Bridging the Granularity Gap: Object-Centric Masking for Contextual Visual Learning

Jike Zhong
University of Southern California
jikezhon@usc.edu

Abstract

In LLMs, emergent capabilities such as in-context learning and chain-of-thought reasoning have been closely associated with learning over discrete prediction units that are often semantically meaningful. In contrast, vision transformers, and multimodal LLMs (MLLMs) built on top of them, have yet to exhibit similarly robust capabilities. We hypothesize that this discrepancy stems in part from vision encoders being typically pre-trained over spatial patch tokens that are only weakly aligned with semantic entities, leading to representations that are less object-aware and less sensitive to global context, and therefore transfer less effectively to downstream tasks such as VQA and spatial reasoning. To bridge this gap, we propose to model objects as a stronger ***semantic unit*** *for visual prediction, encouraging the encoder to learn the global context and semantics among visual elements. Specifically, we conduct a pilot study in the masked image modeling setting, where this hypothesis can be tested cleanly by masking* ***visual objects*** *rather than random patches during pretraining. Across qualitative analyses and quantitative benchmarks, we show that an object-centric objective reduces pixel-averaging shortcuts and yields more globally coherent and context-consistent representations. When used as the vision encoder in the MLLM frameworks LLaVA and BLIP, the resulting representations improve multimodal QA and vision-centric understanding benchmarks, including VQA, GQA, ScienceQA, and CVBench, by up to* ***8.57*** *points, indicating stronger context utilization. Overall, our results highlight object-centric prediction as a simple yet effective design choice for learning more semantic-rich and context-aware vision encoders, offering a promising direction for improving visual and multimodal intelligence.*

1. Introduction

Large language models (LLMs) trained with predictive objectives over discrete tokens have demonstrated robust context-dependent behavior, including in-context learning and chain-of-thought prompting [40, 53, 54]. One practical advantage of language modeling is that prediction is defined over units that often correlate with semantic entities and relations. By contrast, despite extensive progress in architectures and scaling, vision transformers (ViTs)—and the multimodal LLMs (MLLMs) that build on them—are still predominantly pre-trained over fixed spatial patches [5, 14, 23, 46]. These patches are effective computational units, but they are only weakly aligned with objects, parts, and scene relations.

We study this gap through the lens of **representation granularity**. As summarized in Table 1, language modeling uses a token interface that is comparatively dense in semantic content, whereas ViTs operate on structural patch tokens. This mismatch suggests that predictive learning in vision may admit shortcut solutions that do not require modeling objects and relations. Indeed, Figure 1 illustrates a characteristic failure mode in masked image modeling (MIM): when trained with random patch masking [23], reconstruction can be satisfied by local interpolation or low-information averages, with limited reliance on global scene context.

Motivated by this observation, we investigate whether *object-level removal* can serve as a stronger semantic supervision signal than random patch removal. We use masked image modeling (MIM) [24] as a simple and controllable testbed because it lets us intervene directly on the corruption/prediction unit while still allowing qualitative visualization and downstream transfer evaluation. Concretely, we propose an object-centric MIM objective that masks entire objects (via coarse instance masks) rather than random patches, and reconstructs them from the surrounding context. By removing object-specific cues, this objective encourages the model to rely on global scene context and inter-object relationships instead of patch-local statistics.

Scope and non-claims. We intentionally do *not* claim that our method solves visual tokenization in general. The encoder still ingests patch tokens, and not every image admits a clean decomposition into discrete "objects." We instead position object-centric masking as a *first-order semantic probe*: if changing only the masking/prediction unit already improves context use and transfer, then granularity is likely

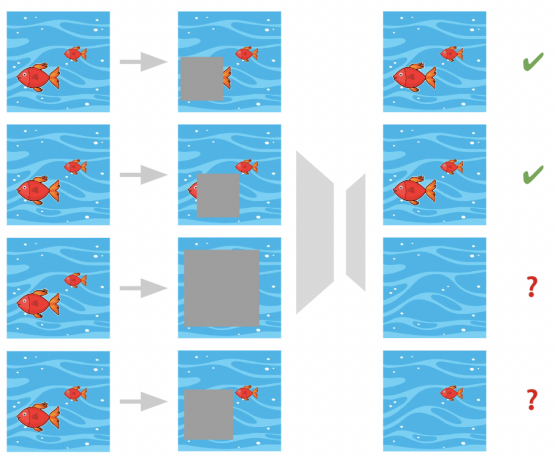

Figure 1. **By masking out random patches, current MIM setup encourages shortcut learning where generation is entirely based on surrounding pixels with limited reference to global context.**

	Token–Semantics Alignment	Semantic Info Density	**Representation Granularity**
Language	High	Dense	**Semantic (word/subword)**
Vision	Low	Sparse	**Structural (patch)**

Table 1. **Language tokens align more directly with semantic entities, while patch tokens are weakly aligned, motivating object-level prediction as a training signal.**

an important part of the broader vision–language mismatch. This framing also addresses an important practical point: our method is a minimal change to standard patch-based MIM pipelines rather than a replacement for them.

We evaluate our method on both multimodal and vision-only tasks. Qualitatively, object-centric masking produces more coherent scene reconstructions and stronger contextual behavior under visual prompting. Quantitatively, it yields especially strong gains on vision-centric spatial reasoning (up to 8.57 points on CVBench) and consistently improves VQA-style benchmarks when used as the vision encoder in both LLaVA and BLIP pipelines. To avoid overselling the phrase "semantic representation," we also make our evaluation target explicit: we measure whether the learned representation is more *context-sensitive*, more *factor-consistent*, and more *transferable* than the patch-masked alternative.

To summarize, our contributions are four-fold:

- We identify **representation granularity** as a concrete and testable source of the vision–language mismatch, and we frame object-centric masking as a controlled probe of that hypothesis.
- We propose a simple object-centric masked image modeling objective that changes the *prediction unit* from random patches to whole objects while keeping the underlying ViT patch encoder unchanged.
- We operationalize "better semantics" through explicit diagnostics—context recovery, factor consistency, and downstream transfer without recognition collapse—rather than relying only on headline benchmark gains.
- We show through qualitative, quantitative, and ablation studies that this simple objective improves context-dependent visual behavior and downstream multimodal reasoning, making it a strong workshop-level baseline and a useful direction for future tokenizer research.

2. Related Works

Vision encoders for MLLMs. The vision encoder is an integral part of modern MLLMs. Earlier designs such as BLIP [30, 31] rely on pretrained ViT backbones, while LLaVA [33] leverages contrastively pretrained encoders (CLIP [42]). Subsequent systems—including MiniGPT-4 [59], InstructBLIP [10], and Qwen-VL series [2, 3, 50]—largely retain patch-based visual interfaces. Our work is complementary to these systems: we do not redesign the MLLM architecture, but instead ask whether a different *pretraining objective* can make the resulting visual representation more context-aware before alignment to language.

