Visual Priming Effect on Large-scale Vision Language Models

Daiki Yoshida[†], **Haruki Sakajo**[†], **Kazuki Hayashi**[†],
Yusuke Sakai[†], **Hidetaka Kamigaito**[†], **Katsuhiko Hayashi**[‡], **Taro Watanabe**[†]
[†]Nara Institute of Science and Technology (NAIST) [‡]The University of Tokyo
{yoshida.daiki.ye6, sakajo.haruki.sd9, hayashi.kazuki.hl4}@naist.ac.jp
{sakai.yusuke.sr9, kamigaito.h, taro}@is.naist.jp
katsuhiko-hayashi@g.ecc.u-tokyo.ac.jp

Abstract

Large-scale Vision-Language Models (LVLMs) integrate linguistic and visual information, demonstrating advanced task-solving capabilities. These models are originally derived from Large Language Models, leading to strong capabilities for language tasks. However, the impact of additional visual information on model responses remains insufficiently understood. In this study, we focus on the priming effect, a psychological phenomenon, to investigate how visual information influences language task processing. We present additional intentionally designed images alongside two types of language tasks with different characteristics and analyze changes in the model's responses. Our experimental results show that model responses shift in the direction intended by the image, suggesting that LVLMs do not simply ignore visual information but actively incorporate it into language processing. Furthermore, the similarity between this behavior and priming effects observed in human cognition suggests that LVLMs may share certain aspects of human cognitive mechanisms.

1 Introduction

Large-scale Vision-Language Models (LVLMs) have demonstrated advanced performance in visual information processing by integrating visual and linguistic information. However, their outputs sometimes show language biases or vision biases (Chen et al., 2024), implying a difference in the relative importance of visual and linguistic information. Yet, the extent to which visual information influences model outputs, as well as the nature of the interaction between visual and linguistic modalities, remains unclear.

Some studies (Kassner and Schütze, 2020; Misra et al., 2020; Michaelov et al., 2023; Sharma et al., 2024; Jumelet et al., 2024; Sakai et al., 2025) investigating the behavior of Language Models (LMs) in terms of cognitive science have revealed that LMs show "priming". Priming is a psychological phenomenon where a preceding stimulus influences the processing of a subsequent stimulus (Tulving and Schacter, 1990). Kassner and Schütze (2020) and Misra et al. (2020) investigated this effect by observing changes in the outputs when placing words at the beginning of the inputs. For example, when the word "iPhone" is given as a primer before the sentence "I work for __", the model's prediction probability for "Apple" increases significantly, as illustrated in Figure 1. A similar effect is expected with a visual primer such as an image of "iPhone".

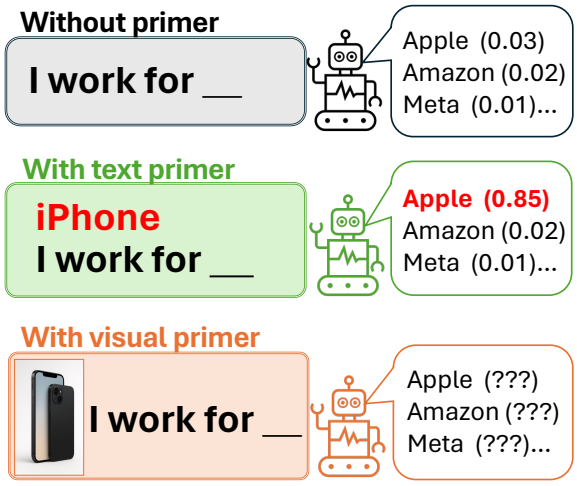


Figure 1: Priming effects in LVLMs. Textual and visual cues modulate prediction probabilities. The likelihood of completing "I work for ___" with "Apple" increases when preceded by a relevant word or image("iPhone").

This raises a question: *Do LVLMs show priming?* While visual and textual inputs are tokenized separately via a vision encoder and tokenizer, it remains unclear whether these tokens function similarly within the model. If LVLM outputs vary depending on preceding visual input, this would indicate priming effects akin to those in humans. We

Proceedings of Recent Advances in Natural Language Processing,pages 1385–1395
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-160

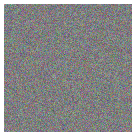
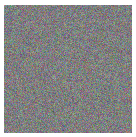
Text	Guiding	Misleading	Unrelated	Noise
Question: What position does ***Roman Turek*** play? (Answer: goaltender)				
Sentence: Equals the original and in some ways even betters it. Is this sentence negative or positive? (Answer: positive)				

Table 1: Examples of visual priming conditions applied to two tasks: Entity Questions (top) and SST-2 (bottom). The **Text** column shows the prompts given to the models for each task (excluding the answer parts).

reinterpret known LVLM biases through the lens of priming to examine how visual input shapes text generation. In this study, we investigate how visual information facilitates the correct answer and influences the confidence of entity-centric question-answering tasks and sentiment classification tasks, which do not require visual information. Our experimental results show that even in tasks solvable with language information alone, LVLMs change their outputs and confidence depending on the type of additional images presented. This finding implies that LVLMs show the priming effect, highlighting the similarity with human cognition.

