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.

Learning to See the Elephant in the Room: Self-Supervised Context Reasoning in Humans and AI

Mengmi Zhang

a0091624@gmail.com

Nanyang Technological University https://orcid.org/0000-0002-2694-7097

Xiao Liu

Nanyang Technological University, Singapore

Soumick Sarker

Nanyang Technological University, Singapore https://orcid.org/0009-0002-6753-1482

Ankur Sikarwar

Nanyang Technological University, Singapore

Bryan Kiely

Nanyang Technological University, Singapore

Gabriel Kreiman

Boston Children's Hospital and Harvard Medical School https://orcid.org/0000-0003-3505-8475

Zenglin Shi

Nanyang Technological University, Singapore

Article

Keywords:

Posted Date: April 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8942453/v1

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

Additional Declarations: There is **NO** Competing Interest.

Learning to See the Elephant in the Room: Self-Supervised Context Reasoning in Humans and AI

Xiao Liu[1, 2, *], Soumick Sarker[1, *] Ankur Sikarwar[1, 2], Bryan Atista Kiely[1, 2], Gabriel Kreiman[3], Zenglin Shi[1, 2], Mengmi Zhang[1]

[1] College of Computing and Data Science, Nanyang Technological University, Singapore

[2] CFAR and I2R, Agency for Science, Technology and Research, Singapore,

[3] Boston Children's Hospital, Harvard Medical School, USA,

[*] Authors contributed equally to this work.

Address correspondence to mengmi.zhang@ntu.edu.sg

Abstract

A tiny object on the table might be a fork, but not an elephant. Humans rarely perceive objects in isolation; instead, they interpret scenes through relationships among co-occurring elements. But how do humans learn these contextual associations? We address this question through a series of human psychophysics experiments. First, we designed a set of context rules by replacing familiar household objects with novel ones. These rules capture different types of associations: global context (e.g., a toothbrush typically appears in bathrooms), local context (e.g., a fork often appears near a plate, even across rooms), and crowding effects (e.g., eggs tend to cluster together). Participants were exposed to short training videos showing these novel objects embedded in naturalistic scenes. We then tested their contextual reasoning using a "lift-the-flap" task, where the central object was hidden, and participants had to infer its identity based on the surrounding context. We also introduced contextual variations in the task by changing the size, resolution, and spatial arrangement of the scene context. Results show that humans can acquire contextual rules in a self-supervised manner without labels or feedback, and can robustly infer the hidden object across a range of contextual variations. To model this human capability, we introduce SeCo (Self-supervised learning for Context reasoning), a novel computational model for learning contextual associations. SeCo first identifies candidate target regions, then encodes the target and surrounding context using separate vision encoders. Inspired by semantic memory in biological brains, SeCo includes a learnable external memory module that stores latent contextual priors. Given a contextual cue, SeCo infers the identity of a hidden object by retrieving a likely object representation from this memory and regressing it toward the actual target. In contrast to existing SSL methods that focus on object-centric learning from single-object images, SeCo explicitly learns contextual relationships from complex scenes. Our results show that SeCo outperforms state-of-the-art SSL methods on the lift-the-flap task. Network analysis reveals that its external memory stores meaningful contextual knowledge, enabling accurate inference. Moreover, we also extend the context reasoning ability of SeCo, state-of-the-art SSL methods, and humans to object priming tasks, where they are asked to place target objects in context-appropriate locations. SeCo predicts object placements most closely aligned with human behavior. In learning to see the elephant in the room, both humans and SeCo reveal that scene understanding arises not from objects alone, but from the contextual associations that bind them together.

Text statistics

6 figures

8 supplementary figures

1 Introduction

A tiny object on a table might reasonably be a fork, but never an elephant. Object recognition relies not just on the visual features of individual objects, but also on learned context regularities: which objects tend to co-occur, their relative sizes, and their typical spatial relationships. For example, a fork often appears with a knife, is smaller than a vase, and is typically placed to the right of the plate.

While the field of visual neuroscience has long been captivated by the computational challenges of object recognition [83, 100, 56], most studies have centered on isolated objects [65, 68, 24]. There has been enduring interest in the statistics of natural scenes and foundational behavioral studies demonstrating the role of context in object recognition [5, 3, 4, 95, 74, 6, 1, 25, 11, 106], detection [10, 51], and search [22, 49]. Yet, how the brain learns to represent high-level contextual priors and integrates them with sensory input to support downstream tasks such as object recognition and detection remains poorly understood.

Prior neuroscience research on contextual modulation has primarily addressed low-level visual processes [36, 50, 71, 58], such as extra-classical receptive fields and surround suppression. While foundational, these mechanisms do not account for how high-level contextual knowledge is acquired, represented, or applied. Recent neurophysiological studies [29] in macaques have started to explore how high-level context shapes object recognition, yet the developmental origins of this knowledge remain largely unknown. These findings also highlight persistent gaps between human and AI performance in using context; however, there are few concrete strategies proposed for bridging this divide through improved computational models.

In recent years, interdisciplinary research in vision science and AI has made progress in modeling cortical processing pathways using artificial neural networks (ANNs) [52, 55, 68, 23, 81, 57, 82, 73, 47, 40]. Some of these ANNs partially explain responses in multiple areas in the visual cortex [12, 62, 101, 78, 8], yet they diverge from real-world object recognition in several ways [106, 38, 39, 76]. Humans actively move their eyes to gather context in a task-dependent manner, even when the target's appearance is uncertain and the surrounding scene varies in structure and resolution [106, 49]. To incorporate contextual understanding into computational modeling, some computer vision studies have employed statistical optimization techniques [42, 102, 61, 17], graph neural networks [9, 21, 27, 53], and transformer-based approaches [15, 11] across tasks such as object recognition [106, 11], object detection [66, 20], semantic segmentation [72], and visual question answering [89]. However, most of these models are trained on millions of labeled images [106, 11], which is misaligned with how humans learn from natural environments with minimal supervisory signals.

Evidence has shown that infants in early months form categorical representations of visually similar natural categories without supervision [80]. Self-supervised learning (SSL) algorithms have recently emerged as promising alternatives [75, 18, 19, 16, 103, 7, 99]. Although some models exhibit closer alignment with neural responses in cortical processing pathways in biological brains [108], they primarily focus on isolated single-object images and pay limited attention to inter-object relationships in naturalistic scenes. Investigating how humans acquire and apply contextual information with little supervision, along with the supporting neural mechanisms in neuroscience, provides valuable constraints that can guide the creation of more human-like and context-aware AI systems.

In our work, to address how humans learn contextual associations, we conducted human psychophysics experiments to quantify context learning and generalization in naturalistic scenes. Participants were first exposed to short training videos showing novel objects embedded in scenes following predefined contextual rules. These rules are based on global (e.g., room-level), local (e.g., co-occurrence), and crowding-based associations. We then tested their reasoning in a "lift-the-flap" task, where the central object was hidden and participants inferred its identity based on context. Despite no labels or feedback, participants rapidly learned these rules and generalized across variations in context area, resolution, and spatial layout.

To model this human capability, we introduce SeCo (Self-supervised learning for Context reasoning), a novel computational model for learning contextual associations. SeCo first identifies candidate target regions, then encodes the target and its surrounding context using separate visual encoders, mirroring the two visual pathways in the brain between ventral regions [93] (for object identity) and dorsal regions [44, 33, 2, 34] (for contextual and spatial layout processing). Inspired by semantic memory systems in the brain [3, 70, 69], SeCo incorporates a learnable external memory module analogous to the hippocampus and medial temporal lobe [87, 32]. This module stores latent contextual priors and retrieves them based on visual cues. In the "lift-the-flap" task, SeCo infers the hidden object from a

contextual cue by retrieving its associated memory representation and regressing it toward the target object, loosely echoing hippocampal pattern completion [84].

SeCo outperforms existing self-supervised baselines on the lift-the-flap task and shows stronger alignment with human behavior. Analysis reveals that its external memory encodes interpretable groupings of contextually related objects. We further evaluate SeCo and humans on an object priming task, where the goal is to place objects in context-appropriate locations. Results show that SeCo predicts placements that closely match human choices. Together, our work demonstrates that contextual scene understanding in both humans and machines emerges from learning the statistical structure that binds objects within its complex environments.

2 Results

We designed a two-stage experiment for humans and AI models: a training stage (learning to reason or LoR) and a test stage to measure their abilities to reason from varying contexts (**fig. 1D4**). For the test stage, we introduced two tasks, lift-the-flap, and object priming (**Fig. 1C**), to address "what" and "where" questions in context reasoning. During the LoR training stage, human subjects and AI models are tasked with learning contextual associations from the visual input of scenes, enabling them to understand relationships between objects and their environments for effective reasoning in subsequent tasks. In the test stage, the lift-the-flap task evaluates how humans and AI models use scene context to infer the category of a hidden object obscured by a black patch, relying solely on surrounding contextual cues. In a complementary manner, the object priming task assesses the ability to predict contextually appropriate regions within a scene to place a target object that is not initially present. These tasks provide a framework for testing contextual reasoning during inference across different scenarios. See **Methods** for details.

To minimize biases stemming from human familiarity with real-world scenes, we curated a scene dataset, *FRIbble in the sceNE* (FRINE). FRINE features novel contextual associations by replacing novel objects named *fribble* [86] (**Fig. 1D1**) with existing objects in the VirtualHome environment [79](**Fig. 1D2**). We defined a set of replacements between a *fribble* and a VirtualHome object as a *context rule*. We designed three distinct *context rules* to evaluate the reasoning abilities of humans (**Fig. 1D4**, see **Methods** for detailed context rules). To assess the robustness of human and AI models' performance on lift-the-flap, we introduced three context manipulations on the FRINE dataset (blur, reduced, and jigsaw context) (**Fig. 2C**). We designed a two-stage human psychophysics experiment: a training stage (LoR) and a test stage (lift-the-flap, object priming) to measure human learning and reasoning from varying contexts (**Fig. 2A**, see **Methods** for details). Specifically, subjects were initially exposed to 40 training video clips, each comprising 15 to 20 frames that rotate around the replaced *fribble* and are presented only once. Each frame features the *fribble* alongside its surrounding objects. After the training phase, subjects view 40 test video clips where the replaced *fribble* is hidden behind a black box mask. Each test video clip is shown for an unbounded duration, allowing subjects to observe the context and make a decision regarding the identity of the hidden fribble. To investigate whether humans can learn to reason from context in a self-supervised manner, we introduced two training modes in LoR (**Fig. 2ABD**). We refer to human participants trained via supervised learning as "SUP humans" and participants trained via self-supervised learning as "SSL humans" for short throughout the rest of the text. In lift-the-flap, we tested 160 human subjects in online experiments recruited through Amazon Mechanical Turk (AMT), randomly assigning participants to different training modes, context rules, and context manipulations, with appropriate quality control measures in place. In the object priming task, we recruited 437 human subjects in online experiments through AMT. See **Methods.** for details.

To investigate whether capturing contextual associations and consequently leveraging them during reasoning is a common capability of self-supervised learning (SSL) methods in AI, we considered several state-of-the-art baselines: Context Encoder [75], SimCLR [18], SimSiam [19], DINO [16], VICReg [7], ORL [99]. We further introduced a self-supervised learning method for context reasoning, named SeCo, which learns associations between objects and their surrounding contexts in scene images (**Fig. 3A**). When humans fixate on an object in a complex scene, they simultaneously perceive its surrounding objects in their peripheral vision (**Fig. 1A**). Similarly, SeCo incorporates an object-context discovery module, using an unsupervised region proposal method (selective search [91]) to locate potential objects of interest and features a two-stream visual processor, where objects are processed in high resolution while their surrounding contexts are processed at a lower resolution. As humans learn and memorize contextual associations after exposure to scenes, SeCo employs an external memory during the pre-training to simulate this underlying mechanism of context modulation. We evaluated SeCo and SSL baselines on both lift-the-flap and object priming tasks after fine-tuning. We curated two sets of natural scene datasets, COCO-OCD and COCO-VOC, to study AI models' reasoning ability in more complex contexts and their robustness against domain shifts. To ensure fair comparisons, all methods are initialized with ImageNet-pretrained weights [26], followed by pre-training and finetuning on the in-domain training set (**Fig. 3B**), then evaluated on the in-domain test set and in a zero-shot setting on the out-of-domain datasets (**Fig. 3C**). See **Methods** for detailed curation process and evaluation protocols.

2.1 Human and AI can learn to reason from context without supervisions

Humans can learn and reason from context. We first evaluated human performance on the curated FRINE dataset. Human learning ability varied substantially across individuals (**Fig. 2E**), consistent with pronounced inter-individual variability. Despite these differences, the top-1 accuracy of humans in the self-supervised learning (SSL) condition significantly exceeded that of a chance-level model, indicating that humans can effectively learn contextual associations and leverage them for reasoning even in the absence of explicit supervision (**Fig. 2**; SSL humans vs. chance, Welch's two-tailed t-test, $p < 0.05$, $t = 229.0$, $df = 1{,}998$). Nevertheless, SSL human performance remained well below ceiling, underscoring the intrinsic difficulty of the task and the challenge of reasoning based solely on contextual cues. The mean reaction time per trial for SSL humans was 13.07 ± 2.52 seconds (**Fig. 2F**). We further examined reaction time distributions separately for correct and incorrect trials (**Fig. 2G**). The mean reaction time was 13.98 ± 3.60 seconds for correct trials and 14.77 ± 4.33 seconds for incorrect trials. A significant difference in response times was observed between correct and incorrect trials (Welch's two-tailed t-test, $p < 0.05$, t = -2.5, df ≈ 621), which agrees with prior object recognition studies in which correct responses are typically faster [48].

To establish an upper bound on human performance under supervised learning (SUP humans), we provided explicit labels for fribbles during the LoR training stage (Fig. 1D4) and subsequently tested participants on the lift-the-flap task. As in the SSL condition, learning and reasoning performance varied across individuals; however, average top-1 accuracy for SUP humans was only marginally higher than that of SSL humans (**Fig. 3D**; SSL vs. SUP humans, Welch's two-tailed t-test, $p < 0.05$, $t = 3.9$, $df = 1{,}998$). This modest performance gain suggests that, for humans, self-supervised learning can be nearly as effective as supervised learning for acquiring and exploiting contextual information for reasoning. The mean reaction time per trial for SUP humans was 13.84 ± 3.70 seconds (**Fig. 2F**), which was significantly shorter than that observed for SSL humans (Welch's two-tailed t-test, p = 0.002, t = -3.1, df = 621). Reaction time distributions for correct and incorrect trials are shown in **Fig. 2G**. As in the SSL condition, correct trials were associated with slightly shorter mean reaction times (14.39 ± 3.78) than incorrect trials (15.30 ± 4.56); however, this difference did not reach statistical significance (Welch's two-tailed t-test, p = 0.057, t = -2.0, df ≈ 332).

To ensure data quality and participant attentiveness, we included a memorization test after each video presentation during the training phase, in which subjects were asked whether they had seen the exact same object instance in the preceding video (**Methods**). As shown in **Fig. S2**, 145 out of 160 human subjects using VirtualHome, and 156 out of 160 using ImageNet passed all data quality control criteria based on the memorization tests. To assess whether attentiveness influenced learning performance in SUP humans and SSL humans, we examined the relationship between memorization test performance and top-1 accuracy in the lift-the-flap task. We did not observe a strong correlation between the two (Spearman $\rho = 0.12$), suggesting that while most participants were attentive during training, learning ability in this task extends beyond mere visual attentiveness.

Second, we compared online participant performance with that of an in-laboratory cohort (**Methods**). To further enhance human performance, we increased training duration and introduced a structured curriculum in the in-lab setting. We report top-1 accuracy separately for online and onsite participants under both SUP and SSL conditions (**Fig. S5**). Onsite participants consistently outperformed the chance model and exceeded online participants by a substantial margin of 9.7% in top-1 accuracy. These results further confirm that humans can learn to reason from contextual information under both SUP and SSL regimes. The superior performance observed in the laboratory setting is likely attributable to extended training and improved curriculum design.

AI models can learn and reason from context with SeCo being the best. Similarly, we evaluated SSL methods on the lift-the-flap task using the FRINE dataset. Notably, none of the SSL methods had prior exposure to the FRINE dataset. Despite this, it is remarkable that these methods can transfer contextual knowledge learned from familiar objects to infer the identities of novel objects that share the same contextual rules, outperforming the chance model in a zero-shot setting (**Fig. 3D**, Welch's two-tailed t-test, $p < 0.05$). This result indicates that SSL methods can generalize learned contextual associations to novel objects governed by the same contextual rules without task-specific fine-tuning. Several SSL models (SeCo, ORL, and SimSiam) outperformed SSL humans (Welch's two-tailed t-test, $p < 0.05$) as well as the supervised baseline (Welch's two-tailed t-test, $p < 0.05$), highlighting their effectiveness in capturing and leveraging contextual structure without explicit labels.

However, similar to SSL humans, the performance of SSL methods remains far from perfect, indicating substantial room for improvement despite their strong generalization capabilities.

Importantly, SeCo outperformed both SUP humans and SSL humans, as well as all baseline methods, demonstrating its superior ability to reason from context (**Fig. 3D**; for all pairwise comparisons between SeCo and humans or other methods, Welch's two-tailed t-test, $p < 10^{-3}$). The Context Encoder [75], trained with pixel-level reconstruction, showed inferior performance compared to SeCo and other baselines, suggesting that pixel-level reconstruction emphasizes visual details while neglecting local contextual associations (for all pairwise comparisons between the Context Encoder and other models, Welch's two-tailed t-test, $p < 0.05$). Contrastive methods such as SimCLR [18] were outperformed by non-contrastive approaches like SimSiam [19], VICReg [7], and DINO [16], likely due to the challenge of selecting effective negative samples in contexts with multiple co-occurring objects (for all pairwise comparisons between SimCLR and other models, Welch's two-tailed t-test, $p < 0.05$).

Moreover, SeCo surpassed ORL [99], a patch-wise SSL method that relies on augmented views of the same object instances from the same input image or different object instances from similar contextual images. This suggests that SeCo is more effective in context reasoning by leveraging an external memory system to retrieve object representations associated with relevant contextual cues. This mechanism guides the context encoder in SeCo to learn accurate context associations, enabling reliable retrieval of target representations from memory. By introducing external memory, SeCo transitions from object-centric to object-context associative representation learning, thereby enhancing context reasoning. Notably, SeCo also outperforms the supervised learning baseline, demonstrating that object-context associative representations can be learned without supervision.

Human and SeCo performance vary across context rules and context associations. Different context rules impose distinct relational and inferential demands. Examining human and model reasoning across these rules allows us to probe how context reasoning strategies adapt under varying levels of structural complexity. Consistent with this, both humans and AI models showed significant variation in context reasoning performance across rules (**Fig. 3E**, one-way ANOVA, $p < 0.05$). In Rule 1, humans and AI models achieved comparable performance (**Fig. S3A**). However, in Rule 2, SUP humans outperformed SSL humans ($p < 0.05$, Welch's two-tailed t-test, $t = 30.6$, $df \approx 1,998$), whereas the reverse trend was observed in Rule 3 ($p < 0.05$, Welch's two-tailed t-test, $t = -22.5$, $df \approx 1,998$). Notably, SeCo consistently outperformed both groups of human subjects and other AI models in Rules 2 and 3 (all comparisons, Welch's two-tailed t-test, $p < 0.05$), demonstrating its robustness in handling the varying complexity of these context rules. Despite these variations, both humans and SeCo maintained above-chance top-1 accuracy across all rules.

