# GROUNDING ON REQUEST: LEARNING MODALITY CONTROL IN VISION-LANGUAGE MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

When an image and its caption disagree, can a user reliably control which one a vision-language model (VLM) follows? Accuracy alone cannot answer this question, since a model might give the correct answer from parametric knowledge without following the requested evidence. We therefore evaluate modality grounding control through conflict-preserving counterfactual interventions. Specifically, FLIP measures whether an initially correct answer follows an edit to the requested modality, while STAY measures whether it remains stable when the competing modality is edited. To learn this control, we build on *neologism learning* by introducing new words for visual and textual grounding, training only their embeddings while keeping the pretrained model fixed. We also introduce Counterfactual Twin-Margin (CTM) supervision, which uses paired interventions to teach how answer probabilities should change when the requested evidence changes. Across three VLMs on three existing benchmarks and our multi-attribute Ground-CF benchmark, we find that learned modality words improve control over natural-language instructions. Surprisingly to us, a single learned embedding per modality provides strong lightweight visual control and remains stable on held-out attribute values. LoRA remains strongest overall, but its visual advantage narrows on the held-out evaluations. Moreover, counterfactual supervision improves control in several settings where standard supervision is weaker. Finally, we test our method on real-world news video, and find that visual control improves when the relevant evidence is made easier to access. Overall, our results show that reliable modality control requires testing not only whether an answer is correct, but whether it continues to follow edits to the modality of choice. [1]

## 1 INTRODUCTION

Vision-language models increasingly receive several modalities at the same time, including images, videos, and accompanying text (Bordes et al., 2024; Fang et al., 2026; Sanders et al., 2024; 2023). Often these sources agree, but what should happen when these modalities do not? In settings with cross-modal conflict Zhang et al. (2025); Zhu et al. (2024), the appropriate answer can depend on which modality the user wants the model to follow Catapang (2026); Hua et al. (2025). In practice, however, VLMs can exhibit *modality dominance*, favoring one modality (often text) even when explicitly asked to use another (Ghatkesar et al., 2025; Hua et al., 2025; Jia et al., 2025). Previous work has therefore focused largely on either strengthening visual grounding or reducing the influence of dominant textual information (Ghatkesar et al., 2025; Shi et al., 2026; Wu et al., 2025). Rather than pushing the model permanently toward one modality, we seek to understand how we can allow the user to choose either modality on demand.

Crucially, measuring this modality control requires more than just checking accuracy. Suppose an image shows a purple frog, but the caption describes it as green (Figure 1). If we ask the model to look at the image and it answers *purple*, standard evaluation marks this as correct. However, correctness alone does not establish that the answer reliably follows the image. To test this, we change only the requested modality (the image) so that the frog becomes orange. An image-grounded answer should then change to *orange*. On the other hand, if we only change the competing caption, an image-grounded answer should remain unchanged. We therefore evaluate modality grounding

[1] Data and code available upon publication



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.

through paired counterfactual interventions. Specifically, FLIP measures whether an initially correct answer follows an edit to the requested modality, while STAY measures whether it remains stable when the competing modality changes.

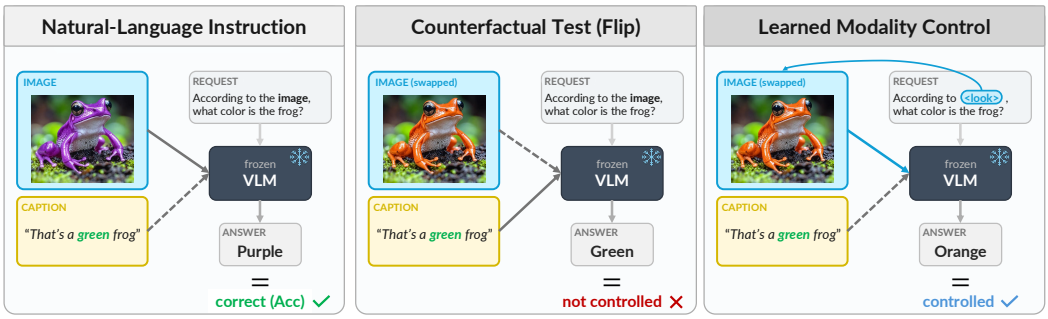


Figure 1: **Accuracy misses modality grounding control.** Left: the image shows a purple frog, the caption says green, and the model correctly answers purple under an image-grounding request. Middle: after the image changes to orange, answering green fails to follow the requested modality. Right: the desired answer is orange, shown with the learned visual grounding word `<look>`.

Our work aims to bridge the gap between measuring modality control and explicitly learning it. In particular, we expand on *neologism learning* (Hewitt et al., 2025a;b), which introduces new words with new word embeddings. These words are placed directly in natural language, while only their embeddings are trained and the pretrained model remains fixed. We introduce `<look>` for visual grounding and `<read>` for textual grounding. Since standard supervised fine-tuning (SFT) optimizes examples independently, it does not explicitly teach how answer preferences should move when the requested evidence changes. To address this, we introduce the *Counterfactual Twin-Margin* (CTM) objective, which uses paired interventions so that the original answer becomes less likely and the replacement answer more likely after an edit to the requested modality.

To study this systematically, we evaluate three VLMs on three existing benchmarks and construct **Ground-CF**, a controlled benchmark with image–text interventions over color, shape, count, and spatial relations. We compare learned modality words with natural-language instructions (Shi et al., 2026), parameter-matched Prompt-tuning (Lester et al., 2021), and higher-capacity LoRA adaptation (Hu et al., 2022). We find that natural-language instructions provide substantially weaker modality control, particularly for visual requests. Surprisingly to us, a single learned embedding per modality already provides strong lightweight visual control and remains stable on average across held-out attribute values. LoRA achieves the strongest absolute control, while its visual advantage becomes smaller on the held-out evaluations. Moreover, counterfactual supervision improves control in several settings where standard supervision is weaker. Finally, we take it one step further and evaluate these methods on real-world news videos using MultiVENT-G (Sanders et al., 2024). Here, Neologisms again provide the strongest lightweight visual control, despite using substantially fewer trainable parameters than LoRA, and improve considerably when the relevant evidence is made easier to access.

Overall, our work contributes: (1) a conflict-preserving counterfactual formulation of modality grounding control using FLIP and STAY, together with Ground-CF; (2) evidence that accuracy and natural-language instructions substantially overstate reliable visual control; (3) adapting neologism learning for user-selectable modality grounding, together with CTM-based counterfactual supervision; and (4) evidence that lightweight learned controls are effective on held-out attributes and extend to realistic news-video frames, where access to relevant visual evidence becomes increasingly important.

## 2 RELATED WORK

**Modality conflict and grounding.** There has been substantial progress in understanding how VLMs combine visual and textual information, particularly when the two modalities disagree Liu et al. (2025); Luo et al. (2025); Deng et al. (2025). Prior work finds that models often rely unevenly on their multimodal inputs, with language priors frequently dominating visual evidence (Ghatke-

sar et al., 2025; Catapang, 2026; Ghosh et al., 2026; Li et al., 2026). This has motivated methods that strengthen visual grounding or reduce the influence of dominant textual information, including contrastive region guidance (Wan et al., 2024), specialized visual perception tokens (Yu et al., 2025), active visual querying (You et al., 2026), perception-focused rewards (Wang et al., 2026), information-flow interventions (Cheng et al., 2026), and explicit user instructions (Shi et al., 2026; Wu et al., 2025). These works largely study which modality dominates and how conflict leads to hallucinations. We approach the problem from a different angle. Rather than deciding which modality should generally dominate, we ask: given a conflict, can the user select *either* modality at inference time, and will that choice remain reliable under intervention?

**Counterfactual dependence and attribution.** Previous work has explored empirically whether model outputs are supported by, or depend on, particular evidence (Song et al., 2026; Yu et al., 2026; Wallat et al., 2024). Since evidence that supports an answer does not establish that the answer depended on it, counterfactual interventions provide a stronger behavioral test by changing the evidence and measuring whether the prediction changes accordingly (Kamahi & Yaghoobzadeh, 2024). VisualSwap, for example, measures whether VLMs respond when visual evidence is replaced (Shi et al., 2026). Our work builds on this idea, but we test both sides of user-selected modality control. Specifically, we measure whether predictions follow changes to the requested modality and remain stable when only the competing modality changes. We also go one step further and use these paired interventions during training rather than only for evaluation.

**Steering, neologisms, and training objectives.** There has also been substantial progress on controlling model behavior, moving from interventions on internal representations (Huben et al., 2024; Rimsky et al., 2024; Tan et al., 2024; Liu et al., 2024), to lightweight adaptations such as prompt-tuning (Lester et al., 2021; Li & Liang, 2021), or LoRA (Hu et al., 2022). In particular, we expand on *neologism learning*, where new words are introduced with new word embeddings while the pretrained model remains fixed (Hewitt et al., 2025a;b). These words have previously been used to learn concepts and behaviors such as response style or simple skills (Hewitt et al., 2025b; Berthon et al., 2026). We instead use them as explicit user-selectable modality controls. Standard SFT can train on both members of a counterfactual pair, but optimizes each context independently. In contrast, CTM directly couples the same candidate answers across paired evidence edits. This allows us to train the directional behavior tested by our evaluation: when the requested modality changes, the old answer should become less likely while the replacement becomes more likely.

## 3 Measuring and Learning Modality Grounding Control

As described in the previous section, a correct answer under a modality request does not establish that the answer reliably follows the requested evidence. To quantify the effectiveness of modality grounding control, we therefore ask how predictions respond to edits of the modality inputs. When the requested modality changes, the answer should follow it. On the other hand, when only the competing modality changes, the answer should remain the same. We first formalize these two behaviors through paired counterfactual interventions, and then use the same interventions to learn stronger modality controls.

### 3.1 Measuring Modality Grounding Control

A modality grounding request $r_s$, such as a natural-language instruction or learned word, selects which source $s$ should ground the answer. We investigate three behaviors: the model should answer correctly from the requested modality, follow it when it changes, and remain stable when only the competing modality changes.

Let $C = (q, s_1, \ldots, s_M)$ contain a question $q$ and $M$ evidence sources, where the requested source $s$ supports answer $v_s$. Since each benchmark provides a fixed set of valid candidate answers, we evaluate the model directly through their sequence log-likelihoods rather than relying on open generation or an external judge. For candidate $y = (y_1, \ldots, y_T)$, we therefore score:

$$m_s(y \mid C) = \sum_{t=1}^{T} \log p(y_t \mid C, r_s, y_{<t}), \qquad \hat{y}_s(C) = \arg\max_{y \in \mathcal{Y}} m_s(y \mid C). \tag{1}$$

Here, $m_s(y \mid C)$ sums the token log-probabilities of candidate $y$ under grounding request $r_s$, and $\hat{y}_s(C)$ is the selected candidate.

**Accuracy.** Accuracy measures whether the model returns the value supported by the requested modality, $A_s = \Pr[\hat{y}_s(C) = v_s]$. This measures correctness on the original input, but not whether the answer continues to follow that modality when the evidence changes.