Masked image modeling and semantic guidance. MIM learns representations by reconstructing corrupted images. Early work used CNNs [41, 48], while MAE [23] established transformer-based random patch masking as a strong baseline. Subsequent methods enriched the target space or objective, including discrete token prediction in BEiT and its variants [4, 5, 12, 32], contrastive/self-distillation variants such as iBOT and Siamese-MIM [45, 58], and semantic-aware masking as in SemMAE [28, 29]. Relative to this literature, our goal is narrower and more diagnostic: we isolate the effect of replacing random patch removal with *whole-object removal* while keeping the rest of the patch-based pipeline largely fixed. This distinction matters because masking semantically important patches or object *parts* can still leave strong identity cues visible, whereas whole-object removal more directly tests whether the encoder can recover missing content from surrounding relations and scene context.

Region- and object-aware visual learning. A growing body of work augments patch-based representations with region, mask, or localization signals [8, 13, 19, 39, 49, 55]. These studies support the broad intuition that objectness and localization are useful inductive biases. Our paper differs in emphasis: rather than improving detection/localization directly, we study whether object-level corruption changes what the encoder must model in order to reconstruct and transfer.

Visual in-context learning and MLLMs as evaluators. Visual prompting and in-context learning unify tasks such as colorization, detection, segmentation, and inpainting into a single generative interface [5, 9, 18, 27, 51, 52, 56, 57]. These settings are particularly useful here because they expose whether a model is relying on local texture completion or on higher-level scene context. We further evaluate the

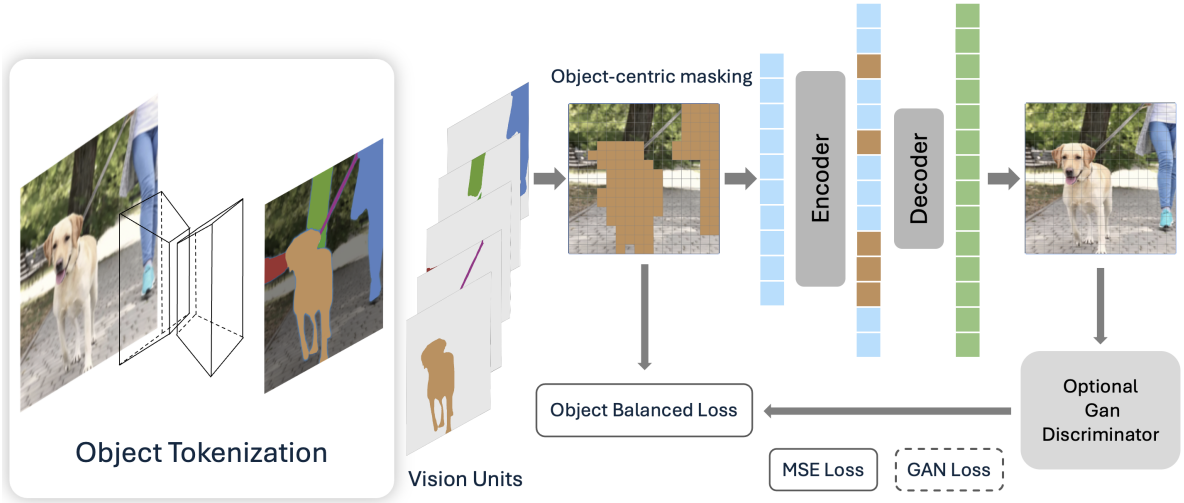

Figure 2. **Overall pipeline for Object-Centric MIM. We utilize a pre-trained segmentation model as an object tokenizer to segment the image into coarse object regions. The masked autoencoder is then trained using object-centric masking, and to further enhance the training of the object-centric encoder, we develop object-balanced loss.**

learned encoder through MLLM fine-tuning, following prior work that treats multimodal QA as a sensitive probe of visual representation quality [31, 33, 46].

3. Learning Global Context with Object-Level Representation

In this section, we introduce our proposed object-level objective using the MIM framework [24], providing insights into future scalable and semantic vision encoder designs for advancing multimodal intelligence.

3.1. Masked Image Modeling

MIM [24] is a simple self-supervised framework that trains a vision encoder by learning to reconstruct corrupted images. In original design, images are corrupted by masking random patches.

Setup. Given an image x, a tokenizer q partitions it into $\mathbb{M}$ non-overlapping patches $\{x_i\}_{i=1}^{\mathbb{M}}$. A random mask $m \in \{0,1\}^{\mathbb{M}}$ selects $\mathbb{N} = \mathbb{M}r_{\text{patch}}$ patches to remove ($m_i = 1$). The encoder receives the visible sequence $\hat{x}_{\text{patch}} = \{\hat{x}_i | (1-m_i)x_i\}_{i=1}^{\mathbb{M}}$. For decoding, removed patches are replaced by a learnable token $e_{[mask]}$ and re-inserted at their original positions. Training uses MSE over masked patches only.

Objective. Let $\mathbb{D}$ be the corpus and θ the model parameters. MIM maximizes:

$$\max_{\theta} \sum_{x \in \mathbb{D}} \mathbb{E}_{\mathbb{N}} \left[\sum_{i \in \mathbb{N}} \log \mathbb{P}_{\theta}(x_i | \hat{x}_{\text{patch}}) \right] \tag{1}$$

where reconstruction is defined over patch tokens. The patch tokenizer is:

$$x_{\text{patch}} \coloneqq q(x; c) \tag{2}$$

where c is typically the patch size. Here q is a spatial divider, distinct from semantic tokenizers in Bao et al. [4], Zhou et al. [58].

3.2. The Object-Centric Objective

We modify MIM by redefining the prediction unit from patches to objects (illustrated in Figure 2).

Objective. Let each image contain $\mathbb{R}$ objects $\{x_j\}_{j=1}^{\mathbb{R}}$, and let $\mathbb{O} = \mathbb{R}r_{\text{obj}}$ denote masked objects under ratio r_{obj}. We rewrite Equation 1 as:

$$\max_{\theta} \sum_{x \in \mathbb{D}} \mathbb{E}_{\mathbb{O}} \left[\sum_{j \in \mathbb{O}} \log \mathbb{P}_{\theta}(x_j | \hat{x}_{obj}) \right] \tag{3}$$

where x_j denotes a removed object and $\hat{x}_{\text{obj}}$ is the image with entire objects masked out. Unlike contiguous patch masking, object masking is defined by semantic instance boundaries, changing the prediction unit rather than only the mask geometry.

Mask generation and expansion. We obtain object masks via an object tokenizer:

$$x_{\text{obj}} \coloneqq q'(x; \phi) \tag{4}$$

where q' produces coarse instance masks. In practice, we decouple q' from training and use off-the-shelf SAM [26] to generate masks. To avoid overfitting to mask shapes, we expand masks to squares (bounding boxes). Additional mask-generation details and alternatives (e.g., unsupervised methods [22]) are deferred to the Appendix.

