Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Visual sequence structure shapes in-context choices in vision–language models

Su Hyoun Park

suhyounpark@gmail.com

Korea University

Min-Shik Kim

Yonsei University

Research Article

Keywords: Vision-language models, Visual statistical learning, Multimodal in-context learning, Sequential processing, Visual sequence structure

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11822807/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Visual sequence structure shapes in-context choices in vision–language models

Su Hyoun Park[1], Min-Shik Kim[2]

[1] BK21 FOUR R&E Center for Psychology, Korea University, Republic of Korea

[2] Department of Psychology, Yonsei University, Republic of Korea

* Correspondence: suhyounpark@gmail.com

Abstract

Vision–language models (VLMs) process ordered images, yet how unlabelled visual context shapes their subsequent choices remains unclear. Using paradigms adapted from human visual statistical learning, we tested three commercial VLM deployments across five experiments without task-specific weight updates. Structured image sequences increased target choices relative to frequency-matched shuffled sequences in two deployments, although invalid responses in the third prevented the preregistered pooled test. Target choices also increased with successor strength. All three deployments preferred previously encountered image orders. Crucially, reversing image order within three-image chunks reduced forward-order choices from 98.7% to 1.9%, with the test question and alternatives unchanged. However, choices involving unseen exemplars of familiar visual categories varied markedly across deployments. These findings show that unlabelled visual sequence structure can systematically influence VLM choices, and underscore the importance of visual context history for understanding and evaluating multimodal models.

Vision–language models (VLMs) are increasingly used in settings where visual inputs accumulate over time or across multiple items. Examples include video frames and long video streams[1,2], successive screenshots in computer-use agents[3,4], visual information distributed across multipage documents[5], and images embedded in multi-turn conversations[6]. In such settings, a response to a final query may depend not only on the query itself but also on which images preceded it and the order in which they appeared. Understanding how visual context history shapes model behaviour is therefore important for interpreting multimodal systems and designing evaluations that account for sequential visual information.

Research on human statistical learning provides a useful experimental framework for investigating sensitivity to sequential structure. Foundational work showed that infants can extract statistical regularities from continuous speech streams[7], and subsequent research demonstrated sensitivity to spatial and temporal regularities in visual scenes and shape sequences[8–12]. Statistical learning can also extend to category-level regularities in natural scenes[13], although its expression and generalisation depend on task demands and stimulus characteristics[14–18]. These paradigms offer methodological tools for isolating the influence of sequence structure by controlling item identity, frequency and serial position while manipulating the relationships among successive items. Importantly, behavioural sensitivity to such regularities in artificial systems need not reflect the implicit learning processes proposed to operate in humans.

Cognitive science approaches have increasingly been used to investigate language models[19], while studies of VLMs have revealed both sophisticated object-concept representations and limitations in visual cognition[20–22]. Multimodal in-context learning has also been studied using labelled image demonstrations or explicit image–text examples[23–25]. However, these approaches do not directly isolate whether the statistical structure of unlabelled visual sequences can influence subsequent choices when model weights remain fixed. In particular, it remains unclear whether VLM choices are sensitive to the presence and strength of sequential regularities, whether they track the experienced direction of image sequences, and whether such sensitivity extends to unseen exemplars of familiar visual categories. Behavioural effects could arise from retrieving previously encountered sequences, recognising local transitions or other forms of context-dependent processing. Computational accounts of transformer in-context learning, including induction-head mechanisms, provide possible explanations[26–30], but behavioural evidence alone cannot establish the underlying computations or their equivalence to human statistical learning[31–33].

Here we investigated these questions across five experiments involving three commercial VLM deployments (Fig. 1). Unlike conventional multimodal in-context learning paradigms based on labelled demonstrations, our experiments presented unlabelled image sequences and systematically manipulated their structure while controlling relevant stimulus and test properties. Experiment 1 compared structured and frequency-matched shuffled sequences to assess sensitivity to repeated visual order, whereas Experiment 2A varied successor strength to examine

graded sensitivity to transition statistics. Experiment 2B-R tested discrimination between previously encountered forward sequences and their reversals, and Experiment 3 examined choices involving held-out exemplars of familiar visual categories. Finally, Experiment 4 directly manipulated exposure direction while holding the test question and alternatives constant, providing a controlled test of how experienced visual order shapes subsequent choices under matched test questions and alternatives. Together, these experiments distinguish sensitivity to sequence structure and transition strength from responses to experienced order, while examining the extent to which sequence-consistent choices extend beyond previously encountered images. By separating these behavioural dimensions without task-specific weight updates, this study provides a controlled framework for assessing how unlabelled visual context history influences VLM decisions.

Results

Experimental framework and inferential scope

We evaluated three fixed API model identifiers—gpt-5.6-terra (OpenAI), gemini-3.8-flash (Google) and claude-sonnet-5 (Anthropic), hereafter OpenAI, Gemini and Claude—using independent multimodal requests. Where exposure was present, each request contained an ordered 144-image stream followed by one two-alternative forced-choice (2AFC) test. No conversation history, feedback or parameter updating occurred between trials. In the abstract-stimulus experiments, each randomized design seed defined an assignment of 12 fixed images and a corresponding sequence schedule. Seeds index stimulus-design randomizations rather than independent networks, stimulus populations or human participants. Incomplete required response cells reduced the number of analysable seed-level contrasts in some deployments.

The final preregistered confirmatory family comprised three contrasts: structured versus shuffled exposure (Experiment 1; H1), graded successor strength (Experiment 2A; H2) and an exposed-versus-no-exposure direction comparison (Experiment 2B-R; H3-R). These were evaluated using two-sided one-sample t-tests with Bonferroni-adjusted $\alpha = 0.05/3 \approx 0.0167$. The third contrast was prospectively specified as a replacement for the unexecuted original Experiment 2B test in an update to the original OSF registration; the Experiment 1 pooled test was withheld under its preregistered response-validity criterion. Experiment 3 was preregistered as exploratory.

Experiment 4 had a separate prospective registration and a distinct two-sided exact sign test ($\alpha = 0.05$). Unless otherwise noted, provider-level intervals are descriptive 95% t intervals; primary pooled Experiment 2A and 2B-R intervals are Bonferroni-adjusted 98.33% t confidence intervals.

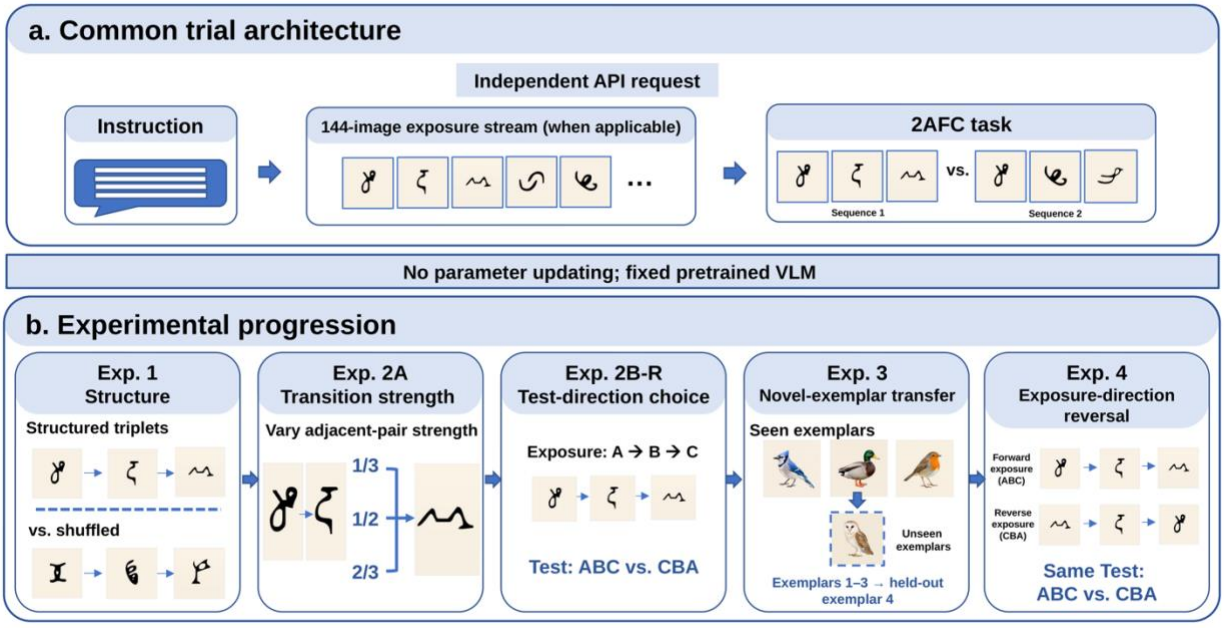


Fig. 1 | Experimental architecture and study progression. a, Each independent API request contained an instruction to attend to the image stream, a 144-image exposure stream where applicable, and a 2AFC test; prior dialogue, feedback and weight updates were absent. b, Schematic overview: structured versus shuffled streams (Experiment 1); graded successor strength (Experiment 2A); direction recognition with forward ABC and reverse CBA alternatives (Experiment 2B-R, the prospectively specified replacement for the original Experiment 2B); preregistered exploratory category-order choices involving held-out exemplars (Experiment 3); and reversal of exposure direction with matched test wording and alternatives (Experiment 4). Examples are schematic rather than full 144-image streams.