**Flip and Stay.** To measure this control directly, we construct two counterfactual twins. For FLIP, $C'_s$ changes only the requested modality, replacing $v_s$ with $v'_s$. For STAY, $C'_u$ changes only the competing modality $u \neq s$, while $v_s$ remains unchanged. We measure:

$$\text{FLIP}_s = \Pr[\hat{y}_s(C) = v_s \wedge \hat{y}_s(C'_s) = v'_s], \qquad \text{STAY}_{s,u} = \Pr[\hat{y}_s(C) = v_s \wedge \hat{y}_s(C'_u) = v_s]. \quad (2)$$

Both metrics are conjunctive at the item level. A successful FLIP requires the answer to remain correct before and after the requested-modality edit, while a successful STAY requires it to remain correct after an edit to the competing modality. Thus, FLIP$s \leq A_s$ and STAY$s, u \leq A_s$. High Accuracy with lower FLIP or STAY therefore exposes initially correct answers that do not reliably behave according to the requested source.

Our result is conceptually related to prior intervention metrics, but differs in jointly testing both sides of control on the same item: sensitivity when the requested modality changes and invariance when the competing modality changes (Shi et al., 2026; Hua et al., 2025; Zhu et al., 2024).

### 3.2 LEARNING MODALITY CONTROL WITH COUNTERFACTUAL SUPERVISION

In this section, we draw from and extend neologism learning (Hewitt et al., 2025a;b) to learn modality-specific controls. At a high level, neologism learning keeps the pretrained model fixed, expands its vocabulary and embedding matrix with new words, and optimizes only these new embeddings on examples that exemplify a desired concept or behavior. We introduce `<look>` for visual grounding and `<read>` for textual grounding, represented by $L$ newly added tokens per word. These words are placed directly in natural language, while only their embeddings are trained. The added token IDs are masked from the output vocabulary so they can solely be used as control inputs but not generated as answers. Since standard SFT optimizes examples independently, it does not explicitly constrain how answer preferences should change across a counterfactual pair. Our work aims to bridge this gap through the *Counterfactual Twin-Margin* (CTM) objective.

Consider a pair $(C, C'_s)$ in which the requested modality supports answer $v$ before the edit and $v'$ afterwards. We seek to make the old answer less likely after the edit and the replacement answer more likely. Specifically, we define:

$$\Delta_{\text{old}} = m_s(v \mid C) - m_s(v \mid C'_s), \qquad \Delta_{\text{new}} = m_s(v' \mid C'_s) - m_s(v' \mid C). \quad (3)$$

Thus, positive values of $\Delta_{\text{old}}$ and $\Delta_{\text{new}}$ indicate that both answers move in the desired direction across the intervention.

We encourage these shifts with a logistic loss:

$$\mathcal{L}_{\text{CTM}} = -\log \sigma(\gamma \Delta_{\text{old}}) - \log \sigma(\gamma \Delta_{\text{new}}), \qquad \gamma > 0. \quad (4)$$

We then train with $\mathcal{L} = \mathcal{L}_{\text{SFT}} + \mathcal{L}_{\text{CTM}}$.

**Proposition 1** (Two-sided margins move relative preference)**.** *Let $D(C) = m_s(v \mid C) - m_s(v' \mid C)$ denote the log-probability ratio of the original answer to the replacement answer. Then: $D(C) - D(C'_s) = \Delta_{\text{old}} + \Delta_{\text{new}}$. Consequently, if both CTM margins are positive, then $D(C) > D(C'_s)$.*

Thus, CTM moves the relative preference toward the replacement answer after the requested modality changes. Importantly, this does not guarantee a successful FLIP: the preference may not cross the decision boundary, or another candidate may still score higher. Moreover, CTM does not explicitly train invariance to edits of the competing modality. Interestingly, as we show later, this distinction matters empirically. Hence, we continue to evaluate FLIP and STAY separately. Appendix C provides the derivation and counterexamples.

# 4 EXPERIMENTAL SETUP

We test modality grounding control across models, benchmarks, and adaptation methods. Our main experiments use controlled image–text conflicts, where we can intervene on one modality while keeping the other fixed. Finally, we go one step further and evaluate the same control methods on more realistic news-video frames.

## 4.1 BENCHMARKS

We evaluate four controlled image–text benchmarks. **Visual-Counterfactual** (Golovanevsky et al., 2025) tests colour changes while preserving object identity. **ColorSwap** (Burapacheep et al., 2024) tests colour–object binding by reassigning colours between objects, while **COCO-CF** (Le et al., 2023) contains controlled entity substitutions in natural image–text pairs. We also introduce **Ground-CF**, inspired by CLEVR (Johnson et al., 2017), which extends these interventions across colour, shape, count, and spatial relations. For all primary evaluations, we construct *conflict-preserving* counterfactual twins. Specifically, when one modality is edited, its replacement value differs from both its original value and the value supported by the competing modality. This allows us to test whether the model continues to follow the requested source while contradictory evidence remains present. Full construction details and a comparison with agreement-resolving interventions are provided in Appendix A. Finally, **MultiVENT-G** (Sanders et al., 2023; 2024) provides real-world news-video frames, text, and human-annotated visual evidence. We use it as a separate application setting to test modality control under substantially more complex visual inputs. Dataset adaptations and evaluation details are given in Appendix E.

**Models and configurations.** We evaluate Gemma-3-4B, Qwen3.5-4B, and Qwen3.5-2B (Team et al., 2025; Yang et al., 2025). We compare learned neologisms with the frozen base model, natural-language grounding instructions (Shi et al., 2026), parameter-matched prompt tuning (Lester et al., 2021), and higher-capacity LoRA adaptation (Hu et al., 2022). Neologism learning and prompt tuning use $L = 1$ learned embedding per request, unless stated otherwise. Prompt tuning prepends the learned vector to the input, while neologism learning adds a new vocabulary token that is used directly inside the natural-language grounding request. LoRA uses rank $r = 16$ and one specialist adapter per requested modality, providing a higher-capacity, source-specific baseline. All learned methods use SFT+CTM supervision in the main comparisons. We evaluate with Accuracy, FLIP, and STAY as defined in section 3. Full prompts, optimization settings, candidate sets, checkpoint selection, and intervention coverage are provided in Appendix A.

# 5 ANALYTICAL RESULTS

We first study modality grounding under controlled image-text conflicts. We begin by asking which modality the unadapted models follow when no grounding request is provided. We then test whether natural-language requests reliably change this behavior, before asking whether modality control can instead be learned.

**Models have different default modality preferences.** Before asking how to control modality grounding, we first examine what models do when image and text disagree. A model may select the value supported by the image, the text, or neither. Table 1 shows that modality preference varies substantially across both models and benchmarks. On COCO-CF, all three models favor textual evidence, whereas on Ground-CF and ColorSwap both Qwen models favor the image. VCF shows further variation across models.

This variation is important for modality control, because if the same modality dominated across models and benchmarks, one could potentially address the problem by strengthening the weaker source or suppressing the dominant one. Instead, we find that there is no single modality that should generally be amplified or suppressed. A reliable modality grounding control method should therefore allow the user to select either modality on demand, irrespective of the base model's default preference.

| | Gemma-3-4B | | Qwen-3.5-4B | | Qwen-3.5-2B | |
|---|---|---|---|---|---|---|
| Benchmark | Image | Text | Image | Text | Image | Text |
| VCF | .40 | **.60** | .50 | .50 | **.67** | .33 |
| ColorSwap | .43 | **.57** | **.68** | .32 | **.74** | .26 |
| COCO-CF | .32 | **.68** | .16 | **.84** | .36 | **.64** |
| Ground-CF | .43 | **.50** | **.70** | .23 | **.62** | .24 |

Table 1: **Default modality preference depends on both model and benchmark.** We report how often each base model selects the value supported by the image or text on conflicting examples. The larger value is shown in bold.

### 5.1 Accuracy Does Not Measure Reliable Modality Grounding Control

We next ask whether natural-language instructions already provide this control. Standard multimodal evaluation focuses on whether a model gives the correct answer (Accuracy). We approach the problem from a different angle, asking whether that answer continues to follow the same modality when the evidence changes. If the requested image changes, the answer should change accordingly. On the other hand, if only the competing caption changes, the answer should remain the same. Flip and Stay measure these two behaviors directly.

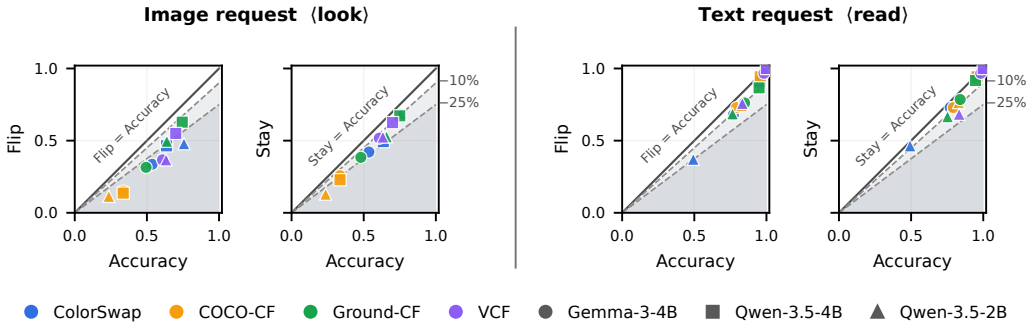


Figure 2: **Accuracy overstates reliable modality control.** for natural-language image requests (left) and text requests (right). If every initially correct answer behaves appropriately under intervention, the corresponding score equals Accuracy. Points below the diagonal therefore expose correct predictions that fail the counterfactual test.

Figure 2 shows that natural-language instructions do not provide reliable modality control, particularly for visual grounding requests. Across all 12 model–benchmark settings, every image request loses at least a 10% of its initially correct answers under the requested-modality intervention, and in 9 settings at least 25% fail to follow the image edit ((Flip)). For example, Gemma-3-4B reaches .336 Accuracy on COCO-CF under an image instruction, but drops to .150 under Flip. On the other hand, text requests are considerably more reliable. Only 5 of 12 settings lose at least 10%, and none loses 25% for (Flip). We find the same asymmetry for Stay, with 11/12 image-request settings showing at least a 10% drop, compared with only 2 of 12 in the text-request settings.

Thus, correctness and control are not equivalent. And overall, the two analyses motivate both parts of our approach. Modality dominance varies across models and benchmarks Table 1, so a general solution cannot simply steer every model toward the same modality. Moreover, natural-language instructions do not reliably provide user-selectable modality grounding control, and Accuracy alone would hide these failures. We therefore want to investigate whether a learned modality grounding request can provide more reliable control over either modality, and evaluate this control through Flip and Stay.

### 5.2 Learning Reliable Modality Grounding Control

Rather than pushing the model permanently toward vision or text, we seek a lightweight modality control that allows either modality to be selected on demand. We compare neologisms with

parameter-matched prompt tuning and a substantially higher-capacity, source-specific LoRA baseline. We evaluate both in-distribution (ID) and on held-out values or concepts.

Table 2 reveals a clear difference between the two modalities. Textual control is comparatively easy, since all learned methods approach the ceiling, with prompt tuning slightly stronger than neologisms and LoRA strongest overall. Visual control, on the other hand, remains substantially harder. Here, neologisms provide the strongest lightweight control, outperforming prompt-tuning on both FLIP and STAY in ID and OOD. This aggregate includes one Gemma-3-4B COCO-CF prompt-tuning failure, while the remaining cells are substantially closer; full per-cell results are reported in Appendix D. We therefore interpret the aggregate result as evidence that neologisms provide competitive lightweight visual control across the full grid while avoiding the benchmark-specific collapse observed for prompt tuning.