2 Background and Related Work

LVLMs. LVLMs (Liu et al., 2024; Bai et al., 2023; Wang et al., 2024; Ye et al., 2024; Radford et al., 2021) are constructed by integrating a pre-trained vision encoder with large language models (LLMs) (Touvron et al., 2023; OpenAI et al., 2024; Chiang et al., 2023). Vision encoders are trained by some training strategies such as contrastive learning, e.g., CLIP (Radford et al., 2021), supervised pretraining on large-scale image classification datasets, e.g., ViT (Dosovitskiy et al., 2021) and BLIP-2 (Li et al., 2023). During LVLM training, datasets typically consist of highly correlated pairs of visual and textual information. Consequently, situations where additional visual input is provided for tasks that can be solved using language alone are rarely considered in the training process. As a result, how visual information contributes in scenarios where it is unnecessary or in tasks where linguistic data play a dominant role, as well as how visual and linguistic modalities interact under such conditions, remains insufficiently understood (Cao et al., 2022; Kawaharazuka et al., 2024; Hayashi et al., 2024; Ozaki et al., 2025b,a; Sakajo et al., 2025).

Priming. Priming is a phenomenon where the presentation of a preceding stimulus (primer) facilitates the processing of a subsequent stimulus (target) (Tulving and Schacter, 1990; Bargh and Chartrand, 2000; Zorzi et al., 2004; Lee et al., 2023; Sharma et al., 2024). However, under certain conditions, the primer may inhibit target processing, a phenomenon known as negative priming (Tipper, 1985; Milliken and Rock, 1997; Reitter and Moore, 2007; Schoch et al., 2020). In this study, to distinguish it from negative priming, we refer to cases where the primer facilitates target processing as positive priming. Here, we assume a situation where the primer corresponds to visual information and the target corresponds to the processing of a language task. Both visual and text-based priming have been shown to shorten human response times.

Priming in LMs. LMs also exhibit priming effects, similar to humans: the output probability of a target word varies depending on the provided text primer, as illustrated in Figure 1. For instance, Misra et al. (2020) investigated the influence of lexical cues on BERT (Devlin et al., 2019) and demonstrated that BERT exhibits a priming effect akin to that of humans. Their study measured how the probability of predicting a target word changed when preceded by either a semantically related or unrelated word, using surprisal as a key metric. The results showed that BERT's priming effect was particularly pronounced in contexts with low constraint but weakened as contextual constraints increased. In addition, they introduced Facilitation as a measure of the priming effect, defining it as the difference in surprisal between conditions where a related versus an unrelated word was used.

3 Task Definition

We investigate how the behavior of the model changes when various types of visual information, containing different intentions, are added to tasks that can originally be solved solely with linguistic information (hereafter, language-only tasks). This investigation aims to examine the priming effect in the context of visual information by evaluating how the addition of visual information affects the accuracy and confidence of models handling linguistic tasks. In the experiment, one of the following four types of images is provided in addition to the text.

Each image type is designed to test whether visual information facilitates, inhibits, or remains neutral to the language understanding process:

Guiding Images. Images designed to induce a *positive* priming effect, making the target word more easily through visual information.

Misleading Images. Images designed to induce a *negative* priming effect, leading the model to recall a word different from the target word.

Unrelated Images. Images containing visual information but unrelated to the task. These test the model's behavior under visually irrelevant stimuli.

Noise Images. Images with randomly assigned pixel values, containing minimal information. These serve as a baseline to isolate biases caused by the presence or absence of a priming effect.

4 Evaluation Metrics

To assess how visual information affects model performance in language-only tasks, we employ three complementary evaluation metrics. Each captures a different aspect of the model's behavior, together providing a comprehensive view of the presence and nature of visual priming effects in LVLMs.

4.1 Accuracy

Accuracy captures the impact of visual information, indicating whether its presence and semantic intent help or hinder task success in a binary sense (correct vs. incorrect).

We use **force decoding** to obtain the log-probabilities of the target word under each input scenario. Let $\mathcal{W} = \{w_1, w_2, \ldots, w_K\}$ denote the set of candidate words. For a given input x, the log-probability of each candidate word w_k is:

$$\log P(w_k \mid x).$$

We convert these log-probabilities into normalized probabilities using the softmax function:

$$p(w_k \mid x) = \frac{\exp(\log P(w_k \mid x))}{\sum_{j=1}^{K} \exp(\log P(w_j \mid x))}.$$

The model prediction is the candidate word $\hat{w}$ that has the highest probability:

$$\hat{w} = \arg\max_{w \in \mathcal{W}} P(w \mid x).$$

Accuracy is then computed as the proportion of cases where the predicted word matches the gold answer word:

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(\hat{w}_i = w_i^*),$$

where N is the total number of inputs and $\mathbb{I}$ is the indicator function that returns 1 if the prediction is correct and 0 otherwise.

