WHAT FORCED-CHOICE HIDES IN VISION-LANGUAGE MODEL MODALITY CONFLICT

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) often favor language-consistent answers over visual evidence under forced-choice image-text conflict. Such evaluation cannot distinguish preference for the language-referenced answer from behavior induced by the absence of an explicit mismatch response. We investigate this distinction using temporal-scale estimation on 500 natural images spanning five ordered temporal bands, comparing a 5-way forced-choice prompt with a 6-way prompt that adds a mismatch option. Five open-weight VLMs select the mismatch option on 56–91% of contradiction items but at most 4% of standard items; without it, they favor language-consistent temporal bands. Paired item-level analysis shows that 50–90% of language-consistent forced-choice responses shift to mismatch even at the smallest contradiction magnitude. Mismatch signaling generally increases with contradiction magnitude, while residual temporal-band responses remain language-aligned. Under the 5-way condition, linear probes recover the vision-consistent band throughout the network, while the model's own unembedding favors the language-consistent answer in later layers. Query-token activation patching strongly shifts outputs toward the vision-consistent answer. This establishes the causal influence of the query representation under the intervention. Removing the image reduces standard-condition accuracy by only 0–18 percentage points. These results show that forced-choice evaluation can obscure the distinction between mismatch signaling and language-aligned temporal selection.

1 INTRODUCTION

Temporal reasoning is an active area of vision-language model (VLM) evaluation (Hasan & Hoang, 2026). Existing benchmarks test temporal ordering, duration estimation, causal change, and temporal direction from images or videos (Wang et al., 2025b; Imam et al., 2025; Shangguan et al., 2025; Weng et al., 2026). Correct answers on these tasks do not distinguish whether the response is grounded in visual evidence or in semantic knowledge associated with the named process.

Temporal-scale estimation provides a setting in which these sources can be placed in direct conflict. Natural processes have characteristic durations, and their names are often associated with strong temporal priors. For example, water boiling with minutes and mountain erosion with millennia. When the image depicts the named process, visual evidence and process knowledge support the same temporal band. Their contributions are difficult to separate from accuracy alone. When the question instead names an unrelated process from a different temporal band, the two sources support different answers and their influence on the model's temporal-band selection can be compared directly.

Recent mechanistic analyses provide further evidence that visual information represented inside a VLM may not determine its final response. Visual attributes can remain decodable even when the final answer follows a linguistic prior (Nooralahzadeh et al., 2026). Visual and textual tasks can also rely on different processing circuits, with visual representations becoming compatible with downstream language computation late in the network (Nikankin et al., 2025). VLMs may further use linguistic labels as an interface for visual comparison (Shahgir et al., 2026). These findings motivate separating information that remains available in hidden states from information that governs the final prediction.

Two aspects of image-text conflict remain unresolved. First, prior mechanistic conflict paradigms generally do not provide an explicit response for cases in which the process named in the question

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.



Standard	Contradiction
Process: Water boiling **Vision band: A** (Seconds–Minutes)	**Process:** Mountain erosion **Vision band: A** (Seconds–Minutes) **Language band: E** (Decades–Millennia)

Question. Look carefully at the image. Approximately how long has the following process been ongoing to reach its current visible state: {*process*}?

5-way prompt: answer options A–E map to (A) Seconds–Minutes, (B) Hours–Days, (C) Weeks–Months, (D) Years–Decades, (E) Decades–Millennia.

6-way prompt: the same five options plus (F) "the described process is not visible in the image."

Figure 1: The contradiction paradigm. The standard question names the depicted process, while the contradiction question substitutes a process from a different temporal band. Each is evaluated under a 5-way forced-choice prompt and a 6-way prompt with an explicit mismatch option (F).

is absent from the image. A language-consistent answer can reflect either selection of the language-referenced temporal band or behavior induced by the absence of a mismatch response. Second, prior tasks do not impose an ordinal structure on conflict magnitude, leaving unclear how model behavior changes as the discrepancy between visual and linguistic evidence increases.

We address both questions using temporal-scale estimation over 500 natural images spanning five ordered bands from seconds–minutes to decades–millennia. Each image is paired with a standard question whose named process matches the image and a contradiction question whose process belongs to a different temporal band. The separation between the vision-consistent and language-consistent bands defines contradiction distance from one to four (Figure 1). We evaluate a 5-way prompt containing the five temporal bands and a 6-way prompt with an additional mismatch option, F: "the described process is not visible in the image." We also introduce a new metric called Language Reliance Score (LRS), which measures language versus vision alignment among responses matching one of the two competing temporal bands.

The behavioral results separate forced temporal-band selection from explicit mismatch signaling. Across the five open-weight VLMs, standard-condition accuracy ranges from 0.69 to 0.80, while option F is selected on at most 4% of standard items. Under the 5-way prompt, LRS ranges from 0.52 to 0.87, showing substantial language alignment when the model is required to choose a temporal band. Once an explicit mismatch response is available, models select F on 56–91% of contradiction items. Among the remaining temporal-band responses, language alignment remains substantial and, for some models, increases with contradiction distance. A text-only baseline further shows that removing the image reduces standard-condition accuracy by only 0 to 18 percentage points across the five models.

We also examine how this behavioral pattern relates to internal representations and output selection. Linear probes recover the vision-consistent temporal band from query-token hidden states with 80–88% accuracy across much of the network under both standard and contradiction inputs, whereas recovery of the language-consistent band under contradiction is substantially weaker. The model's own unembedding follows a different trajectory: language-answer decoding rises in middle-to-late layers, while vision-answer decoding remains low for the more language-aligned models and increases late in some others. Query-token activation patching from the standard trajectory into the contradiction trajectory strongly shifts outputs toward the vision-consistent answer across all five models. Visual temporal information remains recoverable from the hidden state even when the model's later readout favors the process named in language.

Our contributions are:

- **I.** We introduce an image-text contradiction task for temporal-scale estimation with an explicit mismatch response and an ordinal measure of contradiction magnitude.
- **II.** We show that forced-choice evaluation obscures two distinct response patterns under contradiction: mismatch signaling and temporal-band selection. Paired item-level analysis shows that many forced-choice responses shift to the mismatch option when it becomes available.

III. We identify a dissociation between recoverable visual information and output selection: the vision-consistent temporal band remains linearly decodable from hidden states while the model's later-layer readout favors the language-consistent answer.

IV. We show that query-token activation patching strongly shifts outputs toward the vision-consistent answer across all five VLMs, establishing causal influence of the query-token representation under the intervention.

2 Related Work

Temporal Reasoning in Vision-Language Models. Recent benchmarks evaluate several forms of temporal reasoning from visual inputs. TimeCausality studies irreversible processes such as aging and decay from image pairs (Wang et al., 2025b). VQA-Causal and VCR-Causal separate causal reasoning from visual recognition (Weng et al., 2026). TemporalVQA evaluates event ordering and duration estimation (Imam et al., 2025), and TOMATO focuses on temporal direction in video models (Shangguan et al., 2025). These benchmarks measure temporal judgments without separating visual evidence from semantic knowledge associated with the named process. We examine this distinction through temporal-scale estimation, where process names are often associated with strong duration priors.

Language Bias and Visual Grounding. VQA models can exploit regularities in questions while making limited use of images (Liang et al., 2020; Si et al., 2022; Wu et al., 2022). Research on visual grounding has also found heavy dependence on textual information during multimodal inference (Reich et al., 2023; Sim et al., 2025). Under image-text conflict, a language-consistent answer can reflect either preference for the language-referenced interpretation or behavior induced by the absence of an explicit mismatch response. We separate these cases by adding such a response option.

Modality Conflict and Its Mechanisms. Several recent papers examine how visual and linguistic information are represented and used inside VLMs. Deng et al. (2025) and Pezeshkpour et al. (2025) show that modality preference under conflicting evidence varies across tasks and model families. Nooralahzadeh et al. (2026) find that visual attributes remain decodable even when the final output follows a linguistic prior, and that patching visual-token representations can causally shift the output. Nikankin et al. (2025) report largely separate visual and textual processing components at the data-processing stage, followed by greater functional similarity during later query and answer computation. Shahgir et al. (2026) uncover stronger visual correspondence when entities have familiar linguistic labels, while hidden representations remain informative even when such labels are unavailable. These papers point to a distinction between visual information being present in hidden states and that information governing the final response. Our setting adds an explicit mismatch response and an ordered measure of contradiction magnitude. The former separates mismatch signaling from temporal-band selection; the latter tests how conflict behavior changes with the distance between vision- and language-consistent temporal bands.