Different objects are associated with varying contextual cues, rendering some objects easier to reason about than others. To examine how context association types influence reasoning ability, we categorized virtual home objects into three groups—global, local, and crowding—based on the nature of their contextual associations (**Methods**). For example, microwaves (global) typically appear in kitchens, providing a strong and unambiguous contextual cue, whereas cups (local) can occur across diverse environments, making context-based reasoning more challenging (**Fig. 1B**).

Across association types, SUP humans, SSL humans, and SeCo all exhibited higher top-1 accuracy for targets with strong global associations compared to those with local associations (for all comparisons, Welch's two-tailed t-test, $p < 0.05$). A similar pattern was observed across all model baselines (ORL, VICReg, SimSiam, Context Encoder, and Supervised), with each model performing significantly better on global than on local associations (**Fig. S3**; for all within-model comparisons, Welch's two-tailed t-test, $p < 0.05$).

Interestingly, both SUP and SSL humans achieved higher performance under the crowding condition than under global or local association types, suggesting a human advantage in leveraging dense contextual structure. This trend did not generalize to all AI models; nevertheless, SeCo consistently outperformed other AI models in the crowding condition. Notably, despite these variations across association types, both humans and SeCo maintained above-chance top-1 accuracy across all context association categories.

2.2 Contextual variations influence reasoning ability in humans and AI

To evaluate how well humans and AI models can maintain their reasoning abilities when typical visual and spatial cues are disrupted, we analyzed their performance of the lift-the-flap task on the FRINE

dataset with the normal context and the manipulated contexts, including blurred, reduced, or jigsaw context (**Fig. 2C**).

Blurred context is sufficient for reasoning. Blurring the contextual region removes fine-grained visual details while preserving the global scene structure, allowing us to test whether humans and models rely primarily on low-level visual features or higher-level semantic cues for context reasoning. To quantify the effect of context resolution, we applied zero-mean Gaussian blur to the context with standard deviations of $\sigma = 2, 4, 8, 16$, and 32 pixels (video resolution: 1024×1280 pixels). Each subject experienced all blurring conditions, but never for the same target objects within the same scene, preventing learning, memory, and adaptation effects from confounding performance across conditions.

Human top-1 accuracy decreased as contextual blur increased from $\sigma = 0$ (unblurred) to $\sigma = 32$ (**Fig. S4A**; one-way ANOVA, $p < 0.05$). Interestingly, even heavily blurred context remained sufficient for successful reasoning in the lift-the-flap task, with human performance consistently exceeding chance across all blur levels (**Fig. 4A**; Welch's two-tailed t-tests against chance, $p < 0.05$). These results align with prior findings showing that global scene structure can support in-context object recognition despite substantial loss of fine-grained visual detail in the background [106].

Specifically, SSL humans performed on par with SUP humans across almost all blur levels (**Fig. 4A**; Welch's two-tailed t-test, $p < 0.05$). Likewise, SeCo effectively reasoned from blurred contexts across all degrees of blurring (Welch's two-tailed t-tests against chance, t = 800.0, $p < 0.05$, df $\approx$ 1,947), although performance showed a dramatic decline under extreme blurring conditions. Most SSL models (SeCo, ORL, SimSiam, VICReg, and DINO) consistently outperformed the supervised model under mild to moderate contextual blur (**Fig. S4A**, $\sigma \leq 16$; for all comparisons, Welch's two-tailed t-test, $p < 0.05$). Performance degraded substantially under extreme blur ($\sigma = 32$), leaving SeCo as the only SSL method that continued to outperform the supervised model (**Fig. S4A**; for all comparisons, Welch's two-tailed t-test, $p < 0.05$). This robustness can be attributed to SeCo's two-stream visual processing architecture, which encodes contextual information at low resolution while preserving high-resolution representations of target objects. This design mirrors the human periphery–fovea division in visual processing and supports reliable reasoning under extreme blurring by prioritizing global contextual cues over fine-grained visual details.

Contextual areas matter for reasoning. In reduced-context scenarios, portions of the scene are removed or substantially minimized, leaving less information for both humans and models to infer object identity. We quantified the amount of available context using the context–object (CO) ratio [106], with values of 128, 16, 8, 4, and 2, to assess how effectively humans and models reason under limited contextual information (**Fig. 4B**).

Both humans and SSL models were able to reason from limited context, performing significantly above chance across all CO ratios (**Fig. 4B**, **Fig. S4B**; for all sub-conditions, two-tailed t-test against chance, $p < 10^{-3}$). However, progressively reducing contextual information led to a significant decline in performance for both humans and SSL models (for all comparisons, one-way ANOVA, $p < 0.05$). Under extreme contextual reduction (CO = 2 and 4), SSL humans significantly outperformed SUP humans (for all comparisons, Welch's two-tailed t-test, $p < 0.05$). Similarly, SeCo was more robust than all other AI models, including the supervised baseline, under these minimal context conditions (for all comparisons, Welch's two-tailed t-test, $t = 522.0$, $p < 0.05$, df $\approx$ 1,954).

Context reasoning relies on spatial configuration. Jigsaw context disrupts the spatial coherence of a scene by scrambling the locations of surrounding objects while preserving the same set of visual elements. This manipulation challenges the extent to which humans and models rely on spatial relationships and scene layout for context reasoning. Specifically, we partitioned the context into 3×3, 5×5, and 7×7 grids and randomly permuted the pieces, while keeping the central patch containing the target object intact (**Fig. 4C**).

Context reasoning performance for both humans and AI models was strongly affected by spatial configuration (**Fig. 4C**, **Fig. S4C**; one-way ANOVA, $p < 0.05$). However, both SUP and SSL human performance under the 3×3 and 5×5 jigsaw conditions did not differ significantly from the unscrambled context condition, likely because each larger patch retained sufficient local contextual information and because contextual influence decreases with distance from the target [45] (Welch's two-tailed t-test, $p < 0.05$). Overall, humans were more robust to spatial disruption than AI models. Among all the AI

models, SeCo outperformed baseline methods under smaller-scale spatial disruptions (3×3 and 5×5), but its performance degraded under highly fragmented 7×7 contexts (for all comparisons, Welch's two-tailed t-test, $p < 0.05$).

SeCo behaviors align closely with humans across contextual variations. To assess whether SeCo captures not only overall performance but also human-like sensitivity to different forms of contextual manipulation, we examined the alignment between model and human behaviors across multiple context conditions. Various SSL models (SeCo, ORL, and VICReg) exhibited strong linear correlations with SSL human performance across four contextual conditions (all $r \geq 0.8$; **Fig. 4D**). Notably, SeCo achieved higher absolute accuracy than ORL while maintaining comparably high correlations with human performance. Despite these overall similarities, SeCo and SSL humans showed complementary strengths across context manipulations: SeCo outperformed SSL humans under reduced-context conditions, whereas SSL humans achieved higher accuracy under blur and jigsaw contexts.

2.3 SeCo generalizes context reasoning in natural scenes

To further assess the reasoning capability of SeCo and the rest of SSL methods, we considered the real-world scenario where scenes are inherently more complex, containing diverse objects, environments, and spatial relationships. First, we evaluated the performance of SSL methods against the chance model on the in-domain sets of COCO-OCD and COCO-VOC (**Fig. 5A1, A3**). Our results show that all SSL methods successfully reason from complex, real-world scenes (for all pairwise comparisons between SSL methods and the chance model, Welch's two-tailed t-test, $p < 0.05$). Among them, SeCo and ORL consistently outperform the supervised baseline, highlighting their effectiveness in capturing meaningful associations in diverse environments without supervision (Welch's two-tailed t-test, $p < 0.05$). Consistent with observations from the FRINE dataset, SeCo surpasses ORL by a larger margin on both in-domain COCO-OCD and COCO-VOC datasets (Welch's two-tailed t-test, $p < 0.05$). These results suggest that SeCo demonstrates a superior ability to learn and reason from context across diverse and complex naturalistic environments.

Next, we evaluated whether SSL methods remain robust to domain shifts in visual features during context reasoning tasks—for example, recognizing a bird flying in the sky regardless of whether the scene is depicted in Picasso or Monet styles. We evaluated all models on out-of-domain OCD test of COCO-OCD and PASCAL-VOC07 test of COCO-VOC in a zero-shot setting without fine-tuning. Compared to the chance model, all SSL methods retained their reasoning ability under domain shifts, with SeCo consistently outperforming all baselines across both out-of-domain datasets (**Fig. 5A2, A4**; Welch's two-tailed t-test for all SeCo and baseline comparisons, $p < 0.05$). In particular, performance on OCD test was still comparable to in-domain COCO-VOC for all SSL methods (**Fig. 5A4**; one-way ANOVA, $p < 0.05$), likely because the in-domain and out-of-domain COCO-VOC datasets consist of natural images. In contrast, the shift from natural images in in-domain COCO-OCD to synthetic images in out-of-domain OCD test led to a substantial performance drop across all models (**Fig. 5A2**; one-way ANOVA, $p < 0.05$). Importantly, SeCo was the least affected by this domain shift, highlighting its ability to capture contextual associations beyond superficial visual feature correlations.

2.4 Dissecting SeCo reveals critical components for context-based reasoning

SeCo learns meaningful contextual associations via selective search. Objects play a critical role in context reasoning [28, 11]. To facilitate the learning of object–object and object–context associations, we introduce a context–object pair discovery module that leverages selective search [91] to propose regions likely to contain objects of interest (**Methods**).

To assess the effectiveness of this discovery module, we replaced object–context image pairs generated by selective search with pairs derived from annotated ground-truth bounding boxes. Surprisingly, SeCo achieved higher performance with selective search than with ground-truth pairs (**Fig. 5B, Ground Truth**), suggesting that selective search proposals preserve richer or more diverse contextual cues than tightly localized ground-truth annotations. To further test whether selective search yields more informative contextual associations than random pairings, we replaced selective search proposals with randomly generated object–image pairs, differing only in the randomization of target region coordinates. As expected, random generation serves as a lower bound for the discovery module (**Fig. 5B, Random Generation**), highlighting the importance of region "objectness" for learning meaningful contextual

associations. We also compared selective search with the widely used random cropping strategy from the SSL literature, in which the input image is duplicated and randomly cropped. Replacing selective search with random cropping resulted in a clear performance drop, indicating that selective search's object-centric region proposal mechanism is more effective for learning contextual associations (**Fig. 5B, Random Cropping**).

Finally, to examine the role of selective search during pre-training, we analyzed both the quantity and quality of its region proposals (**Methods**). Increasing the Intersection over Union (IoU) threshold with ground-truth bounding boxes did not improve performance, whereas including a larger number of proposals led to modest gains (**Fig. 5C**). This suggests that proposal diversity and coverage, rather than precise spatial localization, are more critical for effective context learning, providing a plausible explanation for why selective search outperforms ground-truth bounding boxes in SeCo.

SeCo stores context associations in the external memory. To study the role of external memory in contextual reasoning, we removed the external memory module from the default SeCo model (see **Methods**). Eliminating external memory resulted in a 12.5% drop in top-1 accuracy, highlighting the importance of external memory for learning object–context associations.

We further varied the number of memory slots (**Fig. S7A**) from 100 to 800, observing a moderate improvement of 2.5% in top-1 accuracy on the lift-the-flap task. Next, we examined the effect of memory slot dimensionality on lift-the-flap performance (**Fig. S7A**) and found a non-monotonic relationship with top-1 accuracy. Performance peaks at a feature dimension of 512, suggesting that increased memory capacity generally facilitates the learning and storage of richer context–object associations. However, excessively large memory capacity may impair contextual reasoning, likely due to overfitting and reduced generalization of learned contextual knowledge.

To probe the representations learned by the external memory, we visualized the pairwise Kullback Leibler (KL) divergence of attention scores across memory slots for object categories in COCO-VOC (**Fig. 5E**). Each matrix entry quantifies the divergence between the memory-slot attention distributions used to retrieve contextual information for a pair of object categories (see **Algo. 2** for details). Darker values indicate greater similarity in the attended memory slots, reflecting a higher likelihood of shared contextual associations between object classes. Across supercategories—including vehicles, animals, and indoor objects, we observe clusters of contextually related object pairs. For example, televisions and potted plants, despite limited visual similarity, exhibit strong overlap in attended memory slots, consistent with their shared indoor context. These patterns suggest that SeCo's external memory encodes meaningful object–context associations that go beyond low-level visual similarity.

Distinct target and context representations enable relational reasoning. SeCo was trained with two separate encoders, $E_t(\cdot)$ and $E_c(\cdot)$, for the target and context streams, respectively. We also evaluated a variant with weight-sharing between the encoders (**Fig. 5B, Shared Encoder**). This shared-encoder baseline achieved lower top-1 accuracy than the full SeCo model, indicating that using identical features for both target and context streams is insufficient for effective context reasoning. Separating the encoders allows the model to extract complementary representations for targets and their context, which facilitates the integration of relational information necessary for accurate inference.

Variance and covariance losses are critical for stable pre-training. SeCo has a joint loss of MSE loss, covariance loss, and variance loss. Here, we remove one loss at a time to analyze its effectiveness on pretraining. We report top-1 accuracy on the in-domain test set of COCO-OCD in **Fig. S7B**. The result demonstrates that without variance loss, SeCo reached information collapse, aligning with the trend in VICReg [7]. Without covariance loss, performance drops 2% in accuracy. Different from the observations made in VICReg [7], without MSE loss, SeCo manages to achieve 41.72% in top-1 accuracy without collapses. One possible reason is that starting from weights obtained on ImageNet [26], the encoder has captured useful visual features. Thus, adding information regularization during pre-training can avoid collapse even without enforcing association between contexts and targets.

2.5 SeCo excels in object priming task

Human vision is characterized by a remarkable ability to recognize spatial relationships and contextual relevance, enabling intuitive expectations about where objects should appear within a scene based on

surrounding cues (**Fig. 1B2**). To test whether SSL models exhibit a similar capacity for context-aware object placements, we designed an object priming task for both humans and AI models.

We curated the *Human Object Priming* (HOP) dataset, which comprises semantically meaningful image–target object pairs drawn from the COCO-Stuff test set [13] (see **Methods**). For each image–target object pair, human priming maps were collected from participants on Amazon Mechanical Turk, who clicked contextually appropriate locations within a scene to place the target object (**Fig. 6A1**; **Methods**). To minimize crowding effects, target objects were removed from the original scenes. Aggregating these mouse-click responses yielded human object priming maps (**Fig. 6A2**; **Methods**). In parallel, AI models performed the same task on the same image–object pairs to generate machine priming maps (**Algo. 3** and **Methods**).

Human and model priming maps were compared using a similarity metric, referred to as the "alignment score" (**Methods**). SeCo achieved the highest alignment with human priming maps among all baselines, indicating superior ability to predict contextually relevant object placement (**Fig. 6B**). Qualitative visualizations further illustrate these differences (**Fig. 6C**): while baseline models often produced diffuse or context-insensitive priming maps, SeCo consistently identified semantically plausible object locations. Importantly, none of the models were trained or fine-tuned using human priming data, highlighting SeCo's remarkable capacity to transfer learned contextual associations to infer semantically appropriate, target-relevant locations within complex scenes.

3 Discussion

We examined how humans acquire and deploy contextual priors to reason about objects embedded in naturalistic scenes and introduced a biologically inspired computational model, SeCo (Self-supervised learning for Context reasoning, **Fig.3A**), that captures essential aspects of this capability. Across psychophysics experiments (**Fig.2A-D**), participants rapidly learned contextual rules from short, unlabelled videos and successfully inferred the identity of hidden objects in a "lift-the-flap" task. These findings demonstrate that humans can rapidly internalize environmental statistics and apply them to guide inference under uncertainty.

Strikingly, participants in the self-supervised condition—who received no labels or feedback—performed competitively relative to those in the supervised condition (**Fig.2E**). This suggests that explicit instruction is not necessary for acquiring structured contextual knowledge. Nevertheless, supervision modulated the efficiency of reasoning: reaction times were shorter in the supervised group (**Fig.2F**), indicating that labeled guidance may accelerate access to contextual priors. Across both learning regimes, correct trials were associated with faster responses than incorrect trials (**Fig.2G**), consistent with successful retrieval of coherent contextual representations rather than guess-based decision-making.

Contextual reasoning performance improved with longer training durations and curriculum-style progression (**Fig.S5**), suggesting that gradual exposure and structured experience enhance the formation of contextual priors. Because humans are already highly familiar with everyday objects and their real-world associations, we curated a novel object dataset, named as FRINE dataset, to minimize the influence of prior semantic knowledge (**Fig.1D1-D4**). Specifically, we introduced four unfamiliar objects ("Fribbles") per rule, each systematically replacing a distinct anchor object from daily environments. Within each rule, these four anchor objects defined a coherent contextual structure, and the associations were expressed through three complementary context types: global scene-level context (e.g., room-level regularities), local co-occurrence relationships, and spatial clustering patterns. This design allowed us to simulate structured yet controlled contextual environments while preventing reliance on pre-existing object knowledge. Across all three rules and all context association types, humans successfully learned and generalized the underlying structure (**Fig.3E,F**). Generalization across variations in spatial layout, context area, and resolution (**Fig.4A,B,C**) further indicates that contextual knowledge is represented at an abstract level rather than tied to specific visual configurations. Together, these behavioral findings suggest that humans learn flexible, transformation-invariant representations of object–context structure that support rapid inference even in the absence of direct object visibility.

SeCo captures these human behaviors through a computational architecture inspired by dual visual processing pathways [93, 44, 33, 2, 34] and hippocampal memory systems [87, 32] (**Fig.3A**). Separate encoders process candidate target regions and surrounding context, loosely paralleling ventral pathways for object identity and dorsal or scene-selective regions for spatial layout. A learnable external memory module stores latent contextual associations and retrieves them when only contextual cues are available. In the lift-the-flap task, SeCo performs inference by retrieving context-linked representations and regressing toward the likely target identity, analogous to hippocampal pattern completion.

Importantly, most existing computational studies of contextual reasoning rely heavily on supervised learning with large-scale annotated datasets to model object co-occurrence and scene structure [5, 3, 4, 95, 74, 6, 1, 25, 11, 106]. Even recent self-supervised learning approaches primarily focus on learning object-centric representations from isolated instances, optimizing invariances within single-object crops or global image embeddings [75, 18, 19, 16, 103, 7, 99]. As a result, these models excel at recognizing what an object is but are not explicitly designed to learn how objects relate to one another within scenes (**Fig.3D, Fig.4D, Fig.5A1-A4, Fig.S3, Fig.S4**). In contrast, SeCo is explicitly structured to learn context–object associations without category labels, embedding relational structure within an external memory system that can be queried for inference. By shifting the focus from object identity learning to structured relational learning, SeCo provides a computational framework for modeling how contextual priors can emerge from natural visual experience with minimal supervision.