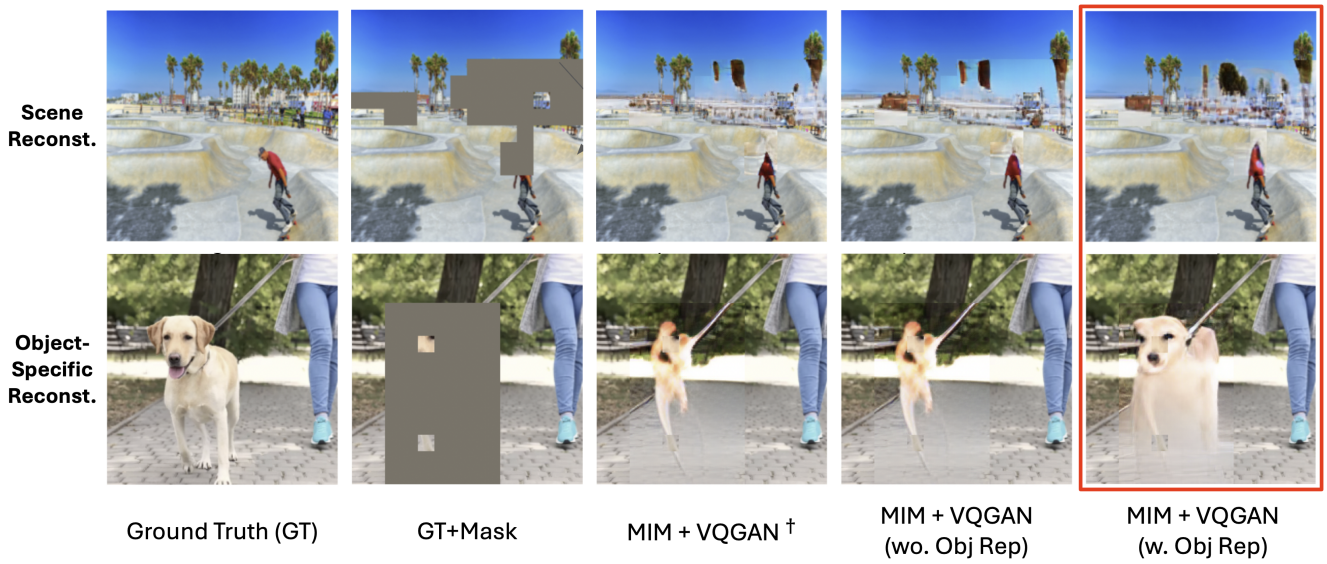

Figure 3. **Reconstruction results of 1) scene reconstruction and 2) object-specific reconstruction.**

Integration. Given object masks, we map each object to its covered patch indices and apply masking at the object level. Because objects vary in size, we cap the total masked-object pixels per image to control computation. Masked objects contribute no pixel tokens to the encoder input; reconstruction is conditioned only on visible regions, forcing prediction to rely on global context.

What changes, and what does not. Our method changes the *corruption and reconstruction unit*, not the encoder input interface. The model still consumes standard patch tokens and therefore should not be interpreted as a new object-native tokenizer. We view this distinction as a feature rather than a limitation for the present study: it lets us test whether object-level removal alone is sufficient to induce more context-dependent representations. A full tokenizer that represents objects, parts, and "stuff" regions directly remains an important direction for future work.

3.3. Learning Context via Object-Centric Objective

Two-stage learning. Directly optimizing Equation 3 can be unstable without basic low-level reconstruction ability. We therefore adopt a two-stage schedule: (i) plain random-patch MIM pretraining with no object information, then (ii) object-centric continued pretraining with an additional size-balancing term.

Optimization. Let $x_j \in \mathbb{R}^{3s_j \times 1}$ and $y_j \in \mathbb{R}^{3s_j \times 1}$ be ground-truth and predicted pixels for masked object j, with object sizes $s = \{s_j\}_{j=1}^{\mathbb{O}}$. The reconstruction loss is:

$$\mathbb{L}_{MIM} = \frac{1}{\Omega(\mathbf{x}_{\mathbb{O}})} \sum_{j \in \mathbb{O}} \sum_{k \in s_j} (y_j^k - x_j^k)^2 \tag{5}$$

where $\Omega(\cdot)$ is the total number of pixels across masked objects. To reduce dominance by large objects, we weight per-object errors using a softmax over normalized inverse sizes:

$$\mathbb{L}_{\text{obj}} = \text{Softmax}(-\frac{s}{||s||})^T \cdot \left[\sum_{k \in s_j} (y_j^k - x_j^k)^2 \right]_{j=1}^{\mathbb{O}} \tag{6}$$

Overall, the second stage minimizes a combined objective:

$$\mathbb{L}_{OBJ-MIM} = \mathbb{L}_{MIM} + \lambda_1 \cdot \mathbb{L}_{\text{obj}} \tag{7}$$

with $\lambda_1 = 0.4$, which we empirically found to be the best.

4. Experiments

In this section, we evaluate whether object-centric masking improves semantic and contextual visual representations. Our comparisons are intentionally *controlled*: we keep the backbone, training data, and downstream adaptation pipeline fixed wherever possible, and change primarily the masking/prediction unit. We primarily consider (i) vision-centric tasks—visual inpainting, detection, and segmentation via visual prompting [5]—and (ii) vision–language tasks such as visual question answering using visual instruction tuning [33]. We additionally include a controlled toy study to isolate the effect of object-level masking on global context modeling.

4.1. Evaluating "Semantic Meaningfulness"

A recurring concern for this line of work is that "better semantics" can easily become a vague claim. We therefore make our evaluation target explicit. Rather than claiming

Property	Probe	Result/Evidence
Context sensitivity	Paired-shape recovery with distractors (Table 7).	93.25% recovery vs. 0% for patch masking; the model infers the correct masked object from surrounding relations.
Factor consistency	Controlled color and shape and minimal-context reconstruction (Figure 5).	Reconstructions preserve the key factor and remain context-consistent without collapsing to ambiguous local averages.
Downstream transfer	Transfer to CVBench, VQA, GQA, ScienceQA, and ImageNet LP/FT (Table 4, Table 5, Table 6, Table 10).	Large gains on the hardest spatial tasks (+7.48 / +8.57). Clear improvement on QAs and no recognition collapse.

Table 2. **Semantic probes used in this work.** Instead of claiming a single "semantic score," we evaluate three concrete properties that a more semantic representation should exhibit.

a single universal scalar for semanticity, we evaluate three necessary properties of a more semantic visual representation: **(i)** *context sensitivity*—can the model recover a missing object from inter-object relations rather than local pixels? **(ii)** *factor consistency*—does the model preserve human-salient factors such as shape/color under controlled cues? and **(iii)** *transferable abstraction*—do these benefits survive transfer to spatial reasoning and multimodal QA without damaging recognition? Table 2 summarizes how each property is measured in this paper.

