functional dependence on pathway c, while higher $\Delta\mathrm{RS}$ indicates stronger redistribution of functional reliance between instructions. Together, GMR and RS characterize complementary aspects of instruction-conditioned evidence modulation: GMR measures changes in pathway activation, whereas RS measures corresponding changes in functional dependence.

4 Evaluation Setup

We evaluate whether GUIDE enables controllable modulation of multimodal evidence usage while preserving task semantics. To assess both controllability and generalization, we consider multimodal reasoning, classification, and generation settings spanning diverse multimodal tasks. All experiments use grouped parameter-efficient adaptation with 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 structured and perturbable evidence categories.

Multimodal Reasoning. We evaluate on GQA (Hudson and Manning, 2019) and TextVQA (Singh et al., 2019). GQA enables analysis of semantic, relational, appearance, and contextual evidence under controlled visual reasoning settings, while TextVQA provides a complementary setting requiring explicit 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 evidence sources and 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 complementary setting in which evidence modulation must persist throughout autoregressive decoding rather than affecting only a single prediction step. We follow 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 fixed lexical formulations. Representative instruction templates are provided in Appendix B.

4.3 Evidence Perturbation Protocol

To measure functional reliance on different evidence sources, we use 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 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-modulation metrics. To quantify instruction-conditioned evidence modulation, we report: (i) Gate Modulation Response (GMR), which measures changes in pathway activations between instructions, and (ii) Reliance Sensitivity (RS), which measures functional dependence on targeted

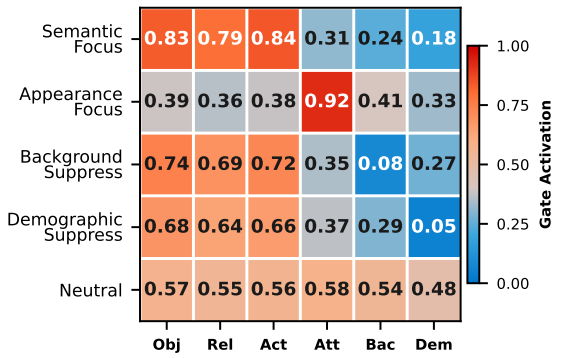

Figure 3: Gate activation heatmap on GQA under different instruction conditions. Obj, Rel, Act, Att, Bac, Dem denote **Obj**ect, **Rel**ation, **Act**ion, **Att**ribute, **Bac**kground, **Dem**ographic respectively

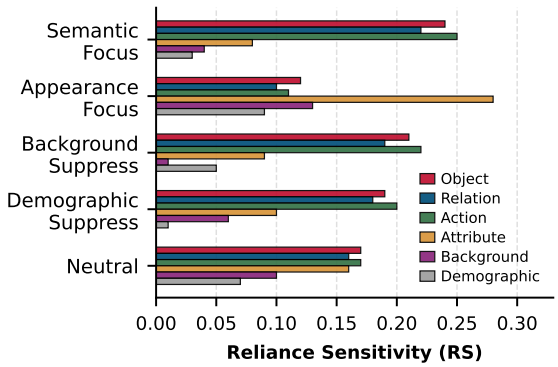

Figure 4: RS under different instruction conditions on GQA. Instructions selectively redistribute functional dependence across evidence pathways.

evidence sources under controlled perturbations. GMR and RS provide complementary measures of pathway activation and functional evidence reliance. We additionally report standard task-specific metrics for each benchmark to evaluate task preservation under instruction-conditioned evidence modulation.

5 Results

5.1 Controlled Evidence Routing

We first evaluate whether instruction-conditioned gating produces structured modulation of 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 increase activation of appearance-related pathways. Similarly, suppression-oriented instructions selectively reduce activation of the targeted pathway while preserving task-relevant pathways. Neutral instructions produce comparatively balanced activation patterns, indicating less pathway-specific modulation in the absence of an explicit evidence preference. These instruction-conditioned patterns remain consistent across multimodal reasoning, classification, and generation settings despite differences in modality and task structure. Overall, GUIDE induces structured and instruction-consistent modulation of evidence pathways rather than only superficial prompt-conditioned output variation.

5.2 Functional Evidence Redistribution

We next evaluate whether instruction-conditioned pathway modulation is accompanied by corresponding changes in functional evidence 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 tasks, RS patterns consistently align with the evidence preferences specified by the instructions. Semantic-focused instructions increase prediction sensitivity to semantic, relational, and interaction-related evidence, whereas appearance-focused instructions increase sensitivity to appearance-related evidence. Suppression-oriented instructions reduce sensitivity to the targeted evidence source while largely preserving dependence on task-relevant pathways. For example, background-suppression instructions consistently reduce functional dependence on contextual evidence without substantially reducing overall prediction sensitivity. These results show that instruction-conditioned pathway modulation is accompanied by systematic changes in the evidence sources on which model predictions functionally depend.

