REVISITING CAPTION GAINS WHEN THE IMAGE REMAINS AVAILABLE

Anonymous authors
Paper under double-blind review

ABSTRACT

In a multimodal conversation, an assistant generates a generic caption to summarize an image and retain it for later visual questions, even when the original image remains available at answer time. Whether such captions help beyond simply placing text in the conversation remains unclear. By generating one question-blind generic caption per model–image pair for seven vision-language models and testing 20 conditions on natural-image and medical-image question-answering datasets, we measure how much of the gain from caption history over a minimal exchange with a fixed, image-independent acknowledgement comes from the image description. To isolate the description's contribution, we compare each frozen caption with length-matched neutral text in the same assistant-history format. Neutral text alone accounts for much of this gain on the natural-image dataset, where replacing it with the caption adds only a small further gain and placing the same text alongside the question yields a caption–neutral difference of similar size. On the medical-image dataset, the caption adds no clear gain over neutral text in history, although on that dataset the average gain of neutral history over the minimal exchange and the average inline caption–neutral difference are each driven by one or two models. We further examine image placement and longer conversations to probe how these comparisons depend on the input sequence. With the original image available, a caption's description adds at most a small average gain beyond neutral text on either dataset. An image-removal replay shows that the captions carry answer-relevant content: when the image is removed, captions preserve most of the accuracy that length-matched neutral text loses on both datasets. A caption's value for later questions must therefore be understood in relation to continued access to the image.

1 INTRODUCTION

Multimodal agents based on a vision-language model (VLM) need information from an image after the turn in which they first saw it. Visual-memory and retrieval systems preserve multimodal evidence or revisit visual sources in response to a later query (Datta et al., 2022; Chen et al., 2022; Majumdar et al., 2024; Wang et al., 2024; Zhang et al., 2024; Liu et al., 2025; Guo et al., 2026). Caption utility has likewise been assessed through downstream question answering (Li et al., 2018; Fisch et al., 2020; Lee et al., 2021; Hu et al., 2023; Yang et al., 2025). These two lines of work meet when an assistant writes a generic caption before the future question is known and later answers with both the caption and original image available. The added value of that stored caption under continued image access remains unclear.

A generic caption carries image information into the later exchange as a fixed selection of visual details made before the question is known. The original image may still contain details the caption leaves unstated (Li & Zhang, 2026), while the caption makes its selected details available as text (Kim & Bansal, 2019). The usefulness of that selection depends on which visual details the later question requires.

However, results from adding captions alongside images do not by themselves show what an earlier caption contributes to a continuing conversation. Adding a caption changes more than the description: it can change the earlier request, the assistant reply, the amount of text, and where the text appears relative to the question. An accuracy difference from direct answering, or from a minimal

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

acknowledgement exchange in which the assistant replies only with a fixed phrase, cannot isolate the contribution of the image description.

We conduct controlled experiments to measure how much a generic caption's image description contributes to later answer accuracy. We compare a frozen caption with length-matched, image-independent neutral text in the same message format, both as an earlier assistant reply (history delivery) and alongside the question (inline delivery). Seven vision-language models each write one question-blind caption per image, which is frozen and reused across 20 conditions on two datasets: Visual7W, with natural images, and SLAKE, with medical images. The conditions also include direct answering, a minimal exchange that ends with the fixed reply "Ready.", and alternative image placements; on both datasets, we extend the conversation for all seven models. To check whether the captions carry answer-relevant content, a paired image-removal replay reruns the caption and neutral-text inputs without the image, all else fixed.

Our contributions are as follows. (i) A controlled design that separates a caption's description from the other changes it brings, by comparing frozen captions with length-matched neutral text in history and inline delivery. (ii) Evidence that, with the image available, the description adds at most a small average gain beyond neutral text on either dataset, and on Visual7W much of the gain from caption history comes from adding text to the conversation; the image-removal replay shows that the captions nonetheless carry answer-relevant content.

2 Related Work

Caption Utility for Question Answering. A caption's usefulness depends on whether it preserves the information needed for later questions. CapWAP trains captioning models to anticipate an intended user population's information needs and evaluates the resulting captions through questions answered from the captions alone (Fisch et al., 2020). CaptionQA similarly measures how much of an image's downstream utility a caption preserves when a language model answers questions without the image (Yang et al., 2025). PromptCap conditions caption generation on the question itself, selecting visual details that help a text-only language model answer that question (Hu et al., 2023). These settings motivate two distinctions: whether the question is known when the caption is generated, and whether the answering model can still access the image. We study a generic caption generated before the question is disclosed and examine its additional utility when the original image remains available.

Captions as Additional Visual Evidence. Captions have also been used alongside images to support question answering. Kim and Bansal (2019) combine an image with a generated paragraph caption in a visual and textual QA model, showing that the textual description can complement visual features. More recently, GEASS analyzes the effects of adding self-generated captions and adjusts how strongly a VLM relies on them for each query (Li & Zhang, 2026). Its analysis includes length-matched neutral text, unrelated captions, and shuffled captions as controls for distinguishing caption content from text length. Related work studies responses to conflicting text and false user claims (Deng et al., 2025; Wu et al., 2025; Zhao et al., 2026; Yuan et al., 2025). We examine naturally generated generic captions, which may contain both errors and omissions, and extend the caption–neutral comparison to open-ended questions across two conversation formats. We compare the same frozen caption with length-matched neutral text separately in history and inline delivery, with the original image available in both. This design measures the caption–neutral contrast within each format and whether that contrast changes with delivery, rather than attributing an absolute difference between formats to caption content.