4.2 Facilitation

Accuracy is a straightforward and intuitive metric to evaluate model performance, as it quantifies the proportion of correct predictions. However, accuracy primarily reflects changes in correctness, making it less sensitive to subtle shifts caused by visual information. In other words, even if a model's prediction is subtly influenced by a visual prime, accuracy can only capture the effect when the predicted label itself changes. This limitation makes it difficult to fully capture the priming effect, especially when visual information affects confidence or probability without altering the final prediction. As a result, the nuanced impact of visual information often remains hidden when relying solely on accuracy as a metric. To address this issue, we introduce the **Facilitation** metric introduced by Misra et al. (2020) with slight modifications. Facilitation directly compares the log-probabilities of the target word under different visual conditions for the same input. By focusing on the difference in prediction probabilities rather than just correctness, Facilitation provides a more nuanced measure of how visual primes influence model outputs. This allows us to evaluate the priming effect even when accuracy remains unchanged.

Facilitation $\mathbb{F}$ is defined as the difference between the log-probability of the target word T when the visual prime Pr_1 and the linguistic context C are presented immediately before T, and the

log-probability of the same T under the identical context C but with an alternative visual prime Pr_2:

$$\mathbb{F} = \log P(T \mid C, \mathrm{Pr}_1) - \log P(T \mid C, \mathrm{Pr}_2).$$

The conditional log-probability of the target word is computed as the mean log-probability over all tokens that constitute $T = [T_1, T_2, \ldots, T_n]$:

$$\log P(T \mid C, \mathrm{Pr}) = \frac{1}{n} \sum_{i=1}^{n} \log P_{\mathrm{LVLM}}(T_i \mid T_{<i}, C, \mathrm{Pr}).$$

where P_{LVLM} denotes the conditional probability distribution defined by the LVLM. A positive value ($\mathbb{F} > 0$) indicates a positive priming effect, where Pr_1 facilitates the prediction of T more than Pr_2; conversely, a negative value ($\mathbb{F} < 0$) indicates a negative priming effect.

To isolate the semantic contribution of the visual cue, rather than the mere presence or absence of a stimulus, we consistently use a *noise* image as the baseline. This ensures that the difference measured captures only the semantic alignment of the visual prime. We define three variants of Facilitation according to image type used (guiding, misleading, and unrelated), as summarized in Table 2. This design enables a more controlled comparison of the priming effect under different semantic conditions, without confounding modality effects.

4.3 Expected Calibration Error (ECE)

The **confidence** of the prediction is then defined as the probability assigned to the predicted word:

$$\text{Confidence} = P(\hat{w} \mid x).$$

However, having high confidence does not always imply correctness. There are cases where a model exhibits high confidence but produces incorrect predictions (overconfidence), as well as cases where it shows low confidence despite being correct (underconfidence). To quantify this discrepancy between the model's confidence and its actual accuracy, we introduce the Expected Calibration Error (ECE) (Pakdaman Naeini et al., 2015; Guo et al., 2017). We first divide the confidence range $[0, 1]$ into M equal-width bins B_m. In this study, we set $M = 10$. The definition of each bin is as follows:

$$B_m = \left\{ x \mid \frac{m-1}{M} < p(w \mid x) \leq \frac{m}{M} \right\}.$$

For each bin, we compute the *accuracy* and *average confidence* as follows:

$$\mathrm{acc}(B_m) = \frac{1}{|B_m|} \sum_{x \in B_m} \mathbb{I}(\hat{w} = w^*),$$

	Pr_1	Pr_2
$\mathbb{F}_{\text{guiding}}$	Guiding Image	Noise Image
$\mathbb{F}_{\text{misleading}}$	Misleading Image	Noise Image
$\mathbb{F}_{\text{unrelated}}$	Unrelated Image	Noise Image

Table 2: Definition of three Facilitation variants based on the type of visual prime (Pr_1).