Mechanistic Interpretability Methods. Our analysis draws on established tools for examining hidden representations and internal computation. Linear probing measures information recoverable from hidden states (Alain & Bengio, 2018; Hewitt & Manning, 2019; Belinkov & Glass, 2019). The logit lens traces intermediate vocabulary predictions through the model's unembedding (nostalgebraist, 2020; Elhage et al., 2021), while the tuned lens learns layer-specific affine transformations to account for changes in intermediate representations (Belrose et al., 2025). Activation patching tests how replacing an activation in one forward pass with the corresponding activation from another changes the output (Meng et al., 2022; Wang et al., 2023; Zhang & Nanda, 2024). We use these methods to compare recoverable visual information with the model's output trajectory and test the causal influence of the query-token representation under contradiction.

3 Methodology

3.1 Task, Dataset, and Two-Condition Contradiction Paradigm

Given an image I depicting a temporal process p_v, the task is to estimate how long the process has been ongoing to reach its current visible state. Answer choices correspond to five ordinal temporal bands, presented in scale order and labeled A through E: *Seconds–Minutes* (A), *Hours–Days* (B), *Weeks–Months* (C), *Years–Decades* (D), and *Decades–Millennia* (E). The benchmark contains 500

natural images, 100 per band, collected from Wikimedia Commons, Flickr, and Unsplash. Dataset construction, annotation procedures, and image-source information are provided in Appendix A.1.

Each image has four prompt variants formed by crossing two factors: the process reference (p_v or a substituted p_l) and the answer space (5-way or 6-way). The *standard* question names the depicted process p_v. The *contradiction* question substitutes an unrelated process p_l from a different temporal band. The 5-way prompt lists the five bands as options A–E; the 6-way prompt appends option F, "the described process is not visible in the image," as a sixth choice. We also evaluate a text-only version of the standard 5-way question for all five models to estimate how much performance depends on the image. In the text-only variant, the process description and answer choices remain unchanged and only the image is removed. Exact question templates are given in Appendix A.2.

A contradiction item has a vision answer y_v defined by the depicted process and a language answer y_l defined by the substituted process. Their absolute band-distance is $d = |b(y_v) - b(y_l)| \in \{1, 2, 3, 4\}$, where $b(\cdot)$ is the ordinal band index. Band-distance measures the separation between the visual and language-consistent answers on the temporal scale. Our sampling assigns one contradiction per image, which yields uneven counts across band-distances: 208, 149, 95, and 48 items at distances 1 through 4. Distance 4 is reachable only when the depicted process's band is at one end of the scale (Seconds–Minutes or Decades–Millennia); distance-4 results are drawn only from endpoint-band items.

3.2 Behavioral Metrics

For standard and text-only questions, we report temporal-band accuracy. Under contradiction, *Vision Accuracy* and *Language Accuracy* measure agreement with y_v and y_l. For the 6-way prompt, *F Rate* measures selection of the mismatch option.

We measure modality alignment among responses matching either reference band with the *Language Reliance Score* (LRS), which is the proportion of responses matching the language-consistent band among cases in which the model selects either the vision-consistent or the language-consistent band. We define it as

$$\mathrm{LRS} = \frac{N_{\text{language}}}{N_{\text{language}} + N_{\text{vision}}} \tag{1}$$

An LRS of 0.5 corresponds to equal counts of vision- and language-consistent responses. Values above 0.5 correspond to greater language alignment and values below 0.5 to greater vision alignment. F selections and responses matching neither reference band are excluded from the denominator.

We stratify contradiction results by band-distance. We also pair each item across the 5-way and 6-way prompts and measure $V \to F$, $L \to F$, $V \to V$, and $L \to L$ transitions, where V denotes a vision-consistent response, L a language-consistent response, and F selection of the mismatch option. Confidence intervals use 1,000 image-level bootstrap resamples, with prompt variants from the same image kept paired.

3.3 Model Selection

We evaluate five open-weight VLMs spanning four language-model families: Qwen3 (Yang et al., 2025), Llama 3 (Grattafiori et al., 2024), Mistral (Jiang et al., 2023), and Phi-4 (Abdin et al., 2024). Qwen3-VL-8B-Instruct (Bai et al., 2025) serves as the **anchor model** for the additional tuned-lens and modality-specific patching analyses. InternVL3.5-8B (Wang et al., 2025a) uses a Qwen3-family language backbone with different visual and multimodal components. We do not treat the comparison between the two models as holding language-model weights fixed. LLaVA-NeXT-8B (Liu et al., 2024) uses Meta-Llama-3-8B-Instruct. Pixtral-12B (Agrawal et al., 2024) uses a Mistral-family decoder, while Phi-4-reasoning-vision-15B (Aneja et al., 2026) belongs to the Phi-4 family. Exact checkpoints, layer counts, numerical precision, decoding settings, software versions, and compute infrastructure are reported in Appendix A.3.

3.4 Mechanistic Protocol

We conduct all mechanistic analyses on the 5-way prompt, where the vision-consistent and language-consistent bands are both available as explicit answer choices. We apply four analyses: linear probing, the raw logit lens, a tuned-lens ablation on the anchor model, and query-token activation patching. Linear probing, raw logit lens, and query-token activation patching are run on

all five models; the tuned-lens ablation and activation-patching specificity controls are run on the anchor model, Qwen3-VL. We analyze the hidden state at the final sequence position immediately preceding generation of the answer letter, which we call the *query token*. This position is identified separately for each model using its native chat template. For transformer block ℓ, we extract the post-block residual state h_ℓ; cross-model plots use normalized LM depth $\ell/(L-1)$.

Linear probing. At each layer, we fit an ℓ_2-regularized multinomial logistic regression probe to h_ℓ using 5-fold cross-validation stratified by target class and grouped by image. Grouping prevents responses from the same image from appearing in both training and test folds. Under the standard condition, the target is the ground-truth temporal band. Under contradiction, we fit separate probes for y_v and y_l. We report held-out decoding accuracy, for which chance accuracy is 0.20 for balanced five-class targets. Probe accuracy measures linear recoverability and does not establish use of the decoded information during generation.

Raw logit lens. We apply the model's final normalization and language-model head to each h_ℓ and extract logits for answer tokens A–E (nostalgebraist, 2020; Elhage et al., 2021). Under the standard condition, we measure whether the highest-logit answer matches the ground-truth band. Under contradiction, we measure vision- and language-answer decoding rates and the logit margin. Positive margins favor the vision-consistent answer while negative margins favor the language-consistent answer. We also stratify the logit margin by band-distance to test whether the layer-wise preference changes with contradiction magnitude.

Tuned lens ablation. For the anchor model Qwen3-VL, we test whether raw logit-lens trajectories depend on intermediate representational calibration using the tuned lens (Belrose et al., 2025). A layer-specific affine translator maps h_ℓ before final normalization and unembedding. Model parameters remain frozen, and translators minimize KL divergence from the model's final next-token distribution. We train independent lenses on COCO captions (Lin et al., 2015) and MMMU (Yue et al., 2024), both disjoint from our temporal benchmark. For per-image analysis, we define *sustained language dominance* as a trajectory in which the language-answer logit exceeds the vision-answer logit from some layer onward through the final layer, and the *onset layer* as the smallest layer at which this ordering begins. Calibration sizes, optimization settings, and multimodal input formats are reported in Appendix A.4.

Query-token activation patching. For each image, we pair its standard and contradiction runs. At layer ℓ, we replace the contradiction query state $h_{\ell,q}^{\text{contradiction}}$ with the corresponding standard state $h_{\ell,q}^{\text{standard}}$ and continue the forward pass through the remaining layers. We report *vision-answer recovery* as the primary measure. The recovery set contains items whose unpatched standard run predicts y_v and whose contradiction run predicts y_l. Recovery at layer ℓ is the fraction for which query-token patching changes the prediction to y_v. A successful patch establishes causal influence of the patched query-token computation under the intervention. It does not establish the query token as the unique location of the relevant information. For the anchor model Qwen3-VL, we additionally patch image-token states and the process-text span (the tokens corresponding to the process-name substring in the question template) as specificity controls. When token counts differ across paired process spans, source states are linearly interpolated to the target span length before replacement.

4 Experiments and Results

We compare behavior across the two prompt formats and against the text-only baseline, then examine hidden-state representations and output routing under the 5-way contradiction condition.

4.1 Behavioral Evidence Across Prompt Conditions

Standard-condition 5-way accuracy ranges from 0.69 to 0.80 across models (Table 1). Adding option F leaves Qwen3-VL, InternVL3.5, and Phi-4 within two percentage points of their 5-way accuracy, while LLaVA-NeXT and Pixtral decrease from 0.74 to 0.57 and from 0.77 to 0.72, respectively. F remains rare on standard items, with rates at or below 0.04 for every model.