Experiment 1: structured versus shuffled exposure

Experiment 1 tested whether structured image order influenced subsequent choices beyond exposure to the same individual images. Structured 144-image streams repeated four intact triplets 12 times each, whereas shuffled streams preserved item frequencies and serial-position classes while disrupting the relevant triplets (Extended Data Fig. 1a). Tests contrasted an intact triplet with a position-matched recombination. The preregistered primary contrast compared structured with shuffled exposure; a separate no-exposure condition was a secondary reference and did not enter that contrast.

Response-format validity determined whether the planned confirmatory test could be conducted (Extended Data Fig. 1b). OpenAI and Claude each had zero invalid outputs among 528 planned calls. Gemini had 173/528 invalid outputs (32.8%), exceeding the prespecified 5% threshold; only two Gemini seeds yielded complete structured–shuffled pairs. The registered three-provider H1 test was therefore not conducted. OpenAI and Claude each provided 44 complete seed contrasts, with descriptive structured-minus-shuffled differences of 0.665 (descriptive 95% t interval, 0.613–0.716) and 0.926 (0.884–0.968), respectively (Extended Data Fig. 1c). These deployment-specific estimates are descriptive and are not substitutes for the withheld pooled test.

Experiment 2A: graded sensitivity to successor strength

Experiment 2A asked whether target choices changed with graded successor statistics. Target image C followed B on 4, 6 or 8 of 12 opportunities, corresponding to conditional rates of 1/3, 1/2 and 2/3. The designated foil successor occurred 2/12 times in every condition, and individual-image frequencies were held constant (Fig. 2a). Test questions and alternatives used the same format across all three exposure strengths. Target-choice rates increased with successor strength across all three deployments (Fig. 2b). Descriptive high-minus-low contrasts were 0.233 for OpenAI (44 complete seeds; descriptive 95% t interval, 0.163–0.303), 0.194 for Gemini (36 seeds; 0.113–0.276) and 0.182 for Claude (44 seeds; 0.083–0.281; Fig. 2c).

The preregistered confirmatory analysis required complete observations at all three exposed levels, including the middle level with zero weight in the primary linear contrast. Thirty-six seeds met this requirement in all three deployments. Their equal-provider high-minus-low contrast was 0.199 (Bonferroni-adjusted 98.33% t confidence interval, 0.113–0.285; $t(35) = 5.844$, two-sided $P = 1.24 \times 10^{-6}$; Fig. 2d), meeting the prespecified confirmatory criterion. The finding supports graded behavioural sensitivity to the manipulated successor strength within this fixed image set. Because target-successor count and conditional probability increased together while B occurred 12 times, the design cannot distinguish sensitivity to observed B→C counts from normalized conditional probabilities. Nevertheless, all target continuations (4–8/12) and the foil (2/12) occurred at nonzero frequencies, so the result cannot be reduced to a present-versus-absent rule.

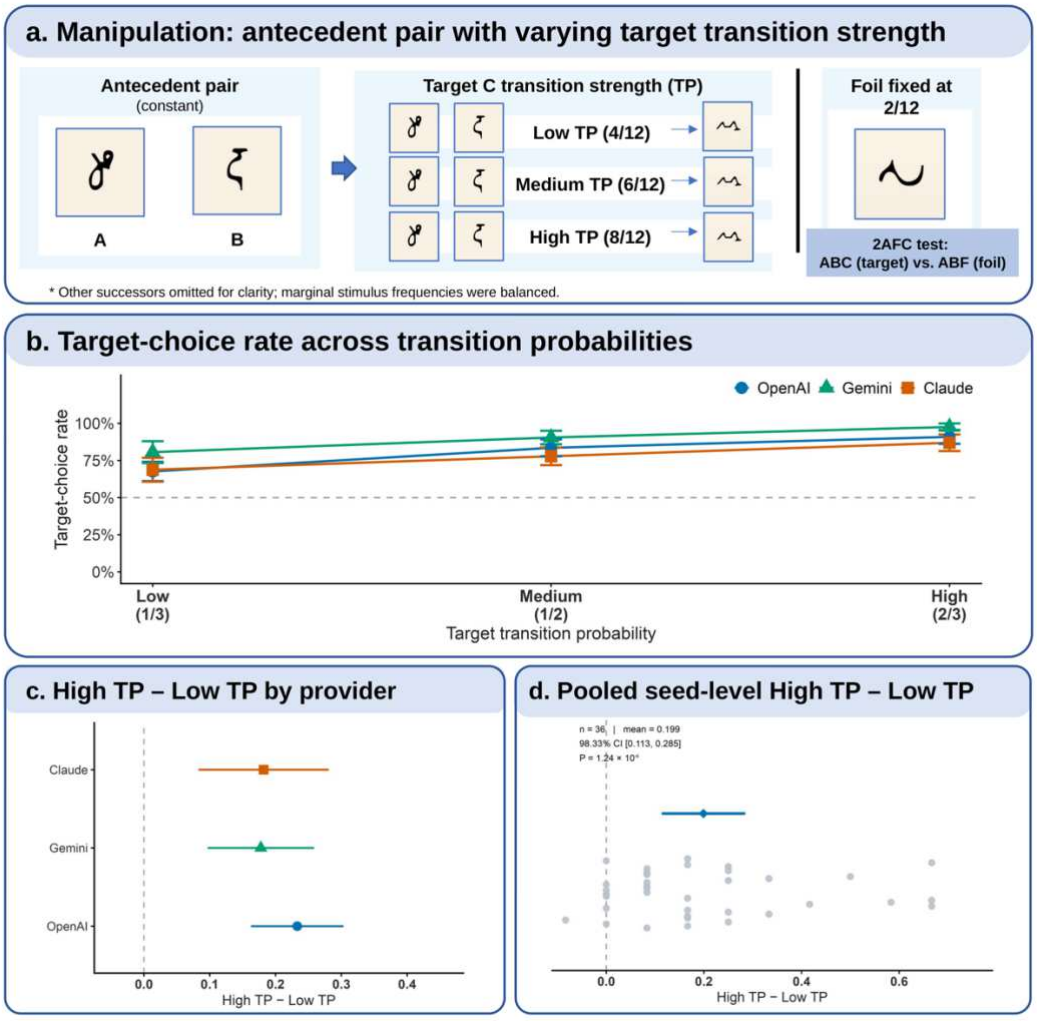


Fig. 2 | Graded target-successor sensitivity in Experiment 2A. a, The A–B antecedent relation remained fixed while B→C target occurrences varied across low (4/12), medium (6/12) and high (8/12) levels; B→F foil occurrences remained 2/12. b, Mean exposed-condition target-choice rates by deployment, with descriptive 95% t intervals across complete seeds. c, High-minus-low contrasts with descriptive 95% t intervals (OpenAI n = 44, Gemini n = 36, Claude n = 44). d, Equal-provider mean over 36 common, fully observed seeds (0.199), with a Bonferroni-adjusted 98.33% t confidence interval (0.113–0.285), t(35) = 5.844 and P = 1.24×10^{-6}. Error bars reflect variation in randomized stimulus designs, not sampled model weights.

Experiment 2B-R: direction recognition and forward-order preference

Experiment 2B-R tested direction recognition using the same three images in forward ABC and reverse CBA order (Fig. 3a), so individual-image familiarity alone could not distinguish the alternatives. This experiment was prospectively specified as a replacement for the original Experiment 2B, which did not proceed to main data collection after Gemini failed its sanity

check. Following high-strength forward ABC exposure, 44 fresh seeds (50–93) yielded forward-order choice rates of 99.4% for OpenAI (175/176), 100% for Gemini (176/176) and 93.2% for Claude (164/176; Fig. 3b). Forward-order preferences were observed across randomized seed assignments (Fig. 3c).

The prospectively specified confirmatory contrast compared exposed forward-order choices with a no-exposure baseline. Descriptive exposed-minus-baseline seed-mean differences were 0.494 for OpenAI (44 complete pairs; descriptive 95% t interval, 0.483–0.506), 0.529 for Gemini (43 pairs; 0.459–0.599) and 0.438 for Claude (44 pairs; 0.376–0.499). Across 43 complete shared seeds, the pooled contrast was 0.486 (Bonferroni-adjusted 98.33% t confidence interval, 0.443–0.530; $t(42) = 28.113$, two-sided $P = 7.25 \times 10^{-29}$), meeting the preregistered confirmatory threshold.

The exposed trials asked which order matched the previously encountered sequence, whereas the no-exposure baseline asked which order seemed more natural. This comparison therefore supported a compound exposure-and-question effect; it could not isolate exposure as the sole causal factor. This question-wording difference motivated the matched reversal in Experiment 4.