Conversational Context and Image Access. Visual-memory and retrieval systems retain or recover visual evidence for later questions (Datta et al., 2022; Majumdar et al., 2024). MuRAG uses an external memory containing image–text pairs and text snippets, retrieving entries by similarity between question and memory embeddings. Its reader combines the retrieved visual and textual inputs with the question to generate an answer, so retrieved images remain available during answering (Chen et al., 2022). This illustrates how the stored representation and subsequent access to visual evidence jointly define the answering setting. Studies of multiple-image contexts further motivate specifying the number and placement of image inputs (Sharma et al., 2024; Das et al., 2026). We examine a conversational setting in which the same frozen caption is retained as an earlier assistant reply or supplied alongside the question, while the original image remains available. These deliveries

differ in message roles and conversation structure as well as text position, so their comparison concerns the format in which stored text is supplied.

3 METHOD

3.1 PROBLEM SETTING

Let $\mathcal{B} = \{(I_i, Q_i, Y_i)\}_{i=1}^{n}$ be a fixed set of n image–question–reference-answer triples, where n is the number of images used from the dataset. For subject model m, a generic caption $C_{mi} = G_m(I_i)$, where G_m denotes model m answering the generic caption request, is generated before Q_i is disclosed and then frozen. Each neutral text $N_{mi}^{(k)}$ is selected from a predeclared pool $\mathcal{P}_k$ as the entry closest to the caption in token count:

$$N_{mi}^{(k)} = \arg\min_{u \in \mathcal{P}_k} \left| \ell_m(u) - \ell_m(C_{mi}) \right|, \quad k \in \{\text{procedure}, \text{notation}, \text{coordination}\} \tag{1}$$

where the pool $\mathcal{P}_k$ uses no image, question, answer, or caption content, ties are resolved by a fixed identifier, and $\ell_m(\cdot)$ counts tokens under the subject tokenizer because the same text can yield different counts under different models' tokenizers. Counting in model m's own input units keeps the caption and its neutral text close in the length that m processes, so the caption–neutral contrast (Section 3.2) changes text content while holding length approximately fixed. Procedure sentences state that the exchange can continue, notation sentences describe the text itself as introductory, and coordination sentences state that the next part of the exchange can begin. Table 1 shows an example caption and the three neutral texts selected for it.

Table 1: Example of neutral-text selection for a Qwen2.5-VL-3B caption. Token counts use the Qwen2.5-VL-3B tokenizer; each family contributes its closest pool entry in token count.

text	content	tokens
caption C	A herd of rhinos and zebras graze in a grassy field.	16
procedure	This exchange can continue, without expressing a preference about the substance of subsequent material.	16
notation	This passage provides a general introduction, rather than an explanation of any particular subject.	16
coordination	The next part of this exchange can begin, with no additional requirement introduced here.	16

For a text T, either the caption C or a neutral text $N^{(k)}$, history delivery H places T in an earlier assistant reply, and inline delivery L places it in the user message that contains the question. Each neutral family is run as a separate condition, and N without a superscript denotes the three families together, with results averaged over them. Crossing text and delivery gives four combinations, (C, H), (N, H), (C, L), and (N, L). In all four, the original image is attached to the user message that contains the question. Section 3.2 adds conditions that attach it to the earlier request instead or to both messages.

3.2 CONDITIONS AND CONTRASTS

From the four text–delivery combinations, we construct 20 conditions (Figure 1): eight conditions realize the four combinations, and three additions supply the remaining 12. Condition IDs join the text with its image placement or delivery. (C, H) is `caption_current` (caption history), (N, H) is the mean of the three neutral `*_current` conditions (neutral history), (C, L) is `caption_inline`, and (N, L) is the mean of the three neutral `*_inline` conditions. First, history delivery also attaches the image to the earlier request instead of the question (`previous`) or to both messages (`both`); these two placements for each of the four texts (the caption and the three neutral families) add eight conditions, for 16 in total. Second, two conditions contain neither the caption nor a neutral text: `direct` supplies the image with the question and no earlier exchange, and `ready` first adds a text-only request answered with the fixed reply "Ready." Third, two diagnostics complete the set: in `okay`, an earlier caption request with the image attached is answered only with "Okay.", and `other_current` replaces the caption in `caption_current` with the same model's caption of