Under 5-way contradiction, LRS ranges from 0.52 for LLaVA-NeXT to 0.87 for Phi-4. With the 6-way prompt, F accounts for 0.56–0.91 of contradiction responses. Among responses matching either reference band, language preference remains similar. LRS changes from 0.77 to 0.74 for Qwen3-VL, 0.70 to 0.66 for InternVL3.5, 0.52 to 0.54 for LLaVA-NeXT, 0.87 to 0.87 for Phi-4,

and 0.64 to 0.70 for Pixtral. The explicit mismatch option redirects many contradiction responses without removing language preference among the residual A–E responses.

Table 1: Behavioral results under 5-way and 6-way prompts for standard and contradiction conditions. V/L = Vision/Language Accuracy, F = F Rate, and LRS = Language Reliance Score. Brackets show 95% bootstrap CIs from 1,000 image-level resamples.

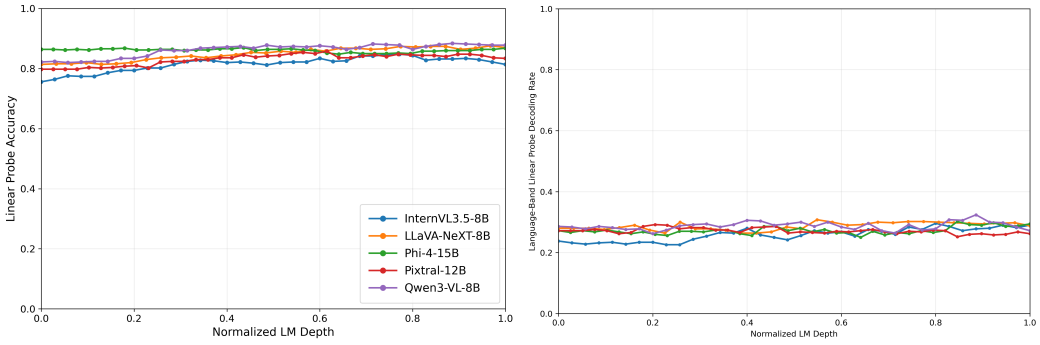
	Standard		**Contradiction (6-way)**			
Model	5-way	6-way	V	L	F	LRS
Qwen3-VL	0.76 [.72,.80]	0.76 [.72,.80]	0.02 [.01,.03]	0.05 [.03,.07]	0.91	0.74 [.59,.89]
InternVL3.5	0.69 [.65,.73]	0.71 [.66,.75]	0.12 [.09,.14]	0.23 [.19,.27]	0.56	0.66 [.58,.74]
LLaVA-NeXT	0.74 [.70,.78]	0.57 [.52,.61]	0.13 [.10,.16]	0.15 [.12,.18]	0.56	0.54 [.46,.63]
Phi-4	0.80 [.77,.84]	0.79 [.76,.83]	0.05 [.03,.07]	0.31 [.27,.35]	0.57	0.87 [.82,.92]
Pixtral	0.77 [.73,.81]	0.72 [.68,.76]	0.09 [.06,.11]	0.21 [.17,.24]	0.64	0.70 [.63,.77]

Band-distance behavior. The four contradiction distances contain 208, 149, 95, and 48 items. Under the 6-way prompt, F generally becomes more frequent with distance. Its rate from distance 1 to 4 changes from 0.45 to 0.81 for LLaVA-NeXT, 0.52 to 0.68 for InternVL3.5, and 0.52 to 0.70 for Phi-4. Qwen3-VL is already high at distance 1 (0.89) and reaches 0.98 at distances 3 and 4. Pixtral is non-monotonic (0.63, 0.65, 0.69, 0.60). Vision-consistent responses are rare at distances 2–4, with Vision Accuracy no higher than 0.07 for any model. Among A–E responses, Qwen3-VL reaches LRS 1.00 at distances 3 and 4 and Phi-4 reaches 1.00 from distance 2 onward. Full distance-stratified results appear in Appendix A.5.

Paired 5-way to 6-way transitions show which forced-choice responses are redirected to F (Table 2). Both vision- and language-consistent 5-way responses transition to F, but with different distance profiles. $L \to F$ is already high at distance 1, ranging from 0.50 to 0.90 across models, and increases only modestly with distance. A large fraction of language-consistent forced-choice responses shift to explicit mismatch signaling even at small contradiction magnitudes. By contrast, $V \to F$ is lower at distance 1 and increases sharply with distance: from 0.77 to 1.00 for Qwen3-VL, 0.32 to 0.80 for InternVL3.5, 0.34 to 0.87 for LLaVA-NeXT, and from 0.48 at distance 1 to 1.00 at distances 2 and 3 for Phi-4 (with no distance-4 estimate because no 5-way vision-consistent responses remain). This pattern indicates that vision-consistent forced-choice responses are more likely to remain vision-consistent at small contradictions, while at larger contradictions they increasingly shift to the explicit mismatch option.

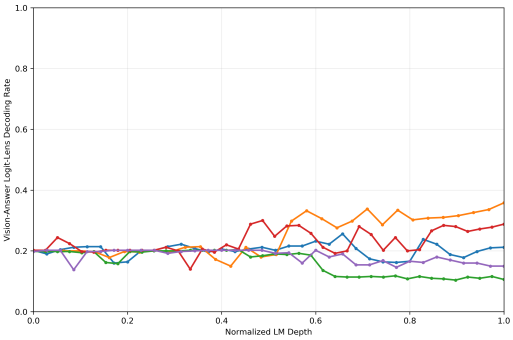
Table 2: Paired response transitions from 5-way to 6-way contradiction, stratified by band-distance d. $V/L \to F$ denotes transition from a vision- or language-consistent 5-way response to the mismatch option F under 6-way, while $V \to V$ and $L \to L$ denote persistence.

Model	d	$V \to F$	$V \to V$	$L \to F$	$L \to L$
Qwen3-VL	1	0.77	0.14	0.90	0.09
	2	0.94	0.06	0.84	0.14
	3	1.00	0.00	0.98	0.02
	4	1.00	0.00	0.98	0.02
InternVL3.5	1	0.32	0.66	0.56	0.44
	2	0.42	0.53	0.55	0.45
	3	0.69	0.31	0.67	0.33
	4	0.80	0.20	0.64	0.36
LLaVA-NeXT	1	0.34	0.47	0.51	0.36
	2	0.64	0.11	0.57	0.30
	3	0.77	0.13	0.72	0.16
	4	0.87	0.04	0.72	0.06
Phi-4	1	0.48	0.41	0.50	0.48
	2	1.00	0.00	0.59	0.39
	3	1.00	0.00	0.59	0.38
	4	—	—	0.68	0.30
Pixtral	1	0.47	0.48	0.63	0.35
	2	0.82	0.18	0.49	0.49
	3	0.68	0.32	0.62	0.36
	4	0.88	0.12	0.27	0.73

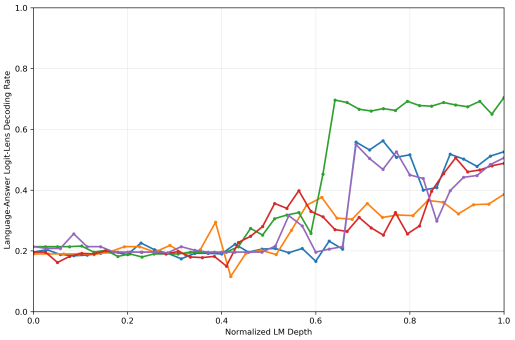


(a) Vision-band linear probe decoding.

(b) Language-band linear probe decoding.



(c) Contradiction vision-answer logit-lens decoding.



(d) Contradiction language-answer logit-lens decoding.

Figure 2: Information availability and output alignment across LM depth. (a) Vision-band and (b) language-band linear-probe decoding under 5-way contradiction; (c) vision-consistent and (d) language-consistent answer decoding under the raw logit lens. The legend in (a) gives the model colors used consistently across all panels.

Text-only baseline. Text-only accuracies are 0.68 for Qwen3-VL, 0.59 for InternVL3.5, 0.63 for LLaVA-NeXT, 0.80 for Phi-4, and 0.59 for Pixtral, compared with standard 5-way image-plus-text accuracies of 0.76, 0.69, 0.74, 0.80, and 0.77. The resulting gaps range from 0 to 18 percentage points. A large portion of standard-condition accuracy can be obtained from the process description alone without the image.