4.2. Qualitative Evaluations

Setup. We evaluate on (i) visual prompting [5], using 4-grid reference/query inputs for copy, inpainting, colorization, and detection, and (ii) scene-context reconstruction, which requires predicting missing content from surrounding context. Following Bar et al. [5], we use a VQGAN tokenizer [15] to produce discrete visual tokens for sharper reconstructions than pixel-regression MAE-style models. To increase scene diversity and contextual complexity, we train/evaluate on the scene-centric SA-1B dataset [26]. To ensure fairness, all baselines use the same additional data and evaluation protocol.

Implementation details. We largely follow Bar et al. [5]. We use a ViT-Large [14] encoder–decoder with 24 encoder blocks and 8 decoder blocks (hidden sizes 1024/512). Image–mask pairs are resized to 224×224 with patch size $p = 16$. For VQGAN, we adopt the ImageNet-pretrained codebook from Bar et al. [5] [11] with vocabulary size $|V| = 1024$. Models are initialized from public checkpoints and trained on 500K images for 50 epochs using AdamW [34] with a cosine learning-rate schedule (base LR 1×10^{-5}). All experiments run on a single NVIDIA A100. Additional details are in the Appendix.

Analysis. Figure 4 evaluates visual in-context learning under the unified prompting template (copy, inpainting, colorization, detection). Across tasks, object-centric masking produces outputs that better preserve semantically meaningful details and adhere to the reference–query relationship. For instance, in the copy task (first column), our prediction recovers the small leaf attached to the orange that the patch-masked baseline omits; similar improvements are observed in the other prompted tasks.

Figure 3 further tests whether the model captures *global* scene semantics beyond local interpolation. In scene reconstruction (top row), patch masking leads to visually inconsistent artifacts (e.g., abrupt sky-color discontinuities and implausible structures), whereas object-centric masking yields more coherent and natural completions consistent with surrounding context. In the harder object-specific reconstruction (bottom row), where only minimal cues remain, our model infers the missing object from relational context (e.g., person + leash $\rightarrow$ dog), while patch masking collapses to unrecognizable content. This behavior is consistent with the "pixel-averaging" shortcut: without object-level guidance, the model can minimize reconstruction loss via low-information averages rather than modeling the underlying object distribution. Finally, important to clarify that these gains are not explained by additional training data: the random-masking model fine-tuned with the same extra data remains visually similar to the original checkpoint (columns 3 vs. 4), indicating that the primary driver is the *object-level* prediction unit.

4.3. Quantitative Evaluations

Setup. We evaluate our method on two task groups: (1) traditional vision tasks via visual prompting, including foreground segmentation and single object detection [5], and (2) visual question answering (VQA) [1]. For vision tasks, we follow the qualitative evaluation setup and report mean IoU ($mIoU$) on Pascal-5i [44] and Pascal VOC 2012 [16]. For VQA, we pair our visual encoder with an LLM tuned following BLIP [30] and LLaVA [33], and evaluate on VQA-V2 [21], GQA [25], ScienceQA [36], and CVBench [46], reporting average multiple-choice accuracy. We use the LLaVA suite as the primary reasoning evaluation and BLIP-VQA as a second architecture check under fixed compute. Further benchmark details are in the Appendix.

Implementation details. For multimodal instruction tuning, we utilize our trained vision encoder and largely follow the same pipeline and procedure as outlined in Li et al. [31], Liu et al. [33].

Model	Foreground Segmentation $mIOU \uparrow$				Single Object Detection $mIOU \uparrow$			
	Split1	Split2	Split3	Split4	Split1	Split2	Split3	Split4
BEiT* [4]	5.38	3.94	3.20	3.29	0.17	0.02	0.14	0.16
MIM* [23]	17.42	25.7	18.64	16.53	5.49	4.98	5.24	5.84
MIM (wo. Obj Rep)	17.58	25.0	19.14	16.13	5.19	5.30	5.24	5.24
MIM (w. Obj Rep)	18.18	25.89	19.23	17.34	5.52	5.23	5.74	5.98
MIM+VQGAN† [5]	27.83	30.64	26.15	24.00	24.20	25.2	25.35	25.12
MIM+VQGAN (wo. Obj Rep)	27.33	29.24	27.15	24.53	24.21	24.88	25.15	25.99
MIM+VQGAN (w. Obj Rep)	**28.32**	**31.02**	**27.34**	**25.13**	**26.21**	**26.41**	**28.19**	**27.43**

Table 3. **Results for foreground segmentation and object detection. "†" denotes direct evaluation using public checkpoints, $*$ denotes entries copied from Bar et al. [5]; *notations apply to all subsquent experiments*. All other methods are trained using the same data.**

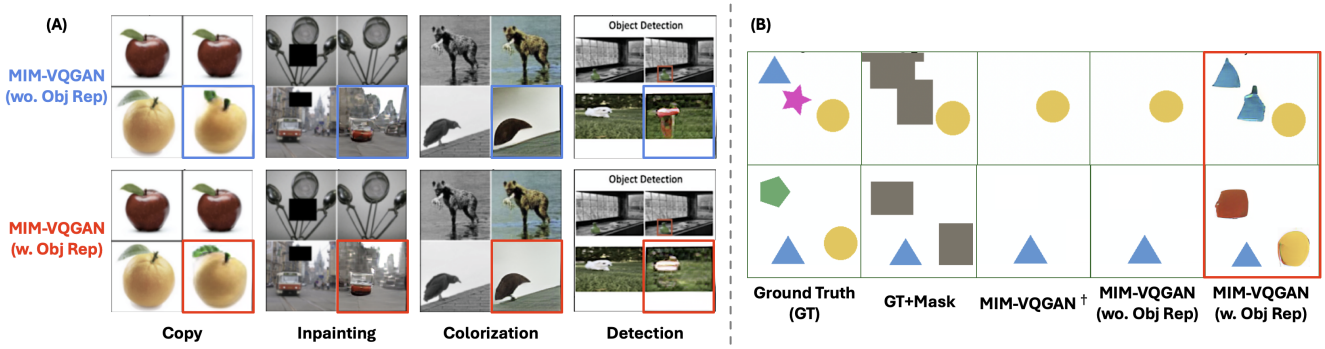

Figure 4. **(a) Visual in-context learning results on various vision tasks. (b) Reconstruction results on toy "context" dataset.**

Model (w. LLaVA)	Spatial Relations	Object Count	Depth Order	Relative Distance
MIM†	53.13	44.28	52.90	48.19
SemMAE†	54.59	44.87	51.70	52.95
MIM (wo. Obj Rep)	53.02	**45.60**	52.64	48.36
MIM (w. Obj Rep)	**60.50**	**45.60**	**55.65**	**56.93**

Table 4. **Performance breakdown on *vision-centric* tasks (CVBench).**

Analysis: Vision-centric spatial reasoning (CVBench). Table 4 shows that object-centric masking yields the largest gains on CVBench, especially on sub-tasks that require *global* scene understanding rather than local texture completion. Compared to patch-based training with the same backbone and protocol (MIM (wo. Obj Rep)), our method improves *Spatial Relations* from 53.02 to 60.50 (+7.48) and *Relative Distance* from 48.36 to 56.93 (+8.57), while also improving *Depth Order* from 52.64 to 55.65 (+3.01). In contrast, patch-masked baselines (including MIM† and SemMAE†) only yield modest changes, and additional data under random masking does not explain the improvements (MIM† vs. MIM (wo. Obj Rep) are nearly unchanged across most columns). These results support our central claim: redefining the prediction unit from patches to objects strengthens context-dependent visual representations.