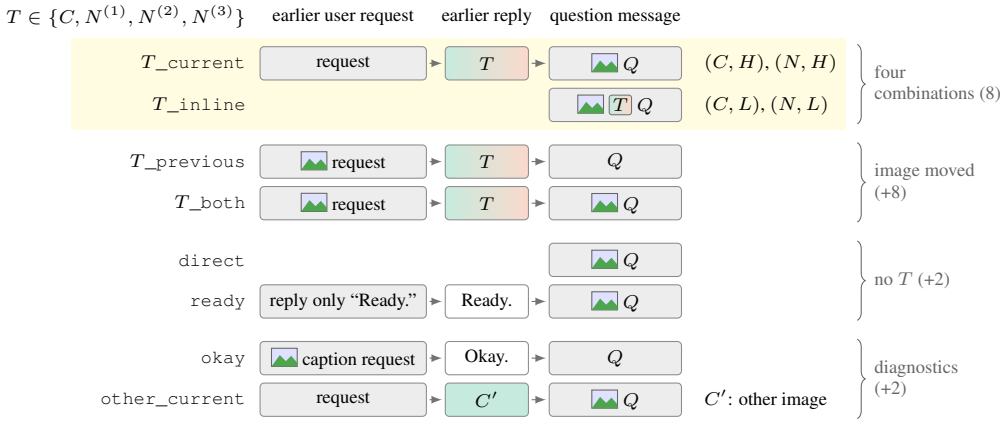

Figure 1: Construction of the 20 conditions. Each row shows the messages that follow the shared system message: an earlier user request, an earlier assistant reply, and the question message. The image icon marks where the image is attached, and Q is the question with the answer instruction. A row with T stands for four conditions, one per text. The shaded band holds the eight conditions of the four combinations, where (N, H) and (N, L) average the three neutral texts. In `other_current`, C' is the same model's caption of another image.

another image. Within each delivery, the caption–neutral contrast measures the effect of replacing length-matched neutral text with the caption; it does not equal the complete caption procedure, $A(\texttt{caption_both}) - A(\texttt{direct})$, which also changes the input sequence.

For stored binary correctness y_{ima} of image i, model m, and condition a, define the per-model accuracy and its average over models

$$A_m(a) = \frac{1}{n}\sum_{i=1}^{n} y_{ima}, \qquad A(a) = \frac{1}{M}\sum_{m=1}^{M} A_m(a) \tag{2}$$

where $M = 7$ is the number of models. For the four combinations, $A_m(C, d)$ and $A_m(N^{(k)}, d)$ denote the corresponding per-model accuracies from Equation 2, where $d \in \{H, L\}$; $A(C, d)$ and $A(N^{(k)}, d)$ denote their averages over models. Neutral accuracy is averaged at the accuracy level:

$$A_m(N, d) = \frac{1}{K}\sum_{k} A_m(N^{(k)}, d), \qquad A(N, d) = \frac{1}{K}\sum_{k} A(N^{(k)}, d) \tag{3}$$

where $K = 3$ is the number of neutral families. The main contrasts are Q1, the accuracy change from the minimal acknowledgement exchange ($\texttt{ready} - \texttt{direct}$); Q2, the history contrast Δ_H; Q3, image position ($\texttt{caption_current} - \texttt{caption_previous}$) and image repetition ($\texttt{caption_both} - \texttt{caption_current}$); and Q4, the difference between the caption–neutral contrasts of the two deliveries. Define the history contrast, inline contrast, and their delivery contrast as

$$\Delta_H = A(C, H) - A(N, H), \qquad \Delta_L = A(C, L) - A(N, L), \qquad \Delta_{\text{del}} = \Delta_H - \Delta_L \tag{4}$$

Three descriptive quantities are reported separately from the main contrasts: the bridge $A(\texttt{neutral_current}) - A(\texttt{ready})$, which compares the minimal exchange with length-matched neutral history; the inline contrast Δ_L; and caption history $- \texttt{ready}$. Here `neutral_current` is the mean of the three neutral `*_current` conditions. The absolute delivery difference $A(C, H) - A(C, L)$, also descriptive, compares caption accuracy between deliveries (Figure 2); unlike Δ_{del}, it does not subtract the neutral text, so it also reflects how delivery affects any text.

We report the average over the seven models, each weighted equally, and pointwise 95% intervals from 10,000 image-level bootstrap resamples with a fixed seed, resampling each image jointly across conditions and models. These intervals reflect image sampling only, with the model roster, frozen captions, stored responses, correctness judgments, and checkpoints fixed.

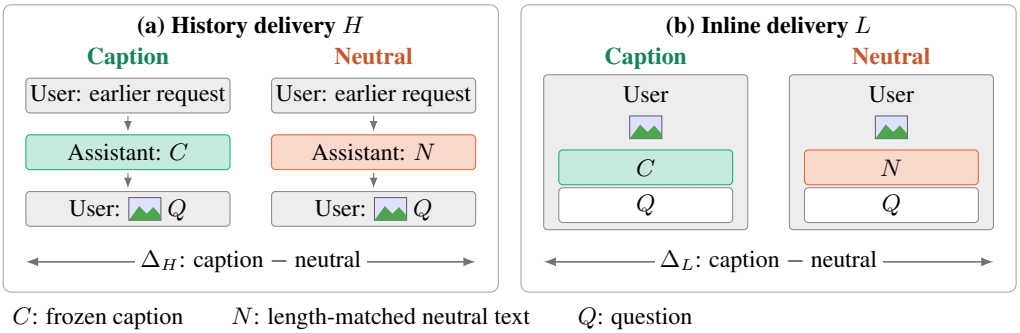

Figure 2: Text–delivery comparison. (a) is history delivery H, and (b) is inline delivery L. A frozen caption (C) or length-matched neutral text (N) appears in the indicated delivery. Within a delivery, the caption–neutral contrast (Δ_H or Δ_L) measures caption replacement, and its difference between deliveries is Δ_{del}. Each condition supplies one image with the question. Neutral results average three text families; message labels are schematic.