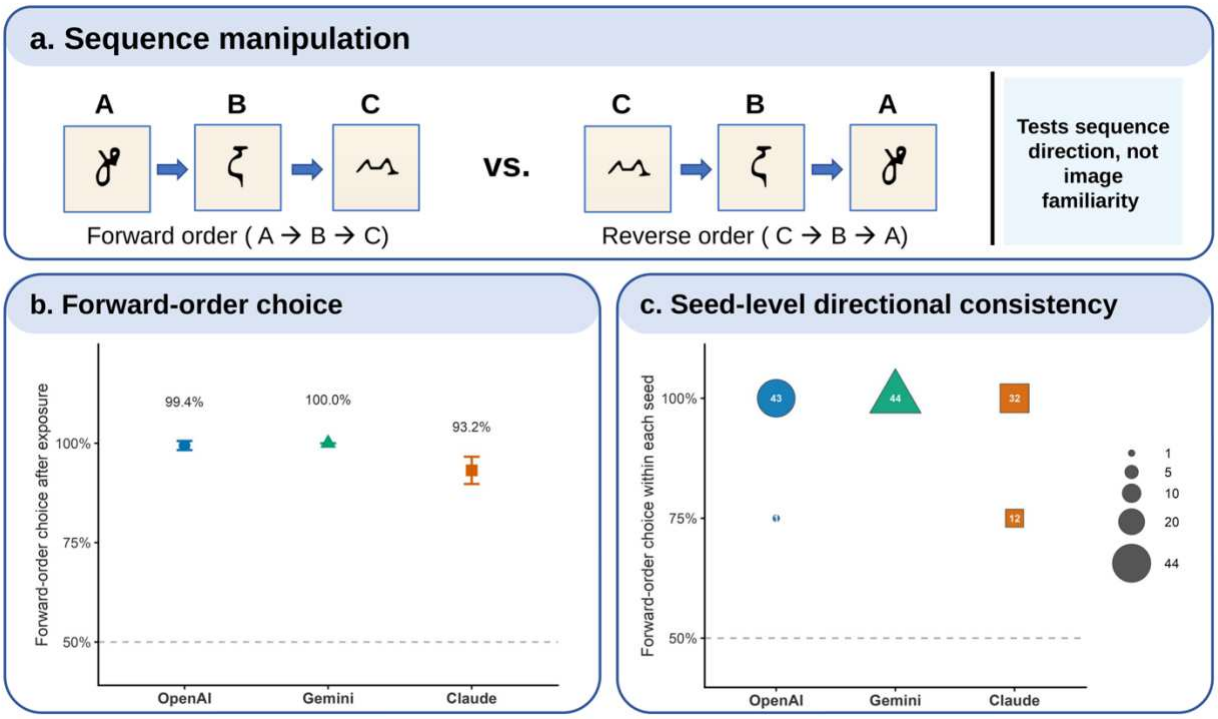


Fig. 3 | Forward-order preferences after exposure in Experiment 2B-R. a, Test alternatives contained the same images in ABC or CBA order following structured forward ABC exposure. b, Forward-order choice rates: OpenAI 175/176 (99.4%), Gemini 176/176 (100%) and Claude 164/176 (93.2%); bars show descriptive 95% t intervals across exposed-condition seed means. c, Distribution of seed-level forward-order choice rates; marker area indicates the number of seeds at each rate. All three deployments contributed 44 exposed-condition seeds; the exposed–baseline contrast included 43 complete paired seeds for Gemini and 44 for OpenAI and Claude.

Experiment 3: preregistered exploratory choices involving held-out category exemplars

Experiment 3 examined choices for visual-category sequences when none of the exact test images had appeared during exposure. Twelve categories each had four exemplars; only exemplars 1–3 appeared in the exposure stream, whereas both test alternatives used held-out exemplar 4 (Fig. 4a). Thus, the model could not retrieve a pixel-identical test image from exposure, although semantic category recognition remained possible. Exposed-condition target choices were 119/176 (67.6%) for OpenAI, 166/168 (98.8%) across 42 complete exposed-condition seeds for Gemini, and 93/176 (52.8%) for Claude (Fig. 4b).

Exploratory exposed-minus-no-exposure seed-level differences were 0.244 for OpenAI (44

complete pairs; descriptive 95% t interval, 0.162–0.327), 0.526 for Gemini (39 pairs; 0.445–0.606) and 0.017 for Claude (44 pairs; −0.067 to 0.101). These results show substantial deployment-dependent variation on the held-out-exemplar task, not uniform transfer across deployments. Because exposed and baseline questions differed, the observed contrasts cannot be attributed solely to exposure.

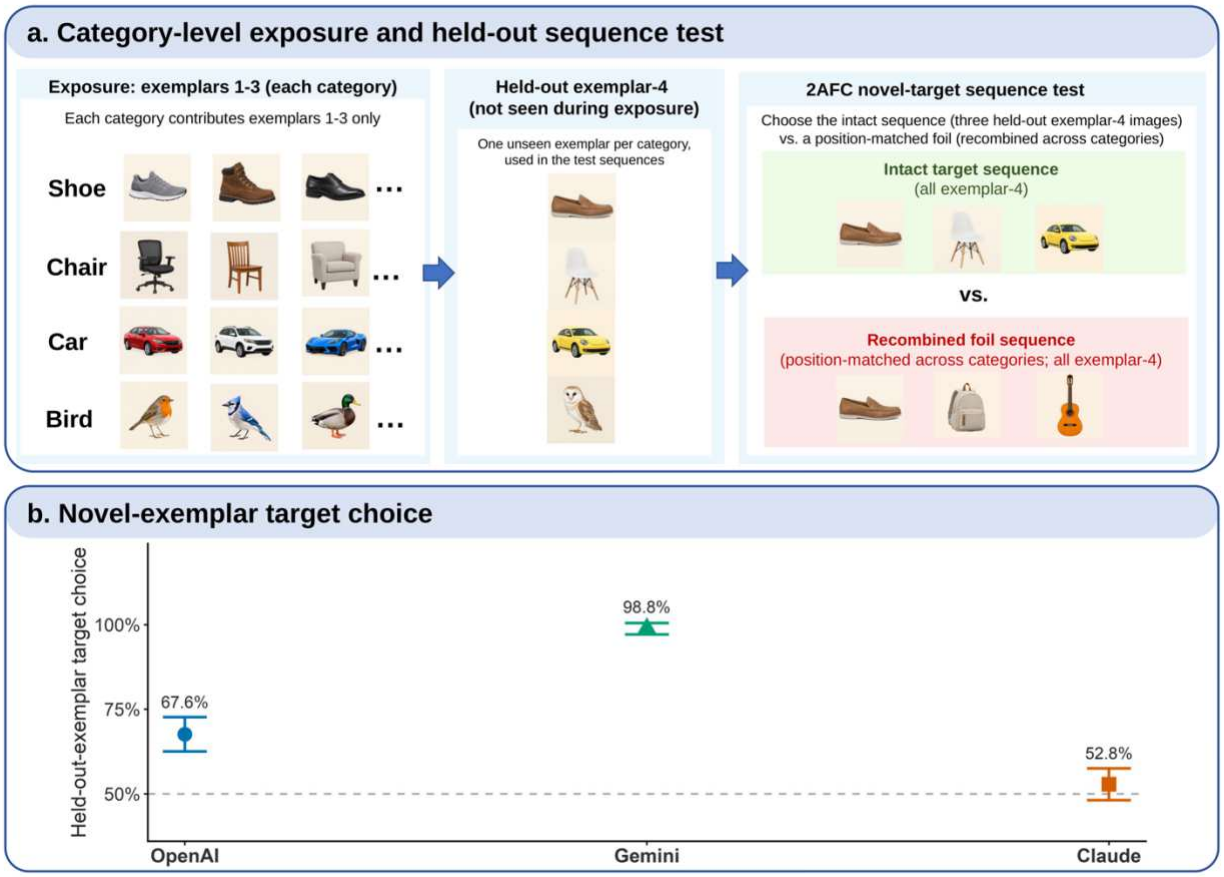


Fig. 4 | Preregistered exploratory category-order choices with held-out exemplars. a, Three visual exemplars per category were used during exposure; the fourth was withheld and used only in the test alternatives. The target preserved the category order of an exposed triplet; the foil used a position-matched recombination. The category images were generated using OpenAI image-generation tools as actual experimental stimuli. b, Mean exposed-condition target-choice rates (OpenAI 67.6%, Gemini 98.8%, Claude 52.8%), with descriptive 95% t intervals over complete exposed-condition seed means. The dashed 50% reference is not the matched no-exposure baseline.

Experiment 4: assigned exposure direction reverses choices

Experiment 4 was a separately preregistered follow-up addressing the question-wording ambiguity in Experiment 2B-R. Forty-four new seeds (94–137) each generated the same 48

three-image chunks for paired forward and reverse streams. The reverse condition inverted the order within every chunk, keeping chunk identity, chunk order, marginal image counts, test alternatives, option positions and test question identical (Fig. 5a). This was a local within-chunk reversal, not a global reversal of the 144-image stream; it also changed transitions between adjacent chunks.