Analysis: Vision-language Question Answering. On standard VQA-style benchmarks, object-centric pretraining consistently improves downstream MLLM performance (Table 5). Our model improves VQA(v2.0) from 53.44 to 56.89 (+3.45), GQA from 36.98 to 40.00 (+3.02), and ScienceQA from 40.46 to 42.98 (+2.52) over the patch-masked baseline trained under the same setup. These gains are consistent with better utilization of scene context and compositional cues, which are often required in GQA/ScienceQA beyond simple recognition. The same trend holds when using BLIP as the MLLM (Table 6). Object-centric masking improves the overall VQA(v2.0) average from 51.79 to 52.97 (+1.18), with consistent gains across question types (e.g., *Num.*: 36.23→37.30; *Other*: 41.30→42.88). We report BLIP on VQA-v2 only as a light-weight cross-architecture sanity check; the full GQA/ScienceQA suite is run with LLaVA, which is our primary reasoning testbed. Together, Table 4, Table 5, and Table 6 indicate that the primary benefit comes from an improved *visual token representation* (object-level prediction) rather than from additional data or model scaling. Finally, we emphasize that our goal is not to push SOTA, but to test whether improving the objective and tokenization interface can advance vision encoders for multimodal reasoning. We also do not directly compare against CLIP [42] as the vision encoder, since integrating object-level masking into CLIP's contrastive training is non-trivial; recent region-aware CLIP variants instead modify alignment/localization mechanisms [8, 13, 39, 49], which is complementary to our findings.

4.4. Toy Study

Setup. To directly test contextual learning in a controlled setting, we construct a synthetic "shape" dataset with five

Model (w. LLaVA)	VQA (v2.0)	GQA	ScienceQA
MIM†	53.02	36.24	39.04
SemMAE†	55.24	37.45	40.62
MIM (wo. Obj Rep)	53.44	36.98	40.46
MIM (w. Obj Rep)	**56.89**	**40.00**	**42.98**

Table 5. **Performance comparison across generic visual question answering tasks: VQA (v2.0), GQA, and ScienceQA.**

Model (w. BLIP)	Fine-Grained Types			Avg
	Num.	Yes/No	Other	
MIM†	36.22	71.00	41.32	51.80
SemMAE†	36.28	70.97	41.37	51.85
MIM (wo. Obj Rep)	36.23	71.15	41.30	51.79
MIM (w. Obj Rep)	**37.30**	**71.69**	**42.88**	**52.97**

Table 6. **Fine-grained results on VQA v2.0.**

shapes. A yellow circle and a blue triangle form an exclusive co-occurrence pair, while the remaining shapes act as distractors; all shapes appear with equal frequency. Given a context pair where one of the paired shapes is visible and the other is missing, the task is to reconstruct the missing paired object from context. We generate 200 training images, train for 100 epochs, and report results in Table 7 (see qualitative examples in the last column of Figure 4).

Model	Context Recovery Rate (%)
MIM+VQGAN† [5]	0.00
MIM+VQGAN (wo. Obj Rep)	0.00
MIM+VQGAN (w. Obj Rep)	**93.25**

Table 7. **Context recovery on the synthetic shape dataset. Only object-centric masking reliably recovers the paired object, indicating explicit context modeling.**

Analysis. As shown in the last column of Figure 4, the object-centric objective learns the co-occurrence rule: it recovers the blue triangle given only the yellow circle (or vice versa) in 93.25% of cases (Table 7), whereas patch-masked baselines fail completely (0%). Importantly, the underlying distribution constrains only the contextual pair; the remaining shapes are unconstrained and may vary, so success is measured solely by whether the paired object is recovered. This makes the probe stricter than unconditional object insertion: merely memorizing object frequency is insufficient, because the model is counted as correct only when it reconstructs the specific paired object implied by the visible cue. The consistent failure of random masking suggests that, without object-level prediction units, the model can minimize reconstruction loss without representing inter-object dependencies—mirroring the shortcut behavior observed in subsection 4.2.

5. Discussion, Limitations, and Practical Considerations

In this section, we directly address the common concerns from readers and clarify the precise scope of the paper.

We propose a masking objective, not a brand new tokenizer. A central point of clarification is that our method does not replace the patch interface of ViTs. The encoder still receives patches; object masks only modify what must be inferred during training. We therefore interpret the gains as evidence that the *prediction unit* matters and a direction that is worth deeper study in future research, not as proof that we have solved semantic tokenization.

Is patch granularity the sole cause of weak visual reasoning? No. We do not claim a single-cause explanation for the gap between vision and language models. Our claim is narrower: when architecture, data, and downstream adaptation are held fixed, intervening on the prediction unit produces the largest gains exactly on tasks that require global scene structure, while recognition remains stable. This makes granularity a plausible *contributing bottleneck*, even if it is not the only one.

Images without clear objects. Objects are an intentionally conservative first approximation to semantic units, not the only possible one. Some scenes are dominated by texture, amorphous "stuff," or attributes rather than discrete countable instances. In those cases, coarse object masks may be imperfect, and future work should extend the idea to parts, stuff regions, and learned segment groupings. Nevertheless, our results show that even this simple approximation is already useful on scene-centric data and downstream reasoning benchmarks.

Isolated semantic metric. We do not believe there is a universally accepted scalar metric of semanticity for visual encoders. Instead, we use multiple diagnostics that probe distinct aspects of semantics: relation recovery in the toy study, factor consistency in controlled reconstructions, and transfer to context-heavy downstream tasks. Our multi-probe framing is closer to how semantic quality is often argued in practice and makes the paper easier to evaluate.

Does removing whole objects make the reconstruction target too ambiguous? Yes—and that ambiguity is partly the point of our study. When an entire object is removed, there can be multiple context-consistent completions. This makes pure pixel regression a less informative end metric, which is why our evaluation does not rely only on reconstruction loss. Instead, we emphasize whether completions are *context-consistent* and whether the resulting encoder transfers better to spatial reasoning and multimodal QA. For qualitative generation, we further use a discrete VQGAN tar-

get, which reduces the blur/averaging issues of direct pixel regression. The minimal-context examples in Figure 5 show exactly this behavior: different valid completions may exist, but object-centric masking steers the model toward plausible completions consistent with the scene.