Interestingly, neologism visual control remains almost unchanged on held-out values: FLIP decreases by only $0.7\%$ and STAY by $1.6\%$. Prompt tuning is similarly stable on average, while LoRA remains strongest in absolute terms but loses more of its visual advantage, with FLIP decreasing by $9.4\%$ and STAY by $8.6\%$. Overall, a single learned embedding per request provides strong visual grounding control with very few trainable parameters.

| Method | Visual request <look> | | | | | | Text request <read> | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ID | | | OOD | | | ID | | | OOD | | |
| | A | F | S | A | F | S | A | F | S | A | F | S |
| Base model | .398 | .208 | .294 | .419 | .251 | .324 | .546 | .422 | .435 | .513 | .417 | .405 |
| Instruction | .555 | .367 | .455 | .573 | .401 | .486 | .848 | .784 | .799 | .825 | .762 | .784 |
| Prompt-tuning† | .742 | .630 | .683 | .734 | .630 | .683 | .979 | .968 | .976 | .981 | .971 | .979 |
| Neologism | .800 | .664 | .733 | .779 | .657 | .717 | .940 | .917 | .930 | .921 | .902 | .912 |
| LoRA | **.911** | **.800** | **.869** | **.833** | **.706** | **.783** | **.999** | **.999** | **.999** | **.998** | **.997** | **.998** |

Table 2: **Learned modality requests substantially improve control, while visual grounding remains harder.** Scores are averaged across three models and four benchmarks. We report FLIP (F) and STAY (S) for ID and held-out evaluation. †The prompt-tuning visual average includes one collapsed Gemma-3-4B COCO-CF cell (image Accuracy .050), the full per-cell results are reported in Appendix D. Bold denotes the best method and underlining the runner-up.

**When does counterfactual supervision help?** The previous comparison fixes SFT+CTM across learned methods. We now isolate the role of counterfactual supervision within neologism learning by comparing the same $L{=}1$ word trained with SFT or SFT+CTM. Since CTM also introduces paired counterfactual contexts, we additionally compare against paired SFT with the same twins but no cross-context coupling. Interestingly, exposure to the twins alone improves caption control but not visual control, whereas CTM improves visual FLIP beyond paired SFT in both ID and held-out evaluation. Thus, the visual-control gain cannot be explained by counterfactual exposure alone and instead depends on coupling the paired contexts. Full results are reported in subsection D.6.

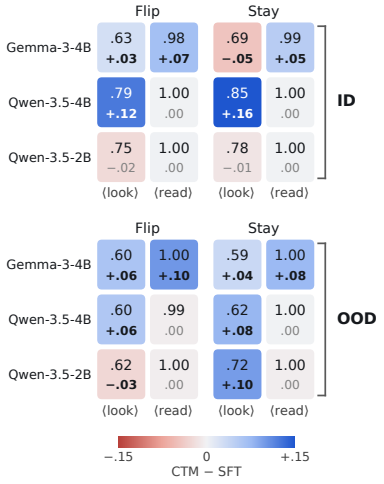


Figure 3: **CTM helps most where SFT is weaker.** Ground-CF results for neologisms trained with SFT or CTM; cells show CTM and CTM−SFT (bold: McNemar $p < .05$).

Figure 3 shows that CTM helps most where SFT is weaker. In ID evaluation, the clearest gain appears for Qwen3.5-4B visual grounding, where FLIP/STAY increase by $+.12/+.16$. On held-out values, improvements are broader, where Gemma-3-4B improves across both requests, Qwen3.5-4B again improves on visual control, and Qwen3.5-2B gains $+.08$ in STAY. The effect is not uniform across every model and metric, with complete results reported in subsection D.6.

Interestingly, explicitly adding STAY supervision does not consistently improve control and can reduce FLIP. On the other hand, CTM alone improves STAY in several settings despite never op-

timizing it directly. This suggests that learning to follow the requested modality can also induce selectivity toward competing evidence, while directly optimizing both behaviors may introduce a sensitivity–invariance trade-off. Overall, paired counterfactual supervision is most useful where standard supervision is weaker, particularly on held-out values.

**Does the model know what the learned words mean?** Since neologisms are actual new vocabulary words, we can ask the model directly what they mean. This is less natural for prompt tuning or LoRA, where the control is stored in learned vectors or adapter weights rather than in a word. As shown in Figure 4, models can often tell that <look> and <read> are different, but are substantially less reliable at saying which modality each word refers to. On the other hand, these same words still strongly control which modality the model follows under intervention. This suggests that the learned words can control behavior without the model being able to clearly explain or self-verbalize their meaning. Since these probes use $L$=1 Ground-CF neologisms, stronger controls or additional learned tokens may make their meaning easier for the model to verbalize.

**Metalinguistic probes**

**1. Distinguishing the learned words**

```
Are <look> and <read> the same or different?
```

Gemma-3-4B: *different*

**2. Identifying the source**

```
What does <look> ask you to use?
A. the image   B. the caption
```

Gemma-3-4B: *B. the caption*

| | Distinguish words | Identify source |
|---|---|---|
| Gemma-3-4B | .750 | .500 |
| Qwen3.5-2B | .625 | .625 |

Figure 4: **Behavioral control does not imply self-verbalizable meaning.** Models often distinguish the learned words, but are substantially less reliable at identifying the modality each word denotes.

## 6 Applied Results

### 6.1 Modality Grounding Control on Real-World News Videos

Controlled benchmarks let us isolate whether a model follows the requested modality, but real-world images are harder because the relevant evidence may be only a small part of a complex scene. We therefore also evaluate on MultiVENT-G (Sanders et al., 2024), a news-video dataset with human-annotated visual evidence for event roles (Sanders et al., 2023). We construct counterfactual pairs by replacing a text role value with another value from the same class, or by selecting another frame from the same video in which that role is linked to a different visual entity. Unlike our controlled benchmarks, these interventions are not perfectly isolated, since changing the frame can also change other parts of the scene. Since this setting is more complex, we use $L$=5 learned embedding rows per request for neologisms and prompt tuning; LoRA remains rank 16, and all learned methods are trained on MultiVENT-G with SFT+CTM. Full details are given in Appendix E.

| | Gemma-3-4B | | Qwen-3.5-4B | | Qwen-3.5-2B | |
|---|---|---|---|---|---|---|
| Method | ⟨look⟩ | ⟨read⟩ | ⟨look⟩ | ⟨read⟩ | ⟨look⟩ | ⟨read⟩ |
| Base model | .214 → .277 | .657 | .145 → .202 | .693 | .087 → .127 | .670 |
| Instruction | .191 → .266 | .727 | .162 → .364 | .820 | .110 → .173 | .760 |
| Prompt-tuning | .173 → <u>.249</u> | <u>.870</u> | .156 → .347 | .807 | .133 → .214 | <u>.863</u> |
| Neologism | .104 → .220 | .863 | .202 → <u>.486</u> | <u>.903</u> | .145 → <u>.295</u> | .850 |
| LoRA | .191 → **.416** | **.983** | .376 → **.595** | **.993** | .353 → **.538** | **.997** |

Table 3: **Making the relevant visual evidence easier to find improves visual control on news frames.** We report FLIP on MultiVENT-G. For ⟨look⟩, scores compare the full frame with a crop centered on the human-annotated evidence region (Full → Crop), making the relevant visual information a larger part of the input. Crops are used only at evaluation time. For ⟨read⟩, we report full-frame performance. Visual pairs have $n$=173 and description pairs $n$=300 per model. Bold denotes the best learned method and underlining the runner-up.

Table 3 shows that textual control remains strong, while full-frame visual control is considerably harder. Most ⟨look⟩ FLIP scores are near or below the .25 chance level, even though per-frame ac-

curacy remains between .50 and .66. Thus, the models can often answer individual frames correctly without reliably changing their answer when the visual evidence changes. Among the lightweight methods, neologisms are particularly strong on the Qwen models. After cropping, they outperform parameter-matched prompt tuning on both Qwen-3.5-4B (.486 vs. .347) and Qwen-3.5-2B (.295 vs. .214), with the largest gain on the larger Qwen-3.5-4B model. Gemma-3-4B is the exception, where prompt tuning remains slightly stronger. This suggests that the benefit of the learned visual word depends in part on the underlying model's ability to use the available visual evidence.

**Making the relevant evidence easier to find.** Since MultiVENT-G provides human-annotated evidence boxes, we can test whether part of the remaining difficulty comes from finding the relevant visual evidence. Cropping around the annotated region improves visual FLIP for every method and model, without retraining. Neologisms improve from .104 to .220 on Gemma-3-4B, from .202 to .486 on Qwen-3.5-4B, and from .145 to .295 on Qwen-3.5-2B. This is consistent with the broader pattern in our controlled experiments. Neologisms are competitive with prompt tuning on average, but their visual advantage becomes clearer in harder visual settings such as COCO-CF and on MultiVENT-G. Overall, these results suggest that learned modality words are especially promising for visual control when the model has sufficient visual capability and the relevant evidence can be accessed. In such settings, a small learned vocabulary interface can provide substantial control over which visual evidence the model follows.

## 7 CONCLUSION

We asked whether a user can reliably control which modality a VLM follows when image and text disagree. Our results show that correctness alone does not establish this control. We define FLIP and STAY to measure whether predictions reliably follow the requested evidence under intervention. We find that this behavior can be learned with very few parameters. Neologisms provide strong lightweight control while keeping the backbone frozen, remain competitive with parameter-matched prompt tuning overall, and are particularly robust in harder visual settings. While LoRA remains strongest overall, its visual control degrades more on held-out values and it uses 5,300–11,600$\times$ more trainable parameters than $L{=}1$ neologisms. Moreover, CTM improves control beyond counterfactual exposure alone. On MultiVENT-G, visual control improves substantially when the relevant evidence is made easier to find, showing that modality selection and evidence access are distinct grounding challenges. Overall, reliable modality control should be evaluated by how predictions respond when the underlying evidence changes, not by correctness alone.

**Limitations.** Our work establishes modality grounding control through behavioral interventions, rather than explaining the internal mechanisms that produce this behavior. Understanding what the learned neologism tokens encode, how they steer the model toward one source over another, and how this changes as their capacity increases are therefore interesting directions for future work. Moreover, we evaluate models using fixed answer choices and held-out values from the same type of task, which allows precise measurement of modality control but does not test arbitrary distribution shifts or fully open-ended generation. MultiVENT-G moves toward more realistic scenes, although its frame interventions are less controlled and the crop analysis relies on human-annotated evidence. Finally, future work should extend these findings to larger models, open-ended generation tasks, and other modalities such as audio.

### AI USE STATEMENT

Language models assisted with writing, literature discovery, research ideation, experiment orchestration, code development, and data generation. In the MultiVENT-G experiments, Fable 5 generated claims from entity–role pairs used to construct dataset captions. The authors verified the generated data, experimental records, claims, citations, and mathematical arguments, and remain responsible for all reported results.

## REFERENCES

Antonin Berthon, Nicolás Astorga, and Mihaela van der Schaar. Skill neologisms: Towards skill-based continual learning. In *International Conference on Machine Learning*, 2026.