The external memory structure in SeCo reveals interpretable clusters of contextually related objects (**Fig.5DE, Fig.S7**), suggesting that contextual knowledge emerges as structured latent organization rather than correlated visual similarities. This parallels evidence from cognitive neuroscience proposing that semantic memory encodes relational structure linking objects within shared environments [3, 70, 69]. The ability of SeCo to retrieve context-conditioned representations demonstrates that contextual reasoning can emerge from self-supervised learning when models are explicitly equipped with mechanisms for structured memory storage and retrieval.

Beyond average task accuracy, it is important to examine the nature of generalization. Humans successfully transferred learned contextual rules across variations in scene appearance and spatial configuration, indicating that contextual knowledge is not tied to superficial image statistics. Similarly, SeCo generalized beyond training layouts (**Fig.3D, and Fig.5A2,A4**), suggesting that its memory module encodes relational structure rather than memorizing individual visual exemplars. In an object priming task where both humans and machines are asked to place the target object in a contextually relevant region in the scene, SeCo's predicted placements closely matched human choices (**Fig.6**), further supporting the idea that both humans and the model learn probabilistic priors governing object-context compatibility.

Even when SeCo approximates human contextual reasoning (**Fig.4D,E** and **Fig.6B**), the mechanisms may differ in important ways. First, the current model assumes explicit candidate region identification, whereas humans may dynamically allocate attention and integrate context over sequential fixations [105, 96, 28, 104]. Second, the memory module in SeCo is implemented as a static external store; biological memory systems likely involve recurrent dynamics, consolidation processes, and interaction with broader cortical networks [46, 85, 98, 30]. Third, learning in SeCo occurs through gradient-based optimization over curated training images, whereas humans acquire contextual knowledge through lifelong multimodal video experience with rich sensory feedback [14, 107, 54]. Fourth, the current experiments focus on predefined contextual rules; natural environments exhibit far more complex and hierarchical statistical dependencies [97]. Finally, SeCo does not yet incorporate high-level semantic reasoning, causal understanding, or world knowledge that may further constrain human contextual inference (e.g., physical plausibility [92] or social conventions [63, 77]).

These limitations suggest several promising directions. Future work could incorporate active visual exploration, allowing models to gather contextual information sequentially. Integrating multimodal cues, such as language or action affordances, may more closely approximate real-world learning. Modeling developmental trajectories, such as how contextual priors emerge over time, could further bridge computational models and developmental psychology literature. Finally, incorporating adaptive memory updating and cross-task transfer may clarify how contextual reasoning interacts with broader cognitive systems.

The present findings advance our understanding of how contextual knowledge is acquired and applied. Rather than viewing context as a post hoc modulation of object recognition, our results suggest that contextual structure is learned as a core component of scene understanding and can guide inference even when direct sensory evidence is absent. SeCo demonstrates that self-supervised learning, when coupled with structured memory mechanisms, can approximate this human capability. Together, our behavioral and computational results support the idea that contextual scene understanding emerges from learning the statistical structure that binds objects within their environments, enabling invariant, rapid, generalizable reasoning about what belongs and what does not within a given scene. In doing so, we move closer to understanding how humans and machines learn to see the elephant in the room.

4 Methods

4.1 Human Psychophysics Experiments

To evaluate human capacity for learning contextual associations and reasoning from context, we conducted three psychophysics experiments. In the first experiment, participants learned novel contextual rules and applied them to infer the identity of a hidden object in synthetic images, a task we refer to as the *lift-the-flap* task. We systematically varied experimental parameters, such as the size, resolution, and spatial configuration of contextual cues, to examine the robustness and invariance of context-based reasoning. In the second experiment, we extended the training duration and introduced a structured curriculum to further enhance the ability to acquire and apply novel context-object associations. The third experiment used real-world scenes in an object priming task, where participants placed a given object in a contextually appropriate location within the scene. Together, these experiments assess human ability to learn contextual rules and apply them to infer what an object is and where it belongs, spanning tasks from simple synthetic setups to complex naturalistic environments. All participants had normal or corrected-to-normal vision, provided written informed consent, and were compensated at a rate of 20 SGD per hour. Experiment durations varied: the first and the second lasted approximately 40-60 minutes, and the third lasted approximately 6 minutes. All studies were approved by our Institutional Review Board.

4.1.1 Learning to Reason from Novel Context Rules

Fribble in the Scene (FRINE) Dataset. Humans accumulate years of experience navigating complex scenes in daily life. To assess how effectively they can learn and apply contextual associations independent of prior knowledge, we introduced novel objects and constructed new contextual associations between these unfamiliar objects and familiar household environments. Specifically, we developed the Fribble in the Scene (FRINE) dataset using novel object families from the NOD dataset [86]. These fribbles, characterized by distinct body structures and appendages, are unfamiliar to humans. We synthesized realistic indoor scenes using the Unity-based VirtualHome environment [79], which contains seven apartments and five types of furnished rooms such as kitchens and bedrooms. By embedding fribbles into these environments, we ensured that the contextual associations encountered by participants were both ecologically valid and entirely novel.

Context associations. To study familiar contextual associations with novel objects, we replaced common household items in VirtualHome with fribbles. Specifically, we selected eight VirtualHome objects (**Fig. 1D**): *keyboard*, *mouse*, *cake*, *cup*, *remote*, *microwave*, *toothbrush*, *pillow*, and *knife*, based on the types of contextual associations they exhibit (**Fig. 1B**, *right*). First, *global association* refers to objects that appear exclusively in a specific place without co-occurrence of other object categories; for example, toothbrushes appear only in bathrooms and microwaves only in kitchens. Second, *local association* refers to objects that either appear in multiple locations or co-occur with other types; for example, keyboards always appear with computer mice, but can be located in the common room, study, or bedroom. Third, the *crowding effect* describes cases where multiple instances of the same object appear together, such as eggs grouped in dozens or knives arranged in a set.

Context rules. To curate context rules, four families of fribbles in the NOD dataset (*Fa1*, *Fb1*, *Fb3*, *Fc1*) were to match with four out of the selected eight household objects (**Fig. 1D4**). Consequently, there were $A_4^8 = 1680$ possible combinations of replacement choices, which is impractical to handle. Therefore, we manually curated three combinations (in the format of the fribble-VirtualHome object to be replaced), where each specific combination is considered as a novel context rule: (1) Rule 1: Fa1-microwave, Fb1-mouse, Fb3-knife, Fc1-cup; (2) Rule 2: Fa1-pillow, Fb1-keyboard, Fb3-yoothbrush, Fc1-mouse; (3) Rule 3: Fa1-mouse, Fb1-cup, Fb3-pillow, Fc1-cake.

Object-centric videos. Humans perceive the world as a continuous stream of visual input in complex scenes [86]. To loosely approximate this visual experience, we generate video clips in which the fribble object of interest remains centered in each frame while the camera rotates around it. In the NOD dataset [86], each fribble family consists of five object instances (e.g., the Fa1 family includes Fa1_1231, Fa1_1322, Fa1_2333, Fa1_3123, and Fa1_3212). To curate video clips, we randomly select

one instance from each family and replace the corresponding household object in the VirtualHome environment, following the specified context rule. A camera is configured to rotate around the fribble while maintaining a fixed attitude angle, capturing a maximum of 20 fixed viewpoints per clip. These viewpoints are spaced by 18 degrees in azimuth, ranging from 0 to 342 degrees, with a constant radius of 1.5 meters. To enrich the range of object views, we vary the camera's elevation angle across four levels: 0, 20, 40, and 60 degrees, generating multiple videos for the same object. Unity's built-in ray casting and collision checking libraries are used to ensure that the camera does not collide with or get occluded by surrounding objects. Videos are recorded at 1 frame per second with a resolution of 1280×1024 pixels, subtending 24.4×19.6 degrees of visual angle. Any frames affected by occlusions or collisions are skipped, resulting in video durations ranging from 10 to 20 seconds.

Training phase of learning to reason. We recruited 160 human participants using Amazon Mechanical Turk (MTurk), with subjects' informed consent. These participants were drawn from a pool of "master workers" who had completed at least 100 approved Human Intelligence Tasks (HITs) with a 95% approval rate. In the training phase, each subject was randomly assigned to one of three context rules. Each subject viewed 40 training videos, with 10 videos sampled per Fribble family. To assess whether participants could learn the context rules without explicit supervision, we introduced two learning modes:

- **supervised learning (SUP)**, where red bounding boxes highlighting the target object and its label (e.g., "fa1") were provided.
- **self-supervised learning (SSL)**, where only the bounding boxes were shown without any labels.

Subjects assigned to the same context rule and learning mode viewed the same 40 training videos. The videos were randomly shuffled and presented only once, with each lasting between 10 and 20 seconds. Video resolution was downsampled to 640×512 pixels subtending 12.8×10.1 degrees in visual angles to minimize internet delays (**Fig. S1A1**).

To ensure data quality and participant attentiveness, we included a memorization test with a 50% probability after each video (**Fig. 1D1**, **Fig. S1A2**). In each test, subjects were shown a visual stimulus and asked whether they had seen the exact same object in the preceding video. These stimuli were sampled from one of the following three categories:

- (1) **VirtualHome-Present**: A positive sample cropped from the last frame of the current training video.
- (2) **VirtualHome-Absent**: A negative sample cropped from an object in a different apartment, ensuring it was not present in the current video.
- (3) **ImageNet-Absent**: A negative sample randomly drawn from ImageNet [26], guaranteed to be absent from all training videos.

Each category was sampled with equal probability (33.3%), and participants were asked: "Did you see the EXACT SAME object instance in the last video?" Subjects made a binary decision. To pass the quality control, participants were required to make fewer than six errors across the VirtualHome-Present and VirtualHome-Absent tests, and fewer than two errors on ImageNet-Absent tests. The memory recall error rates are reported in **Fig. S2A1, B1**. Only data from participants who passed the memorization tests were included in the final analysis.

Test Phase on Lift-the-Flap: What is the target object? Following the training phase, participants entered the test phase involving a lift-the-flap task designed to probe the "what" aspect of context reasoning (**Fig. 1C1**). Each participant was randomly assigned a unique test set consisting of 40 test videos. In this task, human subjects were required to rely solely on contextual information to infer the class identity of a hidden target object in each video. The target object was obscured by a black patch, and the video could be viewed an unlimited number of times. All videos were presented at a resolution of 640×512 pixels, subtending 12.8×10.1 degrees of visual angle, and had a duration ranging from 10 to 20 seconds (**Fig. S1B1**).

Context manipulation. To evaluate the robustness of human reasoning under altered visual contexts, we introduced four types of context manipulations in the test videos:

- (1) **Normal context**: No manipulation applied.
- (2) **Blur context**: Gaussian blur applied to the context (i.e., all video regions excluding the black-patched target area), with one of five kernel sizes (σ = 2, 4, 8, 16, 32 pixels for frame size = 1024×1280 pixels).
- (3) **Context area**: Context was removed according to five different context-object (CO) ratios (128, 16, 8, 4, 2), where the CO ratio refers to the proportion of image area excluding the target object, relative to the size of the target object.
- (4) **Jigsaw context**: Context was scrambled into non-overlapping pieces of size 3×3, 5×5, or 7×7 and then shuffled.

Each manipulation was applied only to the surrounding context, leaving the target region untouched. For the CO=128 condition in the context area manipulation and the 7×7 jigsaw condition, we excluded cases where the fribble object occupied more than 0.5% of the image area. Each test video was randomly assigned to one of the four context manipulation types (normal, blur, area, jigsaw) with equal probability (25%). If a video was assigned to blur, area, or jigsaw, it was then randomly assigned to one of its sub-conditions with a probability of 20%, 20%, or 33.3%, respectively.

To ensure data quality and participant attention during the test phase, we embedded three dummy checks. Specifically, we selected one image each of a dog, cat, and fish from ImageNet [26]. These images were presented after the 10th, 20th, and 30th test videos, with the question "Is this object a *[LABEL]*?", where "*[LABEL]*" was randomly substituted with a label, either "dog", "cat", or "fish". Participants were required to make a binary judgment on whether the shown object matched the given label (**Fig. S1B2**). Only those who correctly answered all three dummy tests passed the quality check. Out of the 160 participants, 108 passed the quality controls for both the training and test phases, resulting in 4320 valid trials. Each context sub-condition within each learning mode included an average of at least 85 trials.

4.1.2 Learning to Reason with Longer Training and Better Curriculum

In education, students follow a structured curriculum where each new concept builds on and reinforces prior knowledge. Thoughtful curriculum design is essential for improving learning in humans [86, 88]. To investigate whether human reasoning from context can be further improved, we extended the training duration and implemented a structured curriculum in a controlled lab setting, free from distractions to ensure participants remained attentive throughout the experiment. We recruited a total of 20 participants. Each was assigned to learn only Context Rule 1 and randomly placed into either a supervised or self-supervised learning mode. All subjects viewed the same set of 40 training videos corresponding to Context Rule 1 but each video was shown multiple times according to the structured curriculum design: (1) in the first stage, 10 videos from each fribble category were shown sequentially; (2) in the second stage, 5 videos from each category were interleaved; (3) in the final stage, 10 videos were presented from the two most difficult fribble categories identified in earlier experiments (Fb3 and Fc1). In total, each participant viewed 80 training videos, each with a resolution of 640×512 pixels subtending 12.8 × 10.1 degrees in visual angles and their durations ranging from 10 to 20 seconds. As in the previous online experiments, memorization tests were included during training.

Participants then proceeded to the same testing phase as in the previous online experiments, performing the lift-the-flap task. For consistency, each subject was randomly assigned a unique test set of 40 videos drawn from the online experiment pool. All four context manipulations, and the dummy checks during testing, were retained. A total of 15 out of 20 participants passed the quality control criteria, resulting in 600 valid trials, with at least 75 trials per context manipulation condition per learning mode on average.

4.1.3 Lift-the-Flap: What Is the Target Object behind the Flap?

COCO-OCD and COCO-VOC Datasets

To further evaluate the reasoning capabilities of AI models, we considered the real-world scenario with inherently complex scenes featuring diverse objects, environments, and unpredictable spatial relationships. We specifically considered one base natural scene dataset COCO-Stuff [13], which originally contains 160K natural images from MSCOCO [64] with 80 thing classes and 91 stuff classes

in total. Importantly, this dataset captures complex relationships between multiple objects and carries rich contextual information. To evaluate whether the learned contextual knowledge from AI models can generalize well in out-of-domain settings, we considered two base datasets: (1) Out-of-Context Dataset (OCD) [11], which originally contains 15,773 synthetic test images of indoor scenes with 36 classes under 6 different contextual conditions. In our work, we only consider the normal context condition with 2,309 test images. (2) PASCAL-VOC07 Dataset [35], which originally contains 9,963 images of realistic scenes with a total of 20 object classes. We designed two custom datasets, curated from COCO-Stuff, OCD and PASCAL-VOC07 as follows (**Fig. S8**).

COCO-OCD. The custom dataset includes one training set derived from COCO-Stuff and two test sets derived from COCO-Stuff and OCD. The training and test images are from 15 overlapping classes shared between the two base datasets: wine glass, cup, knife, bowl, apple, cake, mouse, remote, keyboard, cell phone, microwave, book, toothbrush, pillow, and towel. We refer to test images from COCO-Stuff as in-domain COCO-OCD, and test images from OCD as out-of-domain COCO-OCD.

COCO-VOC. The custom dataset includes one training set derived from COCO-Stuff and two test sets derived from COCO-Stuff and PSACAL-VOC07. The training and test images are from 20 overlapping classes shared between the two base datasets, in a hierarchy of *superclass* and subclass: (1) *Person*: person; (2)*Animal*: bird, cat, cow, dog, horse, sheep; (3) *Vehicle*: aeroplane, bicycle, boat, bus, car, motorbike, train; (4) *Indoor*: bottle, chair, dining table, potted plant, sofa, tv/monitor. We refer to test images from COCO-Stuff as in-domain COCO-VOC, and test images from PASCAL-VOC07 as out-of-domain COCO-VOC.

4.1.4 Object Priming: Where to Put the Target Object in the Scene?

To further evaluate the reasoning capabilities of human subjects, we introduce the object priming task to address the problem of "where" in context reasoning (**Fig. 1C2**) in complex real-world scenes featuring diverse objects, environments, and regular spatial relationships. In this task, human subjects are given an image and a target object as inputs and have to predict contextually correct locations for placing the target object.

To curate a dataset of semantically meaningful image–object pairs for the object priming task, we sampled images from the COCO-OCD test set that contained at least three object classes from the 15 selected categories. For each image, we manually assigned a target class—corresponding to the object to be placed—by selecting a semantically appropriate category from the same 15 classes, while ensuring that the target class was not already present in the image. This procedure resulted in 864 semantically meaningful image–object pairs drawn from 206 unique natural images. We refer to this dataset as the *Human Object Priming* (HOP) dataset.

We recruited a total of 437 human participants via Amazon Mechanical Turk (AMT). For each participant, we randomly sampled 20 image–object pairs and presented each of them to human subjects with an image of size 800×800 pixels, subtending $15.56° \times 15.56°$ of visual angle, together with the question "Where would you put this *[obj]*?", where *[obj]* denotes the sampled target object. Participants were instructed to make 10 non-repeated mouse clicks on semantically relevant regions of the image (**Fig. S1C**). Each image–object pair was completed by 3–7 participants. To ensure equal coverage across all image–object pairs, we randomly subsampled click data from 3 unique participants per pair (10 clicks each; 30 clicks total). These 30 click coordinates were consolidated to generate human priming maps.

Specifically, we first divided each 800×800 stimulus image into a 32×32 grid, yielding 1,024 cells of size 25×25 pixels. Mouse clicks were aggregated within each grid cell such that the pixel intensity reflected the number of clicks. The resulting 32×32 priming map was then smoothed using an 11×11 Gaussian filter, resized to 224×224, and min–max normalized to produce the final human priming maps.

4.2 Models

4.2.1 SeCo

We propose a Self-Supervised Learning Method for Context Reasoning (SeCo) which learns associations between objects and their contexts in natural images (**Fig. 3A)**. SeCo consists of three components: (a) object discovery module, (b) two-stream visual processor, and (c) external memory. First, the target

Algorithm 1: PyTorch-style pseudocode for SeCo

```
# Ec, Et:  context and target encoders
# pc, pt:  context and target projectors
# M: external memory shaped in K-by-H
# pk:  key projection of external memory
# mse:  mean square error loss
# var_loss:  variance loss
# cov_loss:  covariance loss
# alpha, beta, gamma:  weightage of each loss component
# load a batch of N images
for x in loader:
   # randomly augmented target and context
   t, c = augment(x)
   # encode and project context, target stream
   hc, ht = Ec(x), Et(x) # N x D
   sc, st = pc(hc), pt(ht) # N x H
   # compute keys of memory
   m = pk(M) # K x H
   # retrieve memory
   p = softmax(dot(sc, m))/sqrt(H) # N x K
   sc = p * M # N x H
   # calculate loss and update
   loss = alpha * mse(sc,st) + beta * (var_loss(sc) + var_loss(st)) / 2 + gamma *
    (cov_loss(sc)+ cov_loss(st))
   loss.backward()
```

discovery module uses unsupervised region proposal methods to locate potential objects of interest on a full image I_f. Each region proposal together with the full image I_f is subsequently converted to pairs of target images I_t and context images I_c. Second, the two-stream visual processor consists of two independent convolutional neural network (CNN) encoders and projectors, extracting information from I_t and I_c, respectively. Third, SeCo employs a trainable external memory to store knowledge priors about contextual associations learned during training phase. Features from I_c serve as queries to retrieve context-relevant prior knowledge from the external memory with an attention mechanism. The retrieved information provides the complementary signal to the context stream and gets compared with the target features from I_t of the object stream to maximize the agreement between the stored prior knowledge and the context-relevant object in the embedding space (see **Algo. 1** for the PyTorch-style pseudocode of SeCo's pre-training algorithm).