5.3 Counterfactual Evidence Control

We next evaluate whether instruction-conditioned evidence modulation is associated with corresponding changes in robustness under controlled perturbations. Figure 8 and 9 provide qualitative examples showing that different instructions induce distinct evidence emphasis patterns while preserving the same final prediction. Under semantic-focused instructions, the model places greater emphasis on action-relevant and semantically meaningful

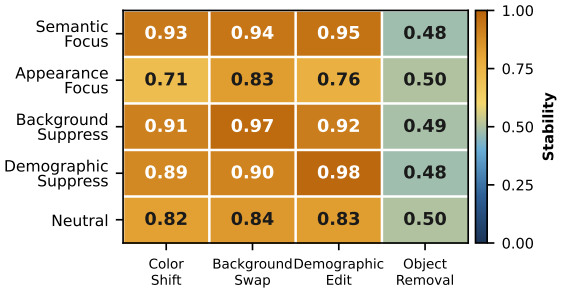

Figure 5: Counterfactual stability under different instruction conditions and perturbation types on GQA. Semantic-focused and background-suppression instructions remain robust under appearance and contextual perturbations, whereas object removal substantially reduces stability across all settings.

regions, whereas appearance-focused instructions emphasize 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 task-critical across routing preferences.

Figure 6 further 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 show that instruction-conditioned pathway modulation is accompanied by systematic changes in functional dependence on targeted evidence sources.

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 pathway activation dynamics throughout generation rather than only at initialization. Under semantic-focused instructions, semantic pathways remain strongly activated across decoding steps, resulting in captions

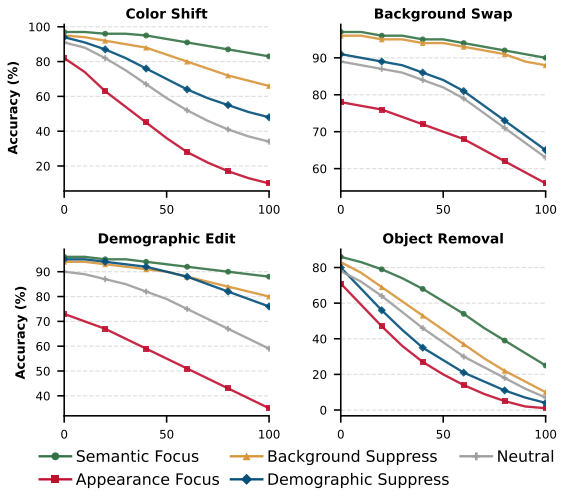

Figure 6: Prediction robustness under increasing perturbation strength on GQA. Appearance-focused instructions degrade most rapidly under color shifts, whereas semantic-focused and background-suppression instructions remain comparatively robust.

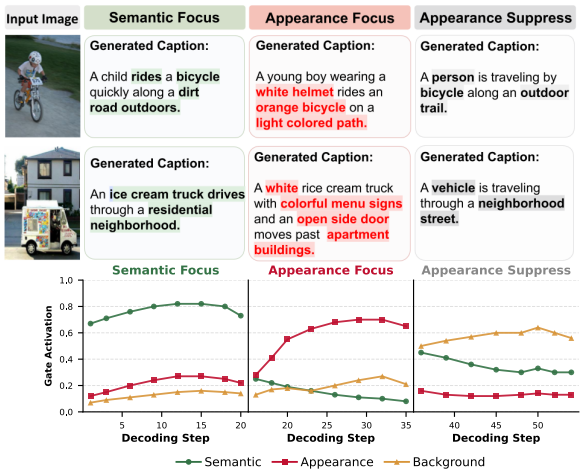

Figure 7: Instruction-conditioned decoding-time evidence routing on Flickr30K. Different instructions induce persistent changes in pathway activations throughout autoregressive generation, resulting in captions emphasizing either semantic/appearance attributes.

that prioritize actions and scene interactions. In contrast, appearance-focused instructions increase activation of appearance-related pathways, leading to captions that emphasize 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 providing fewer appearance-specific details.

Importantly, these instruction-conditioned activation patterns persist throughout autoregressive decoding, indicating that GUIDE modulates path-

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

Dataset	Output Consistency	GMR	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

way contributions during generation rather than only biasing the initial decoding state. The plotted gate trajectories correspond to representative decoding sequences and show 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 (GMR) and changes in functional reliance (ΔRS) across conflicting instruction conditions. Across all benchmarks, GUIDE maintains high output consistency despite substantial changes in pathway activation and functional reliance. 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.

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 changes in functional reliance (ΔRS), indicating that structured pathway decomposition plays an important role in controllable evidence modulation. Replacing dynamic instruction-conditioned gates with static routing also consistently decreases pathway modulation, functional reliance shifts, and perturbation robustness. Prompt-only control exhibits the weakest evidence-modulation behavior across all three settings. Overall, these results show that grouped adaptation pathways and dynamic instruction-conditioned gating jointly contribute to 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). Removing grouped pathways or replacing dynamic instruction-conditioned gates reduces pathway modulation (GMR), changes in functional reliance (ΔRS), and perturbation robustness (Robust.). Full benchmark results are provided in Appendix H.

Dataset	Variant	GMR↑	ΔRS↑	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 multimodal evidence usage through natural language instructions. By combining taxonomy-aligned grouped adaptation pathways with instruction-conditioned gating, GUIDE enables models to modulate different forms of evidence during reasoning and generation while largely preserving task behavior. Across multimodal reasoning, classification, and generation benchmarks, GUIDE produces consistent instruction-aligned pathway modulation, corresponding shifts in functional evidence reliance, and improved robustness under targeted evidence perturbations. We further show that similar outputs can be maintained under substantially different evidence-reliance patterns, and that these instruction-conditioned effects persist throughout autoregressive decoding. Together, these findings suggest that multimodal instruction following can extend beyond controlling what models output toward regulating how different evidence sources contribute to model predictions.