Florian Bordes, Richard Yuanzhe Pang, Anurag Ajay, Alexander C. Li, Adrien Bardes, Suzanne Petryk, Oscar Mañas, Zhiqiu Lin, Anas Mahmoud, Bargav Jayaraman, Mark Ibrahim, Melissa Hall, Yunyang Xiong, Jonathan Lebensold, Candace Ross, Srihari Jayakumar, Chuan Guo, Diane Bouchacourt, Haider Al-Tahan, Karthik Padthe, Vasu Sharma, Hu Xu, Xiaoqing Ellen Tan, Megan Richards, Samuel Lavoie, Pietro Astolfi, Reyhane Askari Hemmat, Jun Chen, Kushal Tirumala, Rim Assouel, Mazda Moayeri, Arjang Talattof, Kamalika Chaudhuri, Zechun Liu, Xilun Chen, Quentin Garrido, Karen Ullrich, Aishwarya Agrawal, Kate Saenko, Asli Celikyilmaz, and Vikas Chandra. An introduction to vision-language modeling, 2024. URL https://arxiv.org/abs/2405.17247.

Jirayu Burapacheep, Ishan Gaur, Agam Bhatia, and Tristan Thrush. ColorSwap: A color and word order dataset for multimodal evaluation. In *Findings of ACL*, 2024.

Jasper Kyle Catapang. When image and text disagree: Cross-modal evidence conflict in multimodal retrieval-augmented generation. In Kenton Murray and Reno Kriz (eds.), *Proceedings of the 2nd Workshop on Multimodal Augmented Generation via Multimodal Retrieval (MAGMaR 2026)*, pp. 1–10, San Diego, USA, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-425-5. doi: 10.18653/v1/2026.magmar-main.3. URL https://aclanthology.org/2026.magmar-main.3/.

Zhe Cheng, Wenyu Chen, Fode Zhang, and Dehuan Shen. Mitigating hallucinations in large vision-language models via causal route gating. In *International Conference on Machine Learning*, 2026.

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, June 2025.

Xiang Fang, Wanlong Fang, Changshuo Wang, Keke Tang, Daizong Liu, Siyi Wang, and Wei Ji. Towards unified vision-language models with incomplete multi-modal inputs, 2026. URL https://arxiv.org/abs/2605.27894.

Aarti Ghatkesar, Uddeshya Upadhyay, and Ganesh Venkatesh. Beyond language priors: Enhancing visual comprehension and attention in MLLMs. In *Second Workshop on Visual Concepts*, 2025. URL https://openreview.net/forum?id=DSAx1oMrYJ.

Rodela Ghosh, Aviral Gupta, and Guangjing Wang. Which source wins? task-dependent reliance in vision-language models. *arXiv preprint arXiv:2608.17205*, 2026.

Michal Golovanevsky, William Rudman, Michael A. Lepori, Amir Bar, Ritambhara Singh, and Carsten Eickhoff. Pixels versus priors: Controlling knowledge priors in vision-language models through visual counterfacts. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 24837–24852, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1262. URL https://aclanthology.org/2025.emnlp-main.1262/.

John Hewitt, Robert Geirhos, and Been Kim. We can't understand AI using our existing vocabulary. In *ICML Position Paper Track*, 2025a.

John Hewitt, Oyvind Tafjord, Robert Geirhos, and Been Kim. Neologism learning for controllability and self-verbalization. *arXiv preprint arXiv:2510.08506*, 2025b.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, 2022. URL https://mlanthology.org/iclr/2022/hu2022iclr-lora/.

Tianze Hua, Tian Yun, and Ellie Pavlick. How do vision-language models process conflicting information across modalities?, 2025. URL https://arxiv.org/abs/2507.01790.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=F76bwRSLeK.

Hongrui Jia, Chaoya Jiang, Haiyang Xu, Wei Ye, Mengfan Dong, Ming Yan, Ji Zhang, Fei Huang, and Shikun Zhang. Symdpo: Boosting in-context learning of large multimodal models with symbol demonstration direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9361–9371, June 2025.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017.

Sepehr Kamahi and Yadollah Yaghoobzadeh. Counterfactuals as a means for evaluating faithfulness of attribution methods in autoregressive language models. In *Proceedings of the 7th BlackboxNLP Workshop*, pp. 452–468, 2024. doi: 10.18653/v1/2024.blackboxnlp-1.28. URL https://aclanthology.org/2024.blackboxnlp-1.28/.

Tiep Le, Vasudev Lal, and Phillip Howard. COCO-counterfactuals: Automatically constructed counterfactual examples for image-text pairs. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=7AjdHnjIHX.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *EMNLP*, 2021.

Sirun Li, Minghao Liu, Ling Dai, Yong Li, Haoxin Lyu, Junting Zhou, and Fan Zhang. SIGNPOST-Bench: Benchmarking text-vision conflict resolution in multimodal large language models. *arXiv preprint arXiv:2608.04244*, 2026.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *ACL*, 2021.

Benlin Liu, Amita Kamath, Madeleine Grunde-McLaughlin, Winson Han, and Ranjay Krishna. Visual representations inside the language model. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=99e72TkWTi.

Sheng Liu, Haotian Ye, Lei Xing, and James Zou. In-context vectors: Making in context learning more effective and controllable through latent space steering, 2024. URL https://arxiv.org/abs/2311.06668.

Grace Luo, Trevor Darrell, and Amir Bar. Vision-language models create cross-modal task representations. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=77ziPGdQct.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Kate Sanders, David Etter, Reno Kriz, and Benjamin Van Durme. MultiVENT: Multilingual videos of events with aligned natural text. In *NeurIPS Datasets and Benchmarks*, 2023. URL https://arxiv.org/abs/2307.03153.

Kate Sanders, Reno Kriz, David Etter, Hannah Recknor, Alexander Martin, Cameron Carpenter, Jingyang Lin, and Benjamin Van Durme. Grounding partially-defined events in multimodal data. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 15905–15927, 2024. doi: 10.18653/v1/2024.findings-emnlp.934. URL https://aclanthology.org/2024.findings-emnlp.934/.

Chufan Shi, Cheng Yang, Yaokang Wu, Linhao Jin, Bo Shui, Taylor Berg-Kirkpatrick, and Xuezhe Ma. Are vlms seeing or just saying? uncovering the illusion of visual re-examination, 2026. URL https://arxiv.org/abs/2605.15864.

Seokwon Song, Minsu Park, and Gunhee Kim. MAVIS: A benchmark for multimodal source attribution in long-form visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. arXiv:2511.12142.

Daniel Chee Hian Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. Analysing the generalisation and reliability of steering vectors. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=v8X70gTodR.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Jonas Wallat, Maria Heuss, Maarten de Rijke, and Avishek Anand. Correctness is not faithfulness in RAG attributions. *arXiv preprint arXiv:2412.18004*, 2024. URL https://arxiv.org/abs/2412.18004.

David Wan, Jaemin Cho, Elias Stengel-Eskin, and Mohit Bansal. Contrastive region guidance: Improving grounding in vision-language models without training. In *ECCV*, 2024.

Haozhe Wang, Qixin Xu, Changpeng Wang, Taofeng Xue, Chong Peng, Wenhu Chen, and Fangzhen Lin. Bad seeing or bad thinking? rewarding perception for multimodal reasoning. In *International Conference on Machine Learning*, 2026.

Huyu Wu, Meng Tang, Xinhan Zheng, and Haiyun Jiang. When language overrules: Revealing text dominance in multimodal large language models, 2025. URL https://arxiv.org/abs/2508.10552.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Weihang You, Qingchan Zhu, David Liu, Yi Pan, Geng Yuan, and Hanqi Jiang. Vithinker: Active vision-language reasoning via dynamic perceptual querying. *arXiv preprint arXiv:2602.02873*, 2026.

Bihui Yu, Caijun Jia, Jing Chi, Xiaohan Liu, Yining Wang, He Bai, Yuchen Liu, Jingxuan Wei, and Junnan Zhu. TRACER: Verifiable generative provenance for multimodal tool-using agents. *arXiv preprint arXiv:2605.09934*, 2026.

Runpeng Yu, Xinyin Ma, and Xinchao Wang. Introducing visual perception token into multimodal large language model. *arXiv preprint arXiv:2502.17425*, 2025.

Zongmeng Zhang, Wengang Zhou, Jie Zhao, and Houqiang Li. Robust multimodal large language models against modality conflict. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=SP43jVv7fJ.

Tinghui Zhu, Qin Liu, Fei Wang, Zhengzhong Tu, and Muhao Chen. Unraveling cross-modality knowledge conflicts in large vision-language models, 2024. URL https://arxiv.org/abs/2410.03659.

# A Data and Evaluation Protocol

**Benchmarks.** We evaluate modality grounding control on Visual-Counterfactual (VCF), ColorSwap, COCO-CF, and Ground-CF. VCF tests object colours, ColorSwap tests colour–object bindings, and COCO-CF tests entity substitutions in natural images. We also construct Ground-CF, a controlled benchmark covering colour, shape, count, and spatial relations. Since our goal is to measure whether a requested source continues to control the answer when the evidence changes, all evaluations use conflicting image–text pairs and counterfactual edits that preserve this conflict. Table 4 summarizes the evaluated subsets and held-out values.

Table 4: **Controlled benchmarks and counterfactual interventions.** "Items" gives the evaluated conflict examples rather than the size of the original dataset. Ground-CF reports all items, with the eligible image-twin subset in parentheses.

| Benchmark | Queried value | Image edit | Text edit | Cand. | Items ID / Held-out | Held out |
|---|---|---|---|---|---|---|
| VCF | object colour | counterfactual-colour image of the same object | colour replaced | 4 | 120 / 140 | 94 objects unseen in training |
| ColorSwap | colour–object binding | colours swapped between objects | colours reassigned | 4 | 140 / 140 | pink, purple |
| COCO-CF | entity noun | counterfactual image | noun replaced | 4 | 140 / 140 | nouns unseen in training |
| Ground-CF | colour, shape, count, spatial relation | queried value changed; count/spatial scenes re-rendered | queried value replaced | 3 | 748 / 800 (707 / 800) | purple, pink; triangle, star; counts 7–9; *above* |

**Conflict-preserving twins.** Each example contains an image and text that support different answers. For FLIP, we change the requested source while keeping the competing source fixed; for STAY, we change only the competing source. Importantly, the two sources continue to disagree after either intervention. The question, grounding request, and candidate set remain fixed across the original example and its twins. This allows us to measure whether the prediction follows the requested evidence rather than whether an edit merely removes the conflict. For VCF, ColorSwap, and COCO-CF, the candidate set contains the original and edited values from both sources, and the same four candidates are scored in every context.

**Ground-CF.** Ground-CF provides the same intervention structure across four attributes. Colour and shape edits change the queried object's attribute, count edits change the number of queried objects, and spatial edits change the queried relation. Text twins replace only the corresponding textual value. We hold out purple and pink, triangle and star, counts 7–9, and the relation *above* as original image answers. Moreover, the image–spatial and text–count request–attribute combinations are not observed during training, allowing us to separately examine whether the learned request carries across attributes. Ground-CF contains 748 ID and 800 held-out evaluation items; valid image twins are available for 707 and 800 of these items, respectively.

**Metrics and denominators.** Accuracy measures correctness on the original input. FLIP requires the model to answer correctly before and after the requested source changes, while STAY requires it to remain correct when only the competing source changes. Since some Ground-CF ID count examples do not have valid image twins, image FLIP and text STAY use the eligible image-twin subset; image STAY and text FLIP use all examples. Any conditional quantity such as FLIP/$A$ therefore recomputes Accuracy on the same eligible items.