Context-Object Pair Discovery. Objects play an important role in context reasoning [31]. To learn object-object and object-context associations, we propose a context-object pair discovery module to exploit regions containing objects of interest. We adopt the selective search algorithm [91] to generate regions of interest (RoI) that potentially contain objects. It is worth noting that selective search is an unsupervised learning algorithm. It performs heuristic searches on hundreds of anchor boxes and proposes RoIs by hierarchically grouping similar regions based on color, texture, size, and shape compatibility. To reduce false positives among many RoIs, we filter out resultant regions according to their area ratio (with a maximum of 0.1) and aspect ratio (within 0.2 and 5). Moreover, we merge RoIs with heavy overlaps by setting the threshold of IoU (intersection over union) as 0.3. For each selected RoI, we generate a pair of target images I_t and context image I_c. I_t is cropped out of full image I_f. The entire image with the RoI blacked out with zeros forms the context image I_c.

Feature Extraction with CNN. Due to eccentricity dependence, human vision has the highest acuity at the fovea and the resolution drops sharply in the periphery with increasing eccentricity [45, 37]. For example, while we are fixating on the mug on the table, the mug is often perceived in high resolution while the context gist of the kitchen scene is processed at low resolution in the periphery. Taking inspiration from this observation, we propose a two-stream visual processor, with one object stream dedicated to encoding the target image I_t and the other context stream dedicated to encoding the context image I_c. The encoded representations are denoted as $h_c = E_c(I_c)$ and $h_t = E_t(I_t)$, where

$E_t(\cdot)$ and $E_c(\cdot)$ are target and context encoders and h_t and $h_c \in \mathbb{R}^D$. Of note, we do not enforce weight sharing between the encoders. We demonstrate the benefit of this approach in the model analysis.

Training With An External Memory. As suggested by cognitive and neuroscience works [106, 83, 90], context processing often happens very fast in the brain. The perceived scene gist serves as a query to retrieve prior knowledge from the semantic memory to modulate object recognition in a top-down manner. To mimic this underlying mechanism of context modulation, we introduce an external memory with trainable parameters, accumulating prior knowledge of contextual associations. Different from the well-established cross-attention mechanism [94], the objective of our external memory focuses on dynamically retrieving and updating the prior knowledge.

We define the external memory as a 2D matrix with trainable parameters, which consists of K memory slots of H dimension, denoted as $M = \{m_1, ..., m_K\}, M \in \mathbb{R}^{H \times K}$. Each memory slot is associated with a key, where $P_k(\cdot) : \mathbb{R}^H \rightarrow \mathbb{R}^H$ defines the linear mapping from the memory content to the keys $P_k(M)$. The encoded representation h_c from the context stream serves as queries to the external memory after a projection operation $P_c(\cdot) : \mathbb{R}^D \rightarrow \mathbb{R}^H$. The retrieved prior knowledge $s_c \in \mathbb{R}^H$ from M can then be represented as

$$s_c = \text{SOFTMAX}(\frac{P_c(h_c)P_k(M)^T}{\sqrt{H}})M \tag{1}$$

where SOFTMAX$(\cdot)$ is the standard softmax operation.

Loss Components. To encourage M to learn rich and meaningful context-object associations, we introduce three types of losses. Ideally, given only the scene gist, the retrieved prior s_c from M should represent useful object information related to the given context (i.e., "what could be the target object given the scene gist" versus "the actual object seen in the scene"). Thus, we apply a mean squared error loss l_{mse} to maximize the agreement between s_c and h_t. To make the vector dimension comparable, h_t is projected to $s_t \in \mathbb{R}^H$ in the embedding space via the projection $P_t(\cdot)$.

As shown by previous works in non-contrastive learning [7, 19], maximizing the agreement between two-stream visual processors alone may lead to model collapses (e.g., where the external memory stores and outputs trivial knowledge of all zeros, while the visual processor encodes images to representations of all zeros). In this case, s_c and s_t align perfectly, but the encoded object representations and content in M are meaningless.

Thus, to prevent model collapses, we follow [7] to enforce covariance L_{cov} and variance L_{var} regularization on both object and context streams. L_{var} maintains the variance of batch-wise representations, encouraging object class diversities, while L_{cov} de-correlates channel-wise variables to diversify attributes of an embedding, i.e., maximize independent attributes to represent objects. SeCo is jointly trained with the total loss:

$$L_{total} = \alpha L_{mse}(s_c, s_t) + \beta[L_{var}(s_c) + L_{var}(s_t)] + \gamma[L_{cov}(s_c) + L_{cov}(s_t)] \tag{2}$$

where $\alpha = 25$, $\beta = 25$ and $\gamma = 1$ are hyper-parameters weighting different loss components.

Data Augmentations. Data augmentation techniques are widely used at image levels in SSL literature. We applied standard image augmentations on both I_t and I_c, including color jitter, grayscale, horizontal flip, gaussian blur, and color normalization using the existing library PyTorch. Moreover, the random resized crop is another effective technique in SSL. However, directly applying this approach is not feasible in our case. Thus, we extended the standard approach to context-object image pairs with context-aware crops by ensuring that the relative locations among objects are preserved and the bounding box encompassing the target object is always intact and present on I_c after geometric transformations.

Model Architecture. We use ResNet-50 [47] with $D = 2048$ output units as encoders E_c and E_t. We set the size of M as $K \times H = 200 \times 8192$ and initialize M by the Xavier uniform initializer [41]. Following VICReg [7], the projectors P_c, P_t, and P_k are implemented as three fully connected layers with 8192 units per layer and ReLU activations between layers.

4.2.2 Baselines

Given that ground truth labels are costly to obtain for supervised learning and that much larger datasets can be used without labels, SSL has become an emerging trend in AI. We considered seven state-of-the-art SSL methods with different types of pretext tasks: (1) Context Encoder[75] is trained to reconstruct missing image regions based on the surrounding pixel context. Since Context Encoder was originally trained with AlexNet [60], we re-implemented it with the standard ResNet-50 backbone (**Fig. S6**) for fair comparisons with SeCo and other SSL methods. (2) SimCLR [18] is a contrastive learning framework that maximizes agreement between differently augmented views of the same image. (3) SimSiam [19] is a non-contrastive method that minimizes the negative cosine similarity between the prediction from one branch and the (stop-gradient) projection from the other branch. (4) DINO [16] is a self-distillation approach where a student network mimics a momentum-updated teacher. (5) VICReg [7] prevents collapse by explicitly regularizing embedding variance and covariance. (6) ORL [99] takes an object-level approach that leverages unsupervised masks to align object representations. (7) The supervised learning baseline was trained using a classification loss with ground-truth labels of the target object. For fair comparisons, all the methods used standard ResNet-50 backbones, with weights pre-trained on ImageNet obtained from their own public checkpoints. We used the same implementations from their original papers.

4.2.3 Experimental Setups for Quantity-quality Analysis of Selective Search

To investigate the role of selective search in SeCo, we compared the default model against three variants that modify the generation of target and context image pairs (I_t, I_c) (**Fig. 5B**). In the **Ground Truth** variant, the context-object discovery module is replaced with human-annotated bounding boxes, ensuring that target images contain precisely localized objects. The **Random Generation** variant preserves the same number of region proposals as the default SeCo model (Selective Search) but replaces them with randomly generated coordinates, thereby testing the importance of object-centric visual signals. The **Random Cropping** variant forms the target–context pair by applying two independent random crops to the same input image, removing the explicit fovea–periphery relationship induced by the discovery module.

To further examine how the quantity and quality of region proposals influence SeCo pre-training, we systematically varied both factors. We first scored each proposal using its intersection-over-union (IoU) with ground-truth bounding boxes. We retained images from the COCO-OCD training set that contained at least 10 proposals with IoU greater than 0.3, resulting in 19.7K images. The following protocols were used to benchmark SeCo with respect to proposal quality and quantity.

[Quality]. We filtered proposals using an IoU threshold γ, yielding a proposal pool $\mathbf{B}_\gamma = \{b \mid IoU(b) > \gamma\}$. To isolate proposal quality, we fixed the number of proposals per image I^i and varied only γ. Specifically, we randomly selected five proposals from $\mathbf{B}^i_\gamma$ while varying $\gamma \in \{0, 0.1, 0.2, 0.3\}$. Training and evaluation followed the procedure described in **Sec. 4.1.1**.

[Quantity]. We fixed the IoU threshold at $\gamma = 0$. For each image I^i, we varied the number of proposals in $\{5, 10, 20, 30\}$ by randomly sampling from $\mathbf{B}^i_{\gamma=0}$. Training and evaluation followed the same procedure described in **Sec. 4.1.1**.

4.2.4 Experimental Setups for Memory Configurations

To examine the role of stored semantic priors in contextual reasoning, we incorporate a learnable external memory module into the SeCo architecture (see **Sec. 4.2.1**). This module is intended to emulate the medial temporal lobe and hippocampal systems of the brain, which encode and retrieve contextual associations acquired from exposure to natural environments.

We investigated the impact of memory configurations by systematically altering the encoding and memory components (**Fig. 5B**). In the **No Memory** variant, we remove the external memory module to evaluate the role of stored contextual priors at inference, forcing the model to rely exclusively on the projection of the encoded context features $s_c = P_c(h_c)$ in contrast to **Equ.1**. Furthermore, we varied memory capacity by adjusting the number of memory slots K and the feature dimensionality H (**Fig. S7A**). Lastly, we introduce the **Shared Encoder** variant, which enforces weight sharing between the target encoder E_t and the context encoder E_c, to evaluate whether a unified feature representation is sufficient for relational reasoning.

To interpret the contextual associations learned by the external memory, we provide a visualization based on attention score distributions. For each sample in the COCO-VOC evaluation set, we identify the memory slot receiving the highest attention weight. These activations are aggregated into category-wise distributions to determine which slots are most frequently associated with specific object classes. To quantify the overlap in contextual knowledge across categories, we compute the pairwise Kullback–Leibler (KL) divergence between the corresponding attention distributions (see **Algo. 2**). This analysis provides a measure of shared contextual associations among different object categories.

Algorithm 2: PyTorch-style pseudocode for calculating pairwise KL divergence of attention score over memory slots for object categories in COCO-VOC.

```
# Ec:  context encoders
# pc:  context projector
# M: external memory shaped in K-by-H
# F: frequency matrix-shaped in C-by-K
# D: pair-wise KL-divergence matrix-shaped in C-by-C
# product:  cartesian product of two sets
# kld:  KL-divergence function
for x, label in loader:
   # obtain erased context
   c = erase(x)

   # encode and project context stream
   hc = Ec(x) # 1 x D
   sc = pc(hc) # 1 x H
   # compute keys of memory
   m = pk(M) # K x H

   # retrieve attention score over memory slots
   p = softmax(dot(sc, m))/sqrt(H) # 1 x K
   # sharpen the distribution
   top1 = p.max(0)[1]
   F[label, top1] += 1
# calculate pairwise KL-divergence
for i,j in product(range(C), range(C)):
   F[i] = (F[i] - F[i].min()) / (F[i].max() - F[i].min())
   F[j] = (F[j] - F[j].min()) / (F[j].max() - F[j].min())
   pi, pj = softmax(F[i]), softmax(F[j])
   D[i,j] = kld(pi, pj)
```

4.2.5 Experimental Setups for Loss Component Configurations

To disentangle the individual contributions of different information constraints to stable pre-training and contextual association, we analyze the joint training objective of SeCo. The overall loss comprises three terms: mean squared error (L_{mse}), variance regularization (L_{var}), and covariance regularization (L_{cov}), with their relative contributions controlled by hyperparameters α, β, and γ (see **Eqn. 2**). We introduce ablated variants that systematically remove one loss component at a time to assess its role in enabling contextual association without representational collapse. The results show that all three components are necessary for optimal performance, while removing any single term leads to degraded performance or training instability, including model collapse (**Fig. S7B**).

4.2.6 Benchmarking AI Models

Training AI models for lift the flap tasks on naturalistic images. SeCo and SSL baselines (**Sec. 4.2.2**) were initialized with ImageNet-pretrained weights, followed by pre-training on the COCO-OCD and COCO-VOC training sets respectively. During the pre-training of SeCo, we set the base learning rate to $lr = 0.05 * \text{batch_size}/256$ [43] for COCO-OCD and $lr = 0.1 * \text{batch_size}/256$ for COCO-VOC. The learning rate grows linearly from 0 to base value during the first 10 epochs and then decays with a cosine scheduler [67] for the rest of epochs with a minimum value of 0.0002. Pre-training

protocols for SSL baselines follow those in their original implementations. As ImageNet-pretrained weights for ORL [99] were unavailable, we directly used weights pre-trained on the full COCO-Stuff dataset [13]. For the supervised learning baseline, we directly used weights pre-trained on ImageNet [26]. After the pre-trained weights for models were obtained, we fine-tuned linear classifiers on the top of frozen encoders for each model on the COCO-OCD and COCO-VOC training sets respectively, using the cross-entropy loss. We conducted the hyperparameter search for each model to find the best learning rate for fine-tuning. The final learning rates used for fine-tuning each model are as follows, decaying with a cosine scheduler [67] for 100 epochs: 0.1 (SeCo), 10 (ORL), 0.1 (VICReg), 0.1 (DINO), 20 (SimSiam), 0.1 (SimCLR), 10 (Context Encoder), 0.1 (Supervised Learning).

Benchmarking AI models for lift-the-flap. We evaluated the models trained in LoR training stage (**Sec. 4.1.1**) on COCO-OCD, COCO-VOC, and FRINE datasets. Specifically, models trained on in-domain COCO-OCD training set were firstly tested on in-domain COCO-OCD test set, then tested on out-of-domain COCO-OCD and FRINE test set in a zero-shot manner; Similarly, models pre-trained on in-domain COCO-VOC training set were firstly tested on in-domain COCO-VOC, then tested on out-of-domain PASCAL-VOC07 test set in a zero-shot manner. Of note, the FRINE test set used for AI models contains static frame images from all the test videos viewed by human subjects. To adapt linear classifiers trained on COCO-OCD (15 classes of real-world objects) to the FRINE dataset (4 classes of novel object fribbles) in a zero-shot manner, we selected the logits corresponding to the 4 VirtualHome objects associated with the fribbles (e.g., Fa1-microwave, Fb1-mouse, Fb3-knife, Fc1-cup) from the 15 logits produced by the classifier (See **Sec. 4.1.1** for detailed context rules). The selected logits, representing the categories of VirtualHome objects before replacement, were then passed through the softmax function. The fribble class corresponding to the highest probability was taken as the final prediction.

Benchmarking AI models for object priming. To predict priming maps for all the models, we converted the object priming task to a series of lift-the-flap tasks with the following steps: (1) we divide the context image into patches. (2) We covered a single image patch with a flap (black pixels) while the remaining patches remained intact. (3) We tested all models fine-tuned on COCO-OCD from the lift-the-flap task in (2) and recorded the predicted classification probability of the model for the given target object class in the object priming task. (4) We iterated through (2) and (3) until we exhaustively performed "lift-the-flap" tasks over all the image patches. (5) For each image patch, we then have a classification score indicating how confidently the model would put the given target object in that patch. We consolidated all the probabilities for all the patches and generated the priming map for each model. As the model predictions were sensitive to the patch sizes, we varied the patch sizes and normalized the final priming map over all patch sizes (see **Algo. 3** for the PyTorch-style pseudocode of object priming algorithm).

4.3 Data Analysis

4.3.1 Behavioral Metrics

Top-1 Accuracy. The lift-the-flap task is formulated as a multi-class classification problem, where the goal is to predict the identity of the hidden target object from a set of candidate categories. Top-1 accuracy is defined as the proportion of trials in which the predicted category matches the ground-truth label of the hidden target. Performance is summarized as the mean trial-wise correctness, with the standard error of the mean (SEM) quantifying uncertainty.

Alignment Score. To quantify spatial reasoning in the object priming task, we measure the agreement between human and AI model predicted priming maps. For each image–object pair, we first compute the root mean square error (RMSE) between the model-predicted and human priming maps. This RMSE is then normalized by the RMSE obtained from a chance model generating random priming maps, resulting in a normalized RMSE (nRMSE) ranging from 0 to 1. Finally, the agreement score is defined as $1 - \text{nRMSE}$, so that higher values indicate greater alignment between the model and human priming maps.

Pearson Correlation. To assess the consistency of behavioral sensitivity between humans and AI

Algorithm 3: PyTorch-style pseudocode for generating priming maps.

```
# Ec:  trained context network with an encoder and a linear classifier
# patch_sizes:  patch sizes when making erased contexts
# load a batch of N images
for x, label in loader:
   maps = []

   # calculate priming maps in multiple scales
   for patch_size in patch_sizes:
      # iteratively erase a patch from image
      contexts = make_context(x, patch_size)

      # retrieve probability w.r.t location for a given object category
      p = softmax(Ec(x)[:,label])

      # normalize so that priming maps in different scales can add up
      p = (p - p.min()) / (p.max() - p.min())

      # upsample to the size of input image
      patch_num = x.size[1] // patch_size
      p = p.view((patch_num,patch_num))
      p = upsample(p)
      maps.append(p)

   # finalize priming maps by averaging and normalizing over different scales
   maps = torch.stack(maps).mean(0)
   maps = (maps - maps.min()) / (maps.max() - maps.min())
```

models, we compute the Pearson correlation coefficient over top-1 accuracy across different contextual manipulations, including blur, reduced area, and jigsaw contexts. This metric captures how closely models replicate human contextual reasoning patterns, with higher correlation values indicating that the model and humans share similar behavioral biases in reasoning from contextual cues.

4.3.2 Statistical Analyses

Bootstrapping and Distribution Analysis. In each experiment, we assessed the statistical robustness of performance metrics by generating empirical distributions of the mean top-1 accuracy using bootstrapping. Specifically, we performed resampling with replacement for 1,000 iterations to obtain a stable distribution of bootstrap means for each condition. Welch's two-tailed t-tests were then used to evaluate differences between conditions and to compare experimental results against chance levels, with statistical significance determined using a consistent threshold of $p < 0.05$. Throughout the manuscript, we used the term "Welch's two-tailed t-test" to assess statistical differences between bootstrapped distributions.