Models	HuggingFace ID
Qwen2.5-VL-7B-Instruct	Qwen/Qwen2.5-VL-7B-Instruct
Llama-3.2-11B-Vision-Instruct	meta-llama/Llama-3.2-11B-Vision-Instruct
mPLUG-Owl3-7B	mPLUG/mPLUG-Owl3-7B-240728
Phi-3.5-vision-instruct	microsoft/Phi-3.5-vision-instruct
Qwen2-7B-Instruct	Qwen/Qwen2-7B-Instruct
Qwen2.5-7B-Instruct	Qwen/Qwen2.5-7B-Instruct

Table 3: Correspondence between the model used and the model name in HuggingFace (Wolf et al., 2020).

$$\mathrm{conf}(B_m) = \frac{1}{|B_m|} \sum_{x \in B_m} P(\hat{w} \mid x).$$

The ECE is then defined as the weighted average of the absolute differences between accuracy and confidence over all bins:

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \left| \mathrm{acc}(B_m) - \mathrm{conf}(B_m) \right|,$$

where $|B_m|$ is the number of samples in bin B_m. This metric quantifies the degree of mismatch between predicted confidence and observed accuracy, thus providing insight into whether the model's confidence is well-calibrated. ECE allows us to evaluate, in a principled way, not only how visual information shifts output probabilities, but also how it affects the *reliability* of those probabilities. By comparing ECE across different visual-prime scenarios, we can capture qualitative differences in the resulting priming effects.

5 Experimental Settings

Entity Questions (EQ). *EQ* (Sciavolino et al., 2021) is a question-answering task that derives answers related to specific entities. As shown in the upper section of Table 1, the questions are constructed by inserting entities into predefined question patterns. In our experiments, we use 13,307 test samples to investigate how visual information influences the outputs.

SST-2. *SST-2* (Socher et al., 2013) is a sentiment classification task in which given sentences are classified as positive or negative (lower section of Table 1). We use 1,821 test samples to examine how image information affects sentiment judgment.

task	model	guiding	misleading	unrelated	noise	text only	LLM
EQ	Qwen2.5-VL	**0.577**	0.433	0.449	0.500	0.491	0.518
	mPLUG-Owl3	**0.569**	0.467	0.474	0.506	0.505	0.581
	Llama-3.2	**0.726**	0.579	0.597	0.632	0.652	-
	Phi-3.5	**0.618**	0.461	0.494	0.520	0.530	-
SST-2	Qwen2.5-VL	**0.930**	0.808	0.885	0.881	0.872	0.924
	mPLUG-Owl3	**0.940**	0.831	0.912	0.920	0.861	0.910
	Llama-3.2	**0.950**	0.796	0.890	0.836	0.921	-
	Phi-3.5	**0.998**	0.164	0.907	0.912	0.909	-

Table 4: **Results of Accuracy for Each Model in EQ and SST-2: blue** values indicate the highest metric among the LVLMs, while red values indicate the lowest metric, both excluding LLM.

task	Facilitaion	Qwen2.5-VL	mPLUG-Owl3	Llama-3.2	Phi-3.5
EQ	$\mathbb{F}_{\text{guiding}}$	-0.046	0.118	0.489	0.918
	$\mathbb{F}_{\text{misleading}}$	-0.556	-0.399	-0.735	-0.756
	$\mathbb{F}_{\text{unrelated}}$	-0.478	-0.350	-0.570	-0.360
SST-2	$\mathbb{F}_{\text{guiding}}$	0.202	-0.196	-0.831	0.199
	$\mathbb{F}_{\text{misleading}}$	0.127	-0.571	-1.299	-2.685
	$\mathbb{F}_{\text{unrelated}}$	0.068	-0.446	-0.483	-0.215

Table 5: **Facilitation scores across different types of visual priming for EQ (top) and SST-2 (bottom) using four LVLMs.** Positive $\mathbb{F}$ values indicate a positive priming, while negative values indicate an negative priming.

5.1 Image Collection Method

Entity Questions. For *EQ*, the guiding images were obtained by searching for the entity mentioned in the question on Wikipedia and selecting an image from the corresponding article. The misleading images were preselected, with two images assigned for each question pattern. Unrelated images were chosen randomly from the set of guiding images for other questions, ensuring no direct relation to the current question.

SST-2. In *SST-2*, one emoji image that appears visually positive, negative, and neutral, respectively, was selected from a Kaggle dataset[1]. Prior to the main experiments, we confirmed that all LVLMs used in this study were able to correctly classify the emotion represented by each emoji. For each sample, the guiding image is an emoji that corresponds to the correct emotion label. In contrast, the misleading image represents the opposite emotion of the correct label. In addition, the unrelated image is an emoji that appears to be emotionally neutral.

5.2 Models

We utilize four LVLMs in our experiments to investigate potential differences in how each model processes visual cues: Qwen2.5-VL-7B-Instruct (Bai et al., 2025), Llama-3.2-11B-Vision-Instruct (Grattafiori et al., 2024), Phi-3.5-vision-instruct (Abdin et al., 2024), and mPLUG-Owl3-7B (Ye et al., 2025). Additionally, to compare the LVLMs with their corresponding base LLMs, we also use Qwen2.5-7B (Qwen et al., 2025) as the counterpart of Qwen2.5-VL and Qwen2-7B (Yang et al., 2024), as the counterpart of mPLUG-Owl3. This comparison provides insight into how the presence or absence of a vision encoder may affect model performance. Although it may not allow for definitive conclusions, it offers a useful indication of the potential impact of incorporating visual processing. Detailed information on the models is provided in Table 3.