# B Training and Evaluation Details

**Configurations.** The controlled experiments use one learned embedding per request ($L$=1) for neologisms and prompt tuning. Both requests are learned in the same model. LoRA uses rank 16 and, in the main comparison, one specialist adapter per requested modality. We additionally evaluate one joint, instruction-conditioned LoRA adapter on Ground-CF. As shown in Table 5, the two lightweight methods use only 4,096–5,120 trainable parameters across both requests, compared with 21.8–59.6 million for the two specialist LoRA adapters.

Table 5: **Trainable parameters in the controlled experiments.** LoRA counts include both source-specific adapters; one is active for each request.

| Method | Gemma-3-4B | Qwen-3.5-4B | Qwen-3.5-2B |
|---|---|---|---|
| Neologism / prompt tuning | 5,120 | 5,120 | 4,096 |
| LoRA, two adapters | 59,604,992 | 42,467,328 | 21,823,488 |
| LoRA, one joint adapter | 29,802,496 | 21,233,664 | 10,911,744 |

**Optimization.** All learned methods in the main comparison use SFT+CTM and are evaluated at checkpoint 800. Table 6 summarizes the training settings. For neologisms and prompt tuning, gradients are restricted to the newly added embedding rows. At evaluation, these learned rows are inserted into the original pretrained embedding matrix, so the evaluated model differs from the pretrained model only through the learned control embeddings.

Table 6: **Training settings for the controlled experiments.**

| Setting | Neologism / prompt tuning | LoRA |
|---|---|---|
| Trainable parameters | added embedding rows | rank-16 adapters, $\alpha = 32$, dropout .05 |
| Adapted modules | — | q, k, v, o, gate, up, down projections; vision tower and projector frozen |
| Optimizer | AdamW, weight decay .01 | AdamW, weight decay .01 |
| Peak learning rate | .05 | .0002 |
| Schedule | 20 warm-up steps, cosine decay to 10% of peak | same |
| Steps / checkpoint | 800 / 800 | 800 / 800 |
| Batch | 1 example $\times$ 8 accumulation steps; gradient clipping 1 | same |
| Objective | $\lambda_{\mathrm{SFT}} = \lambda_{\mathrm{CTM}} = 1$, $\gamma = .1$; CTM ramps from 0 at step 100 to 1 at step 150 | same |
| Initialization | mean embedding of the initialization word's subtokens $+\mathcal{N}(0, .01^2)$; row norms capped at 2$\times$ the mean vocabulary-row norm | library default ($B = 0$) |

**How the request enters the model.** All methods receive the same image, text, question, candidate answers, and scored continuation. They differ only in how the requested modality is represented. The English baseline states the request directly. Prompt tuning inserts a learned vector at the beginning of the textual input. In contrast, a neologism is a new vocabulary word placed inside a short natural-language request. The specialist LoRA baseline selects the requested modality through the corresponding adapter, while the joint LoRA uses an English control string. Table 7 gives representative image-request examples; the text request replaces image/`<look>` with caption/`<read>`.

Table 7: **How each modality request is represented.** Evaluation wordings are not used during training.

| Method | Position | Training example | Evaluation example |
|---|---|---|---|
| English | end | — | *According to the image (and only the image), what is the answer?* |
| Prompt tuning | start | `<pt>` + task sentence | `<pt>` + held-out task sentence |
| Neologism | end | *Give me a `<look>` answer.* | *Answer in a `<look>` manner.* |
| LoRA, separate | adapter | task sentence | held-out task sentence |
| LoRA, joint | end | *Answer using only the image.* | held-out instruction |

On VCF, ColorSwap, and COCO-CF, the image is followed by the conflicting text. Ground-CF additionally includes an explicit question, and image/text order alternates across items while remaining fixed across the original example and its twins. MultiVENT-G uses a shared carrier at the end of the prompt for all learned methods, so neologisms and prompt-tuning vectors occupy the same position.

**Does placement explain the difference between the lightweight methods?** Since prompt tuning is placed before the evidence while the reference neologism follows it, we test whether position explains their difference. We therefore train an additional Ground-CF neologism at the prompt-tuning position. As shown in Table 8, moving the word to the beginning has little effect on its average visual FLIP: .725 to .716 in ID and .608 to .609 on held-out values. The word also retains strong text control at either position. On the other hand, prompt tuning remains stronger on Ground-CF visual control. Thus, placement alone does not explain the difference between the two interfaces. This is consistent with MultiVENT-G, where both controls use the same position and neologisms outperform prompt tuning on visual FLIP for the two Qwen models.

Table 8: **Ground-CF placement control**, averaged across the three models. Words are evaluated at the position in which they were trained.

| Split | Interface | $F_I$ | $S_I$ | $F_T$ | $S_T$ |
|---|---|---|---|---|---|
| ID | Word, end | .725 | .773 | .994 | .996 |
| ID | Word, start | .716 | .767 | .988 | .988 |
| ID | Prompt tuning, start | .767 | .809 | .979 | .983 |
| Held-out | Word, end | .608 | .645 | .995 | .997 |
| Held-out | Word, start | .609 | .647 | .981 | .984 |
| Held-out | Prompt tuning, start | .630 | .667 | .971 | .977 |

**Candidate scoring.** Since each benchmark provides a fixed set of valid answers, we score the candidates directly rather than relying on open generation. For candidate $y = (y_1, \ldots, y_T)$, we sum its token log probabilities while conditioning on the multimodal input, the grounding request, and the preceding candidate tokens. All candidates use the same answer format, and the highest-scoring candidate is selected. Exact ties are resolved using the fixed candidate order. The reported controlled results are therefore forced-choice likelihood results rather than free-generation accuracy.

**Computational resources.** Each run uses one GPU, primarily an NVIDIA L40S 48 GB, with A100, H100, or H200 GPUs used for part of the evaluation. On Ground-CF, 800 training steps take approximately 0.7 hours for SFT, 3.5–7.5 hours for SFT+CTM, and 4.9 hours for SFT+CTM+Stay; a full evaluation takes approximately 20–50 minutes.

## C NEOLOGISM LEARNING AND COUNTERFACTUAL SUPERVISION

**Neologism learning.** Let $E \in \mathbb{R}^{|\mathcal{V}| \times d}$ denote the input-embedding matrix of the pretrained VLM. For each modality request $s \in \{\text{look}, \text{read}\}$, we introduce $L$ new vocabulary entries with learned embeddings $\theta_s \in \mathbb{R}^{L \times d}$. The new token IDs are used only as inputs and are excluded from output scoring. In the controlled experiments, $L{=}1$, so each request is represented by a single learned embedding. At evaluation, the learned rows are inserted into the original pretrained embedding matrix, so the evaluated model differs from the pretrained model only through these new control embeddings.

**Counterfactual Twin-Margin.** Since standard SFT optimizes each example independently, it does not directly specify how answer preferences should change when the requested evidence changes. CTM instead compares the same answers across an original context $C$ and its requested-source twin $C'_s$. If the requested source supports $v$ before the edit and $v'$ afterwards, we define

$$\Delta_{\text{old}} = m_s(v \mid C) - m_s(v \mid C'_s), \qquad \Delta_{\text{new}} = m_s(v' \mid C'_s) - m_s(v' \mid C).$$

We then optimize

$$\mathcal{L}_{\text{CTM}} = -\log \sigma(\gamma \Delta_{\text{old}}) - \log \sigma(\gamma \Delta_{\text{new}}),$$

together with the original-example SFT loss. The first margin encourages the old answer to lose log-likelihood after the requested source changes, while the second encourages the new answer to gain log-likelihood.

To show what these two margins imply, define

$$D(C) = m_s(v \mid C) - m_s(v' \mid C),$$

the model's preference for the original answer over the replacement. Then

$$\Delta_{\text{old}} + \Delta_{\text{new}} = D(C) - D(C'_s).$$

Thus, if both margins are positive, the model's relative preference shifts toward $v'$ after the requested-source edit. On the other hand, a positive sum alone does not imply that both individual margins are positive.

**What CTM does and does not enforce.** CTM provides a directional training signal, but it does not guarantee that $v'$ becomes the highest-scoring answer. We therefore still evaluate FLIP directly. Moreover, CTM only observes edits to the requested source and does not explicitly train invariance to changes in the competing source. For the Stay ablation, we additionally use

$$\mathcal{L}_{\text{Stay}} = -\log \sigma(\gamma[m_s(v \mid C'_u) - m_s(w \mid C'_u)]),$$

where $C'_u$ changes the competing source to value $w$ while the requested answer remains $v$.

**What the objective comparisons test.** SFT trains only on the original examples, whereas CTM introduces the counterfactual twins together with the cross-context margin. To separate these two effects, paired SFT receives the same original and edited examples but optimizes them independently with cross-entropy. Thus, CTM versus paired SFT tests whether explicitly coupling the two contexts provides additional benefit beyond simply seeing the same counterfactual examples. We report this control for Gemma-3-4B and Qwen-3.5-2B and restrict conclusions from this comparison to these two models.

# D Complete Controlled Results

## D.1 Full Model–Benchmark Results

Tables 9–11 report every cell used in the controlled main-grid averages. $I$ and $T$ denote image and text requests, while $A$, $F$, and $S$ denote Accuracy, FLIP, and STAY. Across the full grid, neologisms improve visual FLIP over natural-language instructions in every cell. Prompt tuning and neologisms are closer, and their relative performance depends on the model and benchmark. On the other hand, the substantially larger LoRA adapters provide the strongest average control overall.

## D.2 Ground-CF Results by Attribute

Since Ground-CF covers four different attributes, we additionally separate colour, shape, count, and spatial-relation questions. The largest differences appear for visual grounding. Neologisms remain particularly strong on colour and shape, while count and spatial relations are harder, especially on held-out values. Text control, on the other hand, remains substantially stronger across attributes.

This breakdown also explains why the cross-benchmark held-out average can remain stable even though Ground-CF itself becomes harder. For neologisms, visual FLIP decreases from $.725$ to $.608$ on Ground-CF when moving from ID to held-out evaluation, with much of the difficulty concentrated in count and spatial-relation questions. Moreover, the image–spatial and text–count request–attribute combinations are not observed during training, allowing us to test whether the learned modality request carries across attributes.

## D.3 Control Conditional on an Initially Correct Answer

Since FLIP and STAY require the original prediction to be correct, we also examine control conditional on this initial correctness. Table 12 reports $F/A$ and $S/A$ for the natural-language instruction, with Accuracy computed over the same eligible examples. For visual requests, between $16\%$ and $60\%$ of initially correct answers still fail to follow the image edit. On the other hand, text requests are generally more reliable. Thus, the gap between correctness and control remains even after conditioning on an initially correct prediction.