Does object-level masking simply encourage object hallucination? We do observe that whole-object removal can bias the model toward inserting plausible objects in ambiguous regions; some examples in Figure 5 illustrate this multimodality. To reduce trivial leakage, we expand masks to bounding boxes rather than exposing the exact object silhouette, and we evaluate success using downstream transfer and controlled context recovery rather than reconstruction loss alone. We therefore view false-positive insertion as a genuine limitation of generative reconstruction, not as a complete alternative explanation for the downstream gains.

Does the gain come from object-level masking or SAM? Because we use SAM [26] to obtain object masks, a natural question is whether improvements depend on SAM's mask quality rather than the object-level training signal. Our method only requires coarse masks that roughly localize objects (we use bounding-box expansions in practice), so SAM is a convenient tool but not a necessary component. To isolate this factor, we replace SAM with a fully unsupervised segmentation method [22] to generate masks for the training set and keep the rest of the pipeline unchanged. As shown in Table 8, using unsupervised masks achieves comparable performance, indicating that the gains primarily arise from object-level masking itself rather than the specific mask generator.

Model (w. LLaVA)	VQA (v2.0)	GQA	ScienceQA
MIM (mask w/ SAM)	56.89	**40.00**	**42.98**
MIM (mask w/o SAM)	**57.66**	39.12	42.56

Table 8. **Ablation study: object mask obtained using/without using SAM[26]. Results demonstrate that improvement is NOT tied to SAM.**

Finding 1: Object-level masking encourages explicit context-based semantics in vision models. Qualitative evidence in Figure 3, Figure 4, and Figure 5 shows that object-level masking leads to reconstructions that depend on scene context and inter-object cues rather than purely local texture completion. The same figure also suggests sensitivity to human-salient factors such as *color* and *shape* [43], while the minimal-cue examples remain context-consistent despite ambiguity.

Finding 2: Object-level masking improves reasoning without hurting recognition. To verify that our gains do not come at the expense of recognition, we evaluate the same encoder on ImageNet-1K [11] with linear probing (LP) and full fine-tuning (FT). Appendix Table 10 reports the full numbers. Object-centric masking improves both FT and LP

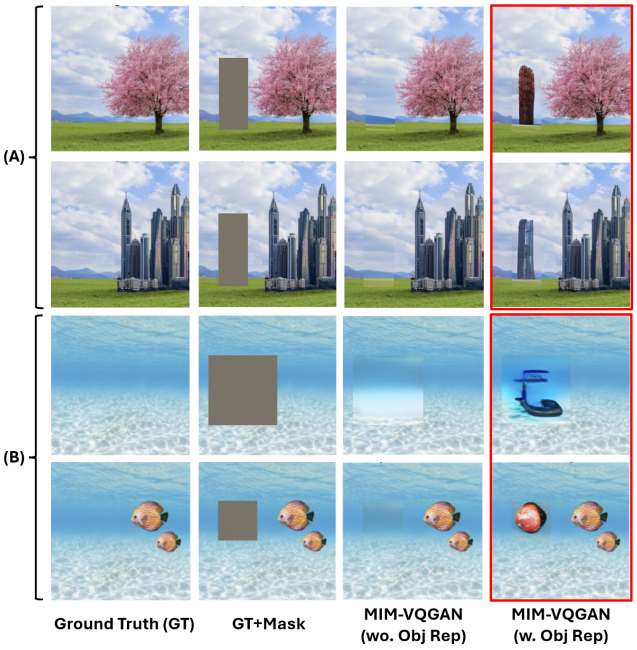

Figure 5. **(A) Object Definition and Context:** The representations of objects are learned based on "Color" and "Shape." **(B) Minimal Context Reconstruction:** Reconstruction with minimal context, with & without object reference.

relative to the matched patch-masked baseline, while random masking with additional scene-centric data degrades substantially, indicating that the benefit comes from the object-level training signal rather than extra data alone.

Finding 3. Random masking encourages learning a "pixel-based shortcut" rather than the true distribution. As shown in Figure 3, Figure 4, and Figure 5, while our approach learns to generate based on a meaningful underlying distribution, random masking results in no object being generated unless it is partially visible. This suggests the model learns a "pixel-based shortcut" akin to interpolation rather than capturing true relationships and semantics.

Practicality and optimization. The method does incur segmentation overhead, but offline preprocessing reduces it substantially; we report the exact cost table and loss ablation in the Appendix. Both support the same conclusion: the main benefit comes from the object-centric corruption signal rather than from loss reweighting alone.

6. Conclusion

In a controlled MIM testbed, masking whole objects—while keeping the ViT patch interface unchanged—consistently yields more context-dependent behavior than random patch masking. The strongest evidence appears on spatial reasoning, controlled context recovery, and multimodal QA transfer, making object-centric masking a strong workshop-level case for taking semantic granularity seriously in vision pretraining.

7. Acknowledgment

This work is supported by an award from the USC and Amazon Center on Secure & Trusted Machine Learning. We also thank Yutong Bai and Alan Yuille for their helpful discussions.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *International Conference on Computer Vision (ICCV)*, 2015. 5

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. 2

[3] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. 2

[4] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEit: BERT pre-training of image transformers. In *ICLR*, 2022. 2, 3, 6

[5] Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei Efros. Visual prompting via image inpainting. *Advances in Neural Information Processing Systems*, 35: 25005–25017, 2022. 1, 2, 4, 5, 6, 7, 13

[6] Clemens G. Bartnik and Iris I. A. Groen. *Visual perception in the human brain: How the brain perceives and understands real-world scenes*. Oxford University Press, 2023. 14

[7] Michael F. Bonner and Russell A. Epstein. Object representations in the human brain reflect the co-occurrence statistics of vision and language. *Nature Communications*, 12(1):4081, 2021. 14

[8] Hong-You Chen, Zhengfeng Lai, Haotian Zhang, Xinze Wang, Marcin Eichner, Keen You, Meng Cao, Bowen Zhang, Yinfei Yang, and Zhe Gan. Contrastive localized language-image pre-training, 2025. 2, 6

[9] Yi-Syuan Chen, Yun-Zhu Song, Cheng Yu Yeo, Bei Liu, Jianlong Fu, and Hong-Han Shuai. Sinc: Self-supervised in-context learning for vision-language tasks, 2023. 2

[10] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 2

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 5, 8

[12] Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Peco: Perceptual codebook for bert pre-training of vision transformers. *arXiv preprint arXiv:2111.12710*, 2021. 2

[13] Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Maskclip: Masked self-distillation advances contrastive language-image pretraining, 2023. 2, 6

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1, 5

[15] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12873–12883, 2021. 5

[16] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015. 5

[17] Zhengcong Fei, Mingyuan Fan, Li Zhu, Junshi Huang, Xiaoming Wei, and Xiaolin Wei. Masked auto-encoders meet generative adversarial networks and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24449–24459, 2023. 14

[18] Thomas Foster, Ioana Croitoru, Robert Dorfman, Christoffer Edlund, Thomas Varsavsky, and Jon Almazán. Flexible visual prompts for in-context learning in computer vision, 2023. 2