6 Experimental Results

Accuracy. From Table 4, guiding images tended to improve accuracy compared to the text-only setting in all tasks and models. On the other hand, misleading images consistently decreased accuracy in all tasks and models compared to the text-only setting, without exception. Notably, in the SST-2 task, the extent of change caused by misleading images varied significantly between models. Specifically, Qwen2.5-VL and mPLUG-Owl3 showed relatively small changes in accuracy (0.03 and 0.06 points), while Llama-3.2 and Phi-3.5 exhibited much larger variations (0.15 and 0.75 points). Regarding unrelated and noise images, no substantial changes in

[1] https://www.kaggle.com/datasets/subinium/emojiimage-dataset

Task	Model	Text-Only	Guiding	Misleading	Unrelated	Noise
EQ	Qwen2.5-VL	0.082 ± 0.008	$0.026 \pm 0.007^{\dagger}$	$0.134 \pm 0.009^{\dagger}$	$0.118 \pm 0.008^{\dagger}$	$0.073 \pm 0.008^{\dagger}$
	mPLUG-Owl3	0.120 ± 0.008	$0.089 \pm 0.008^{\dagger}$	$0.168 \pm 0.008^{\dagger}$	$0.163 \pm 0.008^{\dagger}$	0.177 ± 0.008
	Llama3.2-VI	0.163 ± 0.007	$0.076 \pm 0.007^{\dagger}$	$0.175 \pm 0.008^{\dagger}$	0.162 ± 0.008	$0.125 \pm 0.007^{\dagger}$
	Phi3.5-VI	0.218 ± 0.008	$0.187 \pm 0.008^{\dagger}$	$0.317 \pm 0.008^{\dagger}$	$0.284 \pm 0.008^{\dagger}$	$0.264 \pm 0.008^{\dagger}$
SST2	Qwen2.5-VL	0.231 ± 0.014	$0.309 \pm 0.011^{\dagger}$	0.227 ± 0.018	$0.272 \pm 0.014^{\dagger}$	0.232 ± 0.014
	mPLUG-Owl3	0.016 ± 0.010	$0.097 \pm 0.010^{\dagger}$	$0.047 \pm 0.015^{\dagger}$	$0.076 \pm 0.011^{\dagger}$	$0.099 \pm 0.012^{\dagger}$
	Llama3.2-VI	0.084 ± 0.011	$0.117 \pm 0.009^{\dagger}$	$0.053 \pm 0.016^{\dagger}$	0.088 ± 0.013	$0.066 \pm 0.015^{\dagger}$
	Phi3.5-VI	0.024 ± 0.009	0.017 ± 0.002	$0.664 \pm 0.020^{\dagger}$	$0.039 \pm 0.011^{\dagger}$	$0.042 \pm 0.011^{\dagger}$

Table 6: **Comparison of ECE (↓) between LVLM with Text-Only and LVLM with Images under Four Conditions.** This table shows ECE (↓) for each model in EQ and SST-2. The table compares the Text-Only scenario with various visual priming scenarios. The ECE values are presented with 95% confidence intervals. A bootstrap significance test with 10,000 iterations was performed. † indicates a p-value less than 0.05. Value indicates cases where ECE improved compared to the Text-Only scenario, while Value indicates cases where ECE worsened compared to the Text-Only scenario.

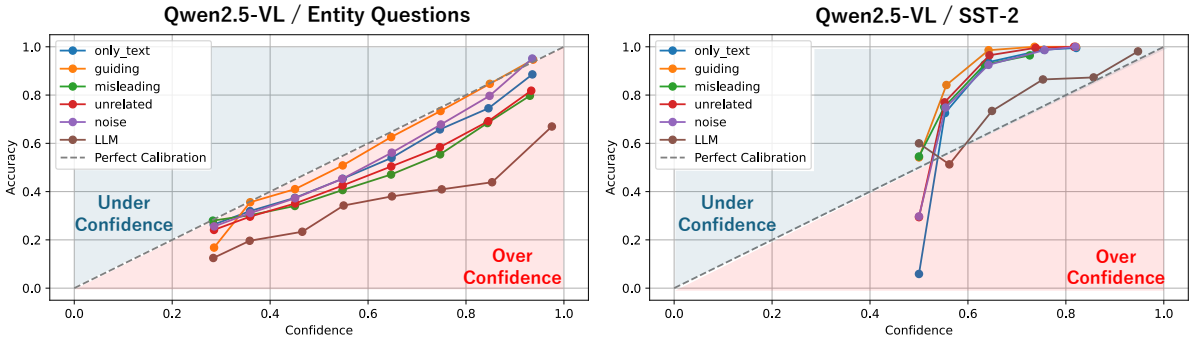


Figure 2: **Confidence Calibration Curves in EQ (Left) and SST-2 (Right)**. This figure visualizes the Confidence Calibration for different models with various visual priming conditions in EQ and SST-2. The x-axis represents the predicted confidence, while the y-axis shows the accuracy. The dashed line indicates the perfect calibration line, where the closer the curve is to this line, the more aligned the predicted confidence is with the actual accuracy. The blue background indicates underconfidence, while the red background indicates overconfidence. Results for Qwen2.5-VL are shown here as a representative example, as similar trends were observed across other models.