The reversal yielded correspondingly reversed choices (Fig. 5b). OpenAI chose ABC in 174/176 forward-exposure trials (98.86%) and 1/176 reverse-exposure trials (0.57%); Gemini in 176/176 (100%) and 0/176 (0%); and Claude in 171/176 (97.16%) and 9/176 (5.11%). Across the three equally represented deployments, ABC was chosen on 521/528 trials (98.67%) after forward exposure and 10/528 (1.89%) after reverse exposure. Provider-level mean seed contrasts were 0.983, 1.000 and 0.920, respectively. Thus, a fixed ABC preference alone cannot explain the exposure-dependent reversal.

All 1,056 planned responses were format-valid, yielding 44 complete cross-provider seed contrasts. Their mean forward-minus-reverse difference in ABC-choice proportion was 0.9678 (descriptive 95% t interval, 0.9521–0.9835). All 44 pooled seed contrasts were positive (0 negative, 0 tied), and the preregistered two-sided exact sign test gave $P = 1.14 \times 10^{-13}$ (Fig. 5c). These contrasts reflect randomized assignments of the same fixed 12-image set, not independently sampled images or models. These findings support an exposure-direction-dependent behavioural effect under matched test questions, without distinguishing implicit statistical learning from explicit recognition or retrieval of repeated triplets. A one-request Gemini acquisition deviation and sensitivity analysis are documented in Methods.

Taken together, the five experiments probe complementary dimensions of visual sequence sensitivity rather than five replications of a single effect (Table 1).

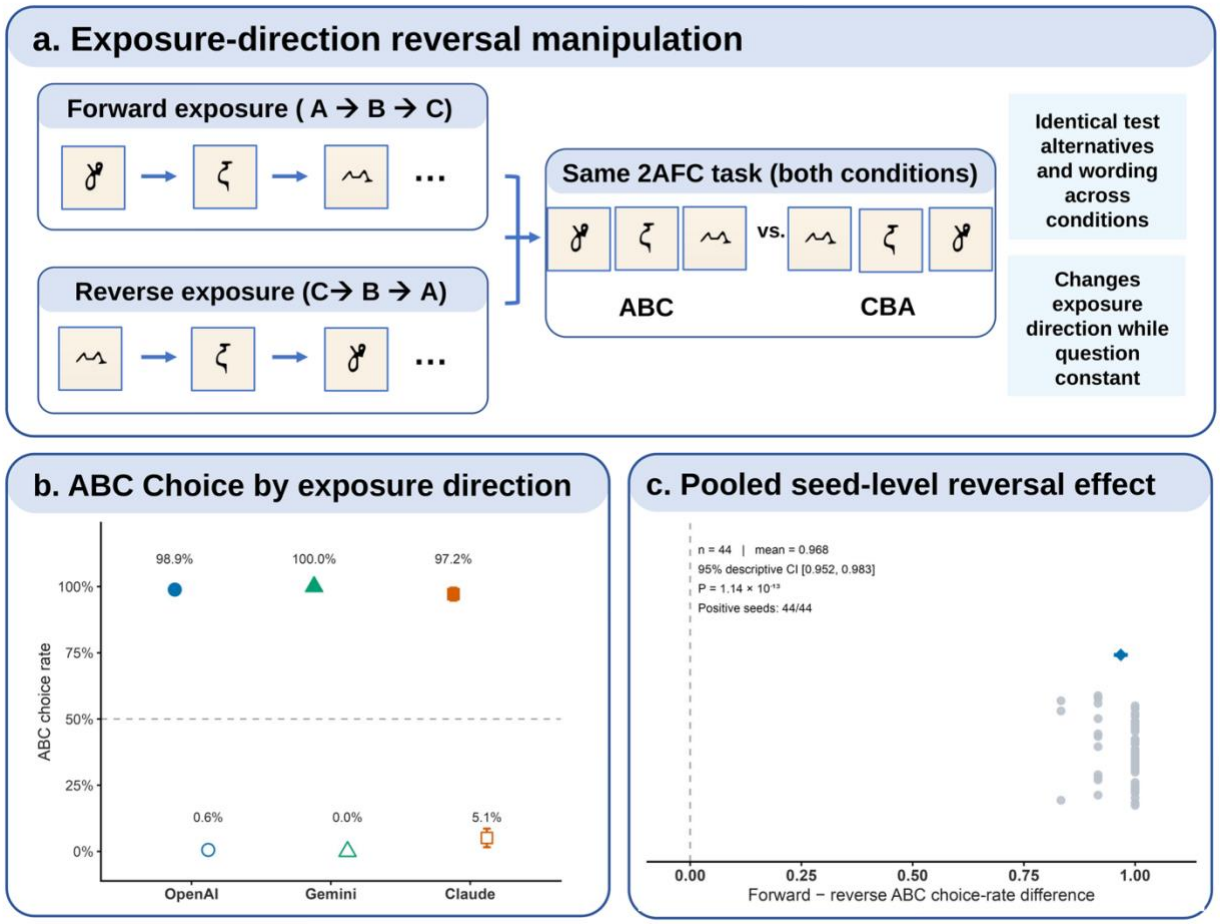


Fig. 5 | Exposure-dependent reversal of sequence preferences. a, Each reverse stream inverted the within-chunk order of the same 48 three-image chunks; chunk identities and chunk order were fixed, as were image frequencies, answer alternatives and test wording. b, Forward-order (ABC) choice rates under forward and reverse exposure, with descriptive 95% t intervals over 44 seeds per model and condition. c, Forty-four equally weighted cross-provider seed contrasts (forward minus reverse ABC-choice rate); vertical displacement of overlapping points is for display only. The pooled mean was 0.9678 (descriptive 95% t interval, 0.9521–0.9835), with 44 positive signs and two-sided exact sign-test $P = 1.14 \times 10^{-13}$.

Table 1 | Experimental findings and inferential status across five experiments.

Experiment	Dimension examined	Key finding	Inferential status
1	Presence of ordered structure	Structured–shuffled contrasts: +0.665 (OpenAI), +0.926 (Claude)	H1 pooled test not conducted because Gemini exceeded the response-validity threshold; other deployment contrasts descriptive.
2A	Successor strength	Pooled high–low contrast = 0.199; $P = 1.24 \times 10^{-6}$	Preregistered H2 supported (36 shared seeds).
2B-R	Sequence-direction discrimination	Forward choices after exposure: 93.2–100%; pooled exposed–baseline contrast = 0.486; $P = 7.25 \times 10^{-29}$	H3-R supported (43 shared seeds; prospective amendment to the original preregistration).
3	Choices involving held-out category exemplars	Exposed-condition target choices: 67.6% (OpenAI), 98.8% (Gemini), 52.8% (Claude)	Preregistered exploratory analysis.
4	Exposure-direction reversal	ABC choices: 98.7% (forward) vs. 1.9% (reverse); pooled $\Delta = 0.9678$; 44/44 positive; $P = 1.14 \times 10^{-13}$	Separately preregistered confirmatory sign test supported (44 seeds).

Notes. H1, H2 and the prospectively substituted H3-R formed a single confirmatory family with Bonferroni-adjusted $\alpha = 0.05/3$; the threshold was retained despite the withheld H1 test. Experiment 3 was preregistered exploratory, and Experiment 4 had a separate prospective registration. Seed-level contrasts reflect randomized assignments of fixed stimuli rather than independent model or image samples.

Discussion

Across five experiments, unlabelled visual sequences influenced subsequent choices in three vision–language model (VLM) deployments. Structured image streams produced higher target-choice rates than frequency-matched shuffled streams in two deployments, although invalid responses in the third prevented the preregistered pooled test (Experiment 1). Target choices also increased with the frequency of the target successor (Experiment 2A). After exposure to forward ABC sequences, all three deployments strongly preferred ABC over its reverse, CBA (Experiment 2B-R). Crucially, reversing image order within each three-image exposure chunk reduced ABC choices from 98.7% to 1.9%, even though both exposure conditions used the same test question and answer options (Experiment 4). When tested with unseen exemplars of familiar categories, however, the three deployments differed markedly in their choices (Experiment 3). Together, Experiments 1, 2A, 2B-R and 4 demonstrate sensitivity to visual sequence structure, successor strength and experienced order, whereas Experiment 3 reveals greater variability

across deployments when test images were not encountered during exposure.

The difference between exact-order tasks and Experiment 3 may reflect the different demands of these tasks. In Experiments 2B-R and 4, test alternatives contained images already presented during exposure. Models could therefore rely on matching previously encountered image orders, retrieving repeated triplets or comparing local transitions. Research on induction heads and statistical sequence induction provides possible computational accounts of such behaviour, although these mechanisms have not been established in the VLMs tested here[26,27]. Experiment 3, by contrast, used new images belonging to categories encountered during exposure. Choosing an exposure-consistent sequence therefore required recognizing these categories and relating them to their previously presented order. Human visual statistical-learning studies have shown that category-level regularities can generalize to new exemplars, although such generalization depends on the task and stimuli[13,17,18]. The differences among deployments could reflect category recognition, the use of previously encountered sequence information or pre-existing preferences.