4.2 Information Is Represented But Not Reflected in Output

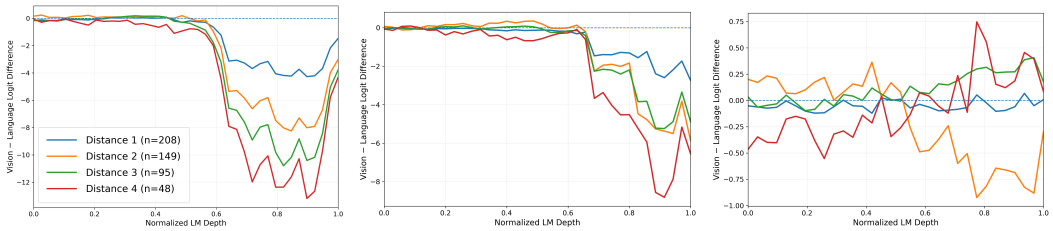
The behavioral evidence in Section 4.1 is compatible with two internal accounts. The visual temporal band could be absent from the model's hidden state under contradiction, or it could be present in the hidden state and not reflected in the answer-selection computation. We distinguish these using linear probing on hidden-state representations and the logit lens on the model's own output under the 5-way condition.

Linear probing. Under contradiction, vision-band decoding remains high across depth and closely tracks the standard condition (Figure 2A). Language-band decoding is much lower across models and layers (Figure 2B). This shows that the depicted temporal band remains linearly recoverable from the hidden state under contradiction even when the final response follows the process named in the question.

Raw logit lens. The model's own readout differs from the linear-probe result. Under the standard condition, top-1 logit-lens accuracy rises sharply in the middle-to-late layers and approaches the model's final behavioral accuracy (See Appendix A.6 for the results). Under contradiction, vision-answer decoding remains near chance for Phi-4, Qwen3-VL, and InternVL3.5, while LLaVA-NeXT and Pixtral show late increases to approximately 0.36 and 0.26, respectively (Figure 2C). Language-answer decoding rises in the middle-to-late layers for all five models and reaches approximately 0.71 for Phi-4 (Figure 2D). Visual temporal information remains recoverable at layers where the model's readout increasingly favors the language-consistent answer.

Tuned lens ablation. We repeat the analysis for the anchor model Qwen3-VL with tuned lenses calibrated independently on COCO captions and MMMU. Sustained language dominance occurs for 75.0% of images under either tuned lens and 73.6% under the raw lens. Median per-image onset layers are 26 for COCO, 24 for MMMU, and 26 for the raw lens. However, the aggregate crossover moves from layer 11 under the raw lens to layers 18 and 17 under the COCO and MMMU lenses. Calibration changes the estimated crossover layer while leaving the prevalence of sustained language dominance nearly unchanged. Complete tuned-lens results are reported in Appendix A.7.

Band-distance mechanism. The vision-minus-language logit difference also varies with contradiction distance (Figure 3). For Phi-4 and Qwen3-VL, the difference becomes increasingly negative in middle-to-late layers as distance grows. Phi-4 reaches approximately -13 at distance 4 compared with -4 at distance 1, while Qwen3-VL reaches approximately -9 and -3. InternVL3.5 and Pixtral follow the same direction with smaller magnitudes (Results are shown in Appendix A.6). LLaVA-NeXT departs from this pattern. Its middle-to-late-layer difference becomes positive at distances 3 and 4 and reaches approximately $+0.35$ at distance 4, consistent with its lower LRS.



(a) Phi-4 (b) Qwen3-VL (c) LLaVA-NeXT

Figure 3: Temporal band distance and the vision-minus-language logit difference under 5-way contradiction. The legend in (a) gives the distance colors and sample sizes used consistently across all panels. Negative values favor y_l, positive values favor y_v, and the dashed line denotes equal preference.

The information-vs-routing asymmetry recurs across the four language-model families in our set: probes recover the vision-consistent temporal band well above the empirical baseline while the model's own unembedding does not favor the vision-consistent answer under contradiction for the most language-aligned models. The asymmetry is not attributable to information being absent from the hidden state; it concerns how the model routes information into its output computation.

4.3 Causal Role of the Query Token

To test the causal role of the query token, we apply the activation patching protocol of Section 3.4. For each contradiction item and each layer ℓ, we replace the contradiction-run query-token hidden state with the standard-run hidden state and continue the forward pass from layer $\ell + 1$.

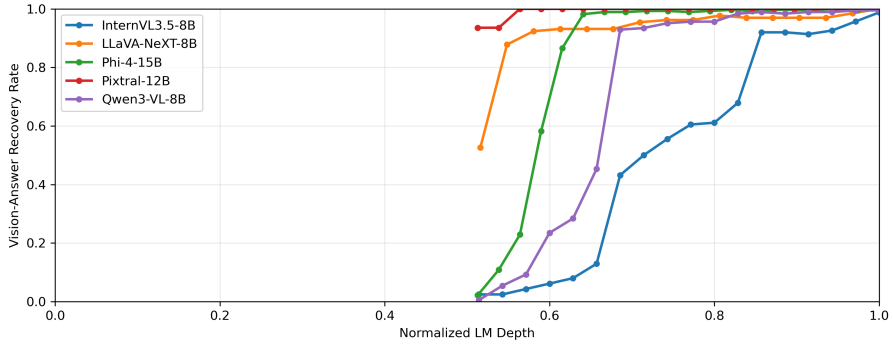


Figure 4: Query-token patching vision-answer recovery across normalized LM depth. Recovery is the fraction of language-consistent contradiction items for which patching restores the vision-consistent answer.

Vision-answer recovery. All five models show high recovery at a middle-to-late layer, although the transition occurs at different depths (Figure 4). Pixtral rises to about 0.94 at normalized depth 0.51 and reaches 1.0 by 0.56. LLaVA-NeXT increases from about 0.53 at 0.52 to 0.88 at 0.55, exceeding 0.9 by roughly 0.58. Phi-4 reaches about 0.98 by 0.64, while Qwen3-VL reaches about 0.93 at

0.68 before approaching ceiling. InternVL3.5 changes later than the other models, remaining below 0.15 through about 0.66 before rising to about 0.92 at 0.86 and near-ceiling values in the deepest layers. These interventions show that the query-token state has a causal influence on recovery of the vision-consistent answer, while not establishing it as the unique site of the computation.

Specificity controls. For Qwen3-VL, image-token patching yields a recovery rate of 0.0055 and process-description-token patching 0.10, compared with about 0.99 at the best query-token layer. The query-token intervention, thus, has a substantially larger effect than either control.

Cross-model observations. The depth at which strong query-token recovery emerges differs substantially across models. Pixtral and LLaVA-NeXT transition earliest, followed by Phi-4 and Qwen3-VL, while InternVL3.5 transitions considerably later. Despite these differences in depth, all five models show a strong query-token patching effect, indicating that the phenomenon recurs across the four LM families represented in our evaluation.

5 Discussion

VLM behavior under image-text contradiction separates into mismatch signaling and temporal-band selection. When option F is available, models frequently indicate the described process is not present in the image. F selection generally increases with band-distance. When models instead select a temporal band, their responses often favor the process named in language. The paired 5-way and 6-way analysis establishes this distinction at the item level. Even at the smallest contradiction magnitude, 0.50 to 0.90 of language-consistent forced-choice responses shift to F when that option becomes available. Forced-choice evaluation assigns many responses to a temporal band that are expressed as mismatch under the expanded answer space. Vision-consistent forced-choice responses also shift to F at rates that increase with contradiction magnitude. These transitions show that forced-choice behavior alone cannot distinguish modality preference from mismatch signaling.

The 5-way mechanistic evidence characterizes temporal-band selection under contradiction. Linear probes recover the vision-consistent band from the query-token hidden state throughout the language model, while the raw logit lens shows the model's readout increasingly favoring the language-consistent answer at middle-to-late layers; the tuned-lens analysis on the anchor model shows the same broad pattern. Query-token activation patching from the standard trajectory into the contradiction trajectory shifts outputs strongly toward the vision-consistent answer, with a much larger effect than image-token or process-text patching on the anchor model. Visual temporal information remains recoverable from the hidden state even when the later readout favors language, and the query-token representation has strong causal influence under the intervention. LLaVA-NeXT, which has the lowest aggregate LRS, differs from the more language-aligned models in that its late-layer readout moves toward the vision-consistent answer at larger contradictions. The correspondence between behavioral and layer-wise readout patterns indicates systematic cross-model differences in the balance between visual and language-aligned selection, without establishing a single shared computation that accounts for them.

The text-only baseline raises a further question about correct performance in the standard condition. Removing the image reduces accuracy by only 0 to 18 percentage points across the five models, and Phi-4 shows no measurable decrease. A substantial portion of standard-condition performance is compatible with retrieving a typical temporal band from the process name alone. Aggregate accuracy alone does not determine how much additional performance reflects visual temporal reasoning. Resolving this question requires tasks in which process-name priors are controlled independently of visual evidence.