accuracy or confidence were observed compared to guiding and misleading images. Specifically, guiding images yielded an average change of 0.071 points and misleading images yielded an average change of 0.151 points; by comparison, unrelated images and noise images produced much smaller average changes of only 0.025 and 0.005 points.

Facilitation. From Table 5, in EQ, when guiding images were given, three out of the four models showed positive $\mathbb{F}$ values, indicating a tendency toward positive priming. Only Qwen2.5-VL showed a slightly negative value (-0.05). On the other hand, when misleading or unrelated images were given, all models exhibited negative $\mathbb{F}$ values, confirming a tendency toward negative priming. Notably, Llama-3.2 and Phi-3.5 showed larger absolute $\mathbb{F}$ values compared to Qwen2.5-VL and mPLUG-Owl3, indicating that even within the same task, the sensitivity to images varied between models. In SST-2, no consistent trend was observed among the models. In Qwen2.5-VL, the $\mathbb{F}$ value remained positive regardless of the image type, whereas mPLUG-Owl3 and Llama-3.2 consistently showed negative values regardless of the image type. Phi-3.5 showed a positive value for guiding images, while negative values were observed for misleading and unrelated images. The absolute $\mathbb{F}$ values in SST-2 tended to be larger compared to EQ.

ECE. From EQ in Table 6, the ECE when providing guiding images was significantly smaller compared to the text-only setting in all models, confirming that calibration was improved. Furthermore, as shown in Figure 2, the presentation of guiding images alleviated the state of overconfidence. On the other hand, when misleading or unrelated images were given, the ECE in all models was significantly

Task	Model	LLM	Guiding	Misleading	Unrelated	Noise
EQ	Qwen2.5-VL (Qwen2.5-7B)	0.305 ± 0.008	0.026 ± 0.007†	0.134 ± 0.008†	0.118 ± 0.008†	0.073 ± 0.008†
	mPLUG-Owl3 (Qwen2-7B)	0.210 ± 0.008	0.089 ± 0.008†	0.168 ± 0.008†	0.163 ± 0.008†	0.117 ± 0.008†
SST-2	Qwen2.5-VL (Qwen2.5-7B)	0.042 ± 0.010	0.309 ± 0.011†	0.227 ± 0.018†	0.273 ± 0.014†	0.231 ± 0.014†
	mPLUG-Owl3 (Qwen2-7B)	0.019 ± 0.009	0.097 ± 0.011†	0.047 ± 0.016†	0.076 ± 0.012†	0.099 ± 0.012†

Table 7: **Comparison of ECE (↓) between LLM (Baseline) and LVLM with Images under Four Conditions.** The ECE values are presented with 95% confidence intervals. A bootstrap significance test with 10,000 iterations was performed. † indicates a p-value less than 0.05. Value indicates cases where ECE decreased compared to the LLM scenario, while Value indicates cases where ECE increased compared to the LLM scenario.

larger compared to the text-only setting, indicating that calibration deteriorated. Moreover, as shown in Figure 2, the presentation of misleading images further worsened the state of overconfidence. Moreover, when comparing the misleading and unrelated scenarios, the ECE values in the misleading condition were consistently higher across all models. Regarding noise images, no consistent trend was observed across models. However, in Qwen2.5-VL and Llama-3.2, the ECE when providing noise images was significantly improved compared to the text-only setting. In SST-2, noise images showed a similar trend to EQ, but guiding images showed a decrease in ECE in almost all models. In addition, in Llama-3.2, the ECE decreased even when misleading images were presented. While EQ exhibited a tendency toward overconfidence, SST-2 generally showed a tendency toward underconfidence. Furthermore, Table 7 shows the comparison between the LVLM and its base LLM when various images are given. From EQ in Figure 2, the LLM exhibits the strongest tendency toward overconfidence, and in all image conditions, the ECE of the LVLM was significantly lower than that of the LLM. On the other hand, in SST-2, the ECE of the LVLM was significantly higher than that of the LLM in all image conditions, indicating a deterioration in calibration.