In humans, statistical learning provides a way to extract regularities from continuous experience without explicit instruction[14,15]. For example, eight-month-old infants can use statistical relations between successive speech sounds to identify word-like units after brief exposure[7,14]. Given the limits of human attention and memory, extracting recurring patterns may offer an efficient way to organize sensory information and guide future behaviour without retaining every encountered event[14,15]. Whether artificial systems benefit from a similar strategy is less obvious. VLMs are trained on vast amounts of data and can process long visual contexts, although doing so entails substantial memory and computational costs[1]. Their choices in our experiments nevertheless depended on the order and local statistics of the immediately preceding images. This behavioural resemblance to human statistical learning does not establish that VLMs extracted or represented those regularities through the same processes; sequence-based in-context computations offer alternative explanations[26,27,31].

A further question is whether detecting statistical structure serves the same purpose in humans and artificial systems. In human learning, regularities acquired from familiar instances can sometimes guide responses to new exemplars[13,16]. In our study, however, choices involving unseen category exemplars varied substantially across deployments despite strong sensitivity to previously encountered image orders. One deployment showed near-ceiling exposed-condition

choices with held-out exemplars, indicating that exposure-consistent responses with new images were possible in this task, although not comparably consistent across deployments. These findings raise a broader question for AI development: could representations that capture statistical regularities at a more abstract level improve generalization and reduce the costs of processing extensive visual histories, or might machine-specific strategies work as well or better? Human-like learning mechanisms need not be the objective; their value should be evaluated against alternative approaches on accuracy, generalization and efficiency[31]. Resolving this question will require direct comparisons under matched memory and computational budgets.

Several aspects of the experimental design define the scope of these findings. The abstract-stimulus experiments used a fixed image set across randomized assignments, and only three commercial VLM deployments were evaluated. Differences among deployments, particularly in Experiment 3, may also reflect differences in reasoning and decoding configurations rather than underlying model capabilities alone. Furthermore, invalid Gemini responses prevented the preregistered pooled test in Experiment 1, leaving the effects in the other two deployments descriptive. The findings therefore do not establish generalization across stimulus populations, model architectures or future deployments.

The interpretation of these behavioural effects also requires consideration of task framing and positional influences. In Experiments 2B-R and 3, exposed and no-exposure conditions used different question wording, making their contrasts sensitive to both exposure and task framing. Experiment 4 addressed this ambiguity by matching test questions, answer options and their positions across exposure conditions. However, the question explicitly directed models to consider the previously encountered order, so the results demonstrate exposure-dependent choices under instructed recognition rather than spontaneous or implicit learning. Reversing the order within each chunk also changed image positions and transitions between adjacent chunks, leaving position-dependent processing as a possible contributor. Previous studies have documented prompt-order and positional biases in language models and multi-image VLMs[34–36]. Although the observed effects establish sensitivity to visual sequence structure, they do not identify the underlying computations or demonstrate advantages in generalization or computational efficiency.

Future research could address these questions by distinguishing the information VLMs use from

the benefits that different processing strategies provide. Manipulating transition counts and antecedent frequencies with appropriate controls could help separate sensitivity to absolute event counts from conditional probabilities. Presenting component pairs without intact triplets could test the contribution of whole-sequence retrieval. For category-level tasks, matching test wording while varying exemplar similarity and category familiarity could clarify when previously encountered sequence information transfers to new images. Beyond these behavioural comparisons, a particularly informative direction would be to examine whether extracting statistical regularities becomes advantageous as visual histories grow longer or contain more irrelevant information. Comparing access to full visual histories with compact representations of recurring patterns, under controlled memory and computational budgets, could reveal trade-offs in accuracy, generalization and efficiency. Studies using open-weight multimodal models could further connect these behavioural effects to internal representations and targeted interventions on context processing[26–30].

These findings also have implications for how multimodal systems are evaluated. In Experiment 4, identical test questions and answer options produced sharply different choices depending on the preceding image order. Evaluating only the final question and response may therefore overlook how preceding visual information shapes a model's choices. This consideration extends to long-video understanding, computer-use agents, multipage-document processing and image-rich conversations, where responses may depend on visual information encountered earlier in the context[1–6]. Evaluations should assess not only final-answer accuracy but also how reliably model choices respond to changes in visual history. Recording the ordered visual inputs, test alternatives, response-validity criteria and model-deployment details would help make such effects reproducible and interpretable. More broadly, controlled manipulations of visual sequence structure provide a way to identify how preceding visual information influences VLM choices, without presuming that the observed sensitivity reflects the same learning mechanisms as in humans[31–33].

In conclusion, our findings show that unlabelled visual sequences can systematically shape VLM choices without task-specific weight updates or feedback. The observed sensitivity to sequence structure, statistical strength and exposure direction suggests that information in the immediate visual context can influence decisions beyond the content of individual images. At the same time, the variation in choices involving unseen category exemplars highlights an important

distinction between recognizing previously encountered patterns and applying sequence information to new visual instances. Whether these behaviours arise from processes resembling human statistical learning or from different computational strategies remains an open question. Understanding when and how artificial systems benefit from extracting regularities in visual experience may help guide the development of multimodal models that generalize more effectively and process complex visual information more efficiently.

Methods

Study design, sequencing and prospective registration

The study adapted visual statistical-learning paradigms to independent black-box vision–language model (VLM) API requests. Human experiments served as methodological inspiration rather than as directly matched participant data. Piloting used design seeds 0–5 and was excluded from confirmatory analyses.

The original preregistration and its prospective amendments are archived at https://osf.io/g3exk/. The original confirmatory hypothesis family comprised H1 (Experiment 1), H2 (Experiment 2A) and H3 (the originally planned Experiment 2B). The original Experiment 2B did not proceed to main data collection after Gemini failed the explicit-order sanity check. Before collecting replacement data, a prospective update to the original OSF registration specified Experiment 2B-R as the replacement task and introduced H3-R using 44 fresh design seeds (50–93). H3-R prospectively replaced H3 in the original three-hypothesis confirmatory family; it was not a separate OSF registration or a retrospective reinterpretation of the failed Experiment 2B protocol.

A subsequent prospective v1.6 amendment raised Gemini's maximum-output ceiling before its previously unstarted Experiment 2A and Experiment 3 main arms were acquired. Experiment 3 was preregistered as an exploratory experiment, with its design and outcome specified before main data collection; it was not included in the confirmatory hypothesis-testing family. Experiment 4 was separately preregistered using 44 new design seeds (94–137), after earlier behavioural results had been inspected but before its own data collection began (OSF registration: https://osf.io/6q9ze/). The previously specified familywise inference and failed

sanity-check records were not modified retrospectively. Version-tagged code and design manifests distinguish these stages.

Models, decoding settings and request independence

We requested OpenAI gpt-5.6-terra, Google gemini-3.8-flash and Anthropic claude-sonnet-5. For brevity, these deployments are referred to below as OpenAI, Gemini and Claude, respectively. Model identity was specified through API model identifiers rather than guaranteed immutable server-side weight snapshots.

OpenAI used the Responses API with reasoning effort set to 'none', temperature set to 0, image detail set to 'original' and a maximum output of 16 tokens. Google used the Interactions API with thinking level set to 'low', high image resolution and no explicitly specified temperature. The original Experiment 2B sanity check used a 128-token maximum-output ceiling. Following prospective amendments, the ceiling was raised to 8192 tokens for Experiment 2B-R and Experiment 4, and for the previously unstarted Gemini Experiment 2A and Experiment 3 main arms. Anthropic used the Messages API with thinking disabled, effort set to 'low', temperature omitted and a maximum output of 64 tokens.

SDK-managed retries were disabled. The original main collection used the frozen dependencies openai==3.14.1, google-genai==2.24.0 and anthropic==1.7.0. Model request IDs, response metadata and executed configurations were logged. No trial reused conversation history, enabled web search or other tools, or provided feedback or parameter updates. Each trial was submitted as an independent, one-turn API request. In exposed conditions, the assigned image stream and its corresponding forced-choice test were presented within the same request; no-exposure conditions omitted the image stream.

Visual stimuli and shared exposure architecture

Experiments 1, 2A, 2B-R and 4 used the same fixed set of 12 abstract images adapted from Afaka (Ndyuka) syllabary material published by Simon Ager on Omniglot[37]. Original sources and reuse terms are documented in the repository's STIMULUS_LICENSES.md; derived stimulus files were identified by SHA-256 hashes and matched across relevant conditions.