One-way Analysis of Variance (ANOVA). To evaluate whether categorical experimental manipulations significantly influenced reasoning performance, we applied one-way analysis of variance (ANOVA). This test was used to compare bootstrapped mean top-1 accuracy across multiple levels of a single context manipulation family, including variations in context blur, context–object size ratios, and jigsaw permutations. The resulting p-value was used to determine whether at least one condition within each multi-level group differed significantly from the others. Statistical significance was determined using a consistent threshold of $p < 0.05$.

4.4 Code and Data Availability

All the source code and raw data are made publicly available in this submission through the following repository: GitHub Link.

Acknowledgments

This work was supported by the National Research Foundation, Singapore under its NRFF award NRF-NRFF15-2023-0001 and Mengmi Zhang's Startup Grant from Nanyang Technological University.

Author Contributions

The tasks were designed by XL, ZS, GK, and MZ. XL, AS, and BAK were involved in collecting the data. The data were analyzed by XL, SS, and AS, under supervision by MZ. The manuscript was written by XL, SS, AS, GK, and MZ, and was approved by all the authors.

Competing Interests

The authors declare no competing interests.

References

[1] Mark E Auckland, Kyle R Cave, and Nick Donnelly. Nontarget objects can influence perceptual processes during object recognition. *Psychonomic bulletin & review*, 14(2):332–337, 2007.

[2] Stephen D. Auger, Sinéad L. Mullally, and Eleanor A. Maguire. Retrosplenial cortex codes for permanent landmarks. *PLOS ONE*, 7(8):1–12, 08 2012.

[3] Moshe Bar. Visual objects in context. *Nature Reviews Neuroscience*, 5(8):617–629, 2004.

[4] Moshe Bar and Elissa Aminoff. Cortical analysis of visual context. *Neuron*, 38(2):347–358, 2003.

[5] Moshe Bar, Karim S Kassam, Avniel Singh Ghuman, Jasmine Boshyan, Annette M Schmid, Anders M Dale, Matti S Hämäläinen, Ksenija Marinkovic, Daniel L Schacter, Bruce R Rosen, et al. Top-down facilitation of visual recognition. *Proceedings of the national academy of sciences*, 103(2):449–454, 2006.

[6] Moshe Bar and Shimon Ullman. Spatial context in recognition. *Perception*, 25(3):343–352, 1996.

[7] Adrien Bardes, Jean Ponce, and Yann Lecun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *ICLR 2022-10th International Conference on Learning Representations*, 2022.

[8] Pouya Bashivan, Kohitij Kar, and James J DiCarlo. Neural population control via deep image synthesis. *Science*, 364(6439):eaav9436, 2019.

[9] Peter Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, et al. Interaction networks for learning about objects, relations and physics. *Advances in neural information processing systems*, 29, 2016.

[10] Irving Biederman, Robert J. Mezzanotte, and Jan C. Rabinowitz. Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, 14(2):143–177, 1982.

[11] Philipp Bomatter, Mengmi Zhang, Dimitar Karev, Spandan Madan, Claire Tseng, and Gabriel Kreiman. When pigs fly: Contextual reasoning in synthetic and natural scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 255–264, 2021.

[12] Santiago A Cadena, George H Denfield, Edgar Y Walker, Leon A Gatys, Andreas S Tolias, Matthias Bethge, and Alexander S Ecker. Deep convolutional models improve predictions of macaque v1 responses to natural images. *PLoS computational biology*, 15(4):e1006897, 2019.

[13] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018.

[14] Yusen Cai, Bhargava Satya Nunna, Qing Lin, and Mengmi Zhang. Learning to see through a baby's eyes: Early visual diets enable robust visual intelligence in humans and machines. *arXiv preprint arXiv:2511.14440*, 2025.

[15] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[16] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

[17] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[18] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[19] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021.

[20] Xinlei Chen, Li-Jia Li, Li Fei-Fei, and Abhinav Gupta. Iterative visual reasoning beyond convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7239–7248, 2018.

[21] Wongun Choi and Silvio Savarese. A unified framework for multi-target tracking and collective activity recognition. In *European Conference on Computer Vision*, pages 215–230. Springer, 2012.

[22] Marvin M Chun and Yuhong Jiang. Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive psychology*, 36(1):28–71, 1998.

[23] Charles E Connor, Scott L Brincat, and Anitha Pasupathy. Transformation of shape information in the ventral pathway. *Current opinion in neurobiology*, 17(2):140–147, 2007.

[24] Inferior Temporal Cortex. Fast readout of object identity from macaque. *science*, 1117593(863):310, 2005.

[25] Jodi L Davenport and Mary C Potter. Scene consistency in object and background perception. *Psychological science*, 15(8):559–564, 2004.

[26] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[27] Zhiwei Deng, Arash Vahdat, Hexiang Hu, and Greg Mori. Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4772–4781, 2016.

[28] Zhiwei Ding, Xuezhe Ren, Erwan David, Melissa Vo, Gabriel Kreiman, and Mengmi Zhang. Efficient zero-shot visual search via target and context-aware transformer. *arXiv preprint arXiv:2211.13470*, 2022.

[29] Sara Djambazovska, Anaa Zafer, Hamidreza Ramezanpour, Gabriel Kreiman, and Kohitij Kar. The impact of scene context on visual object recognition: Comparing humans, monkeys, and computational models. *bioRxiv*, 2024.

[30] Rodney J Douglas and Kevan AC Martin. Recurrent neuronal circuits in the neocortex. *Current biology*, 17(13):R496–R500, 2007.

[31] Dejan Draschkow and Melissa L-H Võ. Scene grammar shapes the way we interact with objects, strengthens memories, and speeds search. *Scientific reports*, 7(1):1–12, 2017.

[32] Howard Eichenbaum. A cortical–hippocampal system for declarative memory. *Nature reviews neuroscience*, 1(1):41–50, 2000.

[33] Russell Epstein and Nancy Kanwisher. A cortical representation of the local visual environment. *Nature*, 392(6676):598–601, 1998.

[34] Russell A Epstein. Parahippocampal and retrosplenial contributions to human spatial navigation. *Trends in cognitive sciences*, 12(10):388–396, 2008.

[35] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[36] Tucker G Fisher, Henry J Alitto, and W Martin Usrey. Retinal and nonretinal contributions to extraclassical surround suppression in the lateral geniculate nucleus. *Journal of Neuroscience*, 37(1):226–235, 2017.

[37] Jeremy Freeman and Eero P Simoncelli. Metamers of the ventral stream. *Nature neuroscience*, 14(9):1195–1201, 2011.

[38] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International conference on learning representations*, 2018.

[39] Robert Geirhos, Carlos RM Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. Generalisation in humans and deep neural networks. *Advances in neural information processing systems*, 31, 2018.

[40] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[41] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

[42] Josep M Gonfaus, Xavier Boix, Joost Van de Weijer, Andrew D Bagdanov, Joan Serrat, and Jordi Gonzalez. Harmony potentials for joint classification and segmentation. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3280–3287. IEEE, 2010.

[43] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

[44] Kalanit Grill-Spector and Rafael Malach. The human visual cortex. *Annu. Rev. Neurosci.*, 27(1):649–677, 2004.

[45] Shashi Kant Gupta, Mengmi Zhang, Chia-Chien Wu, Jeremy Wolfe, and Gabriel Kreiman. Visual search asymmetry: Deep nets and humans share similar inherent biases. *Advances in neural information processing systems*, 34:6946–6959, 2021.

[46] Shuangpeng Han, Ziyu Wang, and Mengmi Zhang. Flow snapshot neurons in action: Deep neural networks generalize to biological motion perception. *Advances in Neural Information Processing Systems*, 37:53732–53763, 2024.

[47] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[48] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[49] John M Henderson, Phillip A Weeks Jr, and Andrew Hollingworth. The effects of semantic consistency on eye movements during complex scene viewing. *Journal of experimental psychology: Human perception and performance*, 25(1):210, 1999.

[50] Christopher A Henry and Adam Kohn. Spatial contextual effects in primary visual cortex limit feature representation under crowding. *Nature communications*, 11(1):1687, 2020.

[51] Andrew Hollingworth. Does consistent scene context facilitate object perception? *Journal of Experimental Psychology: General*, 127(4):398, 1998.

[52] Ha Hong, Daniel LK Yamins, Najib J Majaj, and James J DiCarlo. Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature neuroscience*, 19(4):613–622, 2016.

[53] Hexiang Hu, Guang-Tong Zhou, Zhiwei Deng, Zicheng Liao, and Greg Mori. Learning structured inference neural networks with label relations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2960–2968, 2016.

[54] Yanhao Jia, Ji Xie, S Jivaganesh, Hao Li, Xu Wu, and Mengmi Zhang. Seeing sound, hearing sight: Uncovering modality bias and conflict of ai models in sound localization. *arXiv preprint arXiv:2505.11217*, 2025.

[55] Kohitij Kar and James J DiCarlo. Fast recurrent processing via ventrolateral prefrontal cortex is needed by the primate ventral stream for robust core visual object recognition. *Neuron*, 109(1):164–176, 2021.

[56] Kohitij Kar and James J DiCarlo. The quest for an integrated set of neural mechanisms underlying object recognition in primates. *Annual Review of Vision Science*, 10, 2024.

[57] Kohitij Kar, Jonas Kubilius, Kailyn Schmidt, Elias B Issa, and James J DiCarlo. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature neuroscience*, 22(6):974–983, 2019.

[58] Andreas J Keller, Mario Dipoppa, Morgane M Roth, Matthew S Caudill, Alessandro Ingrosso, Kenneth D Miller, and Massimo Scanziani. A disinhibitory circuit for contextual modulation in primary visual cortex. *Neuron*, 108(6):1181–1193, 2020.

[59] Naitik Khandelwal, Xiao Liu, and Mengmi Zhang. Adaptive visual scene understanding: incremental scene graph generation. *arXiv preprint arXiv:2310.01636*, 2023.

[60] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

[61] Lubor Ladicky, Chris Russell, Pushmeet Kohli, and Philip HS Torr. Graph cut based inference with co-occurrence statistics. In *European conference on computer vision*, pages 239–253. Springer, 2010.

[62] Md Nasir Uddin Laskar, Luis Gonzalo Sanchez Giraldo, and Odelia Schwartz. Deep neural networks capture texture sensitivity in v2. *Journal of vision*, 20(7):21–1, 2020.

[63] Qing Lin, Jingfeng Zhang, Yew-Soon Ong, and Mengmi Zhang. Make me happier: Evoking emotions through image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16367–16376, 2025.

[64] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[65] Hesheng Liu, Yigal Agam, Joseph R Madsen, and Gabriel Kreiman. Timing, timing, timing: fast decoding of object information from intracranial field potentials in human visual cortex. *Neuron*, 62(2):281–290, 2009.

[66] Yong Liu, Ruiping Wang, Shiguang Shan, and Xilin Chen. Structure inference net: Object detection using scene-level context and instance-level relationships. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6985–6994, 2018.

[67] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[68] Najib J Majaj, Ha Hong, Ethan A Solomon, and James J DiCarlo. Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance. *Journal of Neuroscience*, 35(39):13402–13418, 2015.

[69] David Marr. Simple memory: a theory for archicortex. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 262(841):23–81, 1971.

[70] James L McClelland, Bruce L McNaughton, and Randall C O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3):419, 1995.

[71] David A Mély, Drew Linsley, and Thomas Serre. Complementary surrounds explain diverse contextual phenomena across visual modalities. *Psychological review*, 125(5):769, 2018.

[72] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 891–898, 2014.

[73] Keiron O'Shea and Ryan Nash. An introduction to convolutional neural networks, 2015.

[74] tephen E Palmer. The effects of contextual scenes on the identification of objects. *Memory & cognition*, 3(5):519–526, 1975.

[75] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

[76] Vivian C Paulun, Kristine Zheng, and Kohitij Kar. Distributed population activity in the macaque inferior temporal cortex but not current deep neural networks predict the ponzo illusion. *Journal of Vision*, 22(14):3354–3354, 2022.

[77] Maytus Piriyajitakonkij, Rujikorn Charakorn, Weicheng Tao, Wei Pan, Mingfei Sun, Cheston Tan, and Mengmi Zhang. From grunts to grammar: Emergent language from cooperative foraging. *arXiv e-prints*, pages arXiv–2505, 2025.

[78] Dean A Pospisil, Anitha Pasupathy, and Wyeth Bair. 'artiphysiology'reveals v4-like shape tuning in a deep network trained for image classification. *Elife*, 7:e38242, 2018.

[79] Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. Virtualhome: Simulating household activities via programs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8494–8502, 2018.

[80] Paul C Quinn, Peter D Eimas, and Stacey L Rosenkrantz. Evidence for representations of perceptually similar natural categories by 3-month-old and 4-month-old infants. *Perception*, 22(4):463–475, 1993.

[81] Rishi Rajalingham, Elias B Issa, Pouya Bashivan, Kohitij Kar, Kailyn Schmidt, and James J DiCarlo. Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience*, 38(33):7255–7269, 2018.

[82] Rishi Rajalingham, Kailyn Schmidt, and James J DiCarlo. Comparison of object recognition behavior in human and monkey. *Journal of Neuroscience*, 35(35):12127–12136, 2015.

[83] Maximilian Riesenhuber and Tomaso Poggio. Hierarchical models of object recognition in cortex. *Nature neuroscience*, 2(11):1019–1025, 1999.

[84] Edmund T Rolls. The mechanisms for pattern completion and pattern separation in the hippocampus. *Frontiers in systems neuroscience*, 7:74, 2013.

[85] Ankur Sikarwar and Mengmi Zhang. Decoding the enigma: benchmarking humans and ais on the many facets of working memory. *Advances in Neural Information Processing Systems*, 36:74039–74076, 2023.

[86] Parantak Singh, You Li, Ankur Sikarwar, Stan Weixian Lei, Difei Gao, Morgan B Talbot, Ying Sun, Mike Zheng Shou, Gabriel Kreiman, and Mengmi Zhang. Learning to learn: How to continuously teach humans and machines. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11708–11719, 2023.

[87] Larry R Squire, Craig EL Stark, and Robert E Clark. The medial temporal lobe. *Annu. Rev. Neurosci.*, 27(1):279–306, 2004.

[88] Ren Jie Tee and Mengmi Zhang. Integrating curricula with replays: its effects on continual learning. In *Proceedings of the AAAI Symposium Series*, volume 1, pages 109–116, 2023.

[89] Damien Teney, Lingqiao Liu, and Anton van Den Hengel. Graph-structured representations for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2017.

[90] Simon Thorpe, Denis Fize, and Catherine Marlot. Speed of processing in the human visual system. *nature*, 381(6582):520–522, 1996.

[91] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.

[92] Tomer D Ullman, Elizabeth Spelke, Peter Battaglia, and Joshua B Tenenbaum. Mind games: Game engines as an architecture for intuitive physics. *Trends in cognitive sciences*, 21(9):649–665, 2017.

[93] Leslie G Ungerleider. Two cortical visual systems. *Analysis of visual behavior*, 549:chapter–18, 1982.

[94] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[95] Melissa Le-Hoa Vo. The meaning and structure of scenes. *Vision Research*, 181:10–20, 2021.

[96] Bo Wang, Dingwei Tan, Yen-Ling Kuo, Zhaowei Sun, Jeremy M Wolfe, Tat-Jen Cham, and Mengmi Zhang. Gazing at rewards: Eye movements as a lens into human and ai decision-making in hybrid visual foraging. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14810–14823, 2025.

[97] Ziyu Wang, Shuangpeng Han, Mike Zheng Shou, and Mengmi Zhang. Unsupervised prior learning: Discovering categorical pose priors from videos. 2024.

[98] Ziyu Wang, Mike Zheng Shou, and Mengmi Zhang. Object-centric learning with cyclic walks between parts and whole. *Advances in Neural Information Processing Systems*, 36:9388–9408, 2023.

[99] Jiahao Xie, Xiaohang Zhan, Ziwei Liu, Yew Soon Ong, and Chen Change Loy. Unsupervised object-level representation learning from scene images. *Advances in Neural Information Processing Systems*, 34:28864–28876, 2021.

[100] Daniel LK Yamins and James J DiCarlo. Eight open questions in the computational modeling of higher sensory cortex. *Current opinion in neurobiology*, 37:114–120, 2016.

[101] Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014.

[102] Jian Yao, Sanja Fidler, and Raquel Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *2012 IEEE conference on computer vision and pattern recognition*, pages 702–709. IEEE, 2012.

[103] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021.

[104] Mengmi Zhang, Marcelo Armendariz, Will Xiao, Olivia Rose, Katarina Bendtz, Margaret Livingstone, Carlos Ponce, and Gabriel Kreiman. Look twice: A generalist computational model predicts return fixations across tasks and species. *PLoS computational biology*, 18(11):e1010654, 2022.

[105] Mengmi Zhang, Jiashi Feng, Keng Teck Ma, Joo Hwee Lim, Qi Zhao, and Gabriel Kreiman. Finding any waldo with zero-shot invariant and efficient visual search. *Nature communications*, 9(1):3730, 2018.

[106] Mengmi Zhang, Claire Tseng, and Gabriel Kreiman. Putting visual object recognition in context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12985–12994, 2020.

[107] Zhengquan Zhang, Feng Xu, and Mengmi Zhang. Peering into the unknown: Active view selection with neural uncertainty maps for 3d reconstruction. *arXiv preprint arXiv:2506.14856*, 2025.

[108] Chengxu Zhuang, Siming Yan, Aran Nayebi, Martin Schrimpf, Michael C Frank, James J DiCarlo, and Daniel LK Yamins. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3):e2014196118, 2021.