Several constraints limit the scope of these conclusions. Band-distance is unbalanced by design, with 208, 149, 95, and 48 items at distances 1–4, and distance 4 occurs only for endpoint bands. The mechanistic protocol is applied under the 5-way prompt; extending its interpretation to residual A–E responses under the 6-way prompt assumes that the same routing process governs those responses. The present design also does not determine whether F selection depends on visual comparison or can arise partly from language-side incongruity detection; a text-only 6-way condition would directly test this possibility. Finally, the model set contains five open-weight VLMs at 8B to 15B parameters. Whether the same behavioral and mechanistic patterns extend to larger models, closed-weight systems, or non-temporal ordinal tasks remains open.

AI USE STATEMENT

We used generative AI tools for manuscript drafting and editing, including improvements to articulation, grammar, readability, organization, and title and abstract phrasing. We also used these tools to summarize and identify potentially relevant literature and to provide feedback on how the methodology, experimental design, and interpretation of results were presented in the manuscript.

The research questions, hypotheses, conceptual framing, dataset and benchmark design, experimental paradigm, implementation, statistical analyses, and scientific conclusions originated with and were carried out by the authors. Suggestions from generative AI concerning methodology, experiments, or interpretation were treated as external feedback. The authors independently evaluated these suggestions, checked them against the experimental results and primary literature, and made all final research decisions.

Generative AI was not used to generate the dataset, experimental results, reported statistical analyses, or scientific figures, and no AI-generated code was used in the experimental pipeline. Candidate references suggested by AI tools were verified against the original publications before inclusion.

All AI-assisted content was reviewed by the authors. The authors take full responsibility for the final manuscript, including its claims, analyses, citations, and artifacts.

ETHICS STATEMENT

This paper evaluates publicly available open-weight VLMs on a benchmark of 500 natural images collected from Wikimedia Commons, Flickr, and Unsplash. Source provenance and licensing information are retained in the dataset metadata, and image use follows the terms associated with each source item. The dataset release preserves each item's original source and license information and does not assign a single license to the underlying images, whose licensing conditions differ across repositories and individual items.

The benchmark does not contain human-subjects data collected for this research, medical data, or private personal information. Question construction and temporal-band assignments were carried out by the authors with manual review of all items. The benchmark is intended for research on VLM evaluation and mechanistic analysis. The benchmark's composition may reflect biases present in the source repositories and in the types of processes represented.

The paper reports model-specific failure patterns under controlled image-text contradiction. We describe these findings only at the level supported by the experiments and report limitations on their generality. Model selection was restricted to open-weight systems for which the internal representations required by the mechanistic analyses were accessible. The authors declare no conflicts of interest.

REPRODUCIBILITY STATEMENT

We provide the information needed to reproduce the behavioral and mechanistic analyses in the main paper and appendices. Section 3 describes the dataset construction, temporal-band taxonomy, contradiction paradigm, evaluation metrics, linear probing, logit-lens and tuned-lens analyses, and activation-patching protocol. Hyperparameters and implementation details are provided in Appendices A.3 and A.4, and exact model identifiers, versions, and inference configurations are listed in Appendix A.3. Full band-distance-stratified results and per-model layer-wise mechanistic analyses are included in Appendices A.5, A.6, and A.7. The benchmark metadata, paired prompt sets, and code will be released upon publication, subject to the licensing terms of the underlying image sources.

REFERENCES

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril

Zhang, and Yi Zhang. Phi-4 technical report, 2024. URL https://arxiv.org/abs/2412.08905.

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, Albert Q. Jiang, Kartik Khandelwal, Timothée Lacroix, Guillaume Lample, Diego Las Casas, Thibaut Lavril, Teven Le Scao, Andy Lo, William Marshall, Louis Martin, Arthur Mensch, Pavankumar Muddireddy, Valera Nemychnikova, Marie Pellat, Patrick Von Platen, Nikhil Raghuraman, Baptiste Rozière, Alexandre Sablayrolles, Lucile Saulnier, Romain Sauvestre, Wendy Shang, Roman Soletskyi, Lawrence Stewart, Pierre Stock, Joachim Studnia, Sandeep Subramanian, Sagar Vaze, Thomas Wang, and Sophia Yang. Pixtral 12b, 2024. URL https://arxiv.org/abs/2410.07073.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2018. URL https://arxiv.org/abs/1610.01644.

Jyoti Aneja, Michael Harrison, Neel Joshi, Tyler LaBonte, John Langford, and Eduardo Salinas. Phi-4-reasoning-vision-15b technical report, 2026. URL https://arxiv.org/abs/2603.03975.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Yonatan Belinkov and James Glass. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72, 2019. doi: 10.1162/tacl_a_00254. URL https://aclanthology.org/Q19-1004/.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens, 2025. URL https://arxiv.org/abs/2303.08112.

Ailin Deng, Tri Cao, Zhirui Chen, and Bryan Hooi. Words or vision: Do vision-language models have blind faith in text? In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3867–3876, 2025. doi: 10.1109/CVPR52734.2025.00366.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu,

Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent,

Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Mohammad Rashedul Hasan and Chinh Hoang. Position: What cézanne knew about visual intelligence that vision-language models miss. In *Forty-third International Conference on Machine Learning Position Paper Track*, 2026. URL https://openreview.net/forum?id=NmiiK9yQMM.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4129–4138, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1419. URL https://aclanthology.org/N19-1419/.

Mohamed Fazli Imam, Chenyang Lyu, and Alham Fikri Aji. Can multimodal llms do visual temporal understanding and reasoning? the answer is no!, 2025. URL https://arxiv.org/abs/2501.10674.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Zujie Liang, Weitao Jiang, Haifeng Hu, and Jiaying Zhu. Learning to contrast the counterfactual samples for robust visual question answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3285–3292, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.265. URL https://aclanthology.org/2020.emnlp-main.265/.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. URL https://arxiv.org/abs/1405.0312.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 35, 2022.

Yaniv Nikankin, Dana Arad, Yossi Gandelsman, and Yonatan Belinkov. Same task, different circuits: Disentangling modality-specific mechanisms in VLMs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=AD8ksC9bw1.

Farhad Nooralahzadeh, Omid Rohanian, Yi Zhang, Jonathan Fürst, and Kurt Stockinger. Arbitration failure, not perceptual blindness: How vision-language models resolve visual-linguistic conflicts, 2026. URL https://arxiv.org/abs/2604.09364.

nostalgebraist. Interpreting gpt: The logit lens. LessWrong, 2020. Accessed 2026.

Pouya Pezeshkpour, Moin Aminnaseri, and Estevam Hruschka. Mixed signals: Decoding vlms' reasoning and underlying bias in vision-language conflict, 2025. URL https://arxiv.org/abs/2504.08974.

Daniel Reich, Felix Putze, and Tanja Schultz. Measuring faithful and plausible visual grounding in VQA. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 3129–3144, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.206. URL https://aclanthology.org/2023.findings-emnlp.206/.

Haz Sameen Shahgir, Xiaofu Chen, Yu Fu, Erfan Shayegani, Nael Abu-Ghazaleh, Yova Kementchedjhieva, and Yue Dong. VLMs need words: Vision language models ignore visual detail in favor of semantic anchors. In *Third Conference on Language Modeling*, 2026. URL https://openreview.net/forum?id=6DJpzvjWbK.

Ziyao Shangguan, Chuhan Li, Yuxuan Ding, Yanan Zheng, Yilun Zhao, Tesca Fitzgerald, and Arman Cohan. TOMATO: Assessing visual temporal reasoning capabilities in multimodal foundation models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=fCi4o83Mfs.

Qingyi Si, Yuanxin Liu, Fandong Meng, Zheng Lin, Peng Fu, Yanan Cao, Weiping Wang, and Jie Zhou. Towards robust visual question answering: Making the most of biased samples via contrastive learning. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 6650–6662, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.495. URL https://aclanthology.org/2022.findings-emnlp.495/.

Mong Yuan Sim, Wei Emma Zhang, Xiang Dai, and Biaoyan Fang. Can VLMs actually see and read? a survey on modality collapse in vision-language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 24452–24470, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1256. URL https://aclanthology.org/2025.findings-acl.1256/.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao,

Wenhai Wang, and Gen Luo. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025a. URL https://arxiv.org/abs/2508.18265.

Zeqing Wang, Shiyuan Zhang, Chengpei Tang, and Keze Wang. Timecausality: Evaluating the causal ability in time dimension for vision language models, 2025b. URL https://arxiv.org/abs/2505.15435.