4 Experiments

Section 4.1 describes the datasets, models, and judge; Section 4.2 reports the gain from caption history and the part of it that the caption's description contributes, the main contrasts, and an image-removal check; Section 4.3 tests whether the caption–neutral contrast changes as the conversation grows longer.

4.1 Datasets and Models

Datasets. We use two datasets: Visual7W (Zhu et al., 2016), which pairs everyday natural images with multiple-choice telling questions (what, where, when, who, why, and how), whose correct answers we use as open-ended reference answers, and SLAKE (Liu et al., 2021), which pairs medical images (CT, MRI, and X-ray) with questions such as organ position, size, color, and count. From Visual7W, we use 2,900 telling images, sampled with a fixed seed and shared by all seven models and all 20 conditions; one question Q_i and its reference answer Y_i are fixed for each image I_i before any caption or response is generated. From SLAKE, we use 333 images, each paired with one question. The questions are highly repetitive, with only 94 distinct question strings across the 333 images.

Models. We compare seven models: GLM-4.1V (GLM-V Team et al., 2025), InternVL3.5 (Wang et al., 2025), Gemma-4-12B-it (Gemma Team et al., 2026), Qwen3.8-27B (Qwen Team, 2026), Qwen2.5-VL-3B (Bai et al., 2025), InternVL2-8B (OpenGVLab, 2024), and Ministral-3-14B (Liu et al., 2026). Each subject model writes one generic caption from the image without seeing the future question. The request asks for one short sentence describing the image in 200 words or fewer. The resulting C_{mi} is frozen and reused in every condition for its model–image pair; in history delivery, it is inserted as the earlier assistant reply and is not regenerated after the question is shown. Because every contrast is computed within a model, the comparisons do not require equal caption length across models.

We use `gpt-4o-mini` as the judge to score each answer against the reference answer Y_i, because a VLM can express the meaning of the reference answer in different words, which exact string matching would reject. The judge prompt contains the question, the reference answer, and the extracted candidate answer, and asks whether the candidate means the same thing as the reference for that question; the judge must reply YES or NO.

4.2 Effects of Generic Captions

On both datasets, the caption adds little beyond length-matched neutral text in history, and the minimal acknowledgement exchange does not clearly change accuracy (Figure 3; Table 2). The `ready` exchange versus direct answering (Q1) changes accuracy by -0.18 pp $[-0.52, +0.16]$ on Visual7W and -0.51 pp $[-1.54, +0.47]$ on SLAKE. Replacing neutral history with the caption (Δ_H,

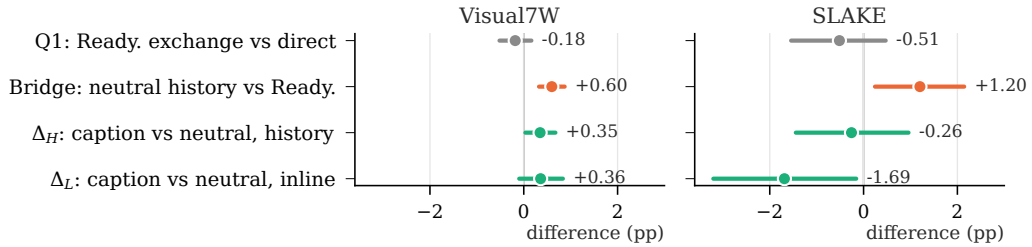

Figure 3: Where the gain from caption history comes from. Points are averages over the seven models with pointwise 95% image-bootstrap intervals. Q1 compares the fixed "Ready." exchange with direct answering; the bridge replaces that exchange with length-matched neutral history; Δ_H and Δ_L replace the neutral text with the caption in history and inline delivery.

Q2) changes accuracy by $+0.35$ pp $[+0.03, +0.67]$ on Visual7W, a small difference whose pointwise interval narrowly excludes zero, and by -0.26 pp $[-1.44, +0.96]$ on SLAKE. The delivery contrast Δ_{del} (Q4) is -0.01 pp $[-0.48, +0.44]$ on Visual7W and $+1.43$ pp $[-0.01, +2.86]$ on SLAKE; both intervals include zero.

The datasets differ in the source of the gain from caption history. On Visual7W, caption history improves on the `ready` exchange by $+0.95$ pp $[+0.56, +1.33]$, and much of this gain is obtained with length-matched neutral text: the bridge from `ready` to neutral history is $+0.60$ pp $[+0.33, +0.87]$. On SLAKE, caption history $-$ `ready` is $+0.94$ pp $[-0.39, +2.32]$, and the bridge is $+1.20$ pp $[+0.24, +2.15]$. The bridge minus Δ_H is $+0.25$ pp $[-0.21, +0.70]$ on Visual7W and $+1.46$ pp $[-0.26, +3.13]$ on SLAKE, so the difference between these contributions is not precisely estimated. The bridge is descriptive rather than a main contrast. On SLAKE, it is carried by Qwen3.8-27B, the only model whose bridge interval lies above zero. Much of Qwen3.8-27B's bridge comes from the repeated body-part question, whose answer is often "Lung" after the "Ready." exchange but "Chest" once neutral history is present. The positive SLAKE average bridge therefore reflects a single model.