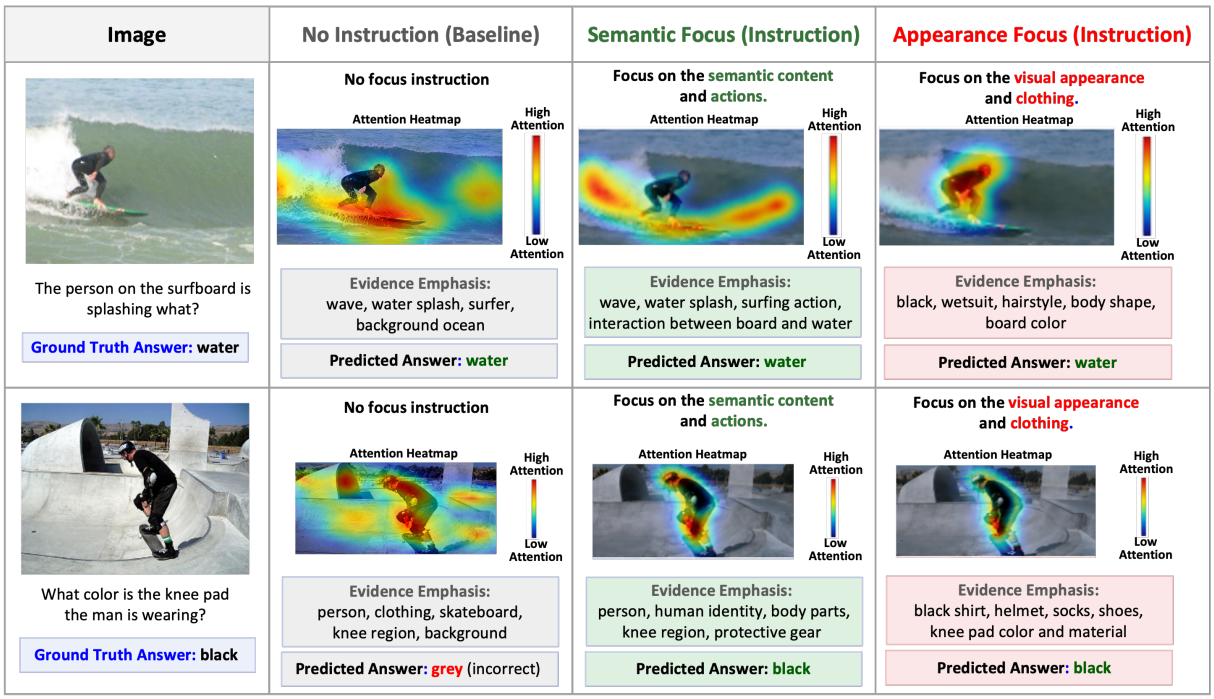

Figure 8: Qualitative examples of instruction-conditioned evidence routing on GQA. Semantic-focused instructions emphasize action-relevant and interaction-based evidence, whereas appearance-focused instructions prioritize visual attribute cues while preserving the same final prediction.

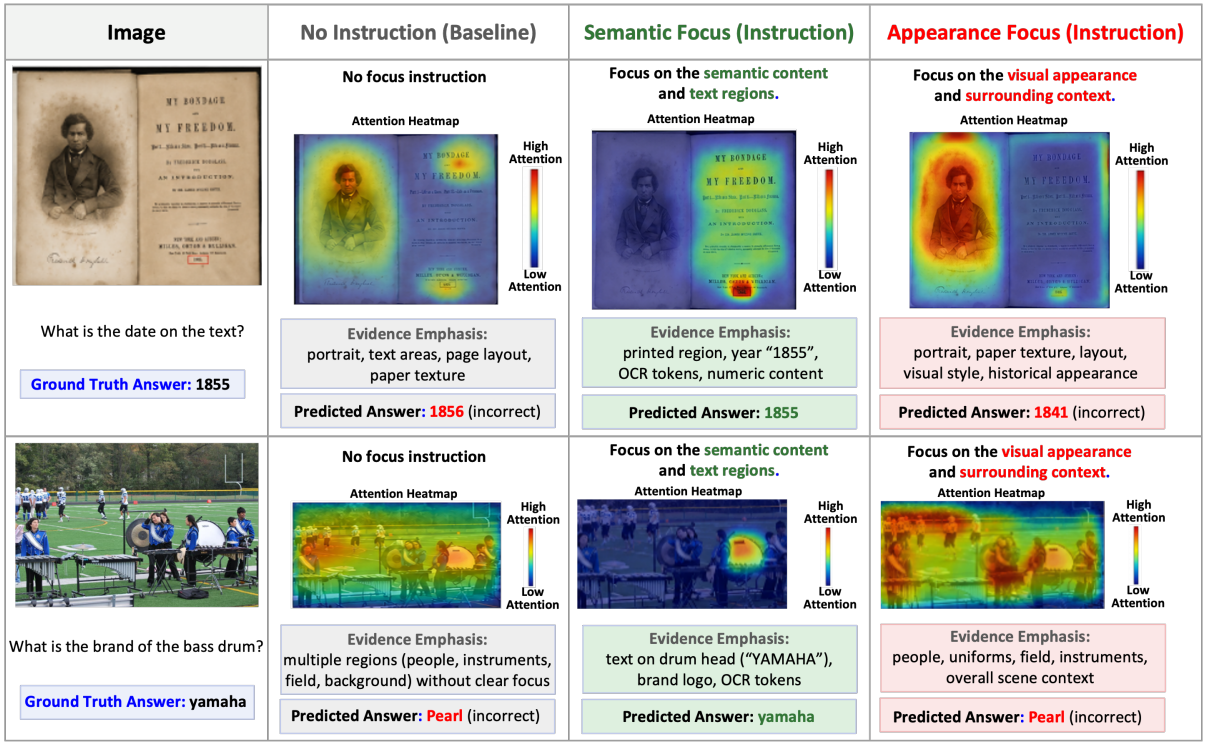

Figure 9: Qualitative examples of instruction-conditioned evidence routing on TextVQA. Semantic-focused instructions increase reliance on OCR and text-relevant regions, whereas appearance-focused instructions emphasize visual appearance and contextual cues, often leading to different prediction behavior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

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.

