When Irrelevant Text Matters: Affine Margin Shifts in Multimodal Large Language Models

Anonymous ACL submission

Abstract

Multimodal large language models (MLLMs) are frequently exposed to auxiliary textual context, the impact of which on visually grounded tasks remains underexplored. In this paper, we investigate the influence of task-irrelevant context by formulating it as a controlled intervention within a binary visual judgment framework. By maintaining an invariant prompt structure while varying auxiliary inputs, we observe that irrelevant text consistently biases model predictions across diverse benchmarks. To move beyond performance metrics, we characterize this sensitivity through a decision margin defined by the log-probability difference between binary candidates. Our analysis reveals a robust geometric regularity: context-conditioned margins follow a consistent affine transformation of their context-free counterparts. This finding demonstrates that irrelevant context does not manifest as unstructured stochastic noise but as a estimable distortion of model preference. We further interpret the fitted affine parameters as metrics for visual commitment preservation and directional answer bias. These findings provide a margin-level diagnostic view of irrelevant-context effects in MLLMs and offer a basis for future studies on noisy-context robustness. We release all the code in https://anonymous.4open.science/r/Irrelevant_Text_Matters-E661.

1 Introduction

Multimodal large language models (MLLMs) have demonstrated strong cross-modal understanding capabilities, enabling complicated vision-language tasks such as visual question answering, multi-turn reasoning, and multimodal agents. In these scenarios, the input image is often accompanied by auxiliary textual context, such as retrieved passages, user-provided side information, or history dialogue. Such context is not always aligned with the current task: retrieved passages may be off-topic (Mortaheb et al., 2025), while instructions and dialogue may also contain irrelevant information (Wang et al., 2024b; Park et al., 2024). This raises a question: *Does irrelevant context affect MLLM predictions on vision-language tasks?*

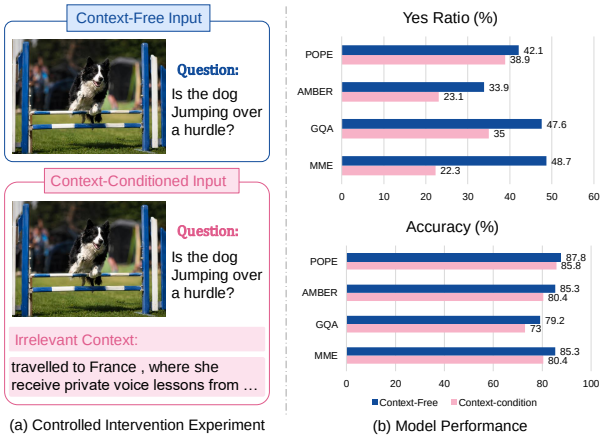

Figure 1: Overview of the controlled intervention and its behavioral effects. (a) For each image-question pair, we compare a context-free input with a paired context-conditioned input, where an irrelevant context is inserted while the prompt structure is kept unchanged. (b) Across multiple benchmarks, irrelevant context not only reduces predictive accuracy, but also shifts model decisions toward negative responses.

Our question is motivated by a broader observation in NLP community: large language models (LLMs) can be sensitive to textual context that is not part of the task-relevant evidence. Prior studies have shown that irrelevant or distracting context can mislead LLMs in question answering and reasoning tasks (Jia and Liang, 2017; Shi et al., 2023; Amiraz et al., 2025). Similar issues arise in retrieval-augmented generation, where retrieved passages containing irrelevant information will affect the model's performance (Yoran et al., 2024; Yan et al., 2024; Cuconasu et al., 2024). In MLLMs, recent work shows that they may over-rely on ad-

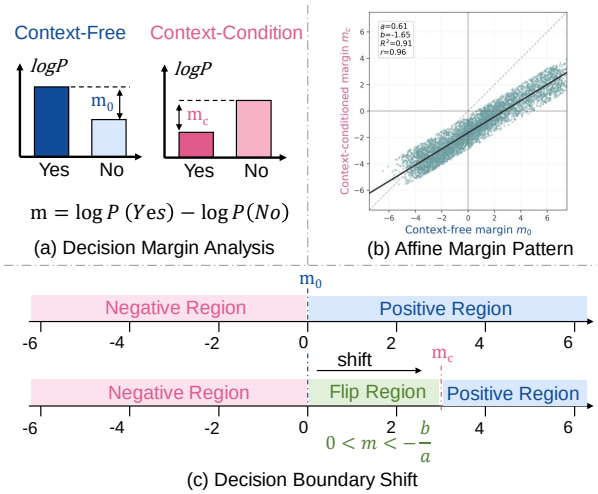

Figure 2: Illustration of the margin-level analysis. (a) The decision margin is defined as the log-probability difference between the positive and negative answers. (b) Adding irrelevant context transforms the context-free margin m_0 into the context-conditioned margin m_c, which follows an approximate affine relation $m_c \approx am_0 + b$. (c) The affine transformation shifts the effective decision boundary from $m_0 = 0$ to $m_0 = -b/a$. This shift expands the Negative Region and compresses the Positive Region. As a result, samples in the Flip Region, i.e., $0 < m_0 < -b/a$, are likely to change from positive to negative predictions, inducing a systematic shift in decision preference.

ditional textual context and resulting in a strong bias (Deng et al., 2025; Chen et al., 2024; Zhang et al., 2024). However, despite these observations, there remains a lack of systematic investigation into how irrelevant context transforms MLLM decision-making in terms of behavioral effects and the underlying mechanisms.

To bridge this gap, we conduct a systematic investigation into the influence of image-irrelevant text, utilizing it as a controlled intervention within binary visually grounded prediction tasks. Specifically, as illustrated in Figure 1(a), for each image-question pair, we treat the original input containing only the image and the question as the *context-free* condition. We then construct a paired *context-condition* input by inserting an irrelevant textual passage randomly sampled from an external corpus, while keeping the overall prompt structure unchanged. As shown in Figure 1(b), evaluation across multiple benchmarks shows that irrelevant context not only degrades the model's overall performance, but also induces a consistent shift in decision preference toward negative responses.

To better characterize this behavior, we introduce the *decision margin*, defined as the log-probability difference between the positive and negative answers, as illustrated in Figure 2(a). Surprisingly, as shown in Figure 2(b), the context-condition margin can be well approximated by a simple affine transformation of the context-free margin, i.e., $m_c \approx am_0 + b$, where m_0 and m_c denote the margins under the context-free and context-conditioned inputs, respectively. This affine structure further explains the observed prediction flips. As illustrated in Figure 2(c), the effective decision boundary shifts from $m_0 = 0$ to $m_0 = -b/a$. Consequently, samples with weak positive margins are likely to cross the shifted boundary and flip to negative predictions, accounting for the systematic tendency toward negative responses.

Empirically, this affine relation exhibits two notable regularities. First, it appears consistently across different MLLMs. Second, for a given model, the fitted parameters remain close across different benchmarks. Inspired by the implicit Bayesian interpretation of in-context learning (Xie et al., 2021), we interpret the slope a as the preservation of the original visual decision structure, and the intercept b as the directional answer offset introduced by irrelevant text. Based on the above observations, we design a lightweight post-hoc calibration, to partially mitigate the effect of irrelevant context at inference time.

Our analysis extends the understanding of irrelevant-context effects in MLLMs. We show that such effects can follow a stable and estimable affine pattern at the margin level, and provide new insight on how textual noise biases multimodal decision-making. These findings suggest that MLLMs remain sensitive to irrelevant context, even when it provides no useful evidence. We hope our work can serve as a building block for future research on noisy-context robustness, and contribute to the development of more trustworthy and robust multimodal systems.

2 Irrelevant Context Affects Prediction

In this section, we first introduce our experiment setup, then empirically show that irrelevant context does change MLLM predictions.

2.1 Problem Formulation

Given an image I, a question Q, and an image-irrelevant textual context C, we compare model

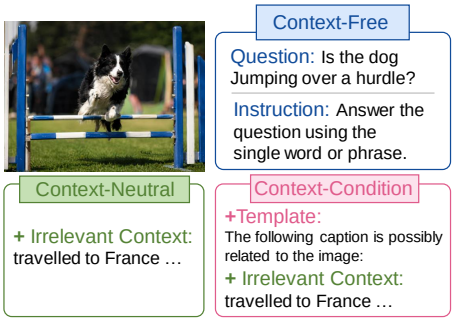

Figure 3: Visualization of three input conditions. These three conditions represent different degrees to which the irrelevant context is framed as potentially image-related.

predictions under two input conditions:

$$x_o = (I, Q), \qquad x_c = (I, C, Q), \tag{1}$$

Here, x_o denotes the context-free input, and x_c denotes the context-conditioned input.

We formulate each sample as a binary visual judgment. The model is asked to decide whether the visual claim implied by the question is supported by the image. We denote the binary label space as

$$\mathcal{Y} = \{y^+, y^-\}, \tag{2}$$

where y^+ denotes a positive judgment that the claim is supported, and y^- denotes a negative judgment that the claim is not supported. For a model f_θ, the corresponding predictions are

$$\hat{y}_o = f_\theta(x_o), \qquad \hat{y}_c = f_\theta(x_c), \tag{3}$$

where $\hat{y}_o, \hat{y}_c \in \mathcal{Y}$.

This formulation adopts two key simplifications for controlled analysis: we cast open-ended visually grounded predictions into binary positive/negative judgments, and we treat auxiliary textual context as a controlled intervention by varying only the irrelevant context. Models typically answer with "Yes" and "No" under the default instruction. See Appendix C.2 for analysis of different verbalizations, and Appendix E.1 for more discussions on problem formulation.

2.2 Experiment Setup

Irrelevant Context Sampling Given an image-question pair (I, Q), we first sample irrelevant contexts C from WikiText (Merity et al., 2017) with a fixed random seed. We control the length of each sampled sentence within a predefined range, and apply filtering process to ensure that the selected context is unrelated to the image. See Appendix A.2 for implementation details.

Input Construction After irrelevant context sampling, we construct inputs under three conditions. A case is shown in Figure 3. For all settings, the question is followed by the same answer instruction:

```
Answer the question using a single word
or phrase.
```

The **context-free** input contains only the image and the question:

```
[Image] [Question]
```

The **context-neutral** input appends the sampled irrelevant context:

```
[Image] [Question] [Context]
```

The **context-condition** input further introduces a relevance template as:

```
The following caption is possibly
related to the image:
```

The template is placed before the context:

```
[Image] [Question] [Template] [Context]
```

These three conditions represent different degrees to which the irrelevant text is framed as potentially image-related. All the samples are then fed into models, and use greedy decoding to get the final answer.