The datasets also differ when the text is placed alongside the question. On Visual7W, the inline caption–neutral contrast Δ_L is $+0.36$ pp $[-0.10, +0.83]$, similar in size to the history contrast. On SLAKE, it is -1.69 pp $[-3.20, -0.16]$, with negative model-level values for InternVL3.5 (-7.51 pp) and InternVL2-8B (-6.21 pp). InternVL2-8B's drop comes almost entirely from the repeated questions asking which part of the body the image shows: with the caption alongside the question, it answers them with a vague region ("Upper part") or with an organ or structure named in the caption instead of the body part. The SLAKE history contrast combines a negative value for InternVL3.5 (-4.60 pp) and a positive value for InternVL2-8B ($+4.10$ pp), whose intervals both exclude zero. InternVL2-8B also has a delivery contrast of $+10.31$ pp $[+6.11, +14.61]$, so the SLAKE caption–neutral contrasts vary by model.

On Visual7W (Table 3), mean accuracy is 57.26% for caption history and 56.91% for neutral history, versus 56.25% for inline captions and 55.89% for inline neutral text. History minus inline accuracy is $+1.01$ pp $[+0.57, +1.45]$ for captions and $+1.02$ pp $[+0.71, +1.34]$ for neutral text. Both intervals

Table 2: Main contrasts and descriptive quantities, averaged equally over seven models (pp, pointwise 95% intervals).

question	contrast	**Visual7W**	**SLAKE**
Q1: acknowledgement exchange	`ready` − `direct`	-0.18 $[-0.52, +0.16]$	-0.51 $[-1.54, +0.47]$
Q2: history contrast	Δ_H from Equation 4	$+0.35$ $[+0.03, +0.67]$	-0.26 $[-1.44, +0.96]$
Q3: image position	`caption_current` − `caption_previous`	$+0.35$ $[-0.06, +0.76]$	$+1.59$ $[+0.47, +2.70]$
Q3: repeated image	`caption_both` − `caption_current`	-0.10 $[-0.43, +0.25]$	-0.26 $[-1.20, +0.69]$
Q4: delivery	Δ_{del} from Equation 4	-0.01 $[-0.48, +0.44]$	$+1.43$ $[-0.01, +2.86]$
bridge	`neutral_current` − `ready`	$+0.60$ $[+0.33, +0.87]$	$+1.20$ $[+0.24, +2.15]$
inline contrast (Δ_L)	Δ_L from Equation 4	$+0.36$ $[-0.10, +0.83]$	-1.69 $[-3.20, -0.16]$
caption history − `ready`	`caption_current` − `ready`	$+0.95$ $[+0.56, +1.33]$	$+0.94$ $[-0.39, +2.32]$

lie above zero, while Δ_{del} is near zero. On SLAKE, the corresponding four accuracies are 71.21%, 71.47%, 67.65%, and 69.34%, respectively. GLM-4.1V is the Visual7W delivery exception: its caption–neutral contrast is -0.25 pp in history and $+2.07$ pp inline, giving $\Delta_{\text{del}} = -2.32$ pp [-3.47, -1.18].

Image placement and model-level variation. Image repetition (Q3) changes caption accuracy by -0.10 pp [-0.43, $+0.25$] on Visual7W and -0.26 pp [-1.20, $+0.69$] on SLAKE; neither interval excludes zero. Attaching the image to the question rather than the earlier request (Q3) changes accuracy by $+0.35$ pp [-0.06, $+0.76$] on Visual7W and $+1.59$ pp [$+0.47$, $+2.70$] on SLAKE. On Visual7W, the other six model-level delivery intervals include zero (Table 3); these pointwise intervals do not test between-model heterogeneity.

Image-removal check. A paired image-removal replay reruns each caption and its three neutral texts in both deliveries with and without the image, all else fixed, for every Visual7W and SLAKE image, and compares the accuracy each caption loses when the image is removed with the loss of its matched neutral texts. Removing the image costs captions far less than their matched neutral texts on both datasets (Figure 4). In Visual7W history delivery, for example, caption accuracy falls by 11.33 pp, whereas matched neutral text loses 40.48 pp. On both datasets every model shows a smaller caption loss in both deliveries, and the interval of the average caption–neutral loss difference lies below zero for each dataset and delivery. The captions therefore carry answer-relevant content that neutral text lacks; the replay measures how accuracy depends on image availability, not how a model uses its inputs.

4.3 Caption Effects Across Conversation Length

We also test whether separating the earlier image–text exchange from the question changes the caption–neutral contrast. For all seven models on both datasets, we insert one to five fixed, image-free user–assistant filler exchanges before the final question, giving 2–6 completed exchanges in total. Every text condition receives the same filler sequence at a given number of completed exchanges, and the original image remains only in the first user message. Thus the extension changes message count and token distance together; it does not identify either mechanism separately. Because the image is attached to the earlier request, as in `previous`, the extension also differs from Δ_H in image placement.