Acknowledgement

We thank Prof. Eduard Hovy and Prof. Sangwook Yi for their support of this research. This research was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2026-25519206, Development of human-like intelligent generative agents based on interactive multimodal reverse prompting), and IITP grant funded by the Korea government (MSIT) (No.RS-2025-02217259, Development of self-evolving AI bias detection-correction-explain platform based on international multidisciplinary governance).

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.

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 may contribute differently to model predictions. The taxonomy is constructed based on three principles: (1) identifying primary evidence sources relevant to successful task completion, (2) separating shortcut-associated or potentially confounding evidence categories, and (3) defining evidence categories that can be targeted through controlled perturbations. The taxonomy distinguishes between *high-level evidence families* and *operational routing groups*. A high-level family may be instantiated by multiple fine-grained routing groups depending on the benchmark. For example, in GQA, semantic evidence is decomposed into Object, Relation, and Action groups, while appearance-related evidence is represented by an Attribute group, together with Background and Demographic groups. This hierarchical organization allows GUIDE to retain a common evidence taxonomy while adapting the operational routing granularity to each task.

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 degradation rather than to simulate naturally occurring corruptions. For visual reasoning benchmarks such as GQA and Flickr30K, the high-level taxonomy includes semantic interaction cues, appearance attributes, environmental context, and person-related or demographic appearance signals. Semantic evidence primarily corresponds to object interactions, action understanding, and relational reasoning, while appearance evidence captures visual attributes such as clothing color, texture, and style. Background evidence captures environmental and scene-level contextual information that may support prediction. Person-related or demographic evidence captures facial and identity-associated appearance cues that may correlate spuriously with task outputs. For TextVQA, the taxonomy additionally includes an OCR/Text family corresponding to explicit textual grounding signals. This provides a distinct evidence source from object appearance or scene context and enables targeted analysis of textual versus visual reliance. For MM-IMDb, the taxonomy includes semantic narrative information, interaction-related character relationships, poster appearance cues, contextual visual style, and demographic-related appearance signals. 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 includes facial-expression signals, acoustic emotion cues, motion dynamics, appearance information, and contextual background signals. This organization enables analysis of how language instructions modulate reliance across visual, acoustic, and temporal evidence while preserving emotion classification behavior. Importantly, these evidence categories are not treated as perfectly independent causal mechanisms. Instead, they provide structured operational partitions for pathway modulation and perturbation-based analysis of functional evidence reliance across diverse multimodal settings.

B Instruction Taxonomy and Pathway Conditioning

GUIDE uses dataset-specific instruction templates to provide semantically consistent conditioning signals for different evidence preferences 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 during training and evaluation. Multiple paraphrased variants were constructed for each evidence preference to reduce sensitivity to fixed prompt wording. Base instruction templates were used during both training and evaluation, while paraphrased variants were additionally evaluated to assess generalization beyond fixed lexical formulations.

C GMR Computation Details

GMR measures the magnitude of instruction-conditioned modulation across taxonomy-aligned routing groups. For each input sample, we compute gate activations under two 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 ver-

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

sus appearance-focused prompts, depending on the benchmark.

For autoregressive generation tasks, gate activations are recorded at each decoding step and across gated transformer layers. Pathway-wise activation differences between the two instruction conditions are first averaged over decoding steps within each layer and then aggregated across gated layers. The resulting pathway differences are averaged over the C operational routing groups, consistent with the GMR definition in Section 3. For classification tasks, gate activations are extracted from the final prediction representation and aggregated across gated transformer layers; no temporal decoding aggregation is required.

All reported GMR values are finally averaged across the evaluation set to reduce variation from individual samples or instruction instances. Higher GMR indicates stronger instruction-conditioned changes in pathway activation patterns.

D Training Objective

GUIDE is trained 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 the grouped low-rank adaptation parameters and gating networks are updated, while the pretrained backbone remains frozen. This preserves the pretrained multimodal capabilities while enabling lightweight instruction-conditioned modulation through the added adaptation pathways.

The gating network is trained jointly with the task objective without explicit supervision over pathway activations. Instead, GUIDE learns instruction-sensitive pathway modulation through task optimization across taxonomy-aligned grouped adaptation pathways, without requiring

Table 4: Representative instruction templates used for dataset-specific evidence-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."

manually annotated evidence labels or explicitly disentangled internal representations.

Because GUIDE uses independent sigmoid gates, each pathway activation $g_c \in [0, 1]$ is treated independently rather than being normalized across pathways. To encourage confident and selective pathway activation, we apply a Bernoulli entropy regularization term:

$$\mathcal{L}_{\text{gate}} = -\lambda_{\text{gate}} \sum_{c=1}^{C} \left[g_c \log g_c + (1 - g_c) \log(1 - g_c) \right],$$

where λ_{gate} controls the strength of gate regularization. The final training objective is:

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

This objective preserves the original task supervision while encouraging selective instruction-conditioned modulation of taxonomy-aligned routing groups.