Table 9: **Complete Gemma-3-4B results.** Ground-CF visual FLIP and text STAY use the eligible image-twin subset; all other metrics use the full evaluation set. Joint LoRA is evaluated only on Ground-CF.

| Split | Dataset | Method | $A_I$ | $F_I$ | $S_I$ | $A_T$ | $F_T$ | $S_T$ |
|---|---|---|---|---|---|---|---|---|
| ID | VCF | No request | 0.167 | 0.017 | 0.100 | 0.833 | 0.800 | 0.683 |
| ID | VCF | Instruction | 0.608 | 0.367 | 0.517 | 0.983 | 0.967 | 0.967 |
| ID | VCF | Prompt tuning | 0.817 | 0.700 | 0.700 | 1.000 | 1.000 | 1.000 |
| ID | VCF | Neologism | 0.933 | 0.867 | 0.900 | 1.000 | 1.000 | 1.000 |
| ID | VCF | LoRA (separate) | 0.892 | 0.783 | 0.842 | 1.000 | 1.000 | 1.000 |
| ID | ColorSwap | No request | 0.407 | 0.129 | 0.250 | 0.571 | 0.471 | 0.493 |
| ID | ColorSwap | Instruction | 0.536 | 0.336 | 0.421 | 0.771 | 0.679 | 0.729 |
| ID | ColorSwap | Prompt tuning | 0.836 | 0.693 | 0.779 | 0.993 | 0.979 | 0.993 |
| ID | ColorSwap | Neologism | 0.771 | 0.586 | 0.707 | 0.786 | 0.693 | 0.757 |
| ID | ColorSwap | LoRA (separate) | 0.921 | 0.400 | 0.857 | 0.993 | 0.993 | 0.993 |
| ID | COCO-CF | No request | 0.214 | 0.057 | 0.157 | 0.579 | 0.371 | 0.500 |
| ID | COCO-CF | Instruction | 0.336 | 0.150 | 0.257 | 0.793 | 0.729 | 0.729 |
| ID | COCO-CF | Prompt tuning | 0.050 | 0.014 | 0.014 | 0.993 | 0.993 | 0.993 |
| ID | COCO-CF | Neologism | 0.607 | 0.379 | 0.479 | 1.000 | 1.000 | 1.000 |
| ID | COCO-CF | LoRA (separate) | 0.943 | 0.886 | 0.929 | 1.000 | 0.993 | 1.000 |
| ID | Ground-CF | No request | 0.457 | 0.300 | 0.349 | 0.477 | 0.320 | 0.376 |
| ID | Ground-CF | Instruction | 0.480 | 0.315 | 0.384 | 0.849 | 0.761 | 0.785 |
| ID | Ground-CF | Prompt tuning | 0.766 | 0.655 | 0.741 | 0.964 | 0.941 | 0.949 |
| ID | Ground-CF | Neologism | 0.759 | 0.632 | 0.693 | 0.996 | 0.983 | 0.987 |
| ID | Ground-CF | LoRA (separate) | 0.862 | 0.755 | 0.840 | 1.000 | 1.000 | 1.000 |
| ID | Ground-CF | LoRA (joint) | 0.825 | 0.741 | 0.802 | 0.837 | 0.807 | 0.833 |
| Held-out | VCF | No request | 0.186 | 0.071 | 0.129 | 0.814 | 0.736 | 0.700 |
| Held-out | VCF | Instruction | 0.529 | 0.271 | 0.457 | 1.000 | 0.986 | 1.000 |
| Held-out | VCF | Prompt tuning | 0.814 | 0.700 | 0.714 | 1.000 | 1.000 | 1.000 |
| Held-out | VCF | Neologism | 0.943 | 0.886 | 0.907 | 1.000 | 1.000 | 1.000 |
| Held-out | VCF | LoRA (separate) | 0.950 | 0.900 | 0.907 | 1.000 | 1.000 | 1.000 |
| Held-out | ColorSwap | No request | 0.557 | 0.300 | 0.379 | 0.371 | 0.286 | 0.286 |
| Held-out | ColorSwap | Instruction | 0.757 | 0.479 | 0.664 | 0.607 | 0.550 | 0.571 |
| Held-out | ColorSwap | Prompt tuning | 0.900 | 0.779 | 0.864 | 0.979 | 0.979 | 0.979 |
| Held-out | ColorSwap | Neologism | 0.886 | 0.700 | 0.821 | 0.600 | 0.507 | 0.564 |
| Held-out | ColorSwap | LoRA (separate) | 0.936 | 0.386 | 0.886 | 0.993 | 0.986 | 0.993 |
| Held-out | COCO-CF | No request | 0.350 | 0.193 | 0.221 | 0.514 | 0.429 | 0.436 |
| Held-out | COCO-CF | Instruction | 0.436 | 0.207 | 0.307 | 0.779 | 0.664 | 0.693 |
| Held-out | COCO-CF | Prompt tuning | 0.079 | 0.014 | 0.036 | 0.993 | 0.979 | 0.986 |
| Held-out | COCO-CF | Neologism | 0.586 | 0.400 | 0.443 | 0.979 | 0.979 | 0.979 |
| Held-out | COCO-CF | LoRA (separate) | 0.871 | 0.757 | 0.857 | 1.000 | 1.000 | 1.000 |
| Held-out | Ground-CF | No request | 0.531 | 0.415 | 0.414 | 0.351 | 0.264 | 0.305 |
| Held-out | Ground-CF | Instruction | 0.537 | 0.440 | 0.449 | 0.770 | 0.695 | 0.726 |
| Held-out | Ground-CF | Prompt tuning | 0.664 | 0.610 | 0.634 | 0.940 | 0.914 | 0.930 |
| Held-out | Ground-CF | Neologism | 0.650 | 0.604 | 0.591 | 0.996 | 0.995 | 0.996 |
| Held-out | Ground-CF | LoRA (separate) | 0.654 | 0.584 | 0.635 | 1.000 | 1.000 | 1.000 |
| Held-out | Ground-CF | LoRA (joint) | 0.581 | 0.560 | 0.565 | 0.853 | 0.809 | 0.849 |

Table 10: **Complete Qwen-3.5-4B results.** Conventions follow Table 9.

| Split | Dataset | Method | $A_I$ | $F_I$ | $S_I$ | $A_T$ | $F_T$ | $S_T$ |
|---|---|---|---|---|---|---|---|---|
| ID | VCF | No request | 0.267 | 0.117 | 0.150 | 0.733 | 0.675 | 0.583 |
| ID | VCF | Instruction | 0.700 | 0.550 | 0.625 | 1.000 | 1.000 | 1.000 |
| ID | VCF | Prompt tuning | 0.850 | 0.733 | 0.758 | 1.000 | 1.000 | 1.000 |
| ID | VCF | Neologism | 0.908 | 0.817 | 0.875 | 1.000 | 1.000 | 1.000 |
| ID | VCF | LoRA (separate) | 0.883 | 0.767 | 0.825 | 1.000 | 1.000 | 1.000 |
| ID | ColorSwap | No request | 0.529 | 0.300 | 0.357 | 0.429 | 0.257 | 0.307 |
| ID | ColorSwap | Instruction | 0.636 | 0.464 | 0.493 | 0.957 | 0.907 | 0.936 |
| ID | ColorSwap | Prompt tuning | 0.886 | 0.721 | 0.850 | 0.979 | 0.964 | 0.979 |
| ID | ColorSwap | Neologism | 0.914 | 0.786 | 0.850 | 0.943 | 0.893 | 0.936 |
| ID | ColorSwap | LoRA (separate) | 0.893 | 0.807 | 0.850 | 1.000 | 1.000 | 1.000 |
| ID | COCO-CF | No request | 0.271 | 0.164 | 0.179 | 0.671 | 0.471 | 0.600 |
| ID | COCO-CF | Instruction | 0.336 | 0.136 | 0.229 | 0.957 | 0.943 | 0.943 |
| ID | COCO-CF | Prompt tuning | 0.464 | 0.307 | 0.386 | 1.000 | 0.993 | 0.993 |
| ID | COCO-CF | Neologism | 0.636 | 0.471 | 0.564 | 1.000 | 1.000 | 1.000 |
| ID | COCO-CF | LoRA (separate) | 0.950 | 0.900 | 0.929 | 1.000 | 1.000 | 1.000 |
| ID | Ground-CF | No request | 0.711 | 0.586 | 0.622 | 0.217 | 0.088 | 0.117 |
| ID | Ground-CF | Instruction | 0.750 | 0.628 | 0.672 | 0.949 | 0.866 | 0.917 |
| ID | Ground-CF | Prompt tuning | 0.888 | 0.846 | 0.873 | 1.000 | 0.997 | 1.000 |
| ID | Ground-CF | Neologism | 0.882 | 0.789 | 0.846 | 1.000 | 1.000 | 1.000 |
| ID | Ground-CF | LoRA (separate) | 0.989 | 0.973 | 0.987 | 1.000 | 1.000 | 1.000 |
| ID | Ground-CF | LoRA (joint) | 0.925 | 0.870 | 0.910 | 1.000 | 0.996 | 0.999 |
| Held-out | VCF | No request | 0.179 | 0.071 | 0.114 | 0.821 | 0.700 | 0.714 |
| Held-out | VCF | Instruction | 0.664 | 0.471 | 0.543 | 1.000 | 1.000 | 1.000 |
| Held-out | VCF | Prompt tuning | 0.864 | 0.757 | 0.757 | 1.000 | 1.000 | 1.000 |
| Held-out | VCF | Neologism | 0.914 | 0.843 | 0.829 | 1.000 | 1.000 | 1.000 |
| Held-out | VCF | LoRA (separate) | 0.921 | 0.843 | 0.850 | 1.000 | 1.000 | 1.000 |
| Held-out | ColorSwap | No request | 0.507 | 0.379 | 0.386 | 0.450 | 0.343 | 0.257 |
| Held-out | ColorSwap | Instruction | 0.721 | 0.643 | 0.629 | 0.986 | 0.979 | 0.986 |
| Held-out | ColorSwap | Prompt tuning | 0.929 | 0.800 | 0.893 | 0.979 | 0.979 | 0.979 |
| Held-out | ColorSwap | Neologism | 0.886 | 0.800 | 0.857 | 0.929 | 0.914 | 0.921 |
| Held-out | ColorSwap | LoRA (separate) | 0.850 | 0.764 | 0.707 | 0.986 | 0.986 | 0.986 |
| Held-out | COCO-CF | No request | 0.350 | 0.136 | 0.250 | 0.600 | 0.550 | 0.443 |
| Held-out | COCO-CF | Instruction | 0.421 | 0.193 | 0.350 | 0.943 | 0.893 | 0.921 |
| Held-out | COCO-CF | Prompt tuning | 0.536 | 0.329 | 0.464 | 0.993 | 0.986 | 0.993 |
| Held-out | COCO-CF | Neologism | 0.686 | 0.471 | 0.636 | 0.979 | 0.979 | 0.971 |
| Held-out | COCO-CF | LoRA (separate) | 0.879 | 0.764 | 0.857 | 1.000 | 1.000 | 1.000 |
| Held-out | Ground-CF | No request | 0.581 | 0.554 | 0.535 | 0.242 | 0.142 | 0.176 |
| Held-out | Ground-CF | Instruction | 0.601 | 0.571 | 0.552 | 0.955 | 0.915 | 0.919 |
| Held-out | Ground-CF | Prompt tuning | 0.627 | 0.585 | 0.601 | 1.000 | 1.000 | 1.000 |
| Held-out | Ground-CF | Neologism | 0.660 | 0.600 | 0.621 | 1.000 | 0.994 | 1.000 |
| Held-out | Ground-CF | LoRA (separate) | 0.786 | 0.718 | 0.767 | 1.000 | 1.000 | 1.000 |
| Held-out | Ground-CF | LoRA (joint) | 0.735 | 0.680 | 0.716 | 0.999 | 0.999 | 0.999 |

Table 11: **Complete Qwen-3.5-2B results.** Conventions follow Table 9.