On both datasets, the caption–neutral contrast depends more on the model than on conversation length (Figure 5). InternVL3.5 has intervals below zero at every length on Visual7W and at all but one on SLAKE, and Gemma-4-12B-it at some lengths on both, though not the same ones; most other model–length intervals include zero. On Visual7W only InternVL2-8B's point estimates rise almost

Table 3: The four text–delivery combinations and their contrasts by model on both datasets. The four A columns are accuracies (%); Δ_H and Δ_L are the caption–neutral contrasts, and $\Delta_{\text{del}} = \Delta_H - \Delta_L$.

model	$A(C,H)$	$A(N,H)$	$A(C,L)$	$A(N,L)$	Δ_H	Δ_L	Δ_{del} [95% interval]
Visual7W							
GLM-4.1V	58.21	58.46	59.21	57.14	-0.25	$+2.07$	-2.32 [-3.47, -1.18]
InternVL3.5	57.79	57.06	57.10	56.41	$+0.74$	$+0.69$	$+0.05$ [-1.26, $+1.36$]
Gemma-4-12B-it	56.28	55.70	54.59	54.52	$+0.57$	$+0.07$	$+0.51$ [-0.49, $+1.54$]
Qwen3.8-27B	63.86	63.20	61.97	61.23	$+0.67$	$+0.74$	-0.07 [-1.23, $+1.07$]
Qwen2.5-VL-3B	57.07	56.75	55.45	55.68	$+0.32$	-0.23	$+0.55$ [-0.44, $+1.56$]
InternVL2-8B	54.69	54.61	54.90	55.00	$+0.08$	-0.10	$+0.18$ [-0.92, $+1.31$]
Ministral-3-14B	52.93	52.62	50.55	51.26	$+0.31$	-0.71	$+1.02$ [-0.09, $+2.15$]
average	57.26	56.91	56.25	55.89	$+0.35$	$+0.36$	-0.01 [-0.48, $+0.44$]
SLAKE							
GLM-4.1V	71.17	70.37	71.77	70.77	$+0.80$	$+1.00$	-0.20 [-3.60, $+3.30$]
InternVL3.5	77.18	81.78	73.27	80.78	-4.60	-7.51	$+2.90$ [-1.60, $+7.51$]
Gemma-4-12B-it	73.57	74.87	74.17	74.37	-1.30	-0.20	-1.10 [-3.70, $+1.50$]
Qwen3.8-27B	76.28	75.28	70.87	69.27	$+1.00$	$+1.60$	-0.60 [-4.40, $+3.10$]
Qwen2.5-VL-3B	65.17	65.17	61.26	59.56	0.00	$+1.70$	-1.70 [-5.71, $+2.30$]
InternVL2-8B	70.27	66.17	59.46	65.67	$+4.10$	-6.21	$+10.31$ [$+6.11$, $+14.61$]
Ministral-3-14B	64.86	66.67	62.76	64.96	-1.80	-2.20	$+0.40$ [-3.00, $+3.70$]
average	71.21	71.47	67.65	69.34	-0.26	-1.69	$+1.43$ [-0.01, $+2.86$]

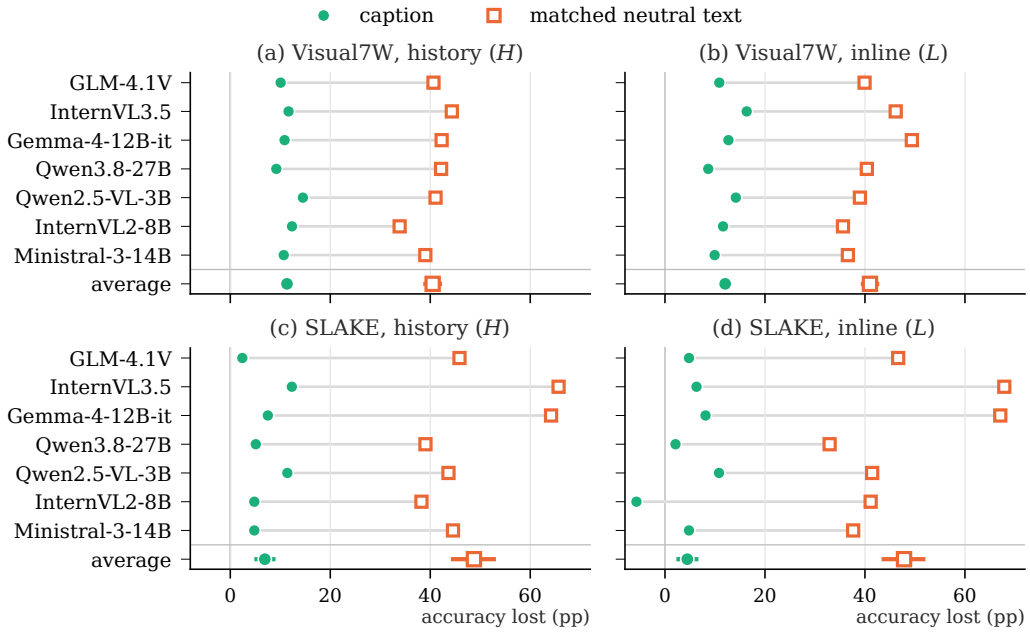

Figure 4: Accuracy lost when the image is removed, for the caption and the mean of its three matched neutral texts, by model on Visual7W (a, b) and SLAKE (c, d) in history (a, c) and inline (b, d) delivery. The last row of each panel is the average with a 95% image-bootstrap interval.

steadily with the number of exchanges, and on SLAKE only GLM-4.1V's rise steadily; every interval of both includes zero. The datasets differ in spread: SLAKE intervals are wider because it has fewer images, yet more of them exclude zero, in both directions, with InternVL2-8B above zero at three of the five lengths, Qwen2.5-VL-3B above zero only at three exchanges, and Ministral-3-14B below zero only at six exchanges.