Zhaotian Weng, Haoxuan Li, Xin Eric Wang, Kuan-Hao Huang, and Jieyu Zhao. What's missing in vision-language models? probing their struggles with causal order reasoning. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5689–5701, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.266. URL https://aclanthology.org/2026.eacl-long.266/.

Yike Wu, Yu Zhao, Shiwan Zhao, Ying Zhang, Xiaojie Yuan, Guoqing Zhao, and Ning Jiang. Overcoming language priors in visual question answering via distinguishing superficially similar instances. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (eds.), *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 5721–5729, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.503/.

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.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi, 2024. URL https://arxiv.org/abs/2311.16502.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Hf17y6u9BC.

A Appendix

A.1 Dataset Construction Details

The benchmark consists of 500 natural images collected from Wikimedia Commons, Flickr, and Unsplash using open licenses (CC-BY, CC-BY-SA, CC BY-NC, CC0, or public domain). Table 3 shows the license distribution across these 500 images.

Each image is assigned to one of the five temporal bands described in Section 3. The bands are *Seconds–Minutes*, *Hours–Days*, *Weeks–Months*, *Years–Decades*, and *Decades–Millennia*. A single temporal process is identified for each image. The assigned temporal band corresponds to the approximate duration required for that process to reach its current visible state. The benchmark contains 100 images for each temporal band. Table 4 shows representative processes in each band.

Table 3: License distribution across the 500 images.

License	Image count
CC0 / Public domain	50
CC-BY	72
CC-BY-SA	324
CC BY-NC	50
Unsplash	4
Total	500

Table 4: Dataset stratification across the five temporal bands. Each band contains representative process types across physical, biological, chemical, and geological domains.

Temporal Band	Images	Representative Process Types
Seconds–Minutes	100	Water boiling, ice melting, water splashing, steam rising
Hours–Days	100	Flowers wilting, puddle drying, food spoilage, candle burning
Weeks–Months	100	Rust formation, vegetation growth, seasonal foliage, construction progress
Years–Decades	100	Tree growth, building weathering, moss growing, road cracking
Decades–Millennia	100	Erosion, ruins, stalactites, volcanoes

A.2 Complete Question Templates

For each image, we construct four prompt variants by crossing two factors: the process referenced in the question (*standard* or *contradiction*) and the answer space (*5-way* or *6-way*). In the standard condition, `process` refers to the process depicted in the image, p_v. In the contradiction condition, it is replaced by an unrelated process p_l sampled from a different temporal band. The 5-way condition contains the five temporal-band options (A–E), whereas the 6-way condition adds option F, "The described process is not visible in the image," to provide an explicit rejection response. The complete templates are shown below.

A.2.1 Standard, 5-way

> Look carefully at the image. Approximately how long has the following process been ongoing to reach its current visible state: *process*?
>
> A. Seconds to minutes
> B. Hours to days
> C. Weeks to months
> D. Years to decades
> E. Decades to millennia
>
> Output only your multiple-choice answer (A, B, C, D, or E) and nothing else.

A.2.2 Standard, 6-way

> Look carefully at the image. Approximately how long has the following process been ongoing to reach its current visible state: *process*?
>
> A. Seconds to minutes
> B. Hours to days
> C. Weeks to months
> D. Years to decades
> E. Decades to millennia
> F. The described process is not visible in the image
>
> Output only your multiple-choice answer (A, B, C, D, E, or F) and nothing else.

Table 5: Model implementation details.

Model	Checkpoint	LM Layers	Precision	Decoding
Qwen3-VL	`Qwen/Qwen3-VL-8B-Instruct`	36	BF16	Greedy
InternVL3.5	`OpenGVLab/InternVL3_5-8B`	36	BF16	Greedy
LLaVA-NeXT	`llava-hf/llama3-llava-next-8b-hf`	32	FP16	Greedy
Phi-4	`microsoft/Phi-4-reasoning-vision-15B`	40	FP16	Greedy
Pixtral	`mistral-experimental/pixtral-12b`	40	BF16	Greedy

A.2.3 CONTRADICTION, 5-WAY

> Look carefully at the image. Approximately how long has the following process been ongoing to reach its current visible state: *process*?
>
> A. Seconds to minutes
> B. Hours to days
> C. Weeks to months
> D. Years to decades
> E. Decades to millennia
>
> Output only your multiple-choice answer (A, B, C, D, or E) and nothing else.

A.2.4 CONTRADICTION, 6-WAY

> Look carefully at the image. Approximately how long has the following process been ongoing to reach its current visible state: *process*?
>
> A. Seconds to minutes
> B. Hours to days
> C. Weeks to months
> D. Years to decades
> E. Decades to millennia
> F. The described process is not visible in the image
>
> Output only your multiple-choice answer (A, B, C, D, E, or F) and nothing else.

A.2.5 TEXT-ONLY BASELINE

For the text-only baseline, we use the standard 5-way prompt but omit the image. The process description and answer options are otherwise unchanged.

A.3 IMPLEMENTATION DETAILS

Table 5 summarizes the checkpoints and model-specific configurations used in our experiments. All evaluations use greedy decoding. Layer indices in our mechanistic analyses refer to transformer layers in the language-model backbone, excluding the embedding layer.

Software. Experiments were conducted using CUDA 12.4.

Compute. Experiments were run on eight NVIDIA A40 GPUs, each with 48 GB of memory.

A.4 TUNED-LENS ABLATION DETAILS

We train two independent tuned lenses for Qwen3-VL, using COCO captions and MMMU as calibration sources. In both cases, the parameters of Qwen3-VL remain frozen and only the layer-specific affine translators are optimized. Qwen3-VL contains 36 language-model transformer layers. We train 35 translators corresponding to the outputs of transformer layers 0–34, while the output of

Table 6: Calibration and optimization settings for the COCO- and MMMU-tuned lenses. The two lenses are trained independently while keeping the underlying Qwen3-VL model frozen.

Setting	**COCO Lens**	**MMMU Lens**
Calibration examples	10,000	10,000
Source	COCO train2017	MMMU
Source splits	train2017	dev + val. + test
Example format	Image + caption	Image + MCQ
Sampling seed	42	42
Epochs	1	1
Optimizer	AdamW	AdamW
Learning rate	1×10^{-3}	1×10^{-3}
Weight decay	1×10^{-4}	1×10^{-4}
Gradient accumulation	8	8
Max. gradient norm	1.0	1.0
Precision	BF16	BF16
Translators	35	35

layer 35 serves as the unmodified final endpoint. Each translated representation is passed through the model's frozen final normalization layer and language-model head.

Calibration data. For the COCO calibration, we use 10,000 images from COCO train2017 set, with one caption selected per image. Eligible image–caption pairs are shuffled deterministically using a random seed of 42 before the 10,000-example calibration subset is selected. Each multimodal calibration sequence contains an image and the user instruction *"Describe this image."*, followed by the corresponding COCO caption as the assistant response. The tuned-lens objective is computed only at positions predicting the assistant caption. Image tokens and user-prompt tokens are excluded from the loss.

For the MMMU calibration, we collect eligible multiple-choice examples across the released dev, validation, and test splits and across MMMU subject configurations. We retain only examples containing at least two answer choices, a valid answer label, and at least one image. The resulting pool is shuffled deterministically with seed 42, from which 10,000 examples are selected. Each calibration sequence contains all images associated with the example, followed by the question and lettered answer choices. The prompt instructs the model to output only the letter corresponding to the correct option, and the ground-truth answer letter forms the assistant response. The loss is evaluated only at positions predicting this assistant response.

Table 6 summarizes the calibration and optimization settings for the two tuned lenses.

Optimization. For both calibration sources, the translators are trained for one epoch using AdamW with a learning rate of 10^{-3} and weight decay of 10^{-4}. Gradients are accumulated over eight successfully processed calibration examples before each optimizer update and are clipped to a maximum norm of 1.0. The model and translator parameters use bfloat16 precision. Both calibrations use minimum and maximum image pixel budgets of $256 \times 28 \times 28$ and $512 \times 28 \times 28$, respectively. Multimodal sequences are processed without sequence truncation.