Evaluation Benchmarks and Models We evaluate on four widely used vision-language benchmarks that cover diverse forms of visually grounded judgment. **POPE** (Li et al., 2023) focuses on object-existence verification and is commonly used to evaluate object hallucination in MLLMs. **AMBER** (Wang et al., 2023) extends this evaluation to multiple hallucination-related dimensions, including object existence, attributes, and relations. **MME** (Fu et al., 2023) provides a broader diagnostic benchmark, covering perception-oriented tasks such as color, count, OCR, and position recognition, as well as cognition-oriented tasks such as commonsense and code reasoning. **GQA** (Hudson and Manning, 2019) evaluates compositional visual reasoning over objects, attributes, relations, and spatial layouts. See Appendix A.3 for data statistics and implementation detail.

For models, we report results on **LLaVA-1.5-7b** (Liu et al., 2023), **Qwen2-VL** (Wang et al., 2024a) family (2b and 7b) and **Intern-VL3-8b** (Zhu et al., 2025).

Table 1: Behavioral effects of image-irrelevant context across datasets and models. Irrelevant context generally degrades performance, induces a negative-response shift, and has stronger effects when presented as potentially image-related. **Free** indicates the context-free condition, while **ΔNeu.** and **ΔCon.** report changes under the context-neutral and context-condition setting relative to context-free condition. **Yes Rate** reports the proportion of positive responses among all answers, indicating the model's tendency toward affirmative predictions. **Flip Rate** reports the percentage of samples whose prediction changes. All values are reported in percentage points.

Dataset	Model	Accuracy			Recall			Yes Rate			Flip Rate	
		Free	ΔNeu.	ΔCon.	Free	ΔNeu.	ΔCon.	Free	ΔNeu.	ΔCon.	Neu.	Con.
MME	LLaVA-1.5-7B	78.5	↑0.1	↓0.9	80.5	↓7.9	↓8.4	51.9	↓7.9	↓7.4	9.0	9.8
	Qwen2-VL-2B	79.9	↓10.6	↓20.8	80.5	↓24.6	↓40.8	48.7	↓16.8	↓26.4	20.5	33.4
	Qwen2-VL-7B	87.6	↑0.2	↓27.5	90.6	↓2.6	↓69.5	52.7	↓2.8	↓41.7	4.3	42.0
	InternVL3	87.9	↓3.1	↓3.7	85.3	↓9.9	↓10.5	47.3	↓7.0	↓7.0	7.8	7.9
GQA	LLaVA-1.5-7B	76.4	→0.0	↓1.9	76.4	↑2.2	↑7.0	50.7	↑2.3	↑9.1	7.6	14.3
	Qwen2-VL-2B	79.2	↓6.3	↓6.2	76.1	↓20.1	↓18.4	47.6	↓14.3	↓12.6	16.0	15.6
	Qwen2-VL-7B	82.2	↓1.4	↓19.8	78.7	↓2.8	↓48.3	47.4	↓1.6	↓29.9	3.9	30.0
	InternVL3	81.9	↓1.5	↓2.0	85.1	↓4.3	↓6.1	54.2	↓2.9	↓4.2	8.0	9.4
AMBER	LLaVA-1.5-7B	81.2	↓0.3	↓1.3	64.8	↓2.4	↓2.3	28.8	↓1.3	↓2.9	6.3	5.7
	Qwen2-VL-2B	85.3	↓0.7	↓4.9	78.5	↓9.8	↓21.6	33.9	↓5.9	↓10.8	7.2	12.3
	Qwen2-VL-7B	86.9	↑0.5	↓9.4	81.2	↓2.0	↓43.1	34.1	↓1.8	↓19.7	2.9	19.7
	InternVL3	88.3	↑0.5	↑0.1	85.7	↓4.7	↓6.0	35.8	↓3.7	↓4.2	5.6	6.0
POPE	LLaVA-1.5-7B	85.5	↓0.7	↓0.3	76.6	↓2.4	↓0.9	41.4	↓1.8	↓0.6	2.3	2.3
	Qwen2-VL-2B	87.8	↓1.8	↓2.0	79.6	↓4.5	↓5.1	42.1	↓2.8	↓3.2	3.3	3.6
	Qwen2-VL-7B	88.3	↓1.3	↓8.4	81.1	↓3.4	↓19.8	43.1	↓2.1	↓11.6	2.6	11.6
	InternVL3	90.9	↓0.8	↓0.3	90.7	↓5.3	↓3.9	50.2	↓4.5	↓3.6	4.8	4.1

Discussion of Confounds We further examine whether the observed behavioral shift can be explained by simple experimental confounds. Firstly, the effect is not tied to a particular context corpus: replacing WikiText with alternative context sources, such as COCO captions (Chen et al., 2015) or shuffled WikiText, preserves the main trend, although image-like captions tend to induce stronger interference. Secondly, it is also not explained by context length alone, since the affine distortion remains observable across different length buckets. Finally, prompt formatting affects the magnitude but not the existence of the phenomenon. Detailed settings and results are provided in Appendix C.1.

2.3 Experimental Results

We report **Accuracy**, **Recall**, and **Yes-rate** under the Context-free condition, together with their changes under the context-neutral and context-condition settings. We further report **Flip-rate** to measure sample-level prediction changes.

The results in Table 1 reveal three main patterns:

Adding irrelevant context hurts performance. Across models and benchmarks, adding image-irrelevant context generally reduces predictive performance. This degradation appears under both context-neutral and context-conditioned settings, indicating that auxiliary text can interfere with visually grounded prediction.

The behavioral shift is directional. The performance drop is accompanied by a systematic change in answer preference. In most settings, the Yes rate decreases after irrelevant context is added, indicating that the induced errors are not merely random.

Semantic framing amplifies the effect. The context-conditioned setting usually produces larger changes than the context-neutral setting, especially in Yes rate and flip rate. This suggests that models are more affected when the same irrelevant text is framed as potentially image-related.

These behavioral patterns suggest that irrelevant context does not act as unstructured noise; instead, it systematically changes the model's decision preferences. In the next section, we investigate this effect in detail through a margin-level analysis. See Appendix B.1 for complement experiment result.

3 Affine Margin Shift

In this section, we analyze the effect of irrelevant context in logit space, and identify a simple regularity: its influence can be approximated by an affine transformation.

3.1 Decision Margin Design

Given an input x, we define the *decision margin* as the log-probability difference between the two candidate answers:

$$m(x) = \log P(y^+ \mid x) - \log P(y^- \mid x), \quad (4)$$

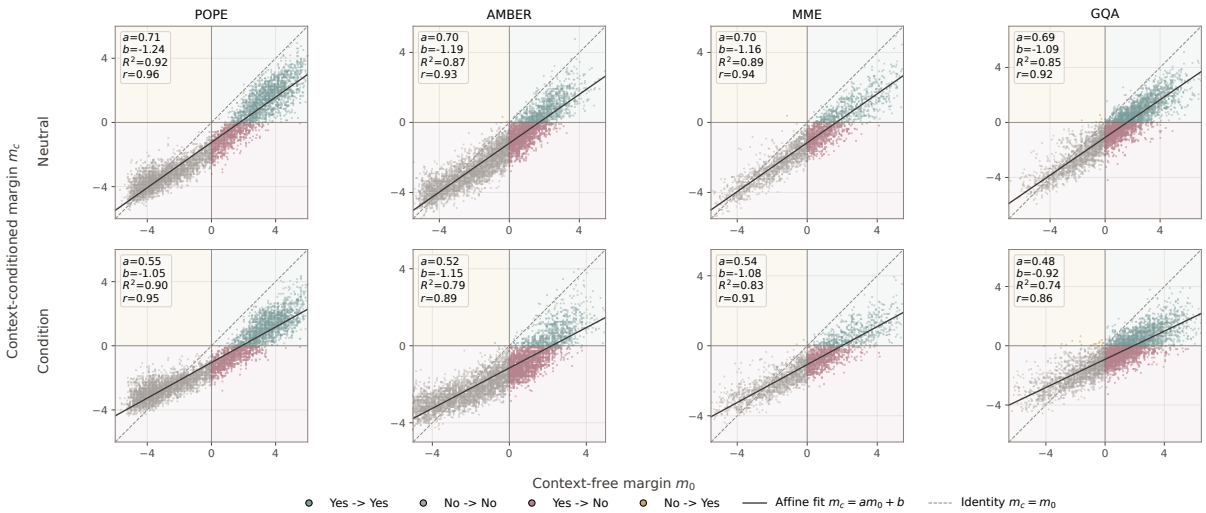

Figure 4: Affine margin patterns for LLaVA-1.5-7B across four benchmarks. Each panel plots the context-free margin m_o against the context-conditioned margin m_c. Points are colored by prediction transitions between the two conditions, and the solid line shows the fitted affine relation $m_c = am_o + b$; the dashed line denotes the identity relation $m_c = m_o$. The consistently high R^2 and correlation values indicate that textual context induces a structured affine transformation of the original decision margin rather than unstructured perturbations. Notably, changing the semantic framing from Neutral to Condition substantially decreases the fitted slope a, while leaving the negative offset b largely stable. Results are shown for LLaVA.

Table 2: Fitted affine parameters across models and benchmarks. The affine relationship exhibits consistently high goodness of fit, indicating that such affine is structural and stable.

Model	Dataset	a	b	R^2	r
LLaVA-1.5-7B	POPE	0.55	-1.05	0.90	0.95
	AMBER	0.52	-1.15	0.79	0.89
	MME	0.54	-1.08	0.83	0.91
	GQA	0.48	-0.92	0.74	0.86
Qwen2-VL-2B	POPE	0.56	-1.58	0.86	0.93
	AMBER	0.59	-1.61	0.86	0.93
	MME	0.56	-1.13	0.74	0.86
	GQA	0.62	-1.55	0.78	0.88

Our design is motivated by DPO (Rafailov et al., 2023): its sign indicates the model preference, while its magnitude reflects the confidence. See Appendix A.1 for implementation details.

3.2 Affine Regularity of Decision Margins

Figure 4 shows the affine margin patterns for LLaVA-1.5-7B across four benchmarks. We identify a regularity: the impact of irrelevant text can be characterized by an affine transformation:

$$m_c \approx a \cdot m_0 + b, \tag{5}$$

where a and b represent model-specific parameters estimated via least squares.