Main Figures

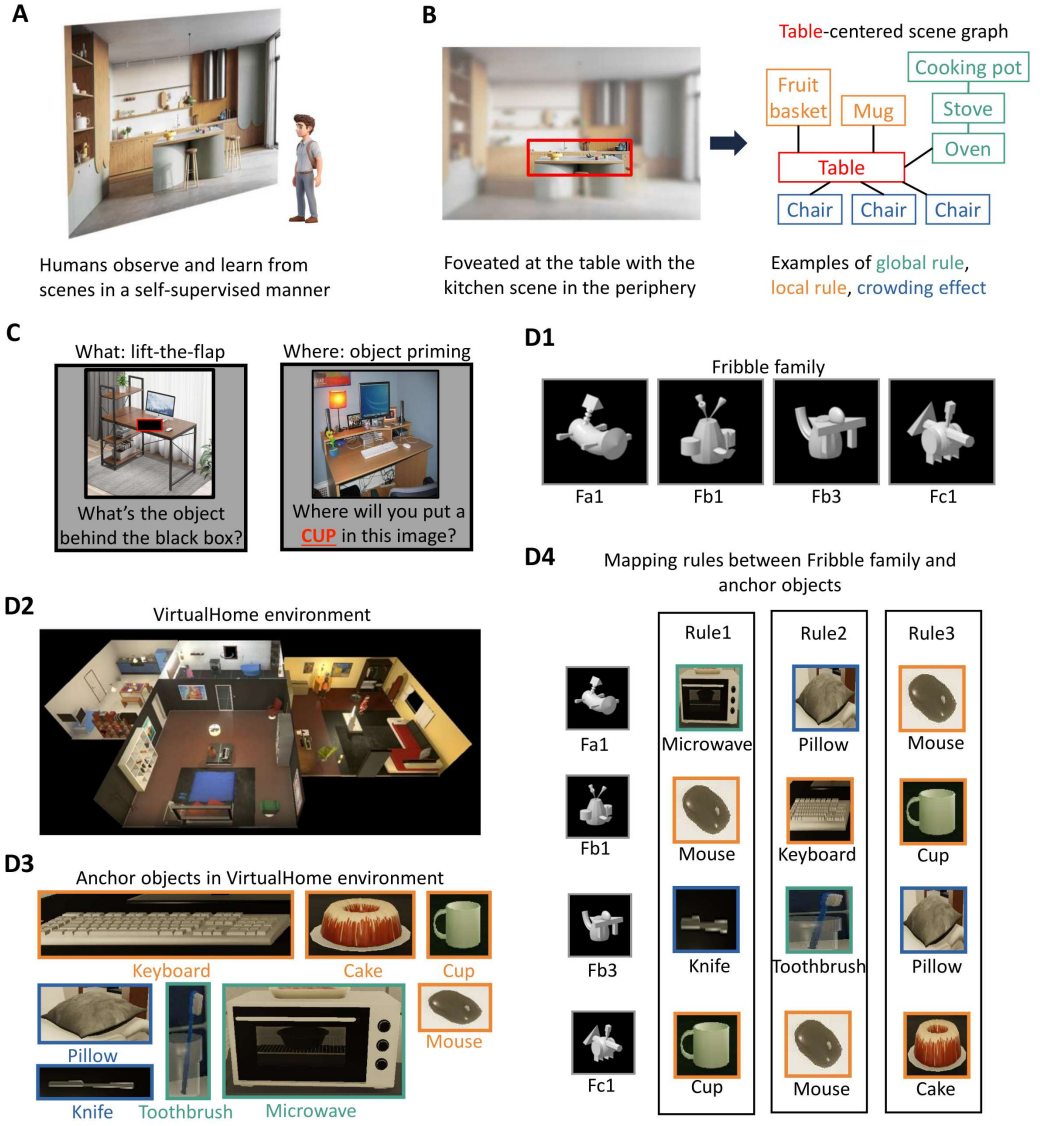


Figure 1: **Humans learn contextual rules from complex natural scenes without explicit instructions or feedback. A.** Humans perceive scenes holistically rather than as isolated objects. Through exposure to rich, multi-object environments, they implicitly learn contextual associations without the need for explicit instruction or supervision. **B.** Foveated vision, with high resolution at the center of gaze and lower resolution in the periphery, supports object-centric representations of complex scenes. When fixating on a table (red), surrounding objects are organized into a table-centered scene graph [59]. Bounding box colors denote contextual associations: local co-occurrence (e.g., mug on table), global context (e.g., stove in kitchen), and crowding (e.g., multiple chairs near the table). **C.** To systematically study contextual reasoning in humans and AI, we introduce two evaluation tasks: lift-the-flap and object priming. In lift-the-flap (left), agents use scene context to infer the identity of a hidden object behind a black patch framed in red. In object priming (right), given a scene and a target object that is not already present, agents predict contextually appropriate locations for placing the object. **D.** FRibble In the sceNE (FRINE) dataset for studying how humans learn to reason from context. Without relying on prior contextual knowledge of familiar objects, we construct the FRINE dataset using novel objects. We begin by selecting four novel object families (Fa1, Fb1, Fb3, Fc1) from the Novel Object Dataset (NOD) [86]. These "fribbles" have distinct body structures and appendages that are unfamiliar to humans as shown in **D1**. Next, we replace eight common household object classes, which serve as anchor objects in the Unity-based indoor scene simulator VirtualHome [79], with fribbles. **D2** shows an example VirtualHome apartment, alongside example images of the eight household objects in **D3**, each color-coded according to the global, local, and crowding associations defined in **B**. The mapping between anchor objects and fribbles defines the three novel contextual rules, as illustrated in **D4**. Each column represents a contextual rule, and each row shows the fribble assigned to the corresponding anchor object class within that rule. See **Methods** for details on FRINE dataset.

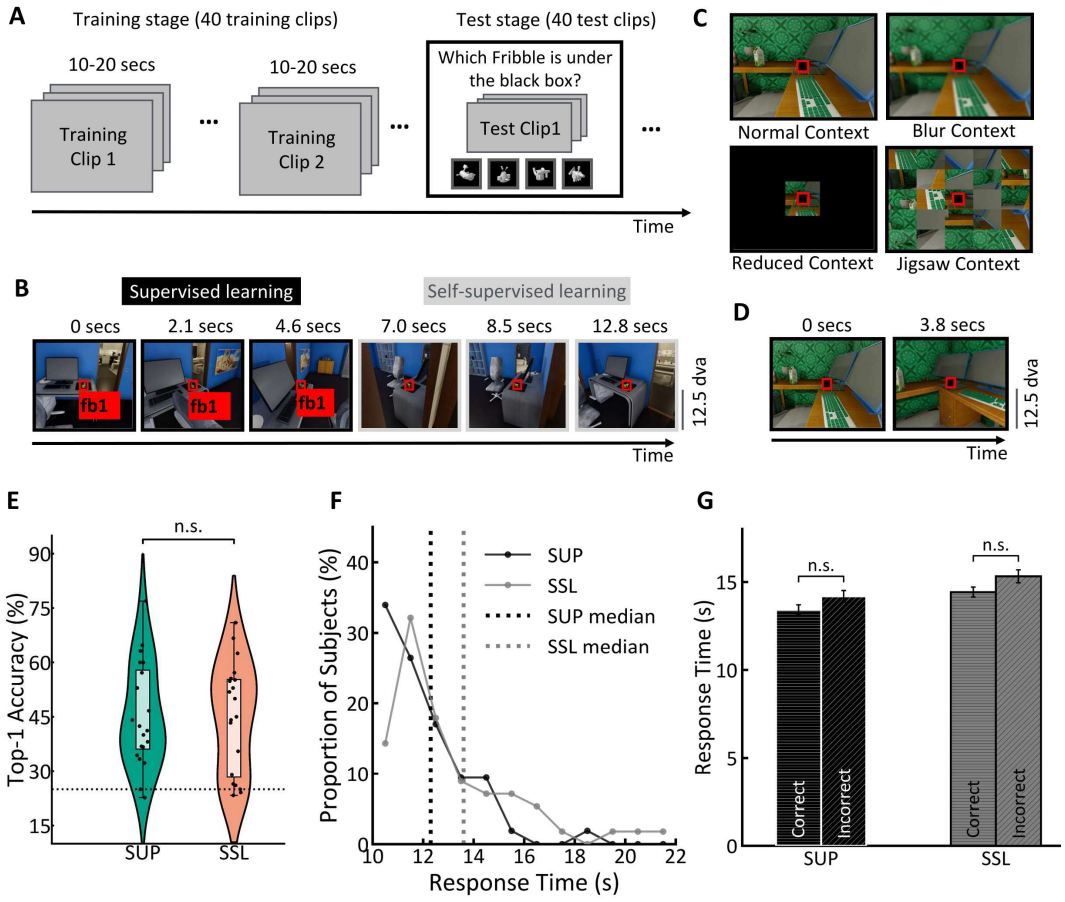


Figure 2: **We introduce human psychophysics experiments in learning to reason on the FRINE dataset.** **A.** Schematic of the Human Psychophysics Experiments. The experiment comprises two phases: a training phase and a testing phase. In the training phase, participants were shown a sequence of 40 training video clips from the FRINE dataset. Each clip lasted 10 to 20 seconds and depicted a novel fribble object centered within a naturalistic scene. The camera rotated around the fribble object in each video. See **B** for an example training clip. Participants were assigned to either a supervised (Sup) or self-supervised learning (SSL) condition. In the Sup condition (clips framed in black), the fribble object was highlighted with a red bounding box and labeled (e.g., "fb1"). In the SSL condition (clips framed in gray), no labels were shown. Importantly, each participant viewed training videos from only one learning condition (either Sup or SSL) throughout the entire training phase. In **B**, the first three frames are shown in the Sup condition and the remaining frames in SSL for illustration only. Frame timestamps are shown above each frame, and a scale bar on the right indicates the frame size in degrees of visual angle. During the test phase, participants viewed 40 test clips from the FRINE dataset. In each clip, the fribble object was occluded by a black patch, and participants were required to infer its identity based solely on contextual information, making a 4-way forced-choice classification. **C.** Context Manipulations in the Lift-the-Flap Task. To investigate the role of context, we introduced four context variations in the test clips: normal context, blur context, context areas, and jigsaw context. **D.** Example Test Clip. We show an example test clip under the normal context condition, where the central fribble object is hidden beneath a black patch while the surrounding context remains intact. Frame timestamps and video frame sizes (in degrees of visual angle) are annotated similarly to **B**. **E.** Individual subject performance differences in the lift-the-flap task. Violin plots show the distribution of top-1 accuracy across individual participants under SUP (green) and SSL (orange) learning conditions, with each black dot representing a single subject. The box spans the interquartile range (25th–75th percentiles); whiskers denote the full range of the data. **F.** Response time across human subjects for SUP and SSL conditions. The plot shows the proportion of SUP (black) and SSL (dark grey) participants as a function of their mean reaction time per trial (in seconds). Vertical dotted lines indicate the median response time for each group. **G.** Reaction times for correct and incorrect trials in SUP and SSL humans. Trials were separated into correct and incorrect responses for SUP and SSL participants, and their average reaction time was computed for each learning condition. Error bars indicate the standard error of the mean (SEM). "n.s." indicates no statistically significant difference between the two distributions ($p > 0.05$).

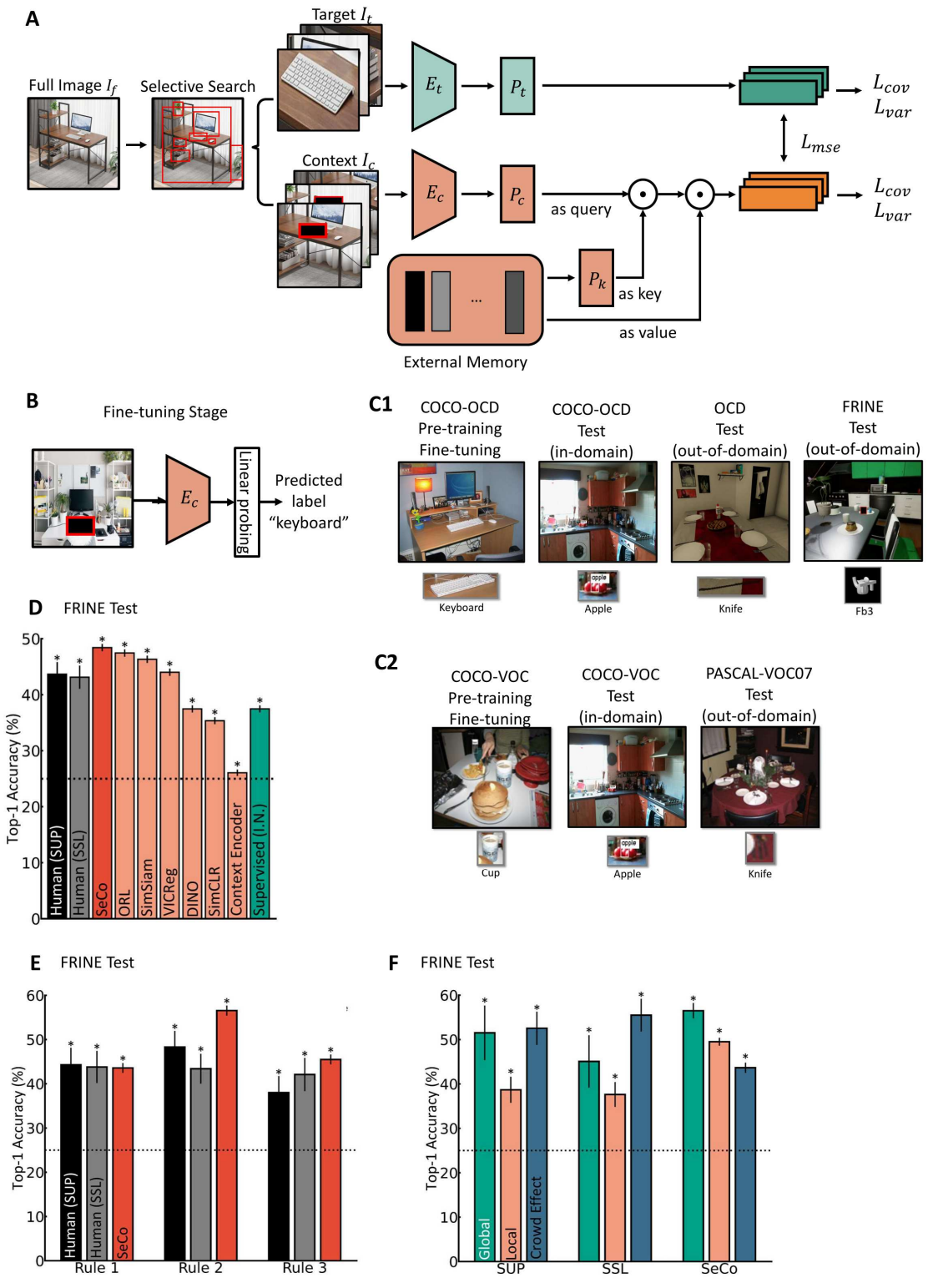

A
Target I_t
Full Image I_f
Selective Search
Context I_c
E_t
P_t
E_c
P_c
L_{cov}
L_{var}
L_{mse}
as query
P_k
as key
as value
External Memory
B
Fine-tuning Stage
E_c
Linear probing
Predicted label "keyboard"
C1
COCO-OCD Pre-training Fine-tuning
COCO-OCD Test (in-domain)
OCD Test (out-of-domain)
FRINE Test (out-of-domain)
Keyboard
Apple
Knife
Fb3
C2
COCO-VOC Pre-training Fine-tuning
COCO-VOC Test (in-domain)
PASCAL-VOC07 Test (out-of-domain)
Cup
Apple
Knife
D
FRINE Test
Top-1 Accuracy (%)
Human (SUP)
Human (SSL)
SeCo
ORL
SimSiam
VICReg
DINO
SimCLR
Context Encoder
Supervised (I.N.)
E
FRINE Test
Top-1 Accuracy (%)
Human (SUP)
Human (SSL)
SeCo
Rule 1
Rule 2
Rule 3
F
FRINE Test
Top-1 Accuracy (%)
Global
Local
Crowd Effect
SUP
SSL
SeCo

Figure 3

Figure 3: **We propose the Self-supervised Method with External Memories for Context Reasoning, named as SeCo. A.** Model schematic of SeCo. The architecture consists of three components: a target discovery module, a two-stream visual processor (trapezoids), and an external memory module (orange squircle). During pre-training, given a full image I_f, SeCo employs an unsupervised method (selective search) to generate potential object proposals. These proposals are then converted into multiple context–target image pairs (I_c, I_t). For each pair, SeCo processes the context and target separately using non-shared encoders (E_c, E_t) followed by non-shared projectors (P_c, P_t). In parallel, SeCo incorporates a trainable external memory to store contextual priors. Ideally, if E_c encodes strong contextual cues, its latent representation should serve as an effective query to retrieve a relevant object representation from the external memory based on learned keys P_k and values. A joint training objective is applied: the mean squared error loss L_{mse} encourages alignment between the retrieved embedding from the memory (based on context) and the target object embedding, while additional regularization losses—the covariance loss L_{cov} and the variance loss L_{var}—promote diversity and prevent collapse of the learned representations (shown as stacked rectangles). **B.** During fine-tuning, SeCo is adapted to the downstream lift-the-flap task. Given a test image with a hidden object, the pre-trained frozen context encoder E_c is used to extract contextual representations, followed by a fully connected layer to predict the label of the hidden object (e.g., "keyboard") behind the black patch via linear probing. **C.** Datasets used for pre-training, fine-tuning, and testing AI models. **C1** The COCO-OCD dataset contains naturalistic images from COCO-Stuff [13] overlapping with object classes in the OCD dataset [11]. **C2** The COCO-VOC dataset contains COCO-Stuff images overlapping with object classes in PASCAL-VOC07 [35]. For each dataset, in-domain and out-of-domain test sets are provided, with target objects listed below each context image. **D.** Top-1 accuracy of human participants and AI models on the lift-the-flap task under normal context conditions using the FRINE dataset. From left to right: SUP humans (black), SSL humans (gray), SeCo (red), self-supervised learning (SSL) baselines including ORL [99], SimSiam [19], VICReg [7], DINO [16], SimCLR [18], and Context Encoder [75] (orange), and a supervised learning baseline (green). A total of 517 SUP human trials, 548 SSL human trials, and 1,926 trials per model were collected. Error bars indicate standard errors computed across all trials for each group. **E.** Top-1 accuracy of human participants and SeCo under normal context conditions for Rules 1–3 of the FRINE dataset. Number of trials per rule is indicated in brackets: Rule 1—SUP humans (165), SSL humans (186); Rule 2—SUP humans (180), SSL humans (187); Rule 3—SUP humans (172), SSL humans (175). SeCo was evaluated on 1,926 trials for each rule. **F.** Top-1 accuracy under normal context conditions for SUP humans, SSL humans, and SeCo across different context associations in the FRINE dataset. For each association type, the number of trials exceeds 640 for SUP humans, SSL humans, and SeCo. See **Methods** for definitions of the three context association types. Across **D**, **E**, and **F**, the chance level (25%) is indicated by a horizontal black dashed line. Error bars indicate the standard error of the mean (SEM). * denotes performance significantly above chance based on Welch's two-tailed t-test ($p < 0.05$).