Training objective. For each non-final transformer layer ℓ, we apply its learned affine translator to the corresponding intermediate hidden state. The translated representation is then decoded using Qwen3-VL's frozen final normalization layer and language-model head. Let $p_{\mathrm{final}}(\cdot \mid x)$ denote Qwen3-VL's final next-token distribution for input x, and let $p_\ell^{\mathrm{TL}}(\cdot \mid x)$ denote the distribution obtained by decoding the translated hidden state from layer ℓ. Each translator is optimized by minimizing

$$\mathcal{L}_\ell = D_{\mathrm{KL}}\left(p_{\mathrm{final}}(\cdot \mid x) \,\|\, p_\ell^{\mathrm{TL}}(\cdot \mid x)\right). \tag{2}$$

The tuned-lens translators are not trained to directly predict the ground-truth COCO caption or MMMU answer. Instead, they learn to map intermediate representations into distributions that approximate the frozen model's own final predictive distribution. For each calibration example, all

Table 7: Band-distance results under the 5-way prompt. Results are reported overall and separately for temporal band-distances 1–4 (Diff 1–4). The corresponding distance strata contain 208, 149, 95, and 48 items, respectively.

VLM	Vision Accuracy	Language Accuracy	Language Reliance Score (LRS)
Qwen3-VL	Overall: 0.15 Diff 1: 0.21 Diff 2: 0.10 Diff 3: 0.12 Diff 4: 0.11	Overall: 0.51 Diff 1: 0.47 Diff 2: 0.50 Diff 3: 0.42 Diff 4: 0.87	Overall: 0.77 Diff 1: 0.69 Diff 2: 0.83 Diff 3: 0.78 Diff 4: 0.89
InternVL3.5	Overall: 0.21 Diff 1: 0.29 Diff 2: 0.12 Diff 3: 0.14 Diff 4: 0.32	Overall: 0.51 Diff 1: 0.51 Diff 2: 0.52 Diff 3: 0.46 Diff 4: 0.60	Overall: 0.70 Diff 1: 0.64 Diff 2: 0.81 Diff 3: 0.77 Diff 4: 0.65
LLaVA-NeXT	Overall: 0.36 Diff 1: 0.44 Diff 2: 0.23 Diff 3: 0.32 Diff 4: 0.49	Overall: 0.39 Diff 1: 0.43 Diff 2: 0.38 Diff 3: 0.34 Diff 4: 0.36	Overall: 0.52 Diff 1: 0.49 Diff 2: 0.62 Diff 3: 0.52 Diff 4: 0.43
Phi-4	Overall: 0.10 Diff 1: 0.22 Diff 2: 0.02 Diff 3: 0.01 Diff 4: 0.00	Overall: 0.71 Diff 1: 0.63 Diff 2: 0.67 Diff 3: 0.82 Diff 4: 0.98	Overall: 0.87 Diff 1: 0.74 Diff 2: 0.97 Diff 3: 0.99 Diff 4: 1.00
Pixtral	Overall: 0.26 Diff 1: 0.29 Diff 2: 0.14 Diff 3: 0.23 Diff 4: 0.53	Overall: 0.47 Diff 1: 0.49 Diff 2: 0.46 Diff 3: 0.43 Diff 4: 0.47	Overall: 0.64 Diff 1: 0.63 Diff 2: 0.76 Diff 3: 0.65 Diff 4: 0.47

35 translators are trained against the same final Qwen3-VL distribution at the relevant assistant-response prediction positions.

The COCO- and MMMU-tuned lenses are initialized and trained independently. The MMMU lens is trained from a fresh tuned-lens initialization instead of being initialized from the COCO-tuned lens. After calibration, the resulting translator parameters are fixed for the corresponding temporal-scale evaluation.

A.5 Band-Distance Results

Tables 7 and 8 report the complete contradiction results stratified by temporal band-distance. The four distance strata contain 208, 149, 95, and 48 items for distances 1–4, respectively. We report Vision Accuracy, Language Accuracy, and Language Reliance Score (LRS) for both prompting conditions. For the 6-way condition, we additionally report the percentage of responses selecting the explicit mismatch option (F).

The 6-way results show that the explicit mismatch response becomes increasingly prominent at larger contradiction distances for several models, although the relationship is not uniformly monotonic across architectures. Comparison with the 5-way condition further shows that the distribution between vision- and language-consistent responses can change when an explicit mismatch response is available.

Table 8: Band-distance results under the 6-way prompt with explicit mismatch option (F). Results are reported overall and separately for temporal band-distances 1–4 (Diff 1–4). F Percentage is the fraction of responses selecting the explicit mismatch option.

VLM	Vision Accuracy	Language Accuracy	F Percentage	Language Reliance Score (LRS)
Qwen3-VL	Overall: 0.02 Diff 1: 0.04 Diff 2: 0.007 Diff 3: 0.00 Diff 4: 0.00	Overall: 0.05 Diff 1: 0.06 Diff 2: 0.07 Diff 3: 0.02 Diff 4: 0.02	Overall: 0.91 Diff 1: 0.89 Diff 2: 0.88 Diff 3: 0.98 Diff 4: 0.98	Overall: 0.74 Diff 1: 0.62 Diff 2: 0.91 Diff 3: 1.00 Diff 4: 1.00
InternVL3.5	Overall: 0.12 Diff 1: 0.19 Diff 2: 0.07 Diff 3: 0.05 Diff 4: 0.06	Overall: 0.23 Diff 1: 0.24 Diff 2: 0.25 Diff 3: 0.16 Diff 4: 0.23	Overall: 0.56 Diff 1: 0.52 Diff 2: 0.54 Diff 3: 0.65 Diff 4: 0.68	Overall: 0.66 Diff 1: 0.56 Diff 2: 0.79 Diff 3: 0.75 Diff 4: 0.79
LLaVA-NeXT	Overall: 0.13 Diff 1: 0.25 Diff 2: 0.03 Diff 3: 0.04 Diff 4: 0.02	Overall: 0.15 Diff 1: 0.23 Diff 2: 0.12 Diff 3: 0.07 Diff 4: 0.02	Overall: 0.56 Diff 1: 0.45 Diff 2: 0.53 Diff 3: 0.70 Diff 4: 0.81	Overall: 0.54 Diff 1: 0.48 Diff 2: 0.78 Diff 3: 0.64 Diff 4: 0.50
Phi-4	Overall: 0.05 Diff 1: 0.11 Diff 2: 0.00 Diff 3: 0.00 Diff 4: 0.00	Overall: 0.31 Diff 1: 0.33 Diff 2: 0.27 Diff 3: 0.33 Diff 4: 0.28	Overall: 0.57 Diff 1: 0.52 Diff 2: 0.57 Diff 3: 0.59 Diff 4: 0.70	Overall: 0.87 Diff 1: 0.75 Diff 2: 1.00 Diff 3: 1.00 Diff 4: 1.00
Pixtral	Overall: 0.09 Diff 1: 0.15 Diff 2: 0.02 Diff 3: 0.07 Diff 4: 0.06	Overall: 0.21 Diff 1: 0.18 Diff 2: 0.23 Diff 3: 0.16 Diff 4: 0.34	Overall: 0.64 Diff 1: 0.63 Diff 2: 0.65 Diff 3: 0.69 Diff 4: 0.60	Overall: 0.70 Diff 1: 0.54 Diff 2: 0.92 Diff 3: 0.68 Diff 4: 0.84

A.6 Additional Logit-Lens Results

This section reports additional logit-lens results under the 5-way probe. We first report the standard-condition layer-wise decoding results and then provide the temporal band-distance results for InternVL3.5 and Pixtral, which complements the models shown in Figure 3.

A.6.1 Standard-Condition Logit-Lens Results

Figure 5 shows the layer-wise top-1 logit-lens accuracy under the standard 5-way condition. Across models, accuracy is low in the earlier layers, rises in the middle-to-late layers, and approaches each model's final behavioral accuracy. These results complement the contradiction-condition analysis in the main paper and show that, when the image and process description agree, the model's own unembedding increasingly decodes the correct temporal band as computation progresses through the language model.

A.6.2 Additional Temporal Band-Distance Results

Figure 6 reports the temporal band-distance analysis for InternVL3.5 and Pixtral, which complements the Phi-4, Qwen3-VL, and LLaVA-NeXT results shown in Figure 3 of the main paper. Both models exhibit the same broad pattern observed for Phi-4 and Qwen3-VL, but with smaller amplitude. The vision-minus-language logit difference is near zero in earlier layers and becomes increasingly negative in middle-to-late layers, with larger temporal band distances generally producing a stronger language-favoring difference.