To step further, we have following observations:

The affine form is consistent across different MLLMs. Table 2 shows the parameters under context-neutral settings across different benchmarks and models. The consistently high R^2 values confirm that the affine relationship is a fundamental score-level response to auxiliary textual input. This regularity persists independently of specific task distributions, suggesting a shared behavioral constraint in MLLMs.

The affine parameters remain stable across benchmarks. For a model, the fitted slope a and offset b remain close across different benchmarks, despite substantial differences in task type and data distribution. This suggests that the affine pattern is not a dataset-specific artifact, but reflects a model-specific response to auxiliary textual context.

Different semantic framings affect the scaling component. Compared with the Neutral, the Condition setting consistently yields a smaller slope a, indicating stronger margin compression. In contrast, the offset term b remains negative and of comparable magnitude across framings, suggesting that the answer-direction shift is relatively stable. We will provide an explanation for this phenomenon in the following section.

These observations suggest that the influence of irrelevant context is a systematic distortion. See

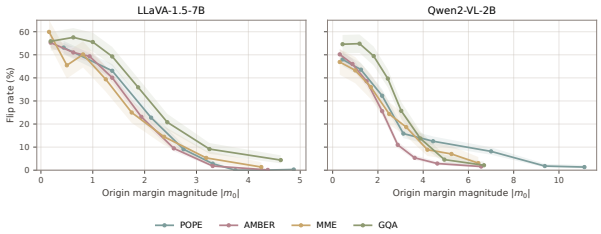

Figure 5: Flip rate across context-free margin bins. Samples with smaller $|m_0|$ are much more likely to change their predictions after adding irrelevant context. This indicates that context-induced flips are concentrated near the original decision boundary.

Appendix B.2 and Appendix D for more results and robust verification.

3.3 Boundary Shift Explains Prediction Flips

Above results also explain the observed behavioral changes. Without context, the decision boundary is located at $m_0 = 0$. After adding context, the margin is approximately transformed as $m_c \approx am_0 + b$, so the context-condition decision boundary satisfies $am_0 + b = 0$. Mapped back to the original margin axis, the effective boundary becomes

$$m_0 = -\frac{b}{a}, \tag{6}$$

Thus, an originally affirmative sample must have a sufficiently large positive margin to remain affirmative after irrelevant context is added; samples with weak margins can otherwise fall below the shifted boundary and flip.

We verify this explanation by grouping samples according to the magnitude of their context-free margin $|m_0|$ and computing the flip rate within each bin. As shown in Figure 5, flip rates are consistently highest for low-margin samples and decrease as $|m_0|$ grows across models and benchmarks. This confirms that context-induced flips are not uniformly distributed over samples, but are concentrated near the original decision boundary.

4 Interpreting the Affine Parameters

In this section, we use a lightweight Bayesian view to interpret the fitted parameters.

4.1 Preliminary

We first introduce a lightweight Bayesian-inspired view to describe how irrelevant context may enter the model's prediction process.

Let $\mathcal{X}$ denote the original input without context. The model defines a context-free predictive distribution $p(Y \mid \mathcal{X})$. When an additional context C is provided, the model may use it as an additional conditioning signal when interpreting the input. Following the prior work (Xie et al., 2021), we introduce a latent interpretation variable $Z \in \mathcal{Z}$, which represents a task-specific "concept". The context-conditioned prediction can then be written as posterior-weighted marginalization:

$$p(Y \mid \mathcal{X}, C) = \int_{\mathcal{Z}} p(Y \mid \mathcal{X}, Z)\, p(Z \mid \mathcal{X}, C)\, dZ, \tag{7}$$

Under this view, adding C updates the model's belief over latent interpretations:

$$p(Z \mid \mathcal{X}, C) \propto p(C \mid Z, \mathcal{X}) p(Z \mid \mathcal{X}), \tag{8}$$

In our experiments, C provides no useful evidence, i.e., $Y \perp C \mid \mathcal{X}$. However, image-text alignment priors may lead the model to treat C as useful evidence rather than irrelevant noise. Consequently, $p(C \mid Z, \mathcal{X})$ may be non-uniform across latent interpretations, shifting $p(Z \mid \mathcal{X}, C)$ away from the original belief $p(Z \mid \mathcal{X})$.

4.2 Explanation of the Affine Parameters

The Bayesian formulation provides a plausible lens. Under this interpretation, the fitted affine map summarizes two observable effects: how much of the original visual decision structure is preserved, and whether the added context introduces a directional answer offset.

Slope as commitment preservation. We interpret the slope a as the degree to which the original visual decision structure is preserved after context-conditioned inference. A larger a indicates that the original margin geometry is better preserved, while a smaller a indicates stronger attenuation of the original visual commitment.

Intercept as answer-direction offset. The intercept b captures the directional offset introduced by the added context. Since the margin is defined such that positive values favor `Yes`, a negative b corresponds to a shift toward `No`. For image-irrelevant context, this term is especially informative because the added text does not provide label-conditioned visual evidence and therefore tends to appear as a global directional bias in the answer space.

This interpretation explains the effect of semantic framing observed in Sec. 3.2. The Neutral and

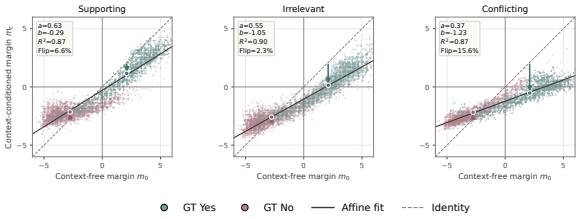

Figure 6: Experiment results for Validating the Parameter Interpretation. Points are colored by ground-truth label. Arrows show the mean label-conditioned margin shift from the identity line to the observed context-conditioned mean. The monotonic trend $a_{\text{sup}} > a_{\text{irr}} > a_{\text{conf}}$ is consistent with the interpretation of a as commitment preservation.

Condition settings contain the same type of image-irrelevant text, but the Condition template frames the text as potentially related to the image: this framing increases the likelihood that the model treats the irrelevant context as useful contextual evidence, thereby reweighting the latent interpretation more strongly and weakening the preservation of the original visual margin. At the same time, because the context is unchanged, the induced answer-direction prior is not fundamentally changed.

4.3 Validating the Parameter Interpretation

To test above implications, we construct a controlled context-type intervention.

Intervention setup Our key idea is that, If a and b respectively capture visual-decision preservation and directional answer offset, then varying the evidential role of the context should change the affine parameters. Therefore, we construct three context types:

Supporting context is consistent with the ground-truth answer, which states that the queried object is present and absent for negative samples. **Irrelevant context** is the same as in the context-conditioned setting. **Conflicting context** states absence for positive samples and presence for negative samples. All context types are inserted using the same template and position in the input.

Results Figure 6 reports the fitted affine parameters under the three context types. The slope changes monotonically with the evidential role of the context:

$$a_{\text{sup}} > a_{\text{irr}} > a_{\text{conf}}, \tag{9}$$

This ordering is consistent with the interpretation of a as commitment preservation. Supporting context preserves the original margin structure the most, irrelevant context introduces stronger attenuation, and conflicting context weakens the original structure most substantially.

However, although supporting context is related, its slope still remains below one. We interpret this behavior as evidence that MLLMs integrate supporting context as imperfect evidence. See Appendix C.3 for more discussion.

5 Post-Hoc Calibration Test

In this section, we use post-hoc calibration as a diagnostic test.

5.1 Calibration Protocol

Given a held-out calibration split $\mathcal{D}_{\text{cal}}$, we compute the context-free margin m_o and the context-conditioned margin m_c for each calibration sample, and fit the affine relation. On the test split, we apply the inverse transformation:

$$m_{\text{cal}} = \frac{m_c - \hat{b}}{\hat{a}}. \tag{10}$$

5.2 Experiment Setup and results

Calibration settings We consider two calibration settings. In the in-domain setting, affine parameters are fitted on a calibration split and applied to the held-out split of the same benchmark.

In the cross-dataset transfer setting, affine parameters are fitted on one source dataset and directly applied to other target datasets. This setting uses no target-dataset calibration samples and tests whether the affine parameters capture a transferable model-level response to irrelevant context.

Experiment Results Table 6 reports results of in-domain calibration, and Table 7 reports the cross-dataset transfer result. These results show that the context-induced margin distortion is structured and partially transferable: inverse-affine correction recovers context-free decisions on held-out data and under cross-dataset transfer. See Appendix B.3 for impelentation details and more experiment results.

6 Related Work

6.1 Context Effects in LLMs

LLMs are known to be sensitive to contextual information that is not part of the task-relevant evidence. Early work on adversarial reading comprehension shows that irrelevant distractor sentences can mislead question-answering models (Jia and Liang,

Table 3: In-domain post-hoc affine calibration under WikiText-after-possibly context. Affine parameters are fitted on calibration splits and evaluated on held-out test splits. We report the Accuracy and Flip rate with percentage.

Dataset	Origin	Con.	Cal.	Con. Flip	Cal. Flip
POPE	86.1	78.0	84.6	16.6	6.1
AMBER	82.7	74.9	79.7	29.2	13.6
MME	78.7	68.5	76.5	30.2	12.6
GQA	73.2	69.2	72.3	36.3	14.8

Table 4: Cross-dataset affine calibration transfer result. Affine parameters are fitted on POPE and directly applied to other benchmarks without target calibration samples. We report the Accuracy and Flip rate with percentage.

Dataset	Origin	Con.	Cal.	Con. Flip	Cal. Flip
AMBER	82.7	75.0	80.1	29.1	14.7
MME	78.7	68.5	76.6	30.4	12.8
GQA	73.2	69.3	71.7	36.2	13.1

2017), and recent studies further demonstrate that large language models can be distracted by irrelevant context in reasoning and question-answering tasks (Shi et al., 2023; Amiraz et al., 2025). Related observations also appear in retrieval-augmented generation, where retrieved passages may contain noisy, weakly related, or irrelevant information that affects the final answer (Yoran et al., 2024; Yan et al., 2024; Cuconasu et al., 2024).