| Split | Dataset | Method | $A_I$ | $F_I$ | $S_I$ | $A_T$ | $F_T$ | $S_T$ |
|---|---|---|---|---|---|---|---|---|
| ID | VCF | No request | 0.417 | 0.150 | 0.317 | 0.583 | 0.483 | 0.317 |
| ID | VCF | Instruction | 0.633 | 0.367 | 0.525 | 0.833 | 0.758 | 0.683 |
| ID | VCF | Prompt tuning | 0.933 | 0.867 | 0.900 | 1.000 | 1.000 | 1.000 |
| ID | VCF | Neologism | 0.875 | 0.767 | 0.792 | 1.000 | 1.000 | 1.000 |
| ID | VCF | LoRA (separate) | 0.850 | 0.733 | 0.742 | 1.000 | 1.000 | 1.000 |
| ID | ColorSwap | No request | 0.557 | 0.143 | 0.471 | 0.400 | 0.279 | 0.357 |
| ID | ColorSwap | Instruction | 0.757 | 0.479 | 0.686 | 0.493 | 0.371 | 0.464 |
| ID | ColorSwap | Prompt tuning | 0.857 | 0.700 | 0.750 | 0.814 | 0.743 | 0.800 |
| ID | ColorSwap | Neologism | 0.914 | 0.793 | 0.843 | 0.550 | 0.443 | 0.479 |
| ID | ColorSwap | LoRA (separate) | 0.836 | 0.771 | 0.786 | 1.000 | 1.000 | 1.000 |
| ID | COCO-CF | No request | 0.164 | 0.079 | 0.086 | 0.786 | 0.679 | 0.721 |
| ID | COCO-CF | Instruction | 0.236 | 0.114 | 0.129 | 0.829 | 0.743 | 0.764 |
| ID | COCO-CF | Prompt tuning | 0.693 | 0.521 | 0.629 | 1.000 | 1.000 | 1.000 |
| ID | COCO-CF | Neologism | 0.579 | 0.329 | 0.464 | 1.000 | 0.993 | 1.000 |
| ID | COCO-CF | LoRA (separate) | 0.957 | 0.914 | 0.893 | 1.000 | 1.000 | 0.993 |
| ID | Ground-CF | No request | 0.620 | 0.458 | 0.488 | 0.271 | 0.164 | 0.167 |
| ID | Ground-CF | Instruction | 0.658 | 0.496 | 0.527 | 0.765 | 0.686 | 0.666 |
| ID | Ground-CF | Prompt tuning | 0.861 | 0.801 | 0.814 | 1.000 | 1.000 | 1.000 |
| ID | Ground-CF | Neologism | 0.826 | 0.754 | 0.779 | 1.000 | 1.000 | 1.000 |
| ID | Ground-CF | LoRA (separate) | 0.957 | 0.917 | 0.953 | 1.000 | 1.000 | 1.000 |
| ID | Ground-CF | LoRA (joint) | 0.917 | 0.853 | 0.910 | 0.814 | 0.793 | 0.829 |
| Held-out | VCF | No request | 0.314 | 0.157 | 0.271 | 0.686 | 0.564 | 0.529 |
| Held-out | VCF | Instruction | 0.464 | 0.243 | 0.414 | 0.864 | 0.786 | 0.757 |
| Held-out | VCF | Prompt tuning | 0.964 | 0.929 | 0.943 | 1.000 | 1.000 | 1.000 |
| Held-out | VCF | Neologism | 0.836 | 0.686 | 0.779 | 1.000 | 1.000 | 1.000 |
| Held-out | VCF | LoRA (separate) | 0.857 | 0.714 | 0.757 | 1.000 | 1.000 | 1.000 |
| Held-out | ColorSwap | No request | 0.721 | 0.279 | 0.629 | 0.264 | 0.157 | 0.207 |
| Held-out | ColorSwap | Instruction | 0.871 | 0.657 | 0.793 | 0.407 | 0.286 | 0.379 |
| Held-out | ColorSwap | Prompt tuning | 0.886 | 0.757 | 0.829 | 0.886 | 0.821 | 0.879 |
| Held-out | ColorSwap | Neologism | 0.929 | 0.843 | 0.900 | 0.586 | 0.479 | 0.536 |
| Held-out | ColorSwap | LoRA (separate) | 0.664 | 0.593 | 0.593 | 0.993 | 0.986 | 0.993 |
| Held-out | COCO-CF | No request | 0.257 | 0.114 | 0.143 | 0.721 | 0.664 | 0.607 |
| Held-out | COCO-CF | Instruction | 0.329 | 0.200 | 0.221 | 0.779 | 0.743 | 0.693 |
| Held-out | COCO-CF | Prompt tuning | 0.764 | 0.600 | 0.693 | 1.000 | 0.993 | 1.000 |
| Held-out | COCO-CF | Neologism | 0.621 | 0.429 | 0.500 | 0.986 | 0.979 | 0.979 |
| Held-out | COCO-CF | LoRA (separate) | 0.864 | 0.743 | 0.836 | 1.000 | 1.000 | 1.000 |
| Held-out | Ground-CF | No request | 0.497 | 0.344 | 0.414 | 0.324 | 0.169 | 0.200 |
| Held-out | Ground-CF | Instruction | 0.544 | 0.434 | 0.454 | 0.812 | 0.651 | 0.762 |
| Held-out | Ground-CF | Prompt tuning | 0.786 | 0.696 | 0.767 | 1.000 | 1.000 | 1.000 |
| Held-out | Ground-CF | Neologism | 0.755 | 0.619 | 0.723 | 0.998 | 0.995 | 0.995 |
| Held-out | Ground-CF | LoRA (separate) | 0.759 | 0.709 | 0.749 | 1.000 | 1.000 | 1.000 |
| Held-out | Ground-CF | LoRA (joint) | 0.708 | 0.671 | 0.698 | 0.800 | 0.790 | 0.794 |

Table 12: **Conditional control for natural-language instructions in ID evaluation.** $F/A$ and $S/A$ give the fraction of initially correct predictions that remain correct after the corresponding intervention.

| Model | Benchmark | $F_I/A_I$ | $S_I/A_I$ | $F_T/A_T$ | $S_T/A_T$ |
|---|---|---|---|---|---|
| Gemma-3-4B | VCF | .603 | .849 | .983 | .983 |
| | ColorSwap | .627 | .787 | .880 | .944 |
| | COCO-CF | .447 | .766 | .919 | .919 |
| | Ground-CF | .639 | .799 | .896 | .934 |
| Qwen-3.5-4B | VCF | .786 | .893 | 1.000 | 1.000 |
| | ColorSwap | .730 | .775 | .948 | .978 |
| | COCO-CF | .404 | .681 | .985 | .985 |
| | Ground-CF | .843 | .897 | .913 | .969 |
| Qwen-3.5-2B | VCF | .579 | .829 | .910 | .820 |
| | ColorSwap | .632 | .906 | .754 | .942 |
| | COCO-CF | .485 | .545 | .897 | .922 |
| | Ground-CF | .778 | .801 | .897 | .885 |

### D.4 Objective Comparison

We compare SFT, paired SFT, SFT+CTM, and SFT+CTM+Stay on Ground-CF in Table 15. CTM provides its clearest gains where standard supervision is weaker. In particular, on the two models with paired-SFT controls, CTM improves held-out visual control beyond exposure to the same counterfactual examples alone: by $+.065/+.059$ FLIP/STAY for Gemma-3-4B and $+.031/+.115$ for Qwen-3.5-2B. On Gemma-3-4B, the visual FLIP gain also appears in ID evaluation.

On the other hand, adding explicit STAY supervision does not consistently improve control. Although it helps some visual settings, it can substantially reduce text control. For example, Qwen-3.5-4B text FLIP falls from $1.000$ to $.512$ in ID evaluation and from $.994$ to $.610$ on held-out values. We therefore use CTM without the additional Stay term as the reference objective.

### D.5 Uncertainty and Aggregation

The main averages give equal weight to each model–benchmark cell. We use paired exact McNemar tests when two methods are evaluated on the same examples, and item-level bootstrap intervals for the aggregate change from ID to held-out evaluation. Table 14 shows that neologism and prompt-tuning visual control change little on average, whereas LoRA loses more of its visual advantage. These are aggregate results and should therefore not be interpreted as uniform behavior in every benchmark.

### D.6 Objective Comparison

We compare SFT, paired SFT, SFT+CTM, and SFT+CTM+Stay on Ground-CF in Table 15. CTM provides its clearest gains where standard supervision is weaker. In particular, on the two models with paired-SFT controls, CTM improves held-out visual control beyond exposure to the same counterfactual examples alone: by $+.065/+.059$ FLIP/STAY for Gemma-3-4B and $+.031/+.115$ for Qwen-3.5-2B. On Gemma-3-4B, the visual FLIP gain also appears in ID evaluation.

On the other hand, adding explicit STAY supervision does not consistently improve control. Although it helps some visual settings, it can substantially reduce text control. For example, Qwen-3.5-4B text FLIP falls from $1.000$ to $.512$ in ID evaluation and from $.994$ to $.610$ on held-out values. We therefore use CTM without the additional Stay term as the reference objective.

## E Real-World News-Frame Evaluation

MultiVENT-G uses separate checkpoints trained directly in this setting; the controlled-benchmark checkpoints are not transferred here. Neologisms and prompt tuning use $L{=}5$, while LoRA uses one joint rank-16 adapter. We evaluate 173 visual-swap pairs from 22 videos and 300 description-swap pairs. None of the evaluation videos appears in the training set used for the learned methods.

Table 13: **Ground-CF objective comparison** for $L$=1 neologisms at checkpoint 800. G3, Q4, and Q2 denote Gemma-3-4B, Qwen-3.5-4B, and Qwen-3.5-2B. Paired SFT is available for Gemma-3-4B and Qwen-3.5-2B.

| Split | Model | Objective | $A_I$ | $F_I$ | $S_I$ | $A_T$ | $F_T$ | $S_T$ |
|---|---|---|---|---|---|---|---|---|
| ID | G3 | SFT | 0.783 | 0.605 | 0.738 | 0.959 | 0.914 | 0.936 |
| ID | G3 | Paired SFT | 0.734 | 0.571 | 0.686 | 0.971 | 0.947 | 0.941 |
| ID | G3 | SFT+CTM | 0.759 | 0.632 | 0.693 | 0.996 | 0.983 | 0.987 |
| ID | G3 | SFT+CTM+Stay | 0.779 | 0.621 | 0.723 | 0.995 | 0.989 | 0.989 |
| ID | Q4 | SFT | 0.755 | 0.666 | 0.687 | 1.000 | 1.000 | 1.000 |
| ID | Q4 | Paired SFT | – | – | – | – | – | – |
| ID | Q4 | SFT+CTM | 0.882 | 0.789 | 0.846 | 1.000 | 1.000 | 1.000 |
| ID | Q4 | SFT+CTM+Stay | 0.896 | 0.842 | 0.861 | 0.650 | 0.512 | 0.547 |
| ID | Q2 | SFT | 0.852 | 0.778 | 0.793 | 1.000 | 0.999 | 1.000 |
| ID | Q2 | Paired SFT | 0.850 | 0.777 | 0.782 | 0.995 | 0.988 | 0.992 |
| ID | Q2 | SFT+CTM | 0.826 | 0.754 | 0.779 | 1.000 | 1.000 | 1.000 |
| ID | Q2 | SFT+CTM+Stay | 0.869 | 0.786 | 0.805 | 1.000 | 1.000 | 1.000 |
| Held-out | G3 | SFT | 0.584 | 0.541 | 0.555 | 0.932 | 0.894 | 0.921 |
| Held-out | G3 | Paired SFT | 0.566 | 0.539 | 0.532 | 0.976 | 0.956 | 0.970 |
| Held-out | G3 | SFT+CTM | 0.650 | 0.604 | 0.591 | 0.996 | 0.995 | 0.996 |
| Held-out | G3 | SFT+CTM+Stay | 0.641 | 0.569 | 0.613 | 0.998 | 0.994 | 0.995 |
| Held-out | Q4 | SFT | 0.585 | 0.542 | 0.544 | 1.000 | 0.999 | 1.000 |
| Held-out | Q4 | Paired SFT | – | – | – | – | – | – |
| Held-out | Q4 | SFT+CTM | 0.660 | 0.600 | 0.621 | 1.000 | 0.994 | 1.000 |
| Held-out | Q4 | SFT+CTM+Stay | 0.704 | 0.677 | 0.686 | 0.710 | 0.610 | 0.634 |
| Held-out | Q2 | SFT | 0.659 | 0.645 | 0.625 | 1.000 | 1.000 | 1.000 |
| Held-out | Q2 | Paired SFT | 0.649 | 0.588 | 0.608 | 0.988 | 0.980 | 0.981 |
| Held-out | Q2 | SFT+CTM | 0.755 | 0.619 | 0.723 | 0.998 | 0.995 | 0.995 |
| Held-out | Q2 | SFT+CTM+Stay | 0.690 | 0.596 | 0.662 | 1.000 | 1.000 | 1.000 |