E Extended Conflict Routing Results

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

Across datasets, GUIDE consistently preserves high output consistency despite substantial changes in pathway activation and functional reliance. Different instruction combinations induce distinct magnitudes of pathway modulation depending on the targeted evidence categories. Contrasts between semantic- and appearance-focused instructions generally produce the largest modulation, whereas neutral or partially aligned instruction pairs yield comparatively smaller changes.

Importantly, GUIDE maintains relatively stable task behavior even under strongly contrasting evidence preferences across visual, textual, acoustic, and generation settings. These results further support that GUIDE enables continuous and compositional evidence modulation 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 accommodate contrasting 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 α controls the relative emphasis

between semantic-focused and appearance-focused evidence preferences. As α shifts from appearance-oriented to semantic-oriented reasoning, GUIDE produces smooth changes in semantic and appearance pathway activations rather than abrupt switching behavior. Balanced conditions yield intermediate activation patterns across the two pathway types, indicating compositional evidence modulation rather than exclusive single-pathway selection. Despite substantial changes in pathway activation, GUIDE maintains relatively stable task behavior, supporting continuous and conflict-aware evidence modulation beyond fixed instruction templates.

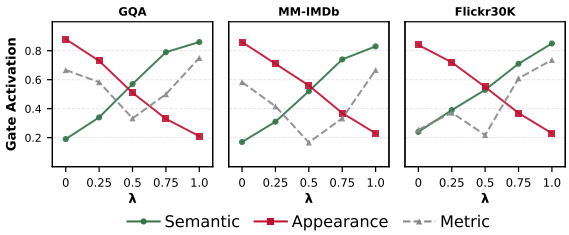

Figure 10: Instruction-conditioned semantic–appearance trade-off curves across GQA, MM-IMDb, and Flickr30K. As instruction emphasis shifts toward semantic reasoning, GUIDE smoothly modulates semantic and appearance pathway activations while maintaining relatively stable task performance. Dashed gray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

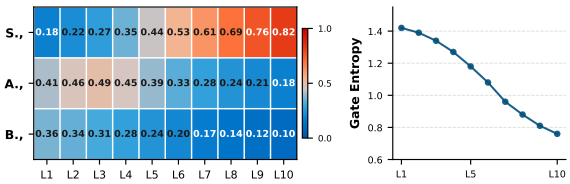

Figure 11: Layer-wise pathway modulation and gate entropy on Flickr30K. Gate entropy denotes the summed Bernoulli entropy across independent sigmoid gates. The heatmap shows Semantic, Appearance, and Background pathway activations. Semantic pathway activation increases in deeper layers, while gate entropy progressively decreases, indicating increasingly selective pathway activation across decoding layers.

G Layer-wise Evidence Routing

We next analyze how instruction-conditioned pathway modulation evolves across transformer depth during autoregressive generation. Figure 11 shows layer-wise pathway activations on Flickr30K. Earlier layers exhibit stronger appearance-related activation, whereas later layers increasingly emphasize semantic pathways. Intermediate layers show more balanced activation patterns, suggesting progressive integration of multimodal evidence before stronger semantic emphasis in deeper layers. In parallel, gate entropy steadily decreases across layers, indicating increasingly selective pathway activation during generation. Overall, these results show that GUIDE produces structured instruction-conditioned modulation across transformer depth rather than isolated changes at a single layer.

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 and appearance-intent instructions form consistently separated clusters across benchmarks. Importantly, this separation remains stable across template, paraphrased, and compositional variants, suggesting that GUIDE captures instruction-conditioned evidence preferences beyond fixed lexical formulations. These results further support that GUIDE generalizes instruction-conditioned evidence modulation to semantically equivalent and compositionally novel instructions.

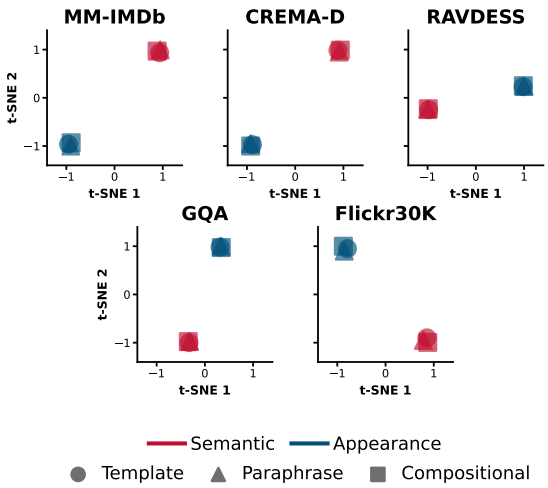

Figure 12: t-SNE of Reliance Sensitivity (RS) vectors for semantic-intent (red) and appearance-intent (blue) instructions across benchmarks and instruction types.

H Full Ablation Results

Table 6 reports full ablation results across all multimodal reasoning, classification, and generation benchmarks. Consistent trends are observed across

Table 5: Same-answer behavior under contrasting evidence-routing instructions across multimodal tasks. GUIDE preserves high output consistency despite substantial pathway modulation (GMR) and changes in functional reliance (ΔRS). Flickr30K reports caption similarity (0–1), while other datasets report same-prediction rate (%).