7 Discussion

7.1 Accuracy

Influence of Visual Information. As shown in Table 4, guiding images improved accuracy, while misleading images decreased it. This suggests that visual information can affect LVLMs' outputs, even in language tasks. In contrast, unrelated and noise images had little impact, indicating that models may partially ignore irrelevant or non-informative visual cues. Interestingly, this aligns with human cognition, where relevant cues aid decision-making, while misleading or unrelated cues can hinder it.

Differences between Models. As shown in Table 4, compared to guiding images, misleading images showed greater variation in accuracy changes across models. Notably, Llama-3.2 and Phi-3.5 exhibited a significant decrease in accuracy when presented with misleading images, indicating that these models actively incorporate visual information even when it is incorrect. In contrast, Qwen2.5-VL and mPLUG-Owl3 did not appear to make much use of visual information.

7.2 Facilitation Effects

EQ. As shown in Table 5, in EQ tasks, positive priming by guiding images was observed in all models except for Qwen2.5-VL. On the other hand, negative priming by misleading images was observed in all models. Furthermore, since the absolute $\mathbb{F}$-values of Llama-3.2 and Phi-3.5 were larger than those of the other two models, it suggests that these models actively utilize visual information.

SST-2. As shown in Table 5, SST-2 did not demonstrate a consistent priming effect comparable to that observed in EQ. Specifically, when guiding images were presented, positive priming was observed in Qwen2.5-VL and Phi-3.5, while the $\mathbb{F}$ values for mPLUG-Owl3 and Llama-3.2 were negative, indicating priming was not observed in these models. Moreover, in the case of misleading images, only Qwen2.5-VL showed a positive $\mathbb{F}$ value, suggesting negative priming was not observed.

Image	Qwen2.5-VL	mPLUG-Owl3	Llama-3.2	Phi-3.5
Guiding	-6.01	-4.83	-12.08	-2.26
Misleading	-6.01	-5.38	-12.59	-4.88
Unrelated	-6.08	-5.24	-11.86	-2.75
Noise	-6.15	-4.79	-11.38	-2.54

Table 8: **Log-probability with various images in SST-2 for LVLMs.** For many models, the highest values occur with a noise image lacking semantic content.

7.3 Calibration Effects (ECE Analysis)

EQ. As shown in Table 6, in EQ tasks, presenting guiding images to models significantly lowered the ECE compared to text-only conditions for all models. This result indicates that guiding images acted as beneficial visual information, enhancing not only the model's performance but also its reliability. However, when misleading images were presented, the ECE significantly increased compared to the text-only conditions for all models, suggesting that misleading images introduced confusion, thus degrading both performance and reliability.

SST-2. As shown in Table 6, in SST-2, even when guiding images were presented, almost all models showed a significant deterioration in ECE compared to the text-only condition. This can be attributed to the fact that most models already exhibited ECE values close to zero with text-only input, and in such situations, adding images introduced redundant information that confused the models, leading to a worsening of ECE even when guiding images were presented. Interestingly, in the case of Llama-3.2, the presence of misleading images resulted in a decrease in ECE. This can be explained by the simultaneous decline in both performance and confidence, which reduced the gap between accuracy and confidence.

Effect of Noise Images. As shown in Table 6, for noise images, some models showed an improvement in ECE. This phenomenon can be attributed to the reduction of overconfidence when presented with meaningless noise images, leading the model to choose more cautiously compared to text-only input, thereby alleviating excessive confidence.

Comparison with LLM. As shown in Table 7, compared to LLM, all image types in EQ tasks demonstrated improved ECE, while SST-2 showed a deterioration in ECE regardless of image type. This can be interpreted as follows: In EQ tasks, deriving correct answers using only text is relatively challenging, leading to a tendency for LLM to exhibit overconfidence. The addition of visual information likely encouraged more cautious responses. In contrast, since the ECE values of LLMs in SST-2 are close to zero, this task is relatively easy to solve using text alone. Adding visual information likely introduced redundant cues, which confused the models.

7.4 Do LVLMs show priming?

In EQ tasks, positive and negative priming were observed in all models except when guiding images were presented to Qwen2.5-VL. In contrast, SST-2 did not exhibit a consistent priming effect. This can be attributed to the fact that SST-2 already shows high accuracy with text-only conditions, indicating that it is a relatively easy task where the correct answer can be derived using only linguistic information. In such a situation, adding images may confuse the model, resulting in a lack of consistent positive $\mathbb{F}$ values when guiding images are presented and a deterioration of ECE. As shown in Table 8, when noise images, which do not contain meaningful content were presented, the log-probability was the lowest among all image types in SST-2. This result suggests that in SST-2, adding meaningful images as supplementary information may introduce redundancy and potentially confuse the model. Therefore, although the visual priming effect is observed in LVLMs, the priming effect does not appear depending on the task.