6.2 Context Effects in MLLMs

Textual bias has long been observed in vision-language tasks. Early VQA studies show that models can exploit language priors and insufficient visual grounding (Goyal et al., 2017; Agrawal et al., 2018). Subsequent work further studies how to reduce such unimodal biases and force models to rely more on visual evidence (Cadene et al., 2019; KV and Mittal, 2020; Deng et al., 2025). In multimodal RAG, prior work shows that irrelevant retrieved passages or conflicting knowledge may reduce overall performance (Chen et al., 2022; Kortukov et al., 2024; Caffagni et al., 2024).

6.3 In-Context Learning and Bias Calibration

In-context learning (ICL) shows that language models can adapt their predictions from contextual examples without parameter updates (Brown et al., 2020; Liu et al., 2022). A theoretical explanation further views ICL as implicit Bayesian inference, where the model predicts by inferring a latent concept from the observed context (Xie et al., 2021).

Despite its ability, prior work indicates that prompts in ICL may introduce systematic biases, which can be mitigated by estimating and calibrating such biases at inference time (Zhao et al., 2021; Han et al., 2023; Nie et al., 2022; Zhou et al., 2023). More broadly, this connects to post-hoc calibration methods such as Platt scaling and temperature scaling, which adjust model scores without updating model parameters (Platt, 1999; Guo et al., 2017).

7 Conclusion

In this paper, we study how image-irrelevant textual context affects visually grounded decisions in MLLMs. By treating irrelevant context as a controlled intervention, we show that it can degrade predictive performance, shift answer preferences, and induce prediction flips across benchmarks and models. To characterize this effect beyond aggregate metrics, we introduce a margin-level analysis based on the log-probability difference between positive and negative answers. Our results show that context-conditioned margins can be well approximated by an affine transformation of context-free margins, indicating that irrelevant context induces a structured distortion of model preference rather than unstructured noise. We further analyze the affine parameters as compact descriptors of margin preservation and directional answer shift, and develop a post-hoc calibration method that partially compensates for context-induced margin shifts without updating model parameters. Our findings provide a diagnostic perspective on irrelevant-context effects in MLLMs and may serve as a useful basis for future work on robust and trustworthy multimodal systems.

8 Limitations

Closed-form evaluation. Our analysis is mainly conducted in binary visual-question settings, and therefore does not fully capture the complexity of open-ended multimodal generation. We provide a preliminary study in Appendix E.3, where we observe similar context-induced effects in open-ended generation. When irrelevant context is introduced, we observe that the model tends to generate shorter responses and focus more on salient objects in the image. One plausible explanation is that the context-induced affine transformation suppresses marginal visual claims with weaker confidence. In

this view, objects or attributes that originally lie near the decision boundary are more likely to be filtered out after the margin shift, while highly salient objects with stronger margins remain in the output. This may explain why irrelevant context reduces response coverage and makes the generation more conservative. However, systematically characterizing whether the same affine margin pattern extends to open-ended outputs remains an important direction for future work.

Limited control experiment settings. Both our experimental setting and calibration method are intentionally simplified. The irrelevant contexts considered in this work are controlled interventions, rather than the heterogeneous auxiliary contexts that may arise in real applications, such as long dialogue histories, partially relevant retrieved passages, tool-use traces, or multi-turn reasoning contexts. As a result, our experiments do not fully capture more complex multimodal interaction scenarios. Moreover, our affine calibration is only a lightweight post-hoc correction for binary decision margins. It can partially compensate for context-induced margin shifts, but it does not remove the irrelevant context from the input or fully eliminate its influence.

Limited theoretical explanation. Our current analysis does not provide a complete theoretical explanation for the observed affine law, nor does it identify the causal mechanism underlying this phenomenon. Instead, we characterize the effect mainly at the behavioral and margin levels, and use the Bayesian-inspired view only as an interpretive lens. In addition, the fitted affine relation can be affected by question format and context presentation, including prompt wording, context position, and input structure. This suggests that the affine law observed in our controlled setting may be a particular manifestation of a broader context-interference phenomenon. Developing a deeper theoretical and causal account of this phenomenon remains an important direction for future work.

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.

Chen Amiraz, Florin Cuconasu, Simone Filice, and Zohar Karnin. 2025. The distracting effect: Understanding irrelevant passages in rag. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18228–18258.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Remi Cadene, Corentin Dancette, Hedi Ben-younes, Matthieu Cord, and Devi Parikh. 2019. RUBi: Reducing unimodal biases in visual question answering. In *Advances in Neural Information Processing Systems*, volume 32.

Davide Caffagni, Federico Cocchi, Nicholas Moratelli, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2024. Wiki-llava: Hierarchical retrieval-augmented generation for multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1818–1826.

Hung-Ting Chen, Michael J. Q. Zhang, and Eunsol Choi. 2022. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. *arXiv preprint arXiv:2210.13701*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Zhanpeng Chen, Chengjin Xu, Yiyan Qi, and Jian Guo. 2024. Mllm is a strong reranker: Advancing multimodal retrieval-augmented generation via knowledge-enhanced reranking and noise-injected training. *arXiv preprint arXiv:2407.21439*.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The power of noise: Redefining retrieval for rag systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 719–729.

Ailin Deng, Tri Cao, Zhirui Chen, and Bryan Hooi. 2025. Words or vision: Do vision-language models have blind faith in text? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji.

2023. MME: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Zhixiong Han, Yaru Hao, Li Dong, Yutao Sun, and Furu Wei. 2023. Prototypical calibration for few-shot learning of language models. In *International Conference on Learning Representations*.

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.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Adam Tauman Kalai, Ofir Nachum, Santosh S Vempala, and Edwin Zhang. 2025. Why language models hallucinate. *arXiv preprint arXiv:2509.04664*.

Evgenii Kortukov, Alexander Rubinstein, Elisa Nguyen, and Seong Joon Oh. 2024. Studying large language model behaviors under context-memory conflicts with real documents. *arXiv preprint arXiv:2404.16032*.

Gouthaman KV and Anurag Mittal. 2020. Reducing language biases in visual question answering with visually-grounded question encoder. *arXiv preprint arXiv:2007.06198*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out: The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114. Association for Computational Linguistics.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *International Conference on Learning Representations*.

Matin Mortaheb, Mohammad A. Amir Khojastepour, Srimat T. Chakradhar, and Sennur Ulukus. 2025. Re-ranking the context for multimodal retrieval augmented generation. *arXiv preprint arXiv:2501.04695*.

Feng Nie, Meixi Chen, Zhirui Zhang, and Xu Cheng. 2022. Improving few-shot performance of language models via nearest neighbor calibration. *arXiv preprint arXiv:2212.02216*.

Dongmin Park, Zhaofang Qian, Guangxing Han, and Ser-Nam Lim. 2024. Mitigating dialogue hallucination for large vision language models via adversarial instruction tuning. *arXiv preprint arXiv:2403.10492*.

John C. Platt. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In Alexander J. Smola, Peter Bartlett, Bernhard Schölkopf, and Dale Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 61–74. MIT Press.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 31210–31227. PMLR.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. 2023. AMBER: An LLM-free multi-dimensional benchmark for MLLMs hallucination evaluation. *arXiv preprint arXiv:2311.07397*.

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

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. 2024b. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15840–15853.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2021. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024. Making retrieval-augmented language models robust to irrelevant context. In *The Twelfth International Conference on Learning Representations*.

Tao Zhang, Ziqi Zhang, Zongyang Ma, Yuxin Chen, Zhongang Qi, Chunfeng Yuan, Bing Li, Junfu Pu, Yuxuan Zhao, Zehua Xie, Jin Ma, Ying Shan, and Weiming Hu. 2024. mr^{2}ag: Multimodal retrieval-reflection-augmented generation for knowledge-based vqa. *arXiv preprint arXiv:2411.15041*.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Han Zhou, Xingchen Wan, Lev Proleev, Diana Mincu, Jilin Chen, Katherine Heller, and Subhrajit Roy. 2023. Batch calibration: Rethinking calibration for in-context learning and prompt engineering. *arXiv preprint arXiv:2309.17249*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A Experimental Setup and Implementation Details

This appendix provides implementation details and supplementary analyses that support the main text. We organize the material into experimental protocols, robustness checks, full results, and formulation discussions.

A.1 Decision-Margin Computation Details

All margin analyses in this paper use the same scoring convention. Given a rendered multimodal input $x = (I, T)$, where I is the image and T is the complete text prompt, we define the binary decision margin as

$$m(x) = \log P_\theta(v_+ \mid I, T) - \log P_\theta(v_- \mid I, T),$$

where v_+ and v_- are the affirmative and negative verbalizers. In the main experiments, $v_+ = \texttt{Yes}$ and $v_- = \texttt{No}$.

We score both candidate continuations under teacher forcing. For a candidate verbalizer $v = (v_1, \ldots, v_K)$ tokenized into K tokens, its sequence log-probability is computed as

$$\log P_\theta(v \mid I, T) = \sum_{k=1}^{K} \log P_\theta(v_k \mid I, T, v_{<k}).$$

We then subtract the sequence log-probabilities of the two verbalizers. This sequence-level scoring is used even when a verbalizer consists of more than one token, so the reported margin is not restricted to a single next-token logit difference.

The final binary prediction induced by the margin is determined by its sign:

$$\hat{y}(x) = \begin{cases} \texttt{Yes}, & m(x) > 0, \\ \texttt{No}, & m(x) \leq 0. \end{cases}$$

A.2 Irrelevant Text Sampling

We construct image-irrelevant textual contexts from an offline WikiText sentence pool. The contexts are not dynamically generated during inference; instead, each sample is assigned a fixed context before evaluation. This makes the context assignment deterministic and reproducible under the same random seed.

We first split corpus into sentence-level candidates and remove duplicate sentences. The candidate pool is then shuffled with a fixed seed, and contexts are assigned to samples by iterating through this shuffled pool.

The filtering process consists of two steps. First, each candidate context is processed by a rule-based script to remove judgment-related words, such as "yes" and "no". Second, we compute the CLIP similarity between the candidate context and the input image, and discard candidates with high similarity scores to avoid selecting visually related sentences.

For CLIP filter, we calculate the similarity score between context and image:

$$\mathrm{sim}(i, t) = \langle \mathrm{CLIP}(i), \mathrm{CLIP}(t) \rangle, \qquad (11)$$

where i is the input image and t is a candidate context.