For each design seed, the 12 glyphs were deterministically reassigned to four ordered triplets. The integer seed initialized a reproducible pseudorandom permutation of the fixed images; the

resulting assignment was reused across the conditions compared for that seed rather than sampling new images. An exposed trial typically presented an ordered 144-image stream with balanced marginal image frequencies, followed by one forced-choice test. The four tests within each condition were submitted as separate API requests, each containing the full exposure stream again. Thus, the four tests did not follow a single shared exposure request; each test had its own independent exposure-and-test context.

The images were transmitted as consecutive image parts within a single multimodal input, preceded by an instruction to attend to the image stream and followed by a closing text marker before the test. Accordingly, 'exposure' refers to visual information presented within a single model request, rather than timed perceptual exposure as used in human participant experiments. Test-option positions and deterministic trial schedules were specified before acquisition.

Experiment 1: matched structured and shuffled streams

Seeds 6–49 specified four image-triplet tests in each of three conditions: structured exposure, shuffled exposure and no exposure. This yielded 12 requests per seed and 528 planned requests per provider.

Structured streams repeated each of four intact triplets 12 times, producing a 144-image input. Shuffled streams preserved marginal individual-image frequencies and serial-position classes but disrupted the intact triplet relationships. Tests compared an intact target triplet (e.g., ABC) with a position-matched recombined foil (e.g., AEI). H1 was defined as the structured-minus-shuffled target-choice contrast; the no-exposure condition served only as a secondary baseline and did not enter the primary contrast.

Each provider-by-seed-by-condition cell required four valid responses. OpenAI and Claude returned no invalid outputs among their 528 planned requests. Gemini returned 173/528 invalid outputs (32.8%), exceeding the prospectively specified 5% threshold. Consequently, the planned three-provider H1 confirmatory test was withheld as registered; the remaining provider-specific contrasts were retained as descriptive results.

Experiment 2A: graded successor statistics

Experiment 2A varied the frequency with which target image C followed antecedent B within four triplets. Target C followed B on exactly 4, 6 or 8 of the 12 B-successor opportunities,

corresponding to conditional transition probabilities of 1/3, 1/2 and 2/3. A designated foil successor occurred twice in every condition. The remaining opportunities were divided equally between two other successor images: 3+3 in the low condition, 2+2 in the medium condition and 1+1 in the high condition. The A→B antecedent structure and marginal individual-image frequencies were controlled.

Each seed (6–49) contained four two-alternative forced-choice (2AFC) completion trials for each of the three target-strength levels and four no-exposure trials, yielding 16 requests per seed and 704 planned requests per provider. At test, the model received an AB context and selected between candidate C and a position-matched foil. The same 'more familiar continuation' question and test format were used across all three exposed target-strength conditions.

H2 was the prespecified linear contrast across low, medium and high target strength, with weights −1, 0 and +1, respectively. This contrast is algebraically equivalent to the high-minus-low difference in target-choice rates. Although the medium condition had zero weight in the primary contrast, valid responses at all three target-strength levels were required for a complete seed-level H2 comparison. No-exposure trials did not enter this primary contrast. Because the number of target B→C transitions and their conditional probability changed together while B occurred 12 times in every condition, the design could not distinguish sensitivity to absolute transition counts from sensitivity to normalized conditional probabilities.

Experiment 2B-R: directional recognition and its baseline

The originally planned Experiment 2B was blocked before main behavioural collection when Gemini returned the text 'Option' with an incomplete-response status during an explicit-order sanity check under a 128-token generation ceiling. This output did not satisfy the response-format requirements. The original response was preserved and was not reinterpreted as a valid choice.

A prospective update to the original OSF registration specified Experiment 2B-R as the replacement task and introduced H3-R. The replacement used 44 fresh design seeds (50–93), a new 20-trial sanity battery and a prospectively raised Gemini maximum-output ceiling of 8192 tokens.

The main design presented 144-image, high-strength ABC-structured streams, with the target

successor occurring at 8/12, and contrasted forward ABC against reverse CBA alternatives containing the same three images. Within each seed, four exposed and four no-exposure trials were paired, yielding eight requests per seed and 352 planned requests per provider.

Exposed tests asked which order matched the previously encountered stream, whereas no-exposure baseline tests asked which order was more natural. H3-R was defined as the forward ABC choice rate under the exposed-question wording minus the forward ABC choice rate under the no-exposure baseline wording. Because exposure condition and question wording differed together, H3-R represents a compound exposure-and-question contrast rather than an isolated causal effect of exposure direction.

All three providers contributed 44 exposed-condition seeds. One Gemini exposed–baseline pair was incomplete, so the H3-R pooled confirmatory analysis used 43 complete shared seeds.

Experiment 3: preregistered exploratory held-out-exemplar test

Experiment 3 examined whether exposure to ordered visual categories influenced choices involving new images of familiar categories. Twelve familiar visual categories (bird, car, chair, shoe, cup, camera, flower, dog, backpack, lamp, clock and guitar) were each represented by four 256 × 256-pixel PNG exemplars generated specifically for this experiment using OpenAI image-generation tools. These images were actual experimental stimuli, not decorative illustrations. Their provenance is documented in STIMULUS_LICENSES.md.

Three exemplars per category appeared in the 144-image exposure streams, while the fourth exemplar was withheld entirely for testing. Category triplets were reassigned by seed; training-exemplar choices were balanced across serial positions, and no exact exemplar triple was repeated. Test sequences contrasted held-out fourth-exemplar images arranged in the category order of an exposed target triplet with position-matched category recombinations.

Four exposed and four no-exposure trials were planned per seed (6–49), yielding 352 requests per provider. As in Experiment 2B-R, exposed questions referred to the preceding visual stream, whereas baseline questions asked which sequence was more natural. Thus, exposed-minus-baseline differences also constituted a compound context-and-question comparison.

Sanity checks did not exhaustively validate every held-out fourth exemplar. Accordingly, Experiment 3 was preregistered as exploratory and was not included in the confirmatory

hypothesis-testing family. Its results cannot uniquely identify a generalization or statistical-learning mechanism. Complete exposed–baseline pairs were available for 44 OpenAI seeds, 39 Gemini seeds and 44 Claude seeds.

Experiment 4: paired within-chunk reversal

Experiment 4 retained the 12 exact abstract images used in Experiment 2B-R and introduced 44 nonoverlapping design seeds (94–137). Each seed assigned the images to four unique ABC triplets and generated 48 three-image chunks. For each antecedent pair A_iB_i, there were 12 successor opportunities: the target successor C_i occurred eight times, a designated foil successor occurred twice, and each of the two remaining successor images occurred once. Each triplet identity therefore appeared in 12 chunks.

In forward exposure, chunks took the form (A_i, B_i, C_j), where i identifies the antecedent triplet and j identifies the successor image. When $j = i$, the chunk contained the target successor; when $j \neq i$, it contained a foil or another alternative successor. In reverse exposure, each of the same 48 chunks was reversed internally to (C_j, B_i, A_i), while the order of the 48 chunks was preserved.

Consequently, forward and reverse streams contained identical image identities and marginal image counts. The target triplet occurred eight times in its assigned orientation in each condition, with the corresponding reverse-oriented target triplet absent as an intact within-chunk sequence. This manipulation reversed the order within individual chunks rather than globally reversing the full 144-image stream. It changed within-chunk image positions and transitions between adjacent chunks.

For each seed, the four triplets were tested once under each exposure direction, yielding eight independent requests per seed and 352 requests per provider. Each tested triplet was paired across exposure directions within its seed; the four binary choices in each direction yielded the corresponding provider-by-seed choice rates used in the primary contrast. Each test contained the 144-image exposure stream followed by the same pair of three-image alternatives, ABC and CBA, giving 150 images per request.

The literal exposure prompt in both conditions was: 'You will be shown a stream of visual stimuli. Attend to the stimuli in the order presented. No response is required.' This was followed by the closing marker 'The visual stream is complete.' The identical test question in both

conditions was: 'Based on the visual stream you just encountered, which sequence is more consistent with the order you encountered? Respond with only 1 or 2.' The alternatives were introduced by the labels 'Sequence 1:' and 'Sequence 2:', followed by the directive 'Answer only: 1 or 2'.

Across the four tests for each seed, ABC appeared in each option position exactly twice. Option positions and question wording were identical across the paired forward- and reverse-exposure conditions. Trial and test order were pseudorandomized deterministically, and condition order alternated by seed parity. Full stimuli, prompts and assignments are retained in the frozen design manifest.

Sanity gates, scoring, completeness and missing data

Each experiment used its prospectively specified, task-specific sanity checks and gating rules. The original Experiment 2B failed the Gemini explicit-order sanity check described above and was not advanced to main collection. Experiment 2B-R used a new 20-trial sanity battery before its main collection. Experiment 4 used an independent 20-trial battery before each provider's main arm, comprising 12 exact-image discrimination tests, four explicitly instructed order tests and four explicitly instructed completion tests. Each provider was required to achieve 20/20 valid and correct responses; all three providers passed.