Table 14: **Change from ID to held-out evaluation.** We report macro-average changes with item-bootstrap 95% intervals.

| Request | Metric | Method | Held-out − ID | 95% interval |
|---|---|---|---|---|
| look | flip | Instruction | +0.034 | [+0.007, +0.062] |
| look | flip | Prompt tuning | -0.000 | [-0.026, +0.025] |
| look | flip | Neologism | -0.007 | [-0.033, +0.020] |
| look | flip | LoRA | -0.094 | [-0.118, -0.069] |
| look | stay | Instruction | +0.031 | [+0.003, +0.060] |
| look | stay | Prompt tuning | +0.000 | [-0.025, +0.024] |
| look | stay | Neologism | -0.015 | [-0.042, +0.010] |
| look | stay | LoRA | -0.086 | [-0.108, -0.064] |
| read | flip | Instruction | -0.022 | [-0.043, -0.000] |
| read | flip | Prompt tuning | +0.003 | [-0.007, +0.013] |
| read | flip | Neologism | -0.015 | [-0.031, -0.000] |
| read | flip | LoRA | -0.002 | [-0.006, +0.001] |
| read | stay | Instruction | -0.015 | [-0.036, +0.007] |
| read | stay | Prompt tuning | +0.003 | [-0.005, +0.012] |
| read | stay | Neologism | -0.018 | [-0.033, -0.003] |
| read | stay | LoRA | -0.001 | [-0.004, +0.001] |

Table 15: **Ground-CF objective comparison** for $L$=1 neologisms at checkpoint 800. G3, Q4, and Q2 denote Gemma-3-4B, Qwen-3.5-4B, and Qwen-3.5-2B. Paired SFT is available for Gemma-3-4B and Qwen-3.5-2B.

| Split | Model | Objective | $A_I$ | $F_I$ | $S_I$ | $A_T$ | $F_T$ | $S_T$ |
|---|---|---|---|---|---|---|---|---|
| ID | G3 | SFT | 0.783 | 0.605 | 0.738 | 0.959 | 0.914 | 0.936 |
| ID | G3 | Paired SFT | 0.734 | 0.571 | 0.686 | 0.971 | 0.947 | 0.941 |
| ID | G3 | SFT+CTM | 0.759 | 0.632 | 0.693 | 0.996 | 0.983 | 0.987 |
| ID | G3 | SFT+CTM+Stay | 0.779 | 0.621 | 0.723 | 0.995 | 0.989 | 0.989 |
| ID | Q4 | SFT | 0.755 | 0.666 | 0.687 | 1.000 | 1.000 | 1.000 |
| ID | Q4 | SFT+CTM | 0.882 | 0.789 | 0.846 | 1.000 | 1.000 | 1.000 |
| ID | Q4 | SFT+CTM+Stay | 0.896 | 0.842 | 0.861 | 0.650 | 0.512 | 0.547 |
| ID | Q2 | SFT | 0.852 | 0.778 | 0.793 | 1.000 | 0.999 | 1.000 |
| ID | Q2 | Paired SFT | 0.850 | 0.777 | 0.782 | 0.995 | 0.988 | 0.992 |
| ID | Q2 | SFT+CTM | 0.826 | 0.754 | 0.779 | 1.000 | 1.000 | 1.000 |
| ID | Q2 | SFT+CTM+Stay | 0.869 | 0.786 | 0.805 | 1.000 | 1.000 | 1.000 |
| Held-out | G3 | SFT | 0.584 | 0.541 | 0.555 | 0.932 | 0.894 | 0.921 |
| Held-out | G3 | Paired SFT | 0.566 | 0.539 | 0.532 | 0.976 | 0.956 | 0.970 |
| Held-out | G3 | SFT+CTM | 0.650 | 0.604 | 0.591 | 0.996 | 0.995 | 0.996 |
| Held-out | G3 | SFT+CTM+Stay | 0.641 | 0.569 | 0.613 | 0.998 | 0.994 | 0.995 |
| Held-out | Q4 | SFT | 0.585 | 0.542 | 0.544 | 1.000 | 0.999 | 1.000 |
| Held-out | Q4 | SFT+CTM | 0.660 | 0.600 | 0.621 | 1.000 | 0.994 | 1.000 |
| Held-out | Q4 | SFT+CTM+Stay | 0.704 | 0.677 | 0.686 | 0.710 | 0.610 | 0.634 |
| Held-out | Q2 | SFT | 0.659 | 0.645 | 0.625 | 1.000 | 1.000 | 1.000 |
| Held-out | Q2 | Paired SFT | 0.649 | 0.588 | 0.608 | 0.988 | 0.980 | 0.981 |
| Held-out | Q2 | SFT+CTM | 0.755 | 0.619 | 0.723 | 0.998 | 0.995 | 0.995 |
| Held-out | Q2 | SFT+CTM+Stay | 0.690 | 0.596 | 0.662 | 1.000 | 1.000 | 1.000 |

For each pair, we score the same two candidates before and after the intervention. Since FLIP requires both predictions to be correct, two independent uniform guesses give an expected score of .25. This provides an independent-guessing reference rather than a general chance threshold.

**Evidence crops.** Since MultiVENT-G provides human-annotated visual evidence, we also evaluate a crop centered on the annotated evidence region. The crop is enlarged by $25\%$ to retain surrounding visual context, and no crop-specific training is used. This allows us to test what happens when the relevant visual evidence is easier to access; it does not test whether the model can find that region itself.

Table 16 reports Accuracy, FLIP, and answer stability $U$. Accuracy averages correctness over both versions of the edited-source pair. $U$ measures whether the answer remains unchanged when the competing source changes; unlike STAY, it is not conditioned on correctness. At full frame, visual Accuracy remains between .50 and .66 across all methods and models, while FLIP is often substantially lower. Thus, the main difficulty is not only answering an individual frame correctly, but reliably changing the answer when the relevant visual evidence changes.

## F RESULT VERIFICATION

All reported tables are generated from the canonical item-level result files. We verify that candidate sets remain fixed across paired contexts, that each intervention preserves the intended conflict, and that the reported aggregate scores reproduce the item-level predictions.

Table 16: **Complete MultiVENT-G results.** We report visual and text Accuracy ($A$), FLIP ($F$), and answer stability ($U$) for the full frame and evidence crop. $U$ is answer agreement rather than correctness-conditioned STAY. Visual pairs have $n$=173 and text pairs $n$=300.

| | | Full frame | | | | | | Evidence crop | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Method | $A_I$ | $F_I$ | $U_I$ | $A_T$ | $F_T$ | $U_T$ | $A_I$ | $F_I$ | $U_I$ | $A_T$ | $F_T$ | $U_T$ |
| Gemma-3-4B | No request | 0.569 | 0.214 | 0.343 | 0.824 | 0.657 | 0.728 | 0.601 | 0.277 | 0.353 | 0.824 | 0.657 | 0.642 |
| Gemma-3-4B | Instruction | 0.573 | 0.191 | 0.330 | 0.860 | 0.727 | 0.751 | 0.593 | 0.266 | 0.327 | 0.862 | 0.727 | 0.613 |
| Gemma-3-4B | Prompt tuning | 0.575 | 0.173 | 0.427 | 0.935 | 0.870 | 0.873 | 0.615 | 0.249 | 0.427 | 0.935 | 0.870 | 0.902 |
| Gemma-3-4B | Neologism | 0.537 | 0.104 | 0.507 | 0.931 | 0.863 | 0.919 | 0.602 | 0.220 | 0.483 | 0.939 | 0.877 | 0.873 |
| Gemma-3-4B | LoRA | 0.575 | 0.191 | 0.430 | 0.992 | 0.983 | 0.827 | 0.702 | 0.416 | 0.450 | 0.994 | 0.987 | 0.769 |
| Qwen-3.5-4B | No request | 0.561 | 0.145 | 0.307 | 0.847 | 0.693 | 0.844 | 0.593 | 0.202 | 0.280 | 0.860 | 0.720 | 0.792 |
| Qwen-3.5-4B | Instruction | 0.578 | 0.162 | 0.303 | 0.910 | 0.820 | 0.855 | 0.679 | 0.364 | 0.297 | 0.916 | 0.833 | 0.751 |
| Qwen-3.5-4B | Prompt tuning | 0.578 | 0.156 | 0.383 | 0.903 | 0.807 | 0.751 | 0.670 | 0.347 | 0.367 | 0.898 | 0.797 | 0.642 |
| Qwen-3.5-4B | Neologism | 0.595 | 0.202 | 0.480 | 0.952 | 0.903 | 0.821 | 0.742 | 0.486 | 0.500 | 0.950 | 0.900 | 0.763 |
| Qwen-3.5-4B | LoRA | 0.659 | 0.376 | 0.567 | 0.997 | 0.993 | 0.821 | 0.792 | 0.595 | 0.567 | 0.995 | 0.990 | 0.740 |
| Qwen-3.5-2B | No request | 0.529 | 0.087 | 0.317 | 0.835 | 0.670 | 0.890 | 0.564 | 0.127 | 0.320 | 0.840 | 0.680 | 0.815 |
| Qwen-3.5-2B | Instruction | 0.549 | 0.110 | 0.323 | 0.880 | 0.760 | 0.896 | 0.584 | 0.173 | 0.317 | 0.880 | 0.760 | 0.838 |
| Qwen-3.5-2B | Prompt tuning | 0.560 | 0.133 | 0.397 | 0.932 | 0.863 | 0.867 | 0.601 | 0.214 | 0.387 | 0.928 | 0.857 | 0.809 |
| Qwen-3.5-2B | Neologism | 0.567 | 0.145 | 0.367 | 0.925 | 0.850 | 0.896 | 0.641 | 0.295 | 0.367 | 0.921 | 0.843 | 0.757 |
| Qwen-3.5-2B | LoRA | 0.659 | 0.353 | 0.543 | 0.998 | 0.997 | 0.896 | 0.763 | 0.538 | 0.543 | 0.998 | 0.997 | 0.827 |