[19] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D. Cubuk, Quoc V. Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation, 2021. 2, 14

[20] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. 14

[21] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 5

[22] Oliver Hahn, Christoph Reich, Nikita Araslanov, Daniel Cremers, Christian Rupprecht, and Stefan Roth. Scene-centric unsupervised panoptic segmentation, 2025. 3, 8

[23] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021. 1, 2, 6, 13

[24] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable

vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 1, 3, 12, 13, 14

[25] Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering, 2019. 5

[26] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 3, 5, 8, 12

[27] Feng Li, Qing Jiang, Hao Zhang, Tianhe Ren, Shilong Liu, Xueyan Zou, Huaizhe Xu, Hongyang Li, Chunyuan Li, Jianwei Yang, Lei Zhang, and Jianfeng Gao. Visual in-context prompting, 2023. 2

[28] Gang Li, Heliang Zheng, Daqing Liu, Chaoyue Wang, Bing Su, and Changwen Zheng. Semmae: Semantic-guided masking for learning masked autoencoders. *Advances in Neural Information Processing Systems*, 35:14290–14302, 2022. 2, 13

[29] Gang Li, Heliang Zheng, Daqing Liu, Chaoyue Wang, Bing Su, and Changwen Zheng. Semmae: Semantic-guided masking for learning masked autoencoders, 2022. 2

[30] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. 2, 5

[31] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. 2, 3, 5

[32] Xiaotong Li, Yixiao Ge, Kun Yi, Zixuan Hu, Ying Shan, and Ling-Yu Duan. mc-beit: Multi-choice discretization for image bert pre-training. *arXiv preprint arXiv:2203.15371*, 2022. 2

[33] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 2, 3, 4, 5

[34] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[35] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 12

[36] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering, 2022. 5

[37] Xiaoxiao Ma, Xinai Lu, Yihong Huang, Xinyi Yang, Ziyin Xu, Guozhao Mo, Yufei Ren, and Lin Li. An advanced chicken face detection network based on gan and mae. *Animals*, 12 (21):3055, 2022. 14

[38] A. Martin. The representation of object concepts in the brain. *Annual Review of Psychology*, 58:25–45, 2007. 14

[39] Muhammad Ferjad Naeem, Yongqin Xian, Xiaohua Zhai, Lukas Hoyer, Luc Van Gool, and Federico Tombari. Silc: Improving vision language pretraining with self-distillation, 2023. 2, 6

[40] OpenAI and GPT 4 team. Gpt-4 technical report, 2024. 1

[41] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *CVPR*, pages 2536–2544, 2016. 2

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 2, 6

[43] Irene Reppa, Kate E Williams, W James Greville, and Jo Saunders. The relative contribution of shape and colour to object memory. *Memory & Cognition*, 48:1504–1521, 2020. 8

[44] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. *arXiv preprint arXiv:1709.03410*, 2017. 5

[45] Chenxin Tao, Xizhou Zhu, Weijie Su, Gao Huang, Bin Li, Jie Zhou, Yu Qiao, Xiaogang Wang, and Jifeng Dai. Siamese image modeling for self-supervised vision representation learning, 2022. 2

[46] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Ziteng Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal llms, 2024. 1, 3, 5

[47] Samyakh Tukra, Frederick Hoffman, and Ken Chatfield. Improving visual representation learning through perceptual understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14486–14495, 2023. 14

[48] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *International Conference on Machine Learning proceedings*. 2008. 2

[49] Bo Wan, Michael Tschannen, Yongqin Xian, Filip Pavetic, Ibrahim Alabdulmohsin, Xiao Wang, André Susano Pinto, Andreas Steiner, Lucas Beyer, and Xiaohua Zhai. Locca: Visual pretraining with location-aware captioners, 2024. 2, 6

[50] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024. 2

[51] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning, 2023. 2

[52] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Segmenting everything in context, 2023. 2

[53] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models, 2022. 1

[54] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. 1

[55] Cheng Zhang, Tai-Yu Pan, Tianle Chen, Jike Zhong, Wenjin Fu, and Wei-Lun Chao. Learning with free object segments for long-tailed instance segmentation, 2022. 2, 14

[56] Jiahao Zhang, Bowen Wang, Liangzhi Li, Yuta Nakashima, and Hajime Nagahara. Instruct me more! random prompting for visual in-context learning, 2023. 2

[57] Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. What makes good examples for visual in-context learning?, 2023. 2

[58] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021. 2, 3

[59] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models, 2023. 2

Appendix

We provide all additional details for our paper in the following sections.

- Border Impact. We discuss the limitations and potential future follow-up work.
- Details of the Implementation. We provide additional details of model setup, training schedules.
- Ablation Studies. We provide additional ablation study results, including masking strategies, model size, and object-mask ratio.
- Discussions. We address additional questions about the usage of additional data, the generalization capability of our proposed tokenization objective, as well as impact of auxiliary Gan loss.

A. Broader Impact

Limitations and future work. While our method improves semantic reasoning, there are still some failure cases (Figure 8). For example, when using fine-grained object masking during pre-training—where the mask follows the exact shape of objects—the model may "cheat" by overfitting to the mask shape. In such cases, it quickly learns to fill in the masked area without acquiring meaningful representations. To resolve this issue, we expand the mask to the bounding box. In future work, we aim to develop a more structured and robust tokenizer to enhance the model's reasoning capabilities. Our object masks are coarse and can be produced by multiple mechanisms; nevertheless, object discovery quality and compute cost remain practical considerations. In addition, we acknowledge the cost of segmentation overhead, but in our respectful opinion, our pipeline should be viewed as a proof-of-concept, and the performance gain is strong enough to justify studying it.

Ethics Statement. We ensure that our approach adheres to all legal and ethical guidelines throughout its development, with no violations. Fair compensation was provided to all annotators and graduate students involved in this work. The problems used in our study were collected from publicly accessible exams[1] and resources licensed under CC Licenses[2][3]. This research is conducted solely for academic purposes, and we strictly prohibit any commercial use of the results. Additionally, the spurious captions generated in Section 4 are limited to problem-solving contexts and pose no harm to individuals.

Reproducibility statement. We are committed to efficient and reproducible research. All code, datasets, and models will be publicly released.

[1] https://gate2025.iitr.ac.in/
[2] https://www.allaboutcircuits.com/worksheets/
[3] https://ocw.mit.edu/

B. Additional Implementation Details

Mask generation and preprocessing. To efficiently generate object masks, we leverage off-the-shelf [26], a popular unsupervised segmentation model, to infer scene-centric images (where many objects are present). This step yields a set of binary object masks, which we then convert into the COCO RLE (Run-Length Encoding) format. Note that this step can be done either online (during the forward pass of each batch) or beforehand. Here we test both and empirically find the pre-processing step crucial as it saves roughly $3\times$ GPU hours as shown in Table 9. This solution is scalable as more data can be generated directly using the pre-trained SAM model.