8 Conclusion

In this study, we investigate whether LVLMs exhibit visual priming effects on language-only tasks. We conducted experiments by presenting four types of images (guiding, misleading, unrelated, and noise) alongside two language-only tasks, and evaluated the model's performance using Accuracy, Facilitation, and ECE. Our results revealed guiding images improved accuracy and calibration, whereas misleading images decreased both metrics. Additionally, the influence of visual information varied across tasks and models, indicating that LVLMs can be both positively and negatively affected by visual cues. These findings suggest that LVLMs do not ignore visual information, but actively integrate it, demonstrating visual priming effects similar to those observed in human cognition.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, and 110 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *Preprint*, arXiv:2308.12966.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

John A. Bargh and Tanya L. Chartrand. 2000. The mind in the middle: A practical guide to priming and automaticity research. In Heinrich T. Reis and Chris M. Judd, editors, *Handbook of research methods in social and personality psychology*, pages 253–285. New York: Cambridge.

Feiqi Cao, Soyeon Caren Han, Siqu Long, Changwei Xu, and Josiah Poon. 2022. Understanding Attention for Vision-and-Language Tasks. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3438–3453, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Meiqi Chen, Yixin Cao, Yan Zhang, and Chaochao Lu. 2024. Quantifying and mitigating unimodal biases in multimodal large language models: A causal perspective. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16449–16469, Miami, Florida, USA. Association for Computational Linguistics.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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.

Kazuki Hayashi, Yusuke Sakai, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. 2024. Towards artwork explanation in large-scale vision language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 705–729, Bangkok, Thailand. Association for Computational Linguistics.

Jaap Jumelet, Willem Zuidema, and Arabella Sinclair. 2024. Do language models exhibit human-like structural priming effects? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14727–14742, Bangkok, Thailand. Association for Computational Linguistics.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.

Kento Kawaharazuka, Tatsuya Matsushima, Andrew Gambardella, Jiaxian Guo, Chris Paxton, and Andy Zeng. 2024. Real-World Robot Applications of Foundation Models: A Review. *Advanced Robotics*, 38(18):1232–1254.

Kyungjun Lee, Abhinav Shrivastava, and Hernisa Kacorri. 2023. Leveraging hand-object interactions in assistive egocentric vision. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(6):6820–6831.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

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.

James Michaelov, Catherine Arnett, Tyler Chang, and Ben Bergen. 2023. Structural priming demonstrates abstract grammatical representations in multilingual language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3703–3720, Singapore. Association for Computational Linguistics.

Bruce Milliken and Adrienne Rock. 1997. Negative priming, attention, and discriminating the present from the past. *Consciousness and Cognition*, 6(2):308–327.

Kanishka Misra, Allyson Ettinger, and Julia Rayz. 2020. Exploring BERT's sensitivity to lexical cues using tests from semantic priming. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4625–4635, Online. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Shintaro Ozaki, Kazuki Hayashi, Miyu Oba, Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2025a. BQA: Body language question answering dataset for video large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 110–123, Vienna, Austria. Association for Computational Linguistics.

Shintaro Ozaki, Kazuki Hayashi, Yusuke Sakai, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. 2025b. Towards cross-lingual explanation of artwork in large-scale vision language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3773–3809, Albuquerque, New Mexico. Association for Computational Linguistics.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1).

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

David Reitter and Johanna D. Moore. 2007. Predicting success in dialogue. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 808–815, Prague, Czech Republic. Association for Computational Linguistics.

Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2025. Revisiting compositional generalization capability of large language models considering instruction following ability. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31219–31238, Vienna, Austria. Association for Computational Linguistics.

Haruki Sakajo, Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2025. Tonguescape: Exploring language models understanding of vowel articulation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12605–12619, Albuquerque, New Mexico. Association for Computational Linguistics.

Stephanie Schoch, Diyi Yang, and Yangfeng Ji. 2020. "this is a problem, don't you agree?" framing and bias in human evaluation for natural language generation. In *Proceedings of the 1st Workshop on Evaluating NLG Evaluation*, pages 10–16, Online (Dublin, Ireland). Association for Computational Linguistics.

Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. 2021. Simple entity-centric questions challenge dense retrievers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6138–6148, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mandar Sharma, Rutuja Taware, Pravesh Koirala, Nikhil Muralidhar, and Naren Ramakrishnan. 2024. Laying anchors: Semantically priming numerals in language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2653–2660, Mexico City, Mexico. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Steven P. Tipper. 1985. The negative priming effect: Inhibitory priming by ignored objects. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, 37A(4):571–590.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Endel Tulving and Daniel L. Schacter. 1990. Priming and human memory systems. *Science*, 247(4940):301–306.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2025. mPLUG-owl3: Towards long image-sequence understanding in multi-modal large language models. In *The Thirteenth International Conference on Learning Representations*.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. 2024. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13040–13051.

Marco Zorzi, Ivilin Peev Stoianov, and Carlo Umilta. 2004. Computational modeling of numerical cognition. *The Handbook of Mathematical Cognition*.