The immutable response whitelist, applied after trimming surrounding whitespace, consisted exactly of '1', '2', 'Option 1' and 'Option 2'. Outputs outside these strings were classified as invalid and were not semantically reinterpreted. SDK-managed automatic retries were disabled. Four valid responses were required to estimate each provider-by-seed-by-condition choice rate. A seed was included in a pooled cross-provider analysis only when all required cells were complete for all three providers. No choice-dependent exclusions or imputation were permitted.

For the preregistered confirmatory tests, invalid outputs exceeding 5% of planned trials in any provider blocked the corresponding three-provider primary test. The first integer counts exceeding this threshold were 27/528 for Experiment 1, 36/704 for Experiment 2A and 18/352 for Experiments 2B-R and 4. Experiment 1 exceeded this threshold in Gemini, so its pooled H1 test was withheld. Incomplete required cells also reduced the number of shared seeds available for pooled analyses, as described below.

Statistical inference and reporting

For each provider, seed and condition, the cell-level choice rate was calculated as the mean of four binary responses. For the confirmatory contrasts, provider-specific seed-level differences were first calculated and then averaged equally across the three providers. Thus, each pooled seed contrast gave equal weight to OpenAI, Gemini and Claude rather than weighting providers by their number of valid responses.

The original confirmatory family comprised H1, H2 and H3. Following the prospective replacement of the unexecuted H3, the final confirmatory family comprised H1, H2 and H3-R. The primary pooled contrasts were evaluated using two-sided one-sample t-tests against zero, with a Bonferroni-adjusted significance threshold of $\alpha = 0.05/3 \approx 0.0167$. Corresponding primary pooled contrasts were reported with 98.33% t confidence intervals. Provider-specific 95% t intervals were descriptive. The Bonferroni threshold was retained when the H1 test was withheld; it was not relaxed retrospectively.

The pooled H2 analysis included 36 complete shared seeds (OpenAI n = 44, Gemini n = 36 and Claude n = 44). The H3-R analysis included 43 complete shared seeds (OpenAI n = 44, Gemini n = 43 and Claude n = 44). For Experiment 3, exploratory exposed–baseline comparisons used 44, 39 and 44 complete pairs for OpenAI, Gemini and Claude, respectively. These differences reflect incomplete required response cells rather than changes to the prospectively specified seed counts. Experiment 3 did not contribute to the confirmatory hypothesis-testing family.

Experiment 4 had a separate preregistered primary analysis. Its primary effect was defined as Δ = Pr(choose ABC | forward exposure) − Pr(choose ABC | reverse exposure), averaged equally across the three providers within each complete seed. The primary test was a two-sided exact binomial sign test applied to the 44 pooled seed contrasts. Zero differences, if present, were excluded from this test alone. Directional support required a positive mean Δ and $P < 0.05$. The reported 95% t interval was descriptive and was not a substitute for the preregistered sign test. Zero-variance provider intervals were reported as point values; an unbounded t interval from a near-ceiling provider could extend beyond the theoretical [−1, 1] contrast range and was interpreted descriptively.

Experiment 4's seed count of 44 was fixed before data collection using the historical study design size rather than a new posthoc power calculation. The 44 seed-level contrasts, rather than the

1,056 individual API responses, were the units of analysis. Because all seeds reused the same 12 images and three VLM deployments, inference concerns randomized assignments of this fixed design rather than independently sampled image or model populations. All computations used recorded categorical choices rather than model confidence scores or log probabilities. For 44 positive, zero negative and zero tied pooled seed contrasts, the exact two-sided sign-test probability was $2 \times (1/2)^{44} = 1.14 \times 10^{-13}$.

Acquisition audit and Experiment 4 protocol deviation

During Experiment 4, one Gemini request was reissued after an interrupted session left its outcome indeterminate. This departed from the preregistered rule against retrying such requests. The original acquisition ledger and a reconciliation note were retained, and all 352 final responses per provider passed response-format checks. A post hoc sensitivity analysis showed that changing any single binary response would leave all 44 pooled seed contrasts positive and the preregistered two-sided exact sign-test result unchanged ($P = 1.14 \times 10^{-13}$). This analysis demonstrates numerical robustness to a single changed response but does not eliminate the procedural deviation.

Outcome visibility, software and reproducibility

During data collection, terminal progress displays reported trial completion and response-format validity without displaying substantive choices or condition contrasts. Experiment 4 was prospectively registered before data collection and used the frozen code snapshot prereg-v1.7-direction-reversal.

Analysis outputs and frozen design manifests preserve trial-level responses, seed assignments, stimulus identities and hashes, sequence orders, test alternatives, parser settings and model configurations. Original API response metadata was also retained to support reproducibility and auditing. Because commercial API deployments may change over time, model identifiers do not guarantee identical server-side model weights in future replications.

Ethics and research assistance

The authors designed the study, formulated and preregistered the hypotheses, and interpreted the findings. Generative AI tools, including ChatGPT (OpenAI; documented model variants included GPT-5.4 Thinking and GPT-5.6 Thinking), assisted with code development and

troubleshooting for stimulus sequencing, API requests, response parsing, quality-control procedures and analysis scripts, as well as manuscript language editing. Experimental workflows were reviewed by the authors and checked through offline preflight validation, deterministic stress tests, task-specific sanity gates and acquisition audits. Preregistered protocols, version-controlled code and frozen design manifests are retained for reproducibility. The authors take full responsibility for all experimental procedures, data analyses and interpretations. OpenAI image-generation tools were also used to produce the real-world-category images in Experiment 3 specifically as experimental stimuli, as disclosed in the corresponding figure legend.

Data availability

The original study preregistration and its prospective amendments are archived at https://osf.io/g3exk/, and Experiment 4 was separately preregistered at https://osf.io/6q9ze/. These registration records are currently embargoed and are not publicly accessible.

Trial-level model responses, response-validity and acquisition logs, frozen design manifests, stimulus identities and hashes, and reproducibility materials are preserved in a restricted, unpublished Zenodo record (record 23934604). These materials are not publicly accessible at the time of this preprint. Public release and redistribution arrangements remain subject to applicable third-party stimulus-use restrictions, as documented in STIMULUS_LICENSES.md.

Code availability

The experimental and analysis code is version-controlled in a private GitHub repository, with the frozen study snapshot identified by the tag prereg-v1.7-direction-reversal (commit 0cd5aa621a011ccb8fa71c8929eef9608ee03209). A corresponding archival copy is retained in the restricted Zenodo record (23934604). The code is not publicly available at the time of this preprint.

References

1. Chen, Y. *et al.* LongVILA: Scaling long-context visual language models for long videos. In Proc. International Conference on Learning Representations (ICLR, 2025).
2. Fu, C. *et al.* Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition 24108–24118 (2025).
3. Koh, J. Y. *et al.* VisualWebArena: Evaluating multimodal agents on realistic visual web tasks. In Proc. 62nd Annual Meeting of the Association for Computational Linguistics 881–905 (2024).
4. Xie, T. *et al.* OSWorld: benchmarking multimodal agents for open-ended tasks in real computer environments. *Adv. Neural Inf. Process. Syst.* **37**, 52040–52094 (2024).
5. Ma, Y. *et al.* MMLongBench-Doc: benchmarking long-context document understanding with visualizations. *Adv. Neural Inf. Process. Syst.* **37**, 95963–96010 (2024).
6. Liu, Z. *et al.* MMDU: a multi-turn multi-image dialog understanding benchmark and instruction-tuning dataset for LVLMs. *Adv. Neural Inf. Process. Syst.* **37**, 8698–8733 (2024).
7. Saffran, J. R., Aslin, R. N. & Newport, E. L. Statistical learning by 8-month-old infants. *Science* **274**, 1926–1928 (1996).
8. Turk-Browne, N. B., Isola, P. J., Scholl, B. J. & Treat, T. A. Multidimensional visual statistical learning. *J. Exp. Psychol. Learn. Mem. Cogn.* **34**, 399–407 (2008).
9. Turk-Browne, N. B., Jungé, J. A. & Scholl, B. J. The automaticity of visual statistical learning. *J. Exp. Psychol. Gen.* **134**, 552–564 (2005).
10. Kirkham, N. Z., Slemmer, J. A. & Johnson, S. P. Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition* **83**, B35–B42 (2002).
11. Fiser, J. & Aslin, R. N. Statistical learning of higher-order temporal structure from visual shape sequences. *J. Exp. Psychol. Learn. Mem. Cogn.* **28**, 458–467 (2002).
12. Fiser, J. & Aslin, R. N. Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychol. Sci.* **12**, 499–504 (2001).