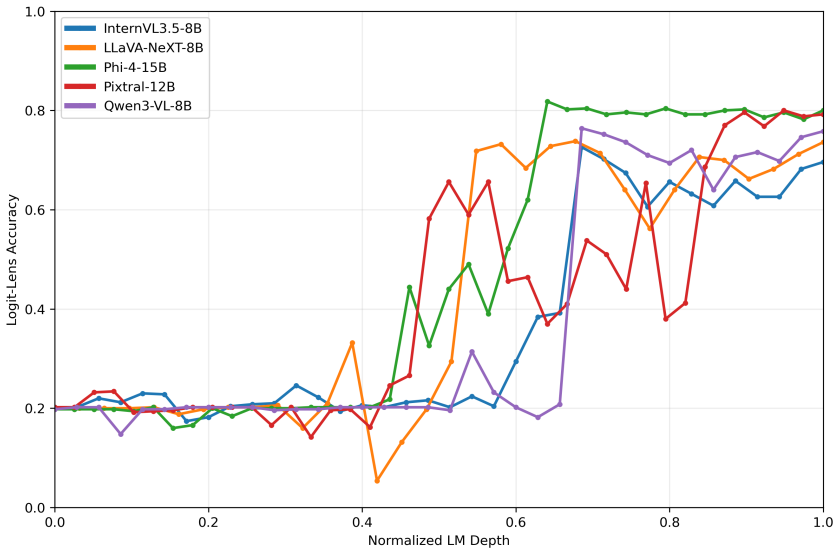
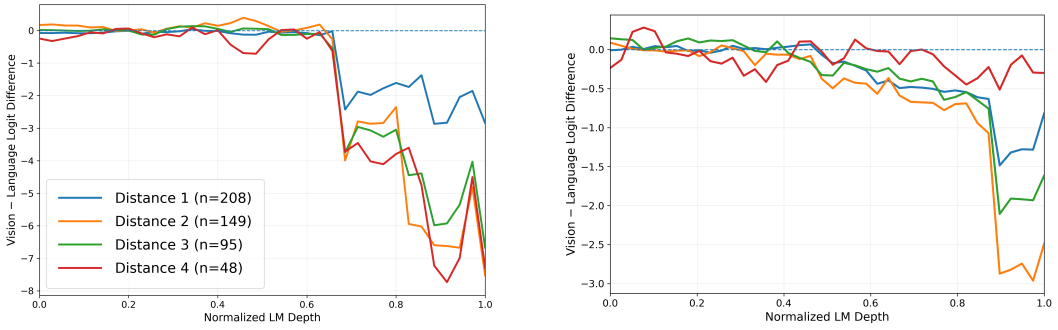


Figure 5: Standard-condition logit-lens decoding. Layer-wise top-1 accuracy of the model's own unembedding under the standard 5-way condition as a function of normalized LM depth. Logit-lens accuracy rises primarily in the middle-to-late layers and approaches final behavioral performance.



(a) InternVL3.5 (b) Pixtral

Figure 6: Additional temporal band-distance results. Layer-wise vision-minus-language logit difference under contradiction for InternVL3.5 and Pixtral, stratified by temporal band distance. The legend in (a) gives the distance colors and sample sizes used consistently across all panels. Negative values favor the language-consistent answer (y_l), positive values favor the vision-consistent answer (y_v), and the dashed horizontal line at zero indicates equal preference between the vision- and language-consistent answers. Both models show the same broad distance-dependent language-favoring pattern as Phi-4 and Qwen3-VL, but with smaller amplitude.

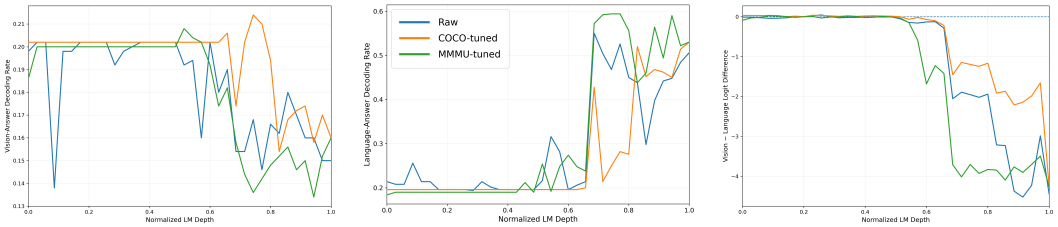
A.7 Tuned-Lens Detailed Results

We compare the raw logit lens with the independently trained COCO- and MMMU-tuned lenses on the same 500 contradiction items and 36 language-model layers of Qwen3-VL. For each lens, we measure the vision-answer decoding rate, language-answer decoding rate, and the logit difference between the vision- and language-consistent answers, $\Delta_{\mathrm{V-L}} = \mathrm{logit}(y_v) - \mathrm{logit}(y_l)$. Negative values of $\Delta_{\mathrm{V-L}}$ indicate that the model's readout favors the language-consistent answer. Layer-wise means and 95% confidence intervals are computed using 1,000 bootstrap resamples.

Sustained language dominance. We characterize the onset of sustained language dominance in two complementary ways. At the individual-image level, the onset is the first layer after which $\Delta_{\mathrm{V-L}} < 0$ for that image at every remaining layer. We report onset statistics only among images

Table 9: Sustained language dominance under raw and tuned lenses. Per-image onset is the first layer after which the vision-minus-language logit difference remains negative for that image through the final layer. Median and mean onset statistics are computed among images with a defined sustained crossover. Brackets give 95% bootstrap confidence intervals. Aggregate crossover is defined analogously using the mean vision-minus-language logit difference across all 500 images. Normalized aggregate depth is computed as layer index divided by 35.

Lens	Sustained	Fraction	Median onset	Mean onset	Aggregate crossover
Raw	368/500	0.736	26 [24, 29]	23.20 [22.29, 24.11]	11
COCO-tuned	375/500	0.750	26 [24, 29]	20.22 [18.84, 21.61]	18
MMMU-tuned	375/500	0.750	24 [24, 24]	20.80 [19.92, 21.68]	17



(a) Vision-answer decoding (b) Language-answer decoding (c) Vision–language logit difference

Figure 7: Raw versus tuned-lens trajectories for Qwen3-VL under 5-way contradiction. The raw logit lens is compared with tuned lenses calibrated independently on COCO and MMMU. (a) Layer-wise vision-answer decoding rate. (b) Layer-wise language-answer decoding rate. (c) Mean vision-minus-language logit difference, $\mathrm{logit}(y_v) - \mathrm{logit}(y_l)$. The legend in (b) gives the lens colors used consistently across all panels. Negative values in (c) favor the language-consistent answer, positive values favor the vision-consistent answer, and the dashed horizontal line denotes equal logit preference.

for which such a sustained crossover occurs. At the aggregate level, the crossover is the first layer after which the mean $\Delta_{\mathrm{V-L}}$ across all 500 images remains negative through the final layer.

Table 9 summarizes these results. Sustained language dominance occurs for 368/500 (73.6%) images under the raw lens and 375/500 (75.0%) images under each tuned lens. The median per-image onset is layer 26 for the raw and COCO-tuned lenses and layer 24 for the MMMU-tuned lens. In contrast, the aggregate crossover shifts from layer 11 under direct unembedding to layers 18 and 17 under the COCO- and MMMU-tuned lenses, respectively. This shows that calibration affects the precise depth assigned to the emergence of language dominance while leaving its prevalence largely unchanged.

Layer-wise trajectories. Figure 7 compares the complete layer-wise trajectories under the three lens choices. The tuned lenses do not make the raw and calibrated trajectories identical, particularly for vision-answer decoding. However, the central contradiction pattern is preserved. Language-answer decoding increases through the network, and the vision-minus-language logit difference becomes increasingly language-favoring in later layers. The calibrated lenses change the precise localization of the crossover without eliminating the broader language-dominant trajectory.

Trajectory agreement. To quantify similarity beyond the crossover location, we compute Spearman correlations between the 36 layer-wise values produced by each pair of lenses. Table 10 reports the resulting correlations. Language-answer decoding trajectories are strongly correlated across lens choices ($\rho = 0.784$–0.808). The same trend is also observed in the vision-minus-language logit-margin trajectories ($\rho = 0.786$–0.889). Vision-answer decoding is more sensitive to calibration, with lower correlations ($\rho = 0.355$–0.594). These results indicate that calibration changes some details of the layer-wise readout while preserving the broader trajectory toward language-consistent decoding.

Table 10: Layer-wise trajectory agreement across lens choices. Values are Spearman correlations (ρ) computed over the 36 layer-wise measurements for each metric.

Metric	Raw–COCO	Raw–MMMU	COCO–MMMU
Vision-answer decoding	0.355	0.594	0.359
Language-answer decoding	0.787	0.784	0.808
Vision–language logit difference	0.889	0.786	0.809

The tuned-lens analysis supports the qualitative conclusion obtained with direct unembedding. The exact layer assigned to the onset of language dominance is calibration-sensitive, but the fraction of images exhibiting sustained language dominance and the overall language-answer and vision–language margin trajectories remain similar across independently calibrated lenses.