Dataset	Instruction Pair	Output Consistency	GMR	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

datasets: removing taxonomy-aligned grouped pathways substantially reduces pathway modulation and changes in functional reliance, while replacing dynamic instruction-conditioned gates with static routing also decreases perturbation robustness. Prompt-only control consistently exhibits the weakest evidence-modulation behavior, further indicating that grouped adaptation and dynamic gating contribute substantially beyond prompt-only control. In the no-grouping variant, the taxonomy-aligned branches are replaced by a single shared LoRA branch with an instruction-conditioned scalar gate ($C = 1$).

I Additional Evidence Routing Visualizations

Figures 13–17 provide additional gate-activation visualizations across multimodal reasoning, classification, and generation benchmarks. Consistent with the main GQA analysis, instruction-conditioned gating produces structured modulation of pathway activations aligned with the intended evidence preference. Semantic-focused instructions increase activation of semantic and interaction-related routing groups, whereas appearance-focused instructions increase activation of appearance-related groups. Suppression-oriented instructions reduce activation of the targeted routing groups while preserving activation of other task-relevant pathways. These patterns are observed across benchmarks with different modalities and task structures.

J Additional Functional Reliance Results

Figures 18–22 report additional Reliance Sensitivity (RS) analyses across multimodal reasoning, classification, and generation benchmarks. Consistent with the main GQA results, different instructions produce systematic shifts in functional dependence across evidence pathways. Semantic-focused instructions increase sensitivity to semantic and interaction-related pathways, whereas

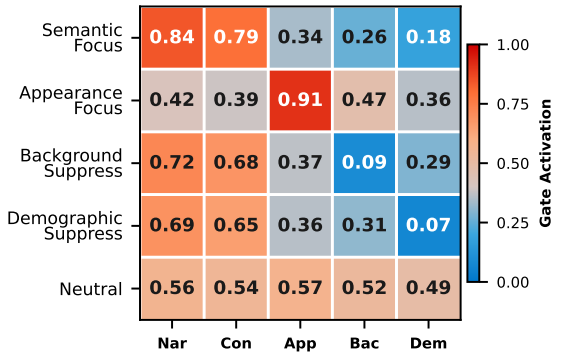

Figure 13: Gate activation heatmap on MM-IMDb under different instruction conditions. X-labels denote **Nar**rative, **Con**ceptual, **App**earance, **Bac**kground, **Dem**ographic respectively.

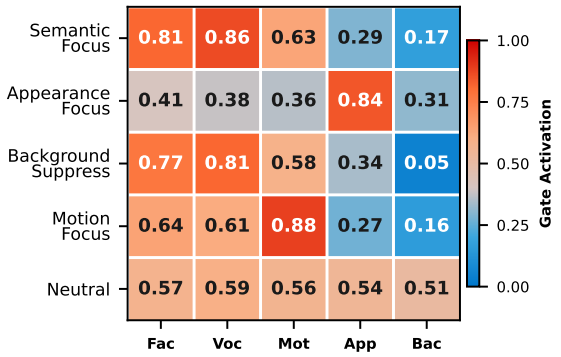

Figure 14: Gate Activation Heatmap on CREMA-D under different instruction conditions. X-labels denote **Fac**ial Expression, **Voc**al Emotion, **Mot**ion, **App**earance, **Bac**kground respectively.

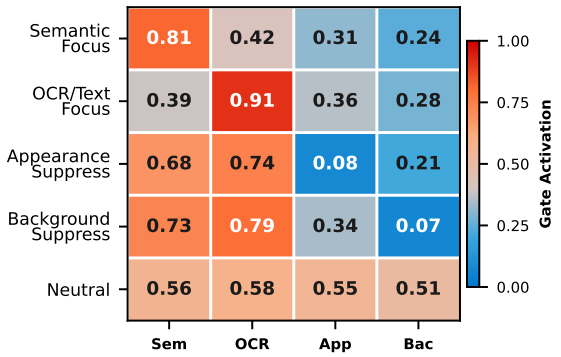

Figure 15: Gate Activation Heatmap on TextVQA under different instruction conditions. X-labels denote **Sem**antic, **OCR**/Text, **App**earance, **Bac**kground respectively

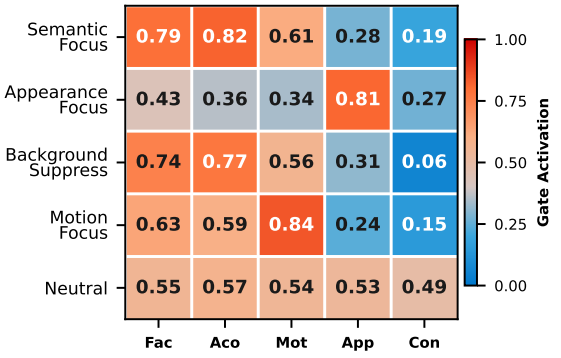

Figure 16: Gate Activation Heatmap on RAVDESS under different instruction conditions. X-labels denote **Fac**ial Emotion, **Aco**ustic Tone, **Mot**ion, **App**earance, **Con**text respectively.