13. Brady, T. F. & Oliva, A. Statistical learning using real-world scenes: Extracting categorical regularities without conscious intent. *Psychol. Sci.* **19**, 678–685 (2008).
14. Aslin, R. N. Statistical learning: A powerful mechanism that operates by mere exposure. *Wiley Interdiscip. Rev. Cogn. Sci.* **8**, e1373 (2017).
15. Frost, R., Armstrong, B. C. & Christiansen, M. H. Statistical learning research: A critical review and possible new directions. *Psychol. Bull.* **145**, 1128–1153 (2019).
16. Parsons, O. & Baron-Cohen, S. Extraction and generalisation of category-level information during visual statistical learning in autistic people. *PLoS One* **18**, e0286018 (2023).
17. Jun, J. & Chong, S. C. Visual statistical learning at basic and subordinate category levels in real-world images. *Atten. Percept. Psychophys.* **80**, 1946–1961 (2018).
18. Emberson, L. L. & Rubinstein, D. Y. Statistical learning is constrained to less abstract patterns in complex sensory input (but not the least). *Cognition* **153**, 63–78 (2016).
19. Binz, M. & Schulz, E. Using cognitive psychology to understand GPT-3. *Proc. Natl. Acad. Sci.* **120**, e2218523120 (2023).
20. Du, C. *et al.* Human-like object concept representations emerge naturally in multimodal large language models. *Nat. Mach. Intell.* **7**, 860–875 (2025).
21. Tangtartharakul, G. & Storrs, K. R. Visual language models show widespread visual deficits on neuropsychological tests. *Nat. Mach. Intell.* **8**, 209–219 (2026).
22. Schulze Buschoff, L. M., Akata, E., Bethge, M. & Schulz, E. Visual cognition in multimodal large language models. *Nat. Mach. Intell.* **7**, 96–106 (2025).
23. Zong, Y., Bohdal, O. & Hospedales, T. VL-ICL Bench: The devil in the details of multimodal in-context learning. In Proc. International Conference on Learning Representations (ICLR, 2025).
24. Xu, N., Wang, F., Zhang, S., Poon, H. & Chen, M. From introspection to best practices: principled analysis of demonstrations in multimodal in-context learning. In *Proc. 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies* 3299–3324 (2025).

25. Ferber, D. *et al.* In-context learning enables multimodal large language models to classify cancer pathology images. *Nat. Commun.* **15**, 10104 (2024).
26. Olsson, C. *et al.* In-context learning and induction heads. *arXiv:2209.11895 (2022)*.
27. Edelman, E., Tsilivis, N., Edelman, B. L., Malach, E. & Goel, S. The evolution of statistical induction heads: in-context learning Markov chains. *Adv. Neural Inf. Process. Syst.* **37**, 64273–64311 (2024).
28. Akyürek, E., Schuurmans, D., Andreas, J., Ma, T. & Zhou, D. What learning algorithm is in-context learning? Investigations with linear models. In Proc. International Conference on Learning Representations (ICLR, 2023).
29. Von Oswald, J. *et al.* Transformers learn in-context by gradient descent. In Proc. 40th International Conference on Machine Learning, Vol. 202, 35151–35174 (PMLR, 2023).
30. Shen, L., Mishra, A. & Khashabi, D. Position: Do pretrained transformers learn in-context by gradient descent? In Proc. 41st International Conference on Machine Learning, Vol. 235, 44712–44740 (PMLR, 2024).
31. Mitchell, M. & Krakauer, D. C. The debate over understanding in AI's large language models. *Proc. Natl. Acad. Sci.* **120**, e2215907120 (2023).
32. Lin, Z. From prompts to constructs: a dual-validity framework for large language model research in psychology. *Annu. Rev. Psychol.* **78**, in press (2026).
33. Abdurahman, S. *et al.* Perils and opportunities in using large language models in psychological research. *PNAS Nexus* **3**, pgae245 (2024).
34. Liu, N. F. *et al.* Lost in the middle: How language models use long contexts. *Trans. Assoc. Comput. Linguist.* **12**, 157–173 (2024).
35. Zhao, Z., Wallace, E., Feng, S., Klein, D. & Singh, S. Calibrate before use: improving few-shot performance of language models. In Proc. 38th International Conference on Machine Learning, Vol. 139, 12697–12706 (PMLR, 2021).
36. Tian, X., Zou, S., Yang, Z. & Zhang, J. Identifying and mitigating position bias of multi-image vision-language models. In Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition 10599–

10609 (2025).

37. Ager, S. Afaka syllabary. Omniglot. https://omniglot.com/writing/afaka.htm (accessed 28 October 2026).

Acknowledgements

This research was supported in part by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. RS-2026-25518317, Development of AI memory mechanism that reflects human cognitive principles), which covered API usage costs for the Google Gemini experiments, and by the BK21 FOUR R&E Center for Psychology at Korea University, which covered API usage costs for the OpenAI GPT and Anthropic Claude experiments.

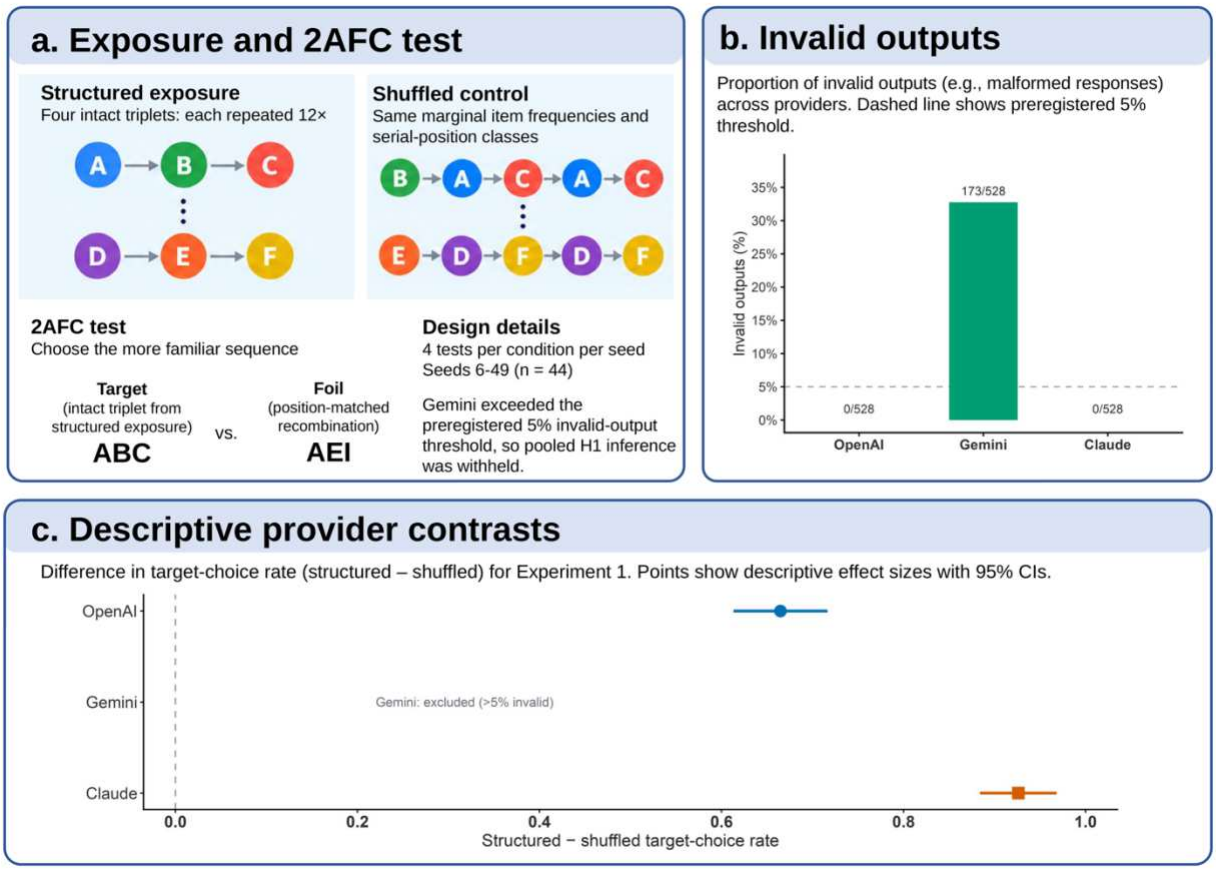
Author contributions

Su Hyoun Park led the conceptualization and design of the study, developed the experimental protocols, stimuli and code, conducted all experiments, collected and analysed the data, prepared the figures, and wrote the original manuscript. Min-Shik Kim provided critical feedback on the interpretation and presentation of the findings and on manuscript drafts. Both authors reviewed and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Extended Data



Extended Data Fig. 1 | Experiment 1 exposure control and response-format quality. a, Structured exposure repeated four intact triplets 12 times each. Shuffled exposure retained marginal item frequencies and serial-position classes but disrupted intact triplet structure; shown sequences are schematic, not literal frozen trial streams. b, Invalid response counts among 528 planned calls per provider: OpenAI 0, Gemini 173 and Claude 0. The dashed line denotes the preregistered 5% invalid-output threshold. c, Descriptive structured-minus-shuffled seed-level target-choice contrasts for OpenAI (n = 44, mean 0.665, 95% t interval 0.613–0.716) and Claude (n = 44, mean 0.926, 95% t interval 0.884–0.968). Gemini was excluded from pooled H1 inference because invalid responses exceeded the threshold; its remaining data were not used in an unregistered replacement confirmatory test.