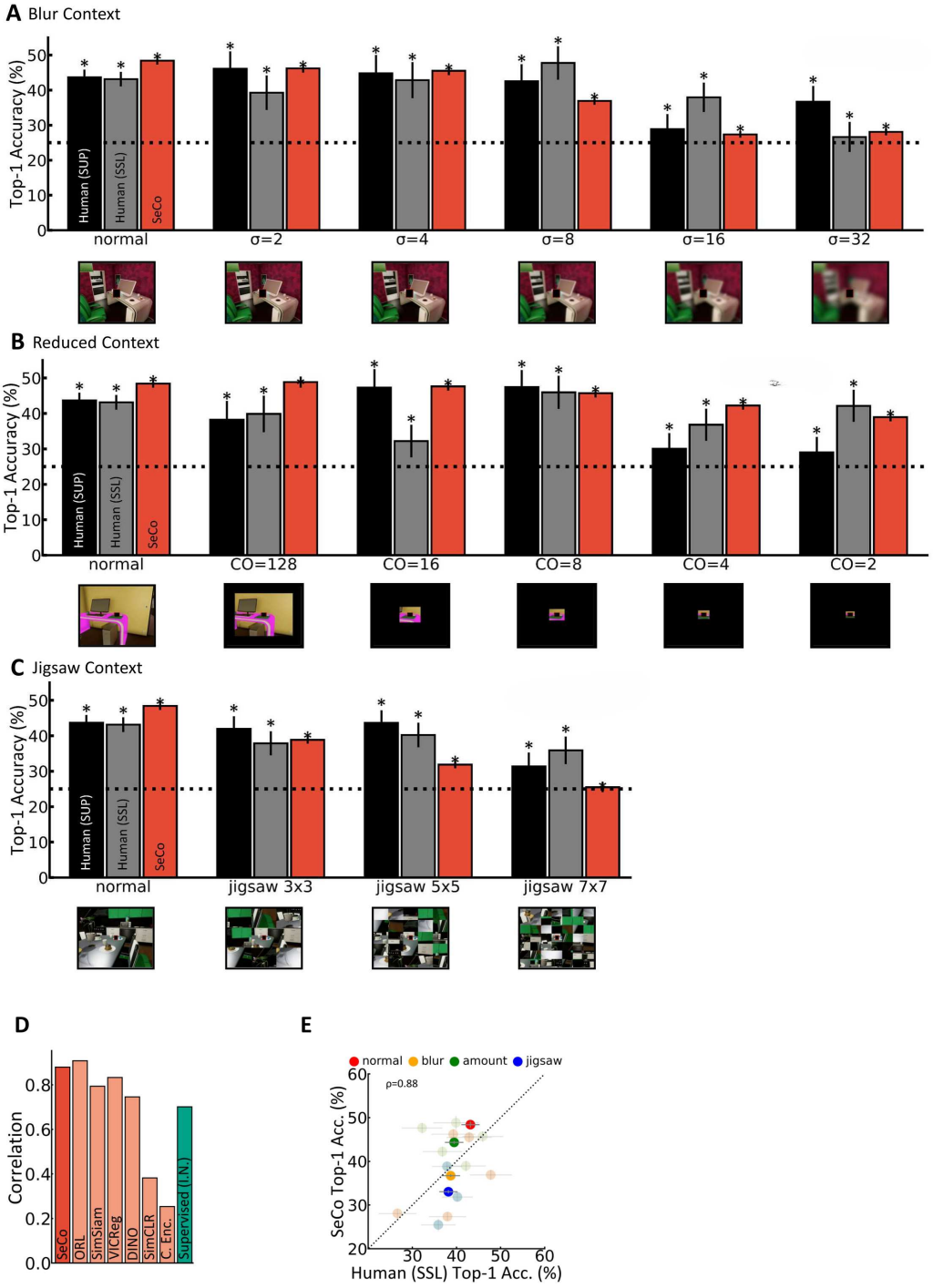
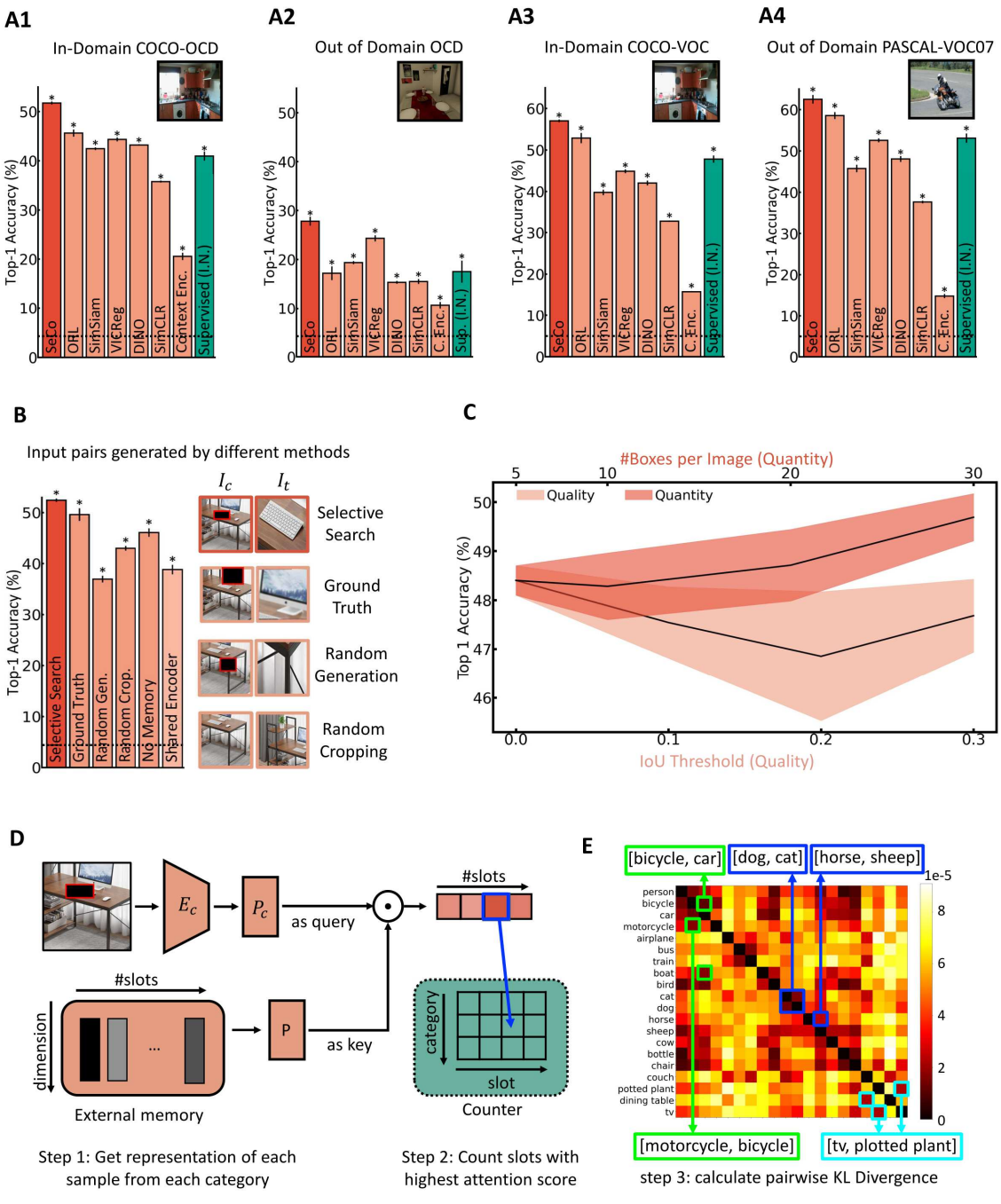

A Blur Context
Top-1 Accuracy (%)
Human (SUP)
Human (SSL)
SeCo
normal
σ=2
σ=4
σ=8
σ=16
σ=32
B Reduced Context
Top-1 Accuracy (%)
normal
CO=128
CO=16
CO=8
CO=4
CO=2
C Jigsaw Context
Top-1 Accuracy (%)
normal
jigsaw 3x3
jigsaw 5x5
jigsaw 7x7
D
Correlation
SeCo
ORL
SimSiam
VICReg
DINO
SimCLR
C. Enc.
Supervised (I.N.)
E
normal
blur
amount
jigsaw
ρ=0.88
SeCo Top-1 Acc. (%)
Human (SSL) Top-1 Acc. (%)

Figure 4

Figure 4: **Contextual variations impact human and SeCo performance on the FRINE dataset in the lift-the-flap task. A.** Top-1 accuracy of humans and SeCo under the blur context condition. Context resolution was manipulated using zero-mean Gaussian blur with standard deviation $\sigma = 2, 4, 8, 16, 32$ pixels. Example images for each sub-condition are shown below the corresponding bar group. Format and design conventions follow **Fig.3E**. **B.** Top-1 accuracy of humans and SeCo under the reduced context area condition. Context amount was varied by adjusting the context-object ratio (CO) = 128, 16, 8, 4, 2. Format and conventions follow **A**. **C.** Top-1 accuracy of humans and SeCo under the jigsaw context condition. Spatial configuration of context was manipulated by randomly scrambling the image into 3×3, 5×5, or 7×7 pieces, while the piece containing the target object remained in place. Format and conventions follow **A**. **D.** Linear correlation between top-1 accuracy of SSL humans and AI models across all four context manipulations. **E.** Top-1 accuracy of SeCo versus SSL humans across four context manipulations: normal context (red), blur context (orange), reduced context area (green), and jigsaw context (blue). High-opacity dots indicate mean accuracy for each manipulation, while low-opacity dots of the same color show mean accuracy for sub-conditions within each manipulation. Cross-shaped error bars denote SEM across all trials. Pearson correlation coefficient: $\rho = 0.88$. Black dotted line indicates the diagonal. Across **A**, and **B**, the chance level (25%) is indicated by a horizontal black dashed line. Error bars indicate the standard error of the mean (SEM). * denotes performance significantly above chance based on Welch's two-tailed t-test ($p < 0.05$).



Step 1: Get representation of each sample from each category

Step 2: Count slots with highest attention score

step 3: calculate pairwise KL Divergence

Figure 5

Figure 5: **SeCo outperforms all competitive baselines in the normal context in the lift-the-flap task, and model analysis reveals key architectural insights.** **A.** Top-1 accuracy on the lift-the-flap task using the COCO-OCD and COCO-VOC datasets. Results are reported separately for in-domain (left) and out-of-domain (right) test sets. Representative examples from each test set are shown for illustration. Format and design conventions follow **Fig. 3E**. **B.** Selective Search identifies effective context–object pairs for pre-training. To evaluate its contribution, we replaced SeCo's context–object discovery module with: ground-truth bounding boxes as targets I_t (*Ground Truth*); randomly sampled crops as targets I_t (*Random Generation*); or two independent random crops, commonly used in self-supervised learning to form positive pairs I_c and I_t (*Random Cropping*). Their performance is evaluated on the in-domain COCO-OCD dataset. Top-1 accuracy is reported alongside example context–object pairs generated by each variant, with frame colors corresponding to the bars in the plot. Across **A**, and **B**, the chance level (6.7%) is indicated by a horizontal black dashed line. Error bars indicate the standard error of the mean (SEM). * denotes performance significantly above chance based on Welch's two-tailed t-test ($p < 0.05$). **C.** Proposal quantity has a slightly greater impact than proposal quality in SeCo. Top-1 accuracy on the in-domain COCO-OCD dataset is shown as a function of proposal quality (light orange; IoU threshold on bottom x-axis) and proposal quantity (dark orange; number of proposals on top x-axis). For the quality analysis, the number of proposals per image is fixed while varying the IoU threshold. For the quantity analysis, the IoU threshold is fixed at 0 while varying the number of proposals. Shaded regions denote standard error. **D.** To examine what SeCo encodes in its external memory, we use test samples from the in-domain test set of COCO-VOC dataset and extract latent representations using the pre-trained context encoder E_c (trapezoid). Attention scores over memory slots (squircles) are computed, and for each sample, the slot with the highest attention weight is identified (blue arrow). **E.** Category-wise distributions of most-activated memory slots are aggregated across all samples, and pairwise Kullback–Leibler (KL) divergence is computed and visualized as a heatmap. Low divergence (dark cells) indicates that categories sharing similar contexts attend to similar memory slots. Colored boxes and arrows denote COCO-VOC supercategories: Vehicle (green), Animal (dark blue), and Indoor (cyan). For example, "bicycle" and "car" within the Vehicle supercategory activate overlapping memory regions, reflecting shared contextual associations.

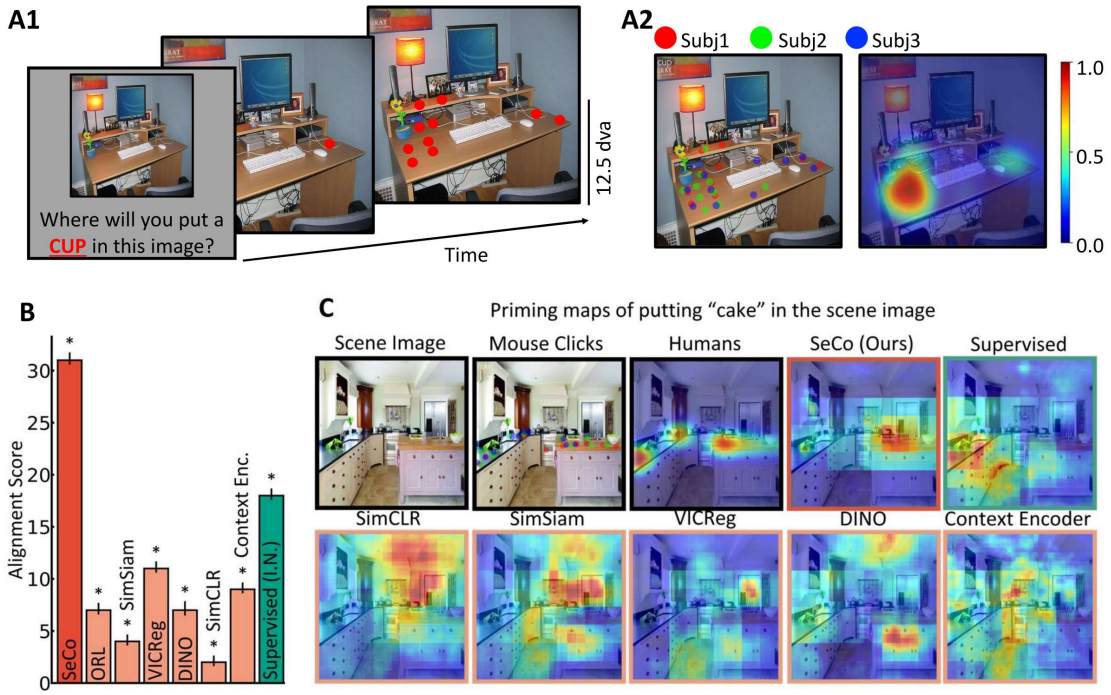


Figure 6: **SeCo excels in the object priming task. A. Schematic for human psychophysics experiments in the object priming task. A1.** Participants were first shown a natural image along with a target object. They were then instructed to place the object in appropriate locations by making 10 non-overlapping mouse clicks (indicated as red dots). There was no time limit for each trial. The scale bar on the right indicates the image size in degrees of visual angle. **A2.** Example priming map based on mouse clicks from humans. The left image shows the mouse click locations made by three human participants (colored dots) for *cup* as the target object. The right panel shows the corresponding normalized probabilistic priming map generated from the consolidated clicks. A higher density of clicks corresponds to a higher probability in the priming map. See the color bar on the right for probability values. **B.** Alignment scores between human priming maps and those predicted by SeCo (red), SSL baselines (orange), and a supervised baseline (green) are presented in the object-priming task. See **Methods** for the definition of alignment scores. A chance model yields a score of 0. Higher scores indicate better alignment with human clicks. Error bars show the standard error of the mean across five independent runs per model. * denotes performance significantly above chance based on Welch's two-tailed t-test ($p < 0.05$). **C.** SeCo priming maps highlight contextually relevant regions of the image and closely approximate human choices in the object priming task. The leftmost column of the first row shows the input scene image where human and computational models are required to put the cake. The rest of the priming map from left to right, up to down are the consolidated click map from 3 human subjects with 10 non-overlapping mouse clicks (dots) each in different colors, ground truth priming maps generated from consolidated human mouse clicks, and priming maps predicted by our SeCo (framed in red) and predicted by supervised baseline (framed in green) and SSL baselines (framed in orange). See the color bar in **A2** for probability values.

Supplementary Figures

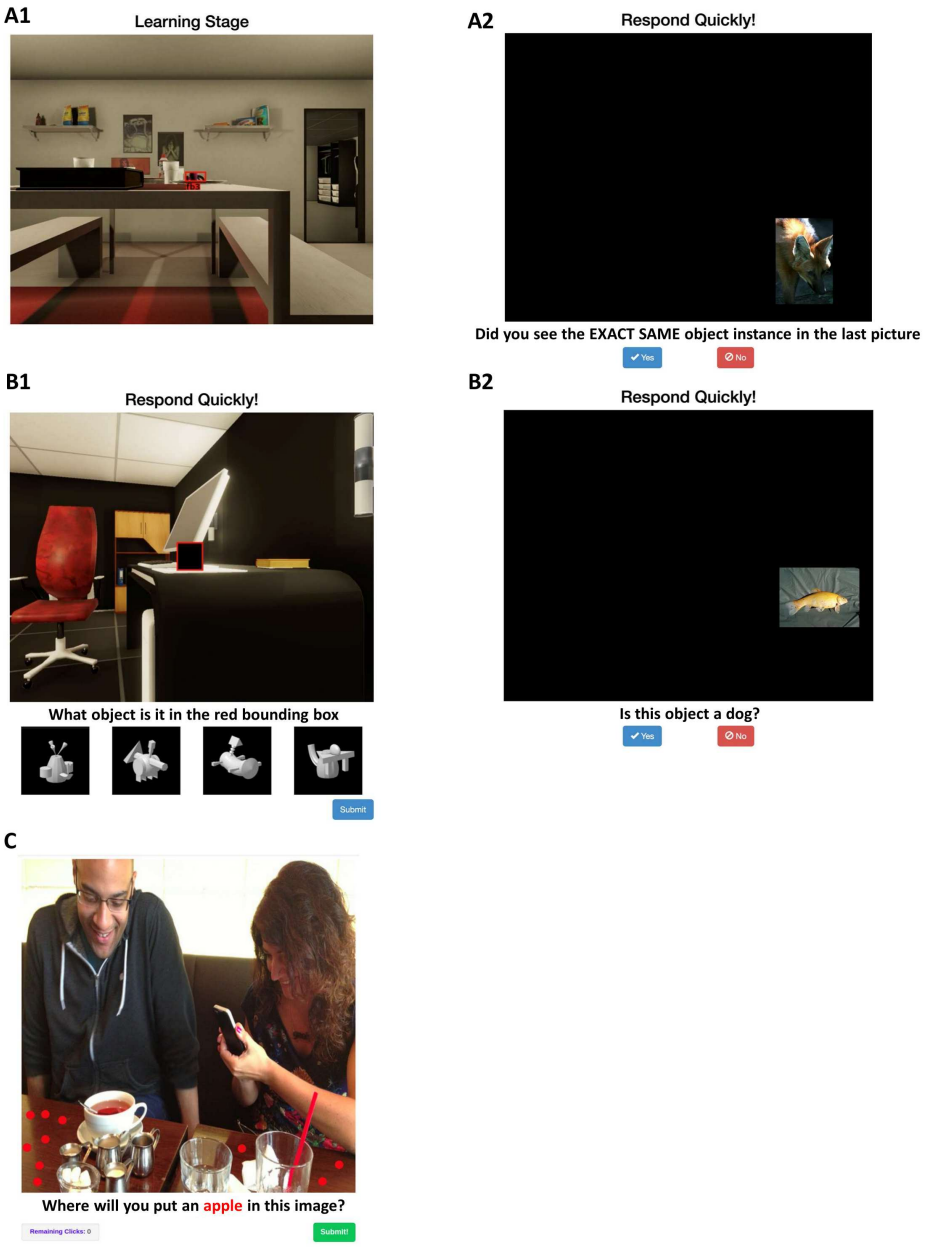


Figure S1: **Amazon Mechanical Turk (AMT) user interfaces of the psychophysics experiments. A1.** AMT user interface of the training stage from FRINE dataset. This example shows a screenshot when the human subject learns a fribble from family *fb3* in the supervised learning condition. **A2.** AMT user interface of memorization test in the training phase. In this example, an image of a fox was presented along with the question. The participant was required to decide whether they had seen this object in the current training video. **B1.** AMT user interface of the lift-the-flap task in the test stage. A test video where the fribble was hidden under the black patch was presented along with the lift-the-flap question and four answer choices. The participant was expected to infer the identity of the hidden target and choose one of four fribble families as their answer. **B2.** AMT user interface of the control trials in the test stage. This example shows an image of a fish and the participant was required to decide whether it is a dog. **C.** AMT user interface of object priming task. Participants are presented with various image-object pairs. For each pair, participants were required to make 10 non-repeated mouse clicks at relevant regions of the image. In this example, the participant was required to place an apple in the given scene. Red dots indicate the past click locations of the participant.