A.3 Benchmark Statistics and Implementation Details

In the main experiments, we use four benchmarks. POPE, MME, and AMBER-discriminative are al-

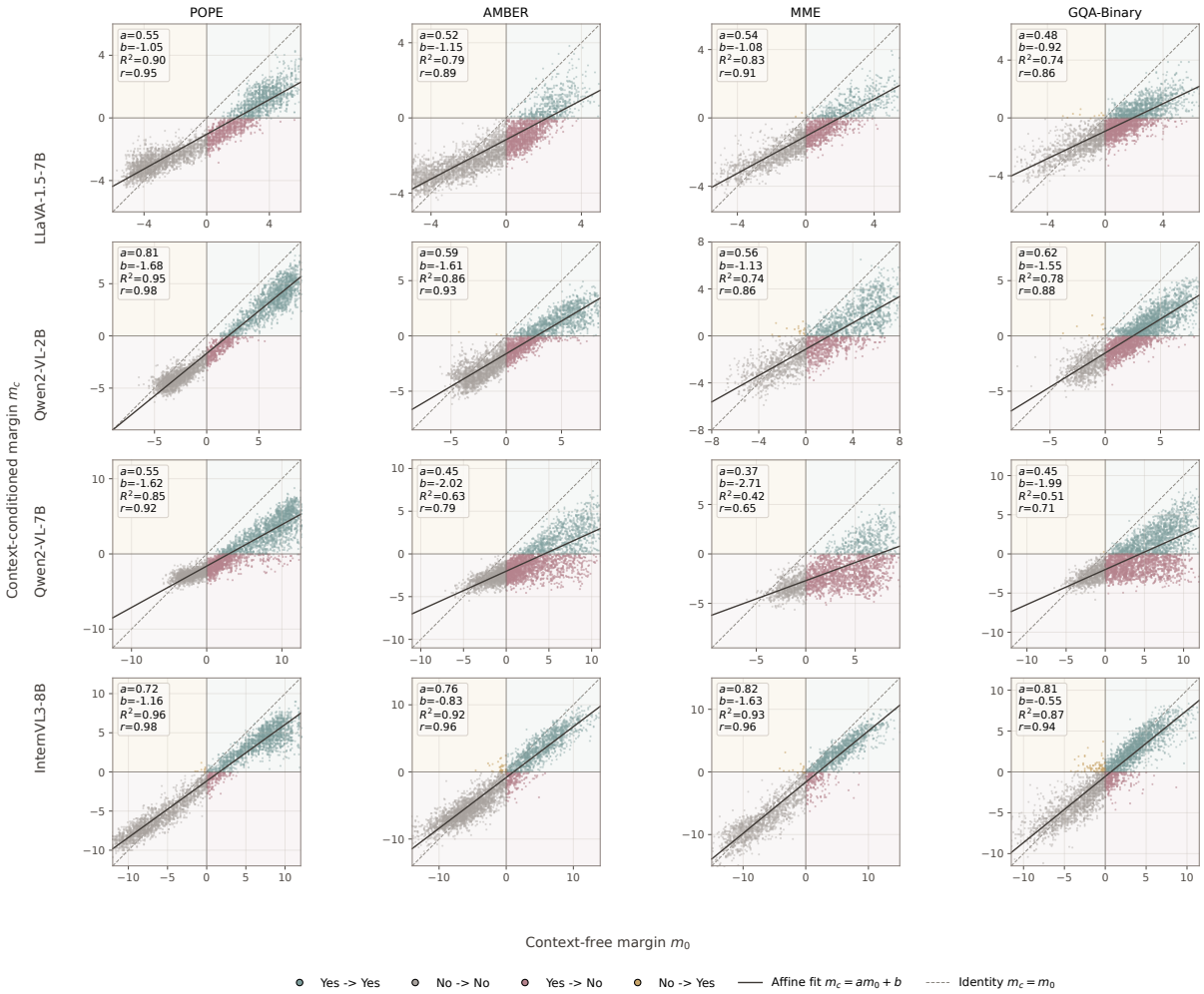

Figure 7: The affine transformation results across all models and benchmarks.

ready binary judgment tasks, so we preserve their original question format and only add a unified answer instruction. POPE contains 8,910 questions, including 4,500 Yes and 4,410 No instances (50.5%/49.5%); its three evaluation categories contain 3,000 adversarial, 3,000 popular, and 2,910 random questions. The AMBER discriminative subset contains 14,216 questions, with 4,789 Yes and 9,427 No instances (33.7%/66.3%), and is therefore the most label-imbalanced benchmark in our main evaluation. The MME input contains 2,374 questions, evenly split into 1,187 Yes and 1,187 No instances across its perception and cognition subtasks.

For GQA, we further divide the benchmark into two subsets: GQA-Binary and GQA-Judgment. GQA-Binary contains the native binary questions in GQA, for which we preserve the original question format and only append the unified answer instruction. It contains 4,525 naturally yes/no questions, with 2,328 Yes and 2,197 No instances (51.4%/48.6%). GQA-Judgment is constructed from the remaining open-ended questions by converting each sample into an answer-verification task: given the image, the original question, and a candidate answer, the model is asked to judge whether the candidate answer is supported by the image. See Appendix E.2 for detailed conversion process and corresponding experiment results.

Our experiments show that this conversion changes the task format from the model's perspective. Although GQA-Judgment still exhibits the affine margin pattern observed in our main analysis, its fitted affine parameters differ from those of native binary tasks. Therefore, we report GQA-Binary in the main text for consistency with the primary binary-judgment setting, and provide GQA-Judgment results as a complementary evaluation.

Table 5: Absolute behavioral metrics across tested models and benchmarks. Free denotes context-free input, Neutral denotes WikiText-after-neutral, and Condition denotes WikiText-after-possibly. Part (b) reports the original-style MME score.

Model	Dataset	Free				Neutral				Condition			
		Acc	F1	Rec.	Prec.	Acc	F1	Rec.	Prec.	Acc	F1	Rec.	Prec.
LLaVA-1.5-7B	POPE	85.5	84.2	76.6	93.5	84.8	83.2	74.2	94.6	85.2	83.8	75.7	93.8
LLaVA-1.5-7B	AMBER	81.2	69.8	64.8	75.8	80.9	68.7	62.4	76.5	79.9	69.2	67.1	71.4
LLaVA-1.5-7B	MME	78.5	78.9	80.5	77.4	78.6	77.2	72.6	82.5	77.6	76.3	72.1	81.0
LLaVA-1.5-7B	GQA-Binary	76.4	76.9	76.4	77.5	76.4	77.4	78.6	76.3	74.5	77.1	83.4	71.7
Qwen2-VL-2B	POPE	87.8	86.8	79.6	95.5	86.0	84.5	75.1	96.4	85.8	84.1	74.5	96.7
Qwen2-VL-2B	AMBER	85.3	78.3	78.5	78.1	84.6	75.1	68.7	82.7	80.4	66.2	56.9	79.1
Qwen2-VL-2B	MME	79.9	80.1	80.5	79.6	69.3	64.6	55.9	76.4	59.1	49.3	39.7	65.0
Qwen2-VL-2B	GQA-Binary	79.2	79.0	76.1	82.1	72.9	68.0	56.0	86.6	73.0	68.7	57.7	84.9
Qwen2-VL-7B	POPE	88.3	87.5	81.1	95.0	87.0	85.8	77.7	95.7	79.9	75.5	61.3	98.4
Qwen2-VL-7B	AMBER	86.9	80.7	81.2	80.2	87.4	80.9	79.2	82.7	77.5	53.3	38.1	88.8
Qwen2-VL-7B	MME	87.6	88.0	90.6	85.5	87.8	87.8	88.0	87.7	60.1	34.6	21.1	95.4
Qwen2-VL-7B	GQA-Binary	82.2	82.0	78.7	85.5	80.8	80.3	75.9	85.2	62.4	45.4	30.4	89.6
InternVL3	POPE	90.9	91.0	90.7	91.3	90.1	89.7	85.4	94.4	90.6	90.3	86.8	94.1
InternVL3	AMBER	88.3	83.1	85.7	80.6	88.8	82.9	81.0	84.9	88.4	82.2	79.7	84.9
InternVL3	MME	87.9	87.6	85.3	90.0	84.8	83.2	75.4	92.7	84.2	82.6	74.8	92.1
InternVL3	GQA-Binary	81.9	82.9	85.1	80.8	80.4	80.9	80.8	81.0	79.9	80.2	79.0	81.4

B Supplement Experient Results

B.1 Absolute Behavioral Metrics

Table 5 reports the absolute behavioral metrics for all tested model families. Neutral uses WikiText-after-neutral, while Condition uses WikiText-after-possibly. All binary metrics are percentages.

B.2 Full Affine Transformation Results

Figure 7 shows the affine transformation results across all models and benchmarks.

B.3 Calibration Experiments

In-domain calibration. First, we test whether affine calibration can recover context-free behavior when the affine parameters are estimated and evaluated on held-out splits from the same benchmark. Specifically, for each model and each dataset, we split the data into a calibration split and a test split. The parameters $\hat{a}$ and $\hat{b}$ are fitted on the calibration split of the same dataset, and then applied to the corresponding test split.

Cross-dataset transfer calibration. Second, we evaluate a stronger transfer setting. Instead of fitting affine parameters on the target dataset, we fit $\hat{a}$ and $\hat{b}$ on one source dataset and directly apply them to other target datasets:

$$m_{\text{cal}}^{s \to t} = \frac{m_c^t - \hat{b}_s}{\hat{a}_s}, \tag{12}$$

where $\hat{a}_s$ and $\hat{b}_s$ are estimated from the source dataset s, and m_c^t denotes the context-conditioned margin on the target dataset t. This setting uses no target-dataset calibration samples.

Table 6 reports the full in-domain calibration on LLaVA-1.5-7B. Inverse-affine calibration consistently recovers a substantial part of the context-induced degradation: it raises accuracy relative to Context by 3.2–7.9 points and reduces flips by 10.5–21.5 points across all four benchmarks. This shows that the fitted affine distortion remains predictive on held-out samples.

Table 7 reports cross-dataset transfer for LLaVA-1.5-7B. In every source–target pair, transferred parameters improve accuracy relative to the uncalibrated Context condition and sharply reduce flips. In particular, POPE-fitted parameters improve target accuracy by 2.4 points on GQA-Binary and by over 5 points on AMBER and MME, without using any target calibration examples. These results support the interpretation that, for LLaVA, irrelevant context induces a substantial model-level distortion rather than an unrelated benchmark-specific artifact.