5 Discussion

Preserving information in a caption and improving answers when the image remains available are distinct outcomes. The image-removal check confirms that captions carry answer-relevant content, but does not explain their small additional gain under continued image access. The complete caption procedure and a caption replacement also answer different questions: on Visual7W, much of the gain from caption history over the minimal exchange is already obtained with length-matched neutral text, although the difference between the two contributions is not precisely estimated. On SLAKE, the positive bridge is concentrated in one model, while two models contribute the negative average inline contrast. The caption–neutral contrast does not clearly differ by delivery on Visual7W; together with the model-level exceptions across the datasets, the results provide insufficient grounds for recommending one delivery format generally.

These findings have several limitations. The evidence comes from seven vision-language models on two datasets, and SLAKE has a small sample with heavily repeated questions, allowing a single question type to account for much of a model-level effect. The pointwise image-bootstrap intervals reflect image sampling with the model roster, checkpoints, questions, captions, responses, and correctness judgments fixed; they do not capture variation across a population of models and are not adjusted for multiple comparisons. This limits the interpretation of individual model and conversation-length intervals. We use one question per image, one frozen caption per model and image from a single caption request, and greedy decoding. Variation from regenerated captions or answers and from the automatic judge, `gpt-4o-mini`, is not measured, nor are cost and latency. The caption–neutral contrasts are relative to the neutral texts tested. The conversation-length extension changes message count and token distance together, and the behavioral comparisons do not identify internal evidence

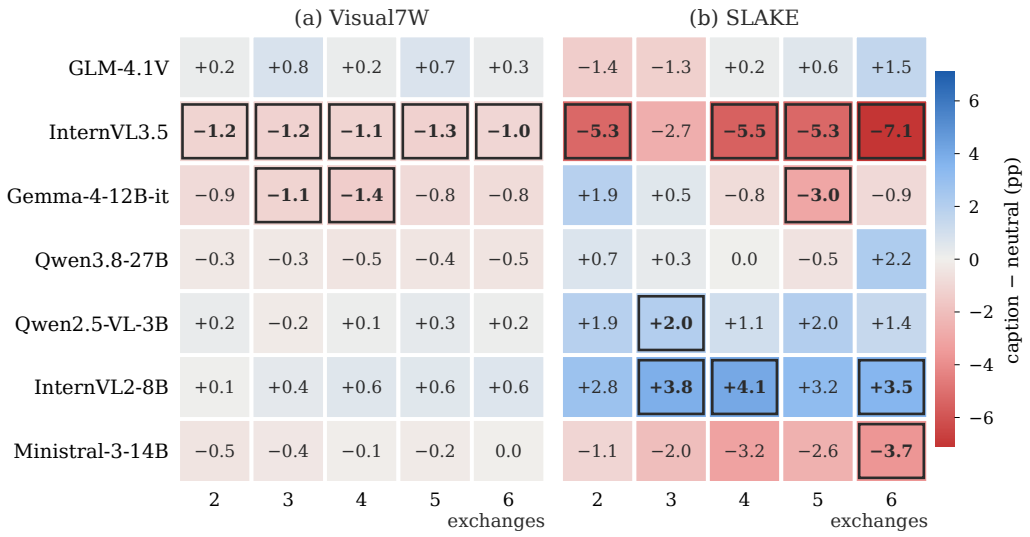

Figure 5: Caption–neutral contrast (pp) by model and the number of completed exchanges before the final question on (a) Visual7W and (b) SLAKE. Blue cells favor the caption and red cells the neutral text on one shared scale; a boxed bold value marks a 95% image-bootstrap interval that excludes zero. The image is attached only to the first user message.

use or spontaneous re-verification. Future work could examine more diverse questions and models and assess sensitivity to caption requests, generation settings, and judges. Controls that separate message count from token distance, together with analyses of model internals, could further examine how image and caption use depend on the conversation.

Averaged over the tested models, captions provide at most a small accuracy gain over length-matched neutral text while the original image remains available. Yet they preserve most of the accuracy neutral text loses once the image is no longer supplied, confirming that they carry answer-relevant content. For an assistant that retains a generic caption, its utility therefore depends on image availability, delivery, the comparison text, and model-level variation. A stored caption's ability to support answers without the image does not by itself establish an additional accuracy benefit when the image remains available.

AI USE STATEMENT

In this work, we used generative AI tools to refine hypotheses, help develop the conceptual framework of the study, design and provide feedback on the research methodology and experiments, implement the methods, generate the length-matched neutral texts used as a control, clean and reformat the datasets, and interpret results. We also used them to draft sections of the paper and to revise and polish its writing. We have not used generative AI tools to assist with translation or to support qualitative and thematic data analysis, and formulating mathematical claims, providing critical ingredients for proving them, and writing proofs are not applicable to this work. We have reviewed all AI-assisted work, and the numbers in the paper were checked against the stored experimental outputs by automated scripts. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Shuai Bai et al. Qwen2.5-VL Technical Report, February 2025.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William Cohen. MuRAG: Multimodal Retrieval-Augmented Generator for Open Question Answering over Images and Text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5558–5570, December 2022. doi: 10.18653/v1/2022.emnlp-main.375.