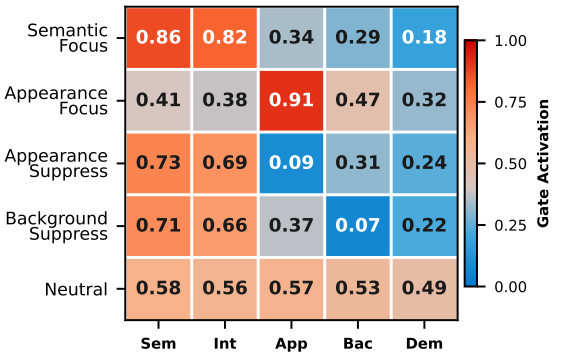

Figure 17: Gate Activation Heatmap on Flickr30K under different instruction conditions. X-labels denote **Sem**antic, **Int**eraction, **App**earance, **Bac**kground, **Dem**ographic respectively

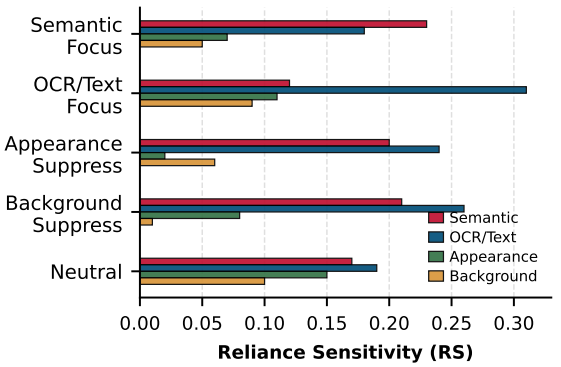

Figure 18: Reliance Sensitivity (RS) across evidence pathways under different instruction conditions on TextVQA. Higher RS indicates greater functional dependence on the corresponding evidence pathway.

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

Dataset	Variant	GMR↑	ΔRS↑	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	**0.57**	**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

appearance-focused instructions increase reliance on appearance-associated evidence. Suppression-oriented instructions reduce functional dependence on the targeted evidence category while preserving sensitivity to other task-relevant pathways. Together with the gate-activation results, these findings indicate that GUIDE modulates not only pathway activations but also the evidence that functionally influences model predictions.

K Additional Perturbation Stability Analyses

Figure 23 reports additional perturbation-stability analyses on GQA, MM-IMDb, and Flickr30K. Consistent with the main GQA results, semantic-focused and suppression-oriented instructions remain comparatively stable under appearance- and context-related perturbations, whereas appearance-focused instructions exhibit larger degradation under appearance-related perturbations. Across the evaluated datasets, perturbations to core semantic evidence produce the largest performance degradation, indicating that task-relevant semantic evidence remains important despite instruction-conditioned modulation of auxiliary pathways.

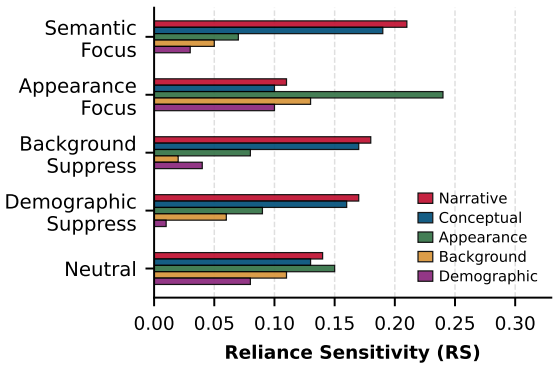

Figure 19: Reliance Sensitivity (RS) across evidence pathways under different instruction conditions on MM-IMDb. Higher RS indicates greater functional dependence on the corresponding evidence pathway.

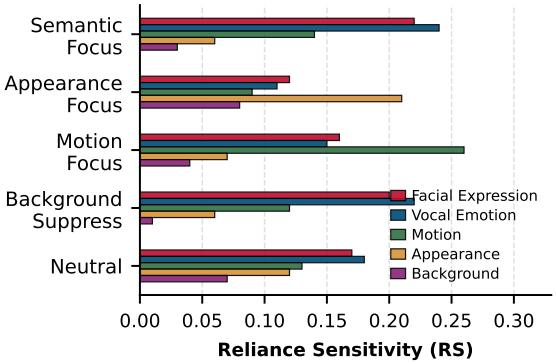

Figure 20: Reliance Sensitivity (RS) across evidence pathways under different instruction conditions on CREMA-D. Higher RS indicates greater functional dependence on the corresponding evidence pathway.

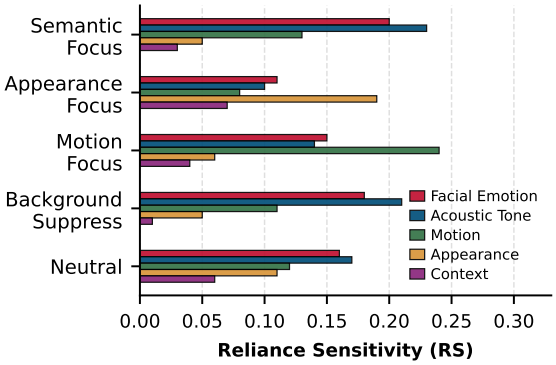

Figure 21: Reliance Sensitivity (RS) across evidence pathways under different instruction conditions on RAVDESS. Higher RS indicates greater functional dependence on the corresponding evidence pathway.

Table 7: Dataset Statistics.

Dataset	Task	Modality	Train	Val	Test	High-Level Evidence Families
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

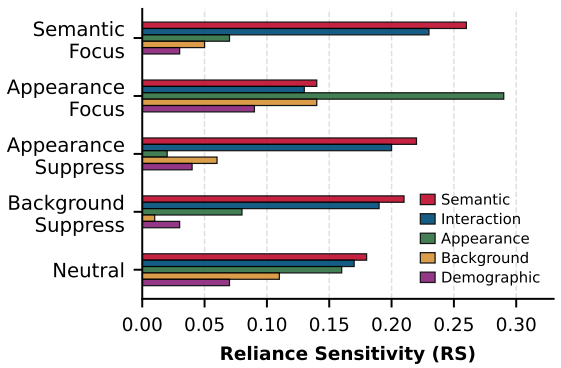

Figure 22: Reliance Sensitivity (RS) across evidence pathways under different instruction conditions on Flickr30K. Higher RS indicates greater functional dependence on the corresponding evidence pathway.