Overall, these results show that for LLaVA, the context-induced margin distortion is structured and partially transferable: inverse-affine correction recovers context-free decisions on held-out data and under cross-dataset transfer, while also consistently improving accuracy and F1. As detailed in Appendix, other model families can exhibit a weaker relation between recovery of clean decisions and improvement of ground-truth accuracy.

Table 6: In-domain post-hoc affine calibration under WikiText-after-possibly context. Affine parameters are fitted on calibration splits and evaluated on held-out test splits. Values are percentages averaged over five splits. Flip is relative to Origin predictions.

Model	Dataset	Origin Acc.	Context Acc.	Cal. Acc.	Origin F1	Context F1	Cal. F1	Context Flip	Cal. Flip
LLaVA-1.5-7B	POPE	86.1	78.0	84.6	85.6	72.6	83.2	16.6	6.1
LLaVA-1.5-7B	AMBER	82.7	74.9	79.7	77.5	47.3	71.0	29.2	13.6
LLaVA-1.5-7B	MME	78.7	68.5	76.5	79.2	56.5	76.4	30.2	12.6
LLaVA-1.5-7B	GQA-Binary	73.2	69.2	72.3	77.7	63.5	75.8	36.3	14.8

Table 7: Cross-dataset affine calibration transfer on LLaVA-1.5-7B under WikiText-after-possibly context. Source-fitted parameters are applied directly to the target benchmark without target calibration samples. Values are percentages.

Source	Target	Origin Acc.	Cal. Acc.	Origin F1	Cal. F1	Origin Flip	Cal. Flip
POPE	AMBER	75.0	80.1	47.7	70.3	29.1	14.7
POPE	MME	68.5	76.6	56.4	76.2	30.4	12.8
POPE	GQA-Binary	69.3	71.7	63.6	75.9	36.2	13.1
AMBER	POPE	78.2	85.2	72.8	84.0	16.5	5.2
AMBER	MME	68.5	76.5	56.4	77.0	30.4	12.4
AMBER	GQA-Binary	69.3	70.3	63.6	75.6	36.2	12.0
MME	POPE	78.2	84.9	72.8	83.6	16.5	5.6
MME	AMBER	75.0	80.2	47.7	70.9	29.1	14.1
MME	GQA-Binary	69.3	71.0	63.6	75.7	36.2	12.4
GQA-Binary	POPE	78.2	84.3	72.8	82.5	16.5	7.2
GQA-Binary	AMBER	75.0	79.9	47.7	68.4	29.1	16.4
GQA-Binary	MME	68.5	76.2	56.4	74.8	30.4	14.0

C Additional Empirical Analysis

C.1 Prompt, Template, and Context-Source Ablations

We further test whether the affine effect depends on a particular prompt rendering choice. All experiments in this subsection use POPE with LLaVA-1.5-7B and reuse the same context-free origin outputs. We vary four factors: the position of the context relative to the question, the wording of the context template, the context length, and the context source. We also include template-only controls, where the template phrase is inserted without any external context text. Table 8 shows that the affine structure is robust across prompt variants, but the fitted parameters are sensitive to how the context is introduced.

Putting WikiText before the question has a much weaker effect than putting it after the question. For the possibly-related template, the before-question variant has $a = 0.726$ and only 3.7% flips, whereas the after-question variant has $a = 0.554$ and 16.5% flips. However, in many practical in-context learning pipelines, especially settings where auxiliary textual information is appended as side information or retrieved evidence before answer generation, the added context is appended after the question. Our after-question setting is designed to reflect this practically relevant regime rather than to claim position-invariant behavior. This position ablation shows that the affine structure remains observable across position variants, therefore does not affect the conclusions in the main paper.

The template-only controls address the concern that the comparison between context-free and context-conditioned prompts may be confounded by the added template itself. Even without external context text, the after-question template changes the margin distribution. However, the neutral template-only setting is substantially weaker than the possibly-related template-only setting: its slope is closer to identity ($a = 0.819$) and its flip rate is lower (6.7% vs. 13.5%). Thus, template wording is not irrelevant, but it does not eliminate the need to model the full context-conditioned transformation.

Length and source controls provide additional checks. Short, medium, and long WikiText contexts all preserve strong affine fits and similar behavioral metrics when placed before the question. Shuffling WikiText also preserves a strong affine relation, suggesting that the effect does not require coherent evidence about the queried object. In contrast, COCO captions placed after the question produce the strongest degradation among these con-

Table 8: Prompt, template, length, and source ablations on POPE with LLaVA-1.5-7B. Acc, Recall, YesRate, and Flip are reported in percentages.

Group	Variant	a	b	R^2	Acc	Recall	Flip
Position/template	Wiki before + possibly	0.726	-0.041	0.943	86.4	82.8	3.7
	Wiki after + possibly	0.554	-1.046	0.896	78.2	57.8	16.5
	Wiki before + neutral	0.755	0.084	0.949	86.2	82.7	3.7
	Wiki after + neutral	0.706	-1.243	0.921	80.1	61.6	14.5
Template only	After + possibly	0.583	-0.910	0.944	80.8	63.3	13.5
	After + neutral	0.819	-0.455	0.976	85.1	74.3	6.7
Length	Short Wiki	0.699	-0.000	0.940	86.3	82.4	4.1
	Medium Wiki	0.726	-0.041	0.943	86.4	82.8	3.7
	Long Wiki	0.743	-0.150	0.934	86.3	82.2	4.2
Source/shuffle	Shuffled Wiki	0.796	0.037	0.965	86.7	82.8	3.2
	COCO caption before	0.617	0.192	0.896	85.7	84.7	6.0
	COCO caption after	0.475	-1.204	0.892	74.2	49.7	20.7

trols, with $a = 0.475$ and a flip rate of 20.7%. Overall, these ablations support the main interpretation: irrelevant text systematically transforms the decision margin, and the strength of the transformation depends on the semantic role and placement of the added text.

C.2 Label Verbalization

Our primary analyses define a binary decision margin using the verbalizers `Yes` and `No`. This raises a potential confound: the observed affine relation might arise from the lexical properties of this particular answer pair, rather than from a structured transformation of the model's visual decision. We therefore evaluate alternative answer verbalizations with LLaVA-1.5-7B. The image, question, label, and fixed WikiText context assignment are kept unchanged; only the response instruction and the two verbalizer sequences used for margin computation are replaced. For an affirmative/negative pair (v_+, v_-), the semantic binary margin is

$$m_v = \log P(v_+ \mid I, Q, C) - \log P(v_- \mid I, Q, C).$$

We evaluate the context-free input (Origin), WikiText introduced through a neutral template (Neutral), and WikiText introduced through the possibly-related template (Possibly).

Table 9 first reports the POPE results. In addition to the standard `Yes`/`No` format, we use `True`/`False` as a natural semantic alternative.

To verify that the result is not restricted to POPE, we further evaluate the natural `True`/`False` alternative on the AMBER. As shown in Table 10, affine compression remains visible on this additional benchmark under both prompt templates.

The natural-verbalizer results support the main conclusion. With `True`/`False`, the estimated slope is below one in every condition: $a = 0.600/0.510$ on POPE and $a = 0.578/0.473$ on AMBER for the Neutral/Possibly settings. Thus, the margin-compression pattern is neither specific to the literal `Yes`/`No` tokens nor confined to one benchmark. However, the intercept is not verbalizer invariant: `Yes`/`No` yields negative shifts on POPE, while `True`/`False` yields positive shifts and substantially higher positive-answer rates, particularly under Possibly. This result suggests that the preference shift induced by irrelevant context is governed by a structured affine transformation of decision margins, which cannot be fully captured by output performance metrics alone.

C.3 Why Supporting Context Still Compresses Margins

The context-role intervention in Sec. 4 introduces a seemingly counter-intuitive result. Since supporting context is constructed to be consistent with the ground-truth answer, one might expect it to strengthen the clean visual decision and therefore yield a slope larger than one. This is not what we observe. Table 11 reports the controlled intervention on POPE with LLaVA-1.5-7B, where the image, question, context position, and rendering template are fixed while only the semantic role of the inserted context is changed.

Supporting context is substantially less destructive than the alternatives: it has the largest slope, the smallest negative intercept, and a much smaller negative shift for ground-truth `Yes` samples than irrelevant or conflicting context. Nevertheless, its slope remains below one ($a = 0.626$). This does

Table 9: Label-verbalization controls on POPE with LLaVA-1.5-7B. For each verbalizer pair, margins score the semantically affirmative token against the negative token. Behavioral metrics and Flip are percentages; affine quantities are fitted between Origin and each context condition.

Verbalizer	Condition	Acc	F1	YesRate	Flip	a	b	R^2	r
	Origin	86.4	86.1	47.1	–	–	–	–	–
`Yes/No`	Neutral	73.5	64.7	24.7	22.4	0.481	-1.229	0.854	0.924
	Possibly	75.3	68.0	26.8	20.4	0.384	-0.871	0.836	0.914
	Origin	85.6	84.7	44.1	–	–	–	–	–
`True/False`	Neutral	80.9	82.5	58.7	16.5	0.600	0.587	0.791	0.889
	Possibly	62.5	72.7	86.6	42.6	0.510	1.124	0.747	0.864

Table 10: Natural verbalizer control on AMBER discriminative with LLaVA-1.5-7B using `True/False`. Behavioral metrics and Flip are reported in percentages.

Condition	Acc	Yes	Flip	a	b	R^2
Origin	77.4	26.9	–	–	–	–
Neutral	68.8	58.7	32.9	0.58	0.61	0.66
Possibly	48.4	84.2	57.4	0.47	1.04	0.57

Table 11: Context-role intervention on POPE with LLaVA-1.5-7B. Δm_+ and Δm_- denote the mean shift $m_c - m_0$ for ground-truth `Yes` and `No` samples, respectively.

Context	a	b	R^2	Δm_+	Δm_-
Irrelevant	0.554	-1.046	0.896	-1.96	0.16
Supporting	0.626	-0.287	0.873	-0.94	0.61
Conflicting	0.367	-1.228	0.869	-2.63	0.59

not contradict the usefulness of supporting evidence, because the fitted slope measures the global dynamic range of the context-conditioned margins relative to clean margins; it does not directly measure whether a context is label-consistent for an individual sample.