Anurag Das, Adrian Bulat, Alberto Baldrati, Ioannis Maniadis Metaxas, Bernt Schiele, Georgios Tzimiropoulos, and Brais Martinez. More Images, More Problems? A Controlled Analysis of VLM Failure Modes. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 7423–7442, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.366.

Samyak Datta, Sameer Dharur, Vincent Cartillier, Ruta Desai, Mukul Khanna, Dhruv Batra, and Devi Parikh. Episodic Memory Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19119–19128, June 2022. doi: 10.1109/CVPR52688.2022.01853.

Ailin Deng, Tri Cao, Zhirui Chen, and Bryan Hooi. 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 (CVPR)*, pp. 3867–3876. IEEE, June 2025. doi: 10.1109/CVPR52734.2025.00366.

Adam Fisch, Kenton Lee, Ming-Wei Chang, Jonathan Clark, and Regina Barzilay. CapWAP: Image Captioning with a Purpose. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8755–8768, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.705.

Gemma Team et al. Gemma 4 technical report, 2026.

GLM-V Team et al. GLM-4.5V and GLM-4.1V-Thinking: Towards Versatile Multimodal Reasoning with Scalable Reinforcement Learning, July 2025.

Minghao Guo et al. MemEye: A Visual-Centric Evaluation Framework for Multimodal Agent Memory, May 2026.

Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A. Smith, and Jiebo Luo. PromptCap: Prompt-Guided Image Captioning for VQA with GPT-3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2963–2975, Paris, France, October 2023. IEEE. doi: 10.1109/ICCV51070.2023.00277.

Hyounghun Kim and Mohit Bansal. Improving Visual Question Answering by Referring to Generated Paragraph Captions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3606–3612, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1351.

Hwanhee Lee, Thomas Scialom, Seunghyun Yoon, Franck Dernoncourt, and Kyomin Jung. QACE: Asking Questions to Evaluate an Image Caption. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 4631–4638, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.395.

Qing Li, Jianlong Fu, Dongfei Yu, Tao Mei, and Jiebo Luo. Tell-and-Answer: Towards Explainable Visual Question Answering using Attributes and Captions. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1338–1346, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1164.

Zeshang Li and Shuoyang Zhang. GEASS: Gated Evidence-Adaptive Selective Caption Trust for Vision-Language Models, May 2026.

Alexander H. Liu et al. Ministral 3, January 2026.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 1650–1654. IEEE, April 2021. doi: 10.1109/isbi48211.2021.9434010.

Zikang Liu, Junyi Li, Wayne Xin Zhao, Dawei Gao, Yaliang Li, and Ji-rong Wen. PAL-UI: Planning with Active Look-back for Vision-Based GUI Agents, October 2025.

Arjun Majumdar et al. OpenEQA: Embodied Question Answering in the Era of Foundation Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16488–16498, June 2024. doi: 10.1109/CVPR52733.2024.01560.

OpenGVLab. OpenGVLab/InternVL2-8B model card. `https://huggingface.co/OpenGVLab/InternVL2-8B`, 2024.

Qwen Team. Qwen3.8-27B model card. https://huggingface.co/Qwen/Qwen3.8-27B, 2026.

Aditya Sharma, Michael Saxon, and William Yang Wang. Losing Visual Needles in Image Haystacks: Vision Language Models are Easily Distracted in Short and Long Contexts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 5429–5451, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.312.

Weiyun Wang et al. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency, August 2025.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. VideoAgent: Long-form Video Understanding with Large Language Model as Agent. In *Computer Vision – ECCV 2024*, volume 15138, pp. 58–76. Springer Nature Switzerland, Cham, October 2024. ISBN 978-3-031-72988-1 978-3-031-72989-8. doi: 10.1007/978-3-031-72989-8_4.

Huyu Wu, Meng Tang, Xinhan Zheng, and Haiyun Jiang. When Language Overrules: Revealing Text Dominance in Multimodal Large Language Models, August 2025.

Shijia Yang, Yunong Liu, Bohan Zhai, Ximeng Sun, Zicheng Liu, Emad Barsoum, Manling Li, and Chenfeng Xu. CaptionQA: Is Your Caption as Useful as the Image Itself?, November 2025.

Botai Yuan et al. EchoBench: Benchmarking Sycophancy in Medical Large Vision-Language Models, September 2025. Comment: 29 pages, 6 figures.

Lu Zhang, Tiancheng Zhao, Heting Ying, Yibo Ma, and Kyusong Lee. OmAgent: A Multi-modal Agent Framework for Complex Video Understanding with Task Divide-and-Conquer. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 10031–10045, November 2024. doi: 10.18653/v1/2024.emnlp-main.559.

Yunpu Zhao, Rui Zhang, Junbin Xiao, Changxin Ke, Ruibo Hou, Yifan Hao, and Ling Li. Sycophancy in vision-language models: A systematic analysis and an inference-time mitigation framework. *Neurocomputing*, 659:131217, January 2026. doi: 10.1016/j.neucom.2025.131217.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei . Visual7W: Grounded Question Answering in Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4995–5004. IEEE, June 2016. doi: 10.1109/CVPR.2016.540.