L Implementation Details

Table 7 summarizes the datasets, task types, modalities, dataset splits, and high-level evidence families used in our experiments. These high-level families provide a shared descriptive taxonomy across benchmarks and should be distinguished from the operational routing groups used by GUIDE. Operational groups may refine a high-level family into multiple task-specific components. For example, the semantic family in GQA is implemented using separate object-, relation-, and action-related routing groups, whereas the appearance family is implemented using an attribute-related routing group. Consequently, the number of operational groups reported in Table 8 may exceed the number of high-level families listed in Table 7.

Across tasks, the shared high-level taxonomy includes semantic, appearance, background, motion, OCR/text, and demographic evidence when applicable. The operational interpretation of these families depends on the modality and benchmark. For example, semantic routing in GQA captures object, action, and relational information, whereas appearance routing captures attribute-related information. CREMA-D and RAVDESS use task-specific facial, acoustic, and motion-related groups. OCR/text routing is used for TextVQA, while motion-related routing is used primarily for audio–visual emotion-recognition tasks.

This hierarchical organization provides a common evidence taxonomy while allowing GUIDE to use task-appropriate operational routing groups.

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.

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

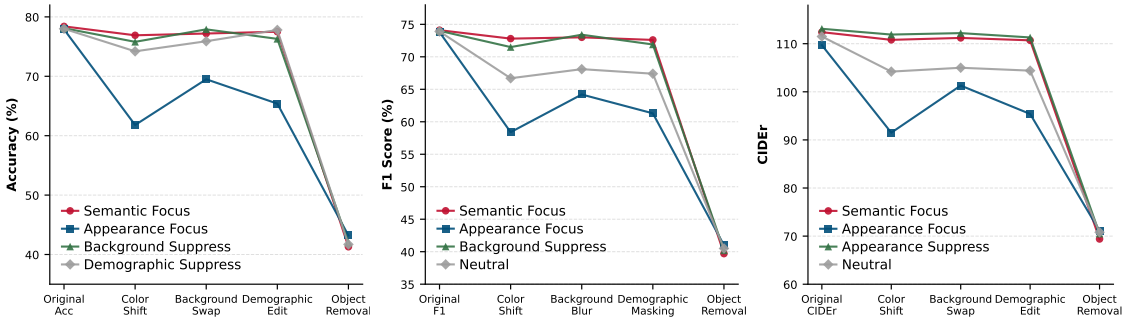

Figure 23: Perturbation-stability results on GQA, MM-IMDb, and Flickr30K under different instruction conditions and evidence perturbations.

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_{gate}
GQA	Qwen3-VL-30B	6	8	2×10^{-4}	32	5	0.01
TextVQA	Qwen3-VL-30B	4	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

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

The pretrained multimodal backbone remains frozen during training, and only the grouped LoRA parameters and instruction-conditioned gating network are updated. We use AdamW 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 experiments. GUIDE uses independent sigmoid gates: each routing-group activation is computed independently and lies in $[0, 1]$, without normalization across groups. Models are trained using mixed precision with batch sizes between 32 and 64, depending on the benchmark and backbone size. The gate-regularization coefficient λ_{gate} is set to 0.01 across experiments. For the compositional-routing analysis, we use a separate trade-off coefficient $\alpha \in [0, 1]$ to interpolate the relative emphasis of semantic- and appearance-focused instruction components during inference. All training hyperparameters are selected using the validation split 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 a human evaluation with eleven annotators: one undergraduate student, one graduate assistant, four Master's students, three PhD students, and two postdoctoral researchers. All annotators had prior experience with multimodal or language-model evaluation. Participation was voluntary and uncompensated, and annotators were in-

formed of the purpose of the study and the intended use of their responses before providing consent. We randomly sampled 50 Flickr30K examples across the evaluated instruction conditions. For each example, annotators were shown the source image, instruction, and generated caption and rated (i) instruction alignment and (ii) semantic preservation using a four-point Likert scale.

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 changes instruction-conditioned evidence emphasis while preserving the underlying input and task. Figures 8 and 9 show representative GQA and TextVQA examples under baseline, semantic-focused, and appearance-focused instructions.

Without an explicit evidence preference, the model exhibits comparatively mixed activation across semantic- and appearance-related routing groups. Under semantic-focused instructions, GUIDE increases activation associated with task-relevant semantic structure. For GQA, this includes actions, object interactions, attributes, and relational cues relevant to answering the question. For TextVQA, semantic-focused instructions emphasize contextual and relational information, while OCR-focused instructions, when applied, emphasize localized textual evidence.

Appearance-focused instructions instead increase activation associated with visual attributes, layout, style, clothing, and other appearance-related cues. These changes illustrate that identical multimodal inputs can produce different pathway-activation patterns under contrasting evidence instructions.

The examples are consistent with the quantitative GMR and RS analyses: gate visualizations show changes in pathway activation, while the perturbation-based RS results establish corresponding changes in functional reliance. Overall, the qualitative examples illustrate how GUIDE modulates evidence processing across reasoning and generation settings.