In particular, the supporting intervention is label-conditioned: positive questions receive a statement that the queried object is present, whereas negative questions receive a statement that it is absent. Its effect is therefore more appropriately summarized as

$$m_c = \alpha m_0 + \delta(y) + \epsilon,$$

where $\delta(y)$ is a label-dependent shift, rather than as a pure margin amplifier $m_c = \gamma m_0$ with $\gamma > 1$. On POPE, the clean mean margins for ground-truth `Yes` and `No` samples are 2.10 and -2.76, giving a mean class separation of 4.86. Under supporting context, the corresponding means are 1.16 and -2.15, reducing the separation to 3.31. Supporting text preserves the decision direction better than other context roles, while still narrowing the overall margin range; an ordinary least-squares fit thus naturally yields $a < 1$.

This compression is also consistent with how the context is presented to the model. The supporting statement is introduced as caption-like natural language under a possibly-related template, rather than as an oracle answer. Consequently, the model may treat it as uncertain evidence, reduce reliance on the original visual signal, or imperfectly process the negated form used for negative samples. The latter possibility is reflected by $\Delta m_- = 0.61$: because positive margins favor `Yes`, a perfectly effective supporting statement for a `No` example would shift the margin in the negative direction, whereas the observed average shift is positive.

We therefore interpret supporting context as reducing harmful context-induced distortion rather than amplifying clean visual margins beyond their original scale. The ordering

$$a_{\text{supporting}} > a_{\text{irrelevant}} > a_{\text{conflicting}}$$

still shows that the semantic role of text systematically changes the affine operator. At the same time, $a_{\text{supporting}} < 1$ indicates that even label-consistent context is integrated as imperfect prompt-conditioned evidence, not as deterministic supervision.

D Affine Robustness Analysis

D.1 Context-Only Prior Subtraction

Another alternative explanation is that irrelevant context only contributes an independent answer prior. Under this view, the context-conditioned margin could be decomposed as the clean margin plus a context-only bias term. To test this, we estimate a context prior margin m_{prior} by removing the original image-question evidence and then subtract it from the context-conditioned margin:

$$m_{\text{sub}} = m_c - m_{\text{prior}}.$$

Table 12: Context-only prior subtraction on POPE with LLaVA-1.5-7B. We report the affine fit between m_0 and the prior-subtracted margin $m_{\text{sub}} = m_c - m_{\text{prior}}$. Acc, F1, YesRate, and Flip are reported in percentages.

Context	Prior estimator	a	b	R^2	r	Acc	F1	YesRate	Flip
neutral	context-only blank	0.705	-0.872	0.898	0.948	81.7	78.5	34.5	11.4
neutral	question+context blank	0.700	1.894	0.904	0.951	79.7	82.4	64.9	19.1
neutral	context-only same image	0.705	-0.653	0.902	0.950	82.7	80.0	35.9	10.0
possibly	context-only blank	0.552	-0.233	0.868	0.932	83.3	81.3	38.5	8.2
possibly	question+context blank	0.549	1.599	0.872	0.934	75.5	79.7	70.0	24.3
possibly	context-only same image	0.551	0.103	0.862	0.928	84.8	83.5	41.9	6.4

We consider three prior estimators: context-only with a blank image, question-plus-context with a blank image, and context-only with the original image.

Table 12 reports the POPE results with LLaVA-1.5-7B. The context-only prior itself is almost uncorrelated with the clean margin: across variants, its fitted slope is close to zero and R^2 is nearly zero. After subtracting this prior, the affine relation remains strong, and the slope remains substantially below one. This is the key observation: subtracting an additive answer prior improves behavior, but it does not remove the margin compression effect.

For the possibly-related context, the uncorrected context margin has $a = 0.554$, $b = -1.046$, and $R^2 = 0.896$. Subtracting the context-only prior changes the intercept and improves behavioral metrics, but the slope remains around 0.55. Similarly, under the neutral context, the prior-subtracted slope remains around 0.70. Therefore, the irrelevant-context effect cannot be fully explained as a standalone Yes/No prior added to the decision margin. Instead, it also changes the effective scaling of the original visual evidence, consistent with the affine interpretation in the main text.

D.2 Nested Margin Models

One possible concern of the affine law is that irrelevant context only induces a simpler transformation, such as a global Yes/No answer-prior shift or a uniform confidence scaling. To verify this concern, we compare four nested margin models:

$$m_c = m_0, \qquad m_c = m_0 + b,$$
$$m_c = am_0, \qquad m_c = am_0 + b.$$

These correspond to identity, shift-only, scale-only, and full affine transformations, respectively. We fit each model on a calibration split and evaluate it on held-out samples. In addition to held-out R^2 and MAE, we also evaluate whether each model predicts which samples will flip after adding context.

Table 13 reports the available result on POPE with LLaVA-1.5-7B under the WikiText-after-neutral setting. The full affine model provides the best held-out fit, reducing MAE from 1.34 for identity and 0.93 for shift-only to 0.52, and increasing held-out R^2 to 0.92. It also gives the strongest flip prediction performance, with F1 increasing from 0.71 for shift-only to 0.85. This supports the conclusion that the context effect is not merely a constant answer-direction shift, nor merely a confidence scaling effect.

For reference, the clean prediction accuracy on the same held-out split is 86.4% with F1 85.9%. Adding context without correction reduces accuracy to 80.0% and F1 to 75.6%, with a flip rate of 14.5%. The full affine correction reduces the flip rate to 5.0% and recovers most of the lost F1. The fact that the full affine model substantially improves held-out margin prediction over both shift-only and scale-only baselines is the key diagnostic: irrelevant context changes both the scale of the original margin and its answer-direction offset.

D.3 Local Fits Across Margin Regions

The global affine relation could be inflated by the dynamic range of high-confidence samples. To check this, we fit the affine relation separately within low-, mid-, and high-margin regions, where the regions are defined by quantiles of $|m_0|$ for each model–dataset pair. We also repeat the global fit after trimming extreme margins. This analysis serves two purposes. First, it tests whether affine regularity remains visible after removing extreme margins. Second, it quantifies where prediction flips occur. Table 14 shows two consistent patterns. First, the affine fit remains strong after trimming the extreme 5–10% of margins. For example, LLaVA on POPE has global $R^2 = 0.896$ and trimmed $R^2 = 0.899$, while Qwen2-2B on AMBER has global $R^2 = 0.863$ and trimmed $R^2 = 0.848$. Thus, the affine law is not solely

Table 13: Nested margin-model comparison on POPE with LLaVA-1.5-7B. The context variant is WikiText-after-neutral. Fit metrics evaluate m_c prediction on held-out samples. Calibration metrics evaluate the corrected margin as a binary prediction.

Model	R^2	MAE	Flip F1	Calib Acc	Calib F1	Calib Flip
Identity	0.522	1.337	0.000	0.800	0.756	0.145
Shift-only	0.760	0.934	0.711	0.854	0.841	0.054
Scale-only	0.642	1.254	0.000	0.800	0.756	0.145
Full affine	**0.920**	**0.518**	**0.849**	**0.855**	**0.842**	**0.050**

Table 14: Local margin analysis under WikiText-after-possibly. For each model and dataset, we report the global affine fit and the flip rate within low-, mid-, and high-margin regions. Flip rates are percentages.

Model	Dataset	Global a	Global R^2	Trim R^2	Low Flip	Mid Flip	High Flip
LLaVA	POPE	0.554	0.896	0.899	42.2	6.4	0.1
LLaVA	AMBER	0.525	0.786	0.785	52.4	31.6	2.6
LLaVA	MME	0.542	0.830	0.805	52.1	33.0	5.3
LLaVA	GQA	0.476	0.743	0.691	56.3	42.3	9.6
Qwen2-2B	POPE	0.557	0.859	0.841	42.6	15.4	3.1
Qwen2-2B	AMBER	0.592	0.863	0.848	46.0	18.7	3.0
Qwen2-2B	MME	0.561	0.740	0.730	42.8	21.7	5.9
Qwen2-2B	GQA	0.615	0.775	0.750	53.6	33.0	5.5

an artifact of a few very large margins. Second, flips are concentrated near the decision boundary. Across both model families and all four benchmarks, low-margin samples have much higher flip rates than high-margin samples. For LLaVA, the low-margin flip rate ranges from 42.2% on POPE to 56.3% on GQA-Binary, while the corresponding high-margin flip rates range only from 0.1% to 9.6%. Qwen2-2B shows the same pattern. This supports the analysis in the main text: irrelevant context changes the margin distribution globally, but the final prediction changes primarily when the original decision is close to the boundary. It also motivates gated calibration, where correction is focused on near-boundary examples instead of being applied uniformly to all samples.

D.4 Random Seed Robustness

The main experiment assigns one fixed irrelevant WikiText context to each POPE example. To test whether the affine relation is an artifact of a particular random assignment, we independently rebuild the WikiText assignment with seeds 13 and 21, while keeping the model, samples, context position, and possibly-related rendering template fixed. We reuse the context-free margins and rerun only the context-conditioned inference.

The estimated parameters are nearly invariant across the two additional context assignments. Moreover, they closely match the main seed-42 fit ($a = 0.5538$, $b = -1.0465$, $R^2 = 0.8964$, and $r = 0.9468$). Thus, the reported affine transformation is not attributable to an idiosyncratic WikiText sampling seed.

Table 15: Random-context-seed robustness on POPE with LLaVA-1.5-7B under the WikiText-after-possibly condition. Behavioral metrics and Flip are reported in percentages.

Seed	a	b	R^2	r
13	0.5541	-1.0441	0.8968	0.9470
21	0.5536	-1.0492	0.8952	0.9461
Mean	0.5539	-1.0466	0.8960	0.9466
Std.	0.0003	0.0036	0.0012	0.0006

D.5 Residual Analysis

We further examine whether the fitted affine transformation leaves systematic structure in the prediction error. For each model–benchmark pair under the WikiText-after-possibly condition, we compute the residual

$$\epsilon_i = m_c^{(i)} - \left(a m_0^{(i)} + b\right), \tag{13}$$

where a and b are fitted on all samples of the corresponding setting. Figure 8 plots these residuals against the context-free margin for four model families and four binary benchmarks. The red curve shows the mean residual within bins of m_0 and therefore exposes systematic departures from a single global affine map.