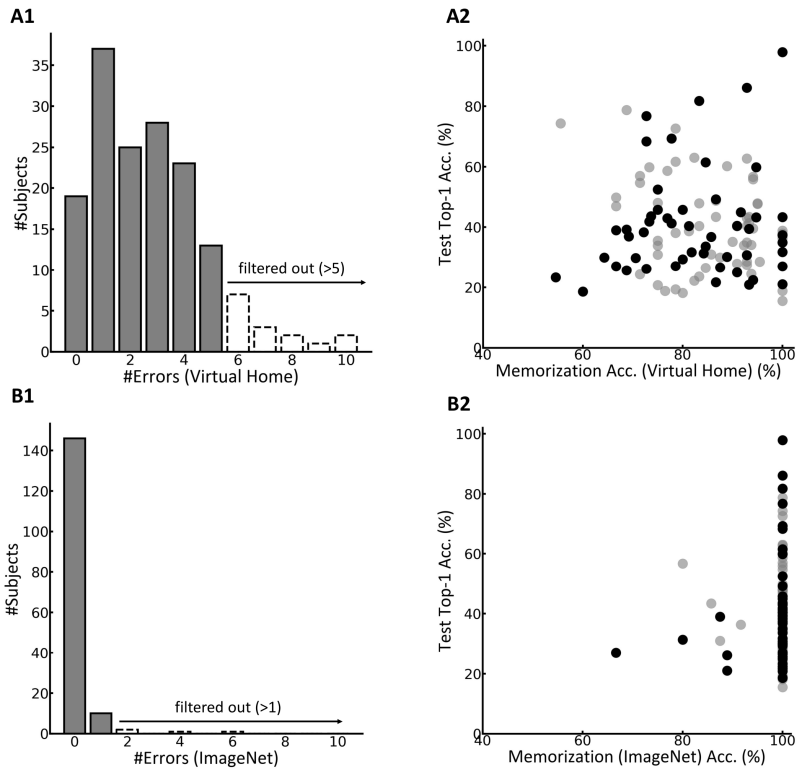


Figure S2: **Quality of human subject data is high, with most participants passing control trials with high accuracy.** We report performance on the memorization test in control trials during the training phase of the human psychophysics experiment on the FRINE dataset. **A1.** Error distribution in the memorization tests where test objects were either positive samples from the current training video clip of FRINE (VirtualHome-Present) or negative samples from other FRINE training videos but absent from the current clip (VirtualHome-Absent). See **Methods** for details. Data from participants who made more than five errors (indicated by dashed white bars) were excluded. **A2.** Memorization test accuracy (VirtualHome-Present and VirtualHome-Absent) versus top-1 accuracy on the lift-the-flap task on the FRINE dataset. Each dot represents one participant trained via supervised learning (black) or self-supervised learning (gray). **B1.** Error distribution in the memorization test where test objects were always negative and they were sampled from ImageNet images whose class labels never appear in FRINE (ImageNet-Absent). See **Methods** for details. Data from participants who made more than one error (indicated by dashed white bars) were excluded. **B2.** Memorization test accuracy (ImageNet-Absent) versus top-1 accuracy on the lift-the-flap task using the FRINE dataset. Each dot represents one participant trained via supervised learning (black) or self-supervised learning (gray).

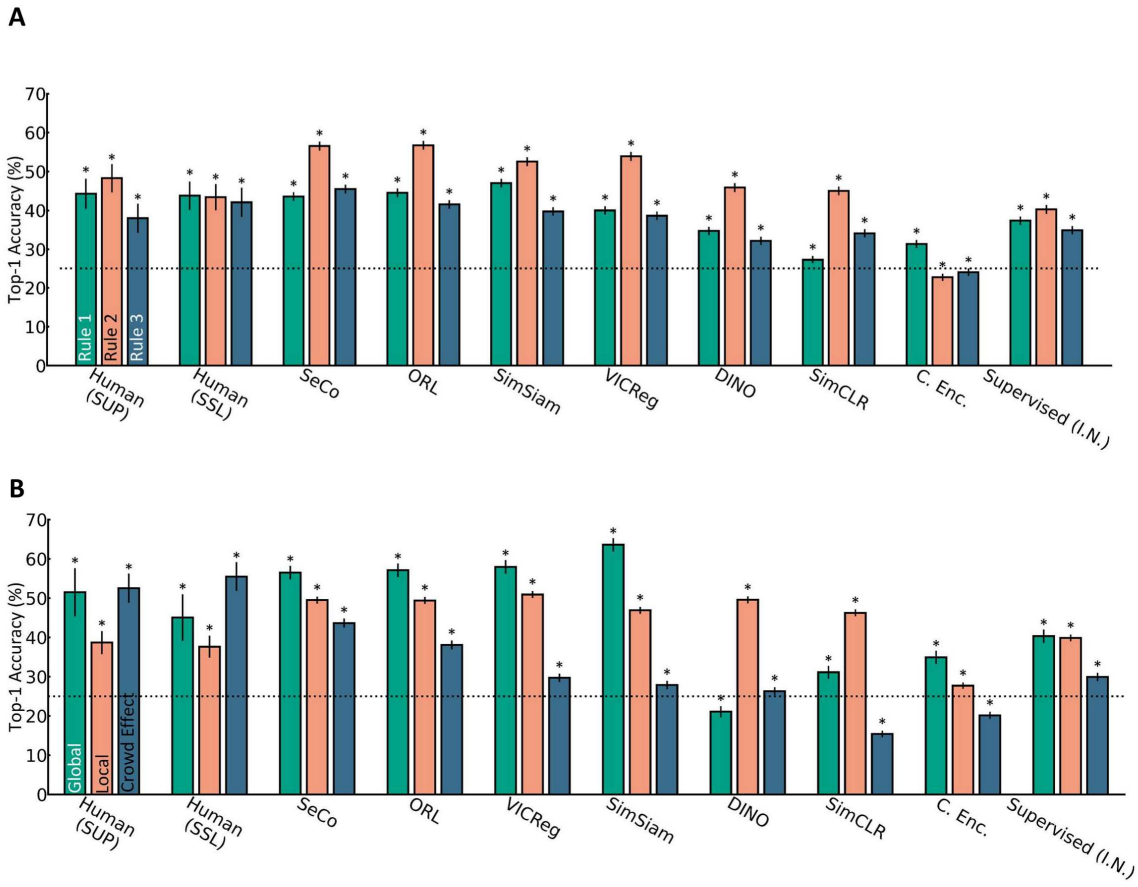


Figure S3: **The performance of humans and AI models on all three context rules and all three contextual associations on FRINE dataset.** Top-1 accuracy of human participants and AI models in the lift-the-flap task under each **A.** contextual rule and **B.** type of contextual association on the FRINE dataset. From left to right, the models are ordered as follows: human participants trained via supervised learning, human participants trained via self-supervised learning, SeCo, ORL, VICReg, SimSiam, DINO, SimCLR, Context Encoder (C. Enc.), and the supervised model (Supervised (I.N.)). See **Methods** for details on all baselines. The chance-level performance is indicated by a horizontal black dotted line. The design convention follows that of **Fig. 3F**.

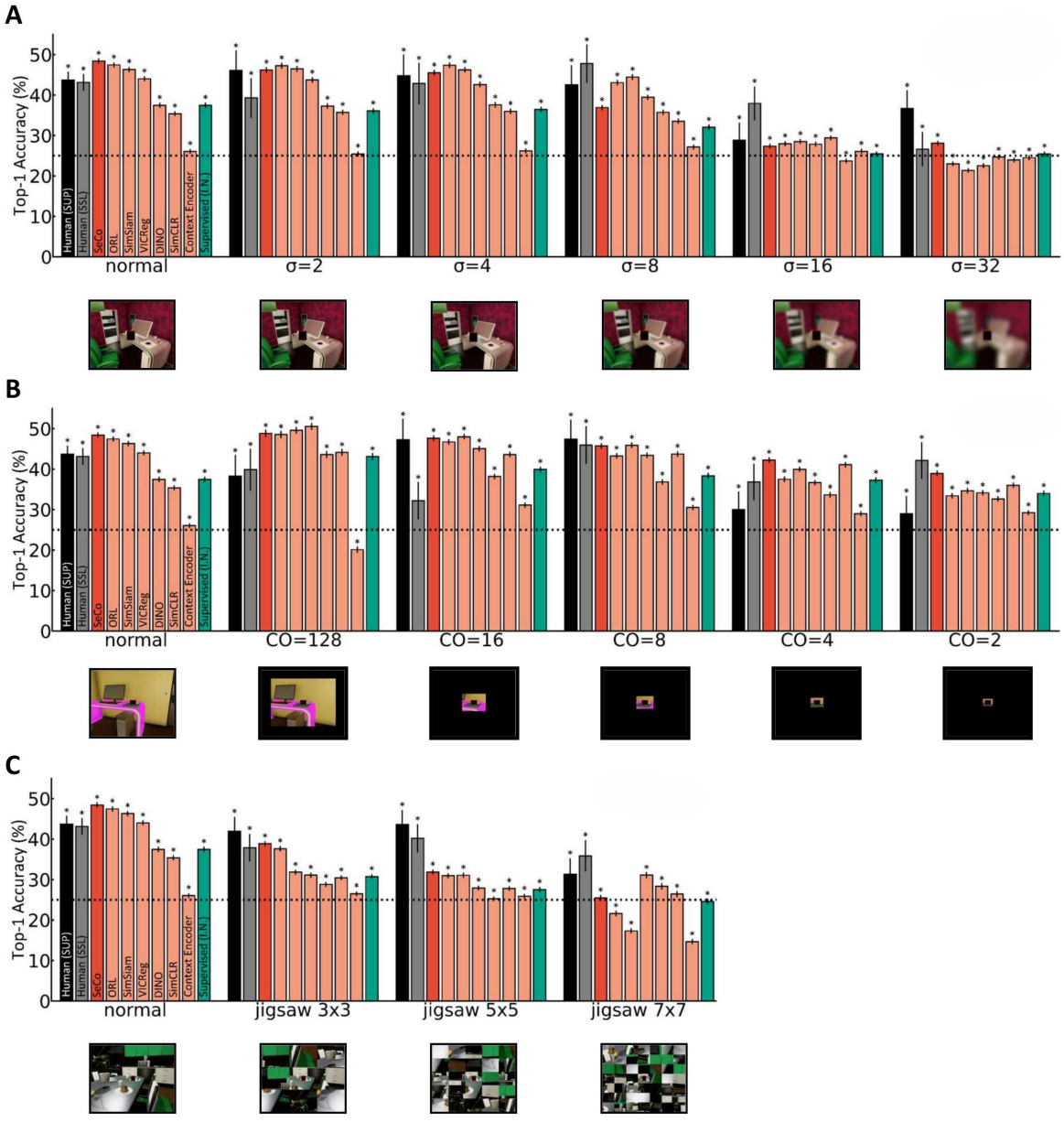


Figure S4: **The effect of context on human and AI models in the lift-the-flap task on the FRINE dataset.** The design conventions in **A**–**C** follow those in **Fig. 3C**–**E**, respectively. See **Methods** for details on all baselines.

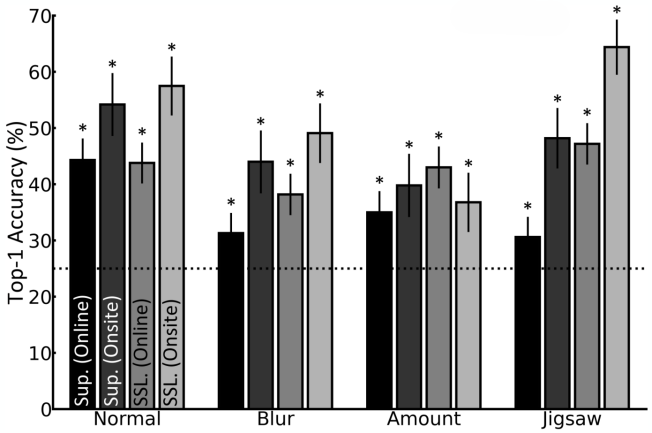


Figure S5: **Comparison of human performance in online and onsite experiments.** From left to right: top-1 accuracy of SUP humans in the online experiments (Sup. Online), SUP humans in the onsite experiment (Sup. Onsite), SSL humans in the online experiment (SSL. Online), SSL humans in the onsite experiment (SSL. Onsite). Results were grouped according to four context manipulations: normal, blur, amount, and jigsaw. The format and conventions follow those in **Fig. 3D**. We recruited 20 onsite subjects (17 subjects left after the data quality control) to study the effect of prolonged exposure in LoR and the design of the curriculum on human performance of the lift-the-flap task on the FRINE dataset. See **Methods** for details on the onsite experiments.

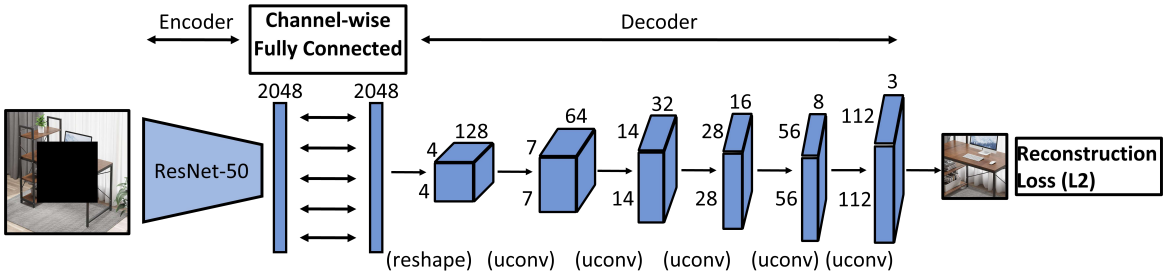


Figure S6: **The architecture of Context Encoder [75] with ResNet-50 [47] as the backbone encoder.** To ensure fair comparisons with other SSL baselines, we replaced the AlexNet backbone used in the original Context Encoder with a ResNet-50 (trapezoid) architecture (see **Methods**). Following the original design, the encoder output is passed through a channel-wise fully connected layer (shown as double-headed arrows), followed by a five-layer decoder (represented by cuboids of varying sizes) that reconstructs the masked central region. We employed the same pixel-level L2 reconstruction loss as in the original configuration.

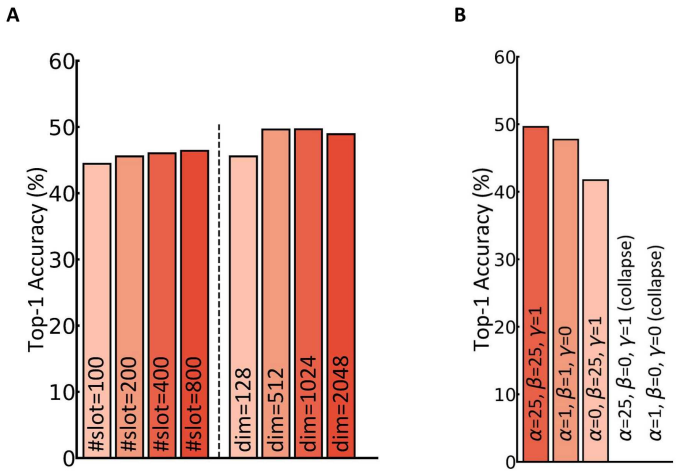
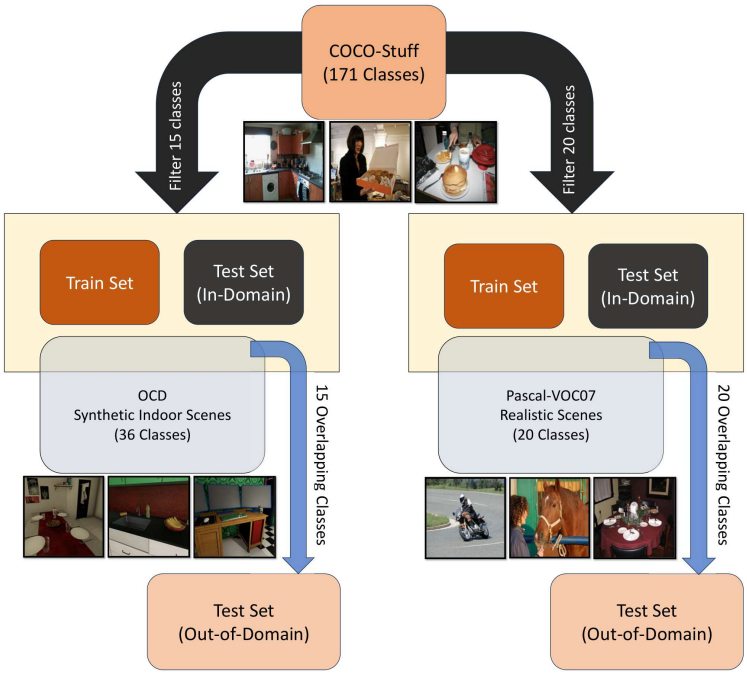


Figure S7: **Impact of memory capacity (number of slots and memory vector dimension) and loss components on SeCo performance in the normal context on the COCO-OCD dataset in the lift-the-flap-task**. **A.** We analyze top-1 accuracy as a function of the number of memory slots (K) in the external memory module of SeCo and the vector dimension (H) for each memory slot. **B.** We also quantify the impact of individual loss components on SeCo's reasoning performance (top-1 accuracy). The hyperparameters α, β, and γ represent the weights assigned to the Mean Squared Error (MSE) loss, variance loss, and covariance loss, respectively (Refer to **Eqn. 2**). Removing the variance loss component $\beta = 0$ results in model collapse. Hence, top-1 accuracy is not applicable in those cases.



(A) COCO-OCD Curation (B) COCO-VOC Curation

Figure S8: **Schematic overview of the systematic curation of the COCO-OCD and COCO-VOC datasets**. **(A)** COCO-OCD curation: Training and in-domain test sets were derived from COCO-Stuff [13] by filtering for the 15 object classes shared with the Out-of-Context Dataset (OCD) [11]. An out-of-domain test set was curated from OCD synthetic indoor scenes, restricted to the same 15 classes under normal context conditions. **(B)** COCO-VOC curation: Following a procedure analogous to (A), training and in-domain test sets were derived from COCO-Stuff [13] by filtering for the 20 object classes shared with the PASCAL-VOC07 dataset [35]. The out-of-domain test set was curated from realistic PASCAL-VOC07 scenes containing the same 20 target classes.