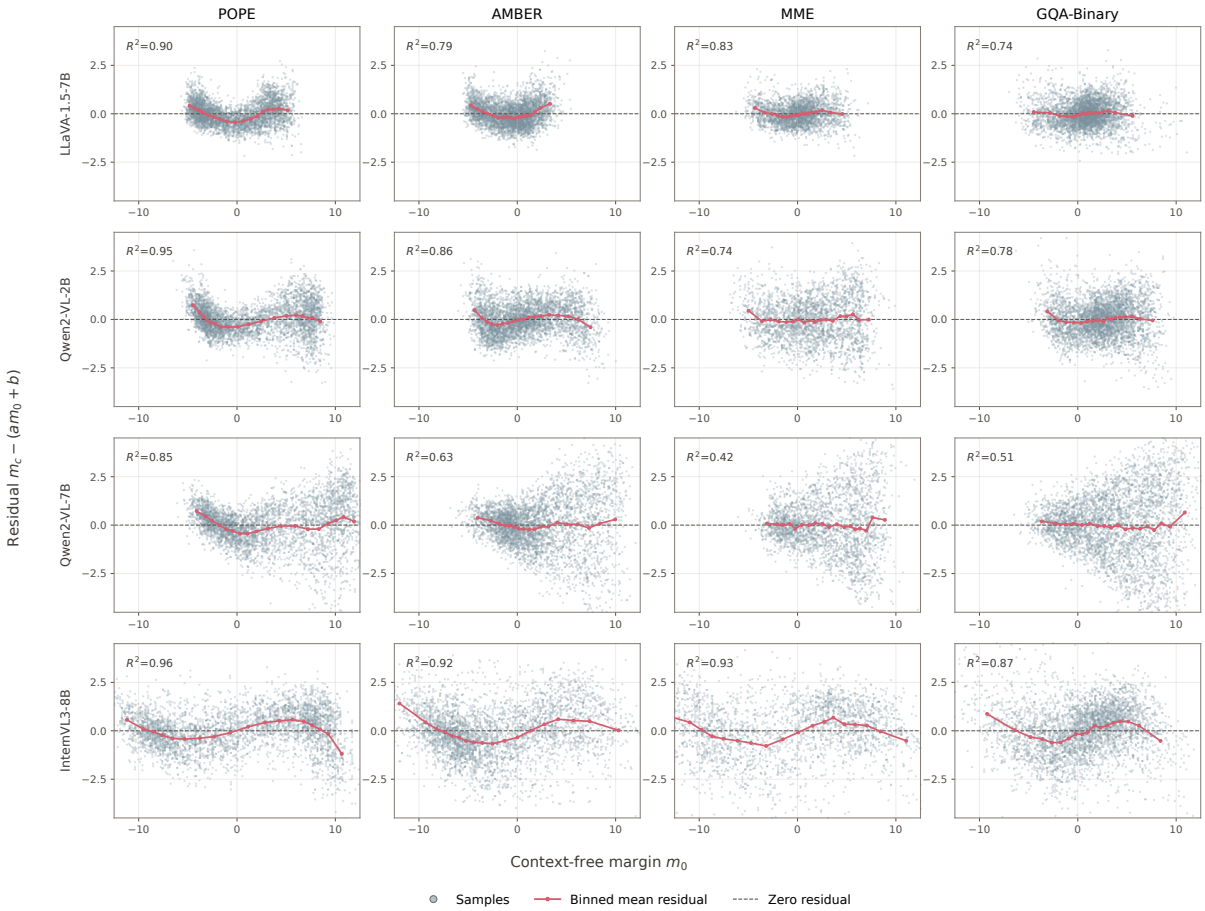

Figure 8: Residual diagnostics for the affine fit under the WikiText-after-possibly context. Columns denote benchmarks and rows denote models. Each point is a sample residual $m_c - (am_0 + b)$; the red curve is the binned mean residual. All panels use shared symmetric axes so that residual dispersion is directly comparable.

For LLaVA-1.5-7B, the residual cloud remains comparatively compact across all benchmarks, with R^2 ranging from 0.74 on GQA-Binary to 0.90 on POPE. InternVL3-8B also retains high explanatory power ($R^2 = 0.87$–0.96), although its larger margin scale produces a wider absolute residual spread. Qwen2-VL-2B is broadly compatible with the affine approximation, with reduced fit quality on MME and GQA-Binary. In contrast, Qwen2-VL-7B exhibits visibly larger and margin-dependent residual dispersion on AMBER, MME, and GQA-Binary, where R^2 drops to 0.63, 0.42, and 0.51, respectively. Thus, the residual analysis supports the affine law as a strong first-order description across models, while also identifying settings in which a single global affine operator is incomplete. This qualification is consistent with the Qwen2-VL calibration results above: calibration can reduce context-induced decision changes even when the remaining residual structure prevents uniform accuracy recovery.

E Additional Discussion and Experiments

E.1 Why Binary Judgment?

We adopt a binary judgment formulation to obtain a controlled and comparable measure of model preference. Open-ended multimodal generation involves many confounding factors, including answer phrasing, verbosity, decoding strategy, and automatic evaluation noise. These factors make it difficult to determine whether a change in the output reflects a genuine change in visual judgment or merely a change in surface generation. By constraining the answer space to $\{\texttt{Yes}, \texttt{No}\}$, we isolate the model's binary decision-making from the confounding effects of open-ended generation and uncontrolled answer formats.

This formulation also enables margin-level analysis. For each input, the model assigns probabilities to the same two candidate answers, allowing us to define a decision margin as the log-probability difference between Yes and No. Because the candi-

Table 16: GQA-Judgment behavior and affine fit for LLaVA under the WikiText-after-possibly context. Origin accuracy is evaluated without irrelevant context, while context metrics are evaluated after inserting irrelevant context.

Affine Fit		Behavior	
Metric	Value	Metric	Value
Slope a	0.336	Origin Acc.	81.3
Intercept b	-1.377	Context Acc.	61.6
R^2	0.580	Context F1	40.4
Pearson r	0.762	Yes Rate	14.5
		Flip Rate	32.1

date set is fixed across the context-free and context-conditioned inputs, the margin provides a direct score-level probe of how the same image-question pair is affected by the added context. This is essential for identifying the affine relation between the original and context-conditioned margins.

The binary setting is not intended to replace open-ended evaluation. Rather, it serves as a controlled abstraction for studying whether image-irrelevant context changes visually grounded decisions. Several benchmarks, such as POPE and MME, are already naturally formulated as yes/no questions; for open-ended benchmarks, we convert them into answer-verification questions so that the same analysis can be applied consistently. This design allows us to compare predictions, flip rates, and margin shifts under a unified evaluation protocol.

E.2 GQA-judgment Conversion and Results

For non-binary GQA questions, Following (Kalai et al., 2025),, we convert them into answer-verification questions using scene-graph annotations. Given an original question and its ground-truth answer, we construct a positive verification sample by using the ground-truth answer as the candidate answer. To construct a negative sample, we sample an incorrect candidate from the GQA scene graph while preserving the semantic type of the answer whenever possible. The final question is converted into following format:

```
Question: [Question] Candidate answer:
[Answer]. Is the candidate answer
correct according to the image?
```

Table 16 shows that GQA-judgment behaves differently from the cleaner binary existence benchmarks. We use LLaVA as the testing model. Without irrelevant context, the converted task is solvable: LLaVA reaches 81.3% accuracy. adding irrelevant context can strongly suppress affirmative verification judgments. Under WikiText-after-possibly, the fitted slope is only $a = 0.336$ with $R^2 = 0.580$, and the YesRate drops to 14.5%. Additionally, the affine parameter is different, indicating that the prompt template or task will change the affine parameter.

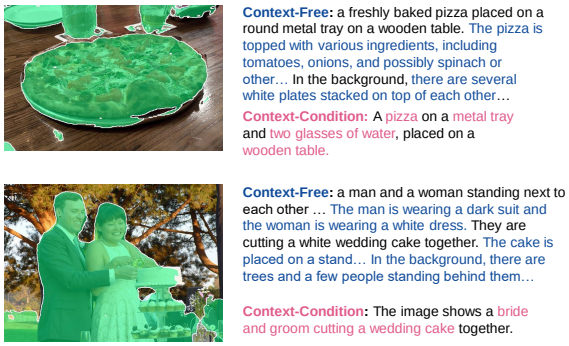

Figure 9: A case of our open-ended generation experiment. We use the green mask to indicate the region covered by the model's prediction under the context-condition setting. After adding irrelevant context, the model tends to describe only salient objects in the image, while peripheral objects are no longer mentioned.

E.3 Preliminary Study on Open-Ended Generation

To further examine whether irrelevant context affects open-ended generation, we conduct a preliminary study using AMBER generative set. We follow the setting of context-free and context-condition, and ask the model to describe each given image. To be specific, for all settings, the question is:

```
Describe this image in detail.
```

The **context-free** input contains only the image and the question:

```
[Image] [Question]
```

The **context-condition** input appends the sampled irrelevant context and the template:

```
[Image] [Question] [Template] [Context]
```

The experiment result is in Table 17. We find that adding irrelevant context substantially changes the model's behavior: the generated responses become significantly shorter, accompanied by a large drop in coverage.

Table 17: AMBER generative evaluation with LLaVA-1.5-7B. `Free` and `Condition` denote context-free and context-conditioned generation, respectively. Hal., Cov., Cog., and Len. denote hallucination rate, object coverage, cognitive error, and average output length.

Setting	CHAIR ↓	Cov. ↑	Hal. ↓	Cog. ↓	Len.
`Free`	7.9	49.4	33.3	4.0	68.1
`Condition`	3.8	39.3	9.5	0.7	16.1

The example in Figure 9 provides an intuitive illustration of this behavioral shift: the model tends to describe only salient objects in the image, while peripheral objects are no longer mentioned.

One possible interpretation is that irrelevant context changes the model's effective decision margin rather than simply injecting random noise into generation. Under the context-induced affine transformation, many visual claims are shifted toward a more conservative decision region. Consequently, objects or attributes that originally receive only weak or moderate support are more likely to fall below the implicit threshold for verbalization. These marginal claims often correspond to small objects, peripheral regions, fine-grained attributes, or visually ambiguous details. By contrast, highly salient objects tend to have stronger visual evidence and larger margins, making them more robust to the margin shift and more likely to remain in the output. This explains why the model still describes the dominant objects in the scene while omitting many peripheral details.

From this perspective, the reduced coverage is not merely a side effect of shorter responses, but reflects a systematic change in the model's generation policy: irrelevant context makes the model more selective about which visual claims to express, suppressing uncertain details and producing more conservative descriptions.

However, further analysis is needed to verify the causal link between margin suppression and reduced coverage, and to better understand which visual claims are most affected.