Model	Pre-Processing	Training Cost
MIM (w. Obj Rep)	✓	3.6 ($-2.7\times$)
MIM (w. Obj Rep)	×	9.8
MIM+VQGAN(w. Obj Rep)	✓	5.1 ($-2.5\times$)
MIM+VQGAN(w. Obj Rep)	×	13.2

Table 9. Comparison of training costs in GPU hours with and without pre-processing for 1 epoch training using $500K$ data and a single A100 GPU.

Implementation details on downstream tasks. Following He et al. [24], we first discard the decoder after pre-training is complete. For end-to-end FT, we use AdamW [35] optimizer with base learning rate $blr = 1.0 \times 10^{-3}$, weight decay 0.05, layer decay 0.75 and train for 20 epochs with 5 rounds of warmup epochs. Additionally, we use drop path 0.1 with mixup 0.8 and ensure the effective batch size is 1024 by accumulating SGD iters. For LP, we use base learning rate $blr = 1.0 \times 10^{-1}$ and an effective batch size of 16384 while keeping other settings the same. In our model. each self-attention layer includes $\alpha = 16$ attention heads.

Implementation details on pertaining. For the first stage, we use AdamW [35] optimizer with a base learning of $blr = 1.5 \times 10^{-4}$, weight decay $wd = 0.05$, and the cosine learning rate decay scheduler. We accumulate iterations to emulate the recommended batch size of 4096 and pre-train the model for 25 epochs with 5 warmup epochs. During this stage, the mask ratio is set for $mr_{patch} = 75\%$. For the second stage, we start from the saved checkpoint from stage one. We apply an object ratio of $mr_{\text{obj}} = 50\%$ which randomly masks out 25 objects in each image by hiding the patches spatially covering them. To enable batch processing, we apply an additional mask ratio constraint of $mr_{patch} = 60\%$ on all images. The mask ratio is set 15% lower to accommodate increased difficulty in the objective.

Due to constraints in computing resources, we use publicly available pre-trained checkpoints[4][5] as the starting

[4] https://github.com/facebookresearch/mae
[5] https://github.com/amirbar/visual_prompting

Model	FT (%)	LP (%)
MIM† [24]	83.66	70.80
SemMAE† [28]	83.73	71.25
MIM (wo. Obj Rep)	67.72 ↓15.94	58.75 ↓12.05
MIM (w. Obj Rep)	**84.43** ↑0.77	**71.91** ↑1.11

Table 10. **Linear probing (LP) and finetuning (FT) results on ImageNet-1K.**

model for both stages of pre-training, unless otherwise specified. Importantly, using pre-trained checkpoints does not undermine our objective, as they are trained with a patch-level objective, which aligns with the first stage of our framework for learning low-level representations (Two Stage Learning Section). Essentially, we retrain these models on a different dataset with some adaptations.

Loss function for MIM-VQGAN. MIM-VQGAN was proposed by Bar et al. [5] to study the effectiveness of visual prompting, which effectively shifted the MIM evaluation paradigm from fine-tuning on downstream tasks to direct output generation via prompting. This can be seen as a unified framework for vision tasks. Unlike He et al. [23], which computes the MSE loss by directly regressing on pixel values, MIM-VQGAN instead computes the cross-entropy (CE) loss on the corresponding patch value in the quantized codebook. This design effectively alleviates ambiguity in generation, as the codebook is discrete, unlike pixel values. Notably, the underlying objective—masked autoencoding—remains unchanged. Hence, MIM-VQGAN provides an effective way to directly compare our proposed method. In our experiments, we follow the implementation of Bar et al. [5].

C. Additional Ablation Study.

Influence of different object masking strategies: As shown in Figure 9 and Figure 10, we evaluate reconstruction performance using three masking strategies: masking strictly based on the object shape, masking the square region of the object, and a combination of both. While these visualizations demonstrate the superiority of object-based masking compared to random masking strategies, they also reveal certain limitations. Specifically, relying solely on object shape masking can lead to the model overfitting to the mask shape ("cheating"), while using only square masking results in sub-optimal performance on details. By combining these two strategies, we achieve more realistic and effective reconstruction.

Study on how the model captures context: We investigate and visualize if our model has learned to capture the context during the pretraining process. Here we focus on learning the "shape" and "color", two of the most important ingredients to human learning. As we have addressed learning the "shape" in Figure 5 and Discussion Section, we showcase the learning of color in Figure 7. In this example, when the

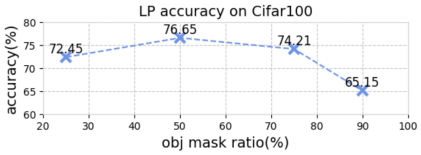

Figure 6. **Effect of object mask ratio:** The number of objects masked out during masked image modeling.

Model	Backbone	Cifar100 Top-1 Acc (%)	
		FT	LP
MIM†	ViT-B	89.98	75.01
MIM†	ViT-L	92.67	76.20
MIM (w. Obj Rep)	ViT-B	90.08	72.44
MIM (w. Obj Rep)	ViT-L	93.77	76.65

Table 11. Comparison of different model sizes. Results show our approach is able to scale with model size.

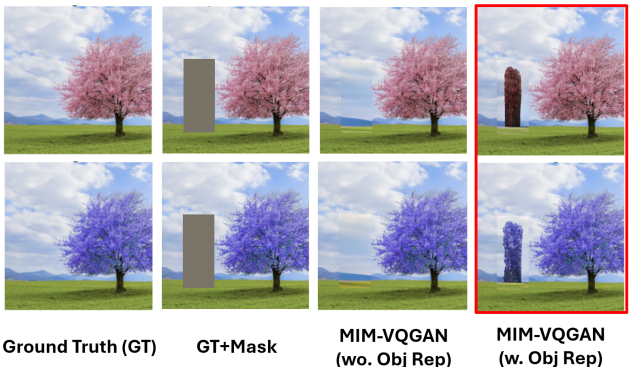

Figure 7. **Extend of color learning example**

same pair of examples but with different colors is given to the model, it is able to reconstruct objects of colors similar to the example, meaning that it does not infer color based on memorization but rather from the context that is given.

Study on model sizes: Table 11 shows the LP and FT results on different vit base models, and the result shows our observations and findings in Quantitative Evaluation and Discussion sections hold for different model sizes.

Obj-Mask Ratio. To determine the influence of the masking strategy, we train our model with different mask ratios, as shown in Figure 6. Unlike traditional random patch-level masking, as in He et al. [24], object-level masking becomes less effective when obj-mask ratios exceed 50%. This decline occurs because random masking often leaves portions of objects visible, which can help guide reconstruction, while object-level masking requires the model to learn the semantic relationships between objects only from other objects. We note that a 50% obj-mask ratio effectively masks out around 75% of the image.

Loss functions. We further ablate the effect of object balance loss defined in Equation 7. Results in Table 12 shows that combining both $\mathbb{L}_{MIM}$ and $\mathbb{L}_{\text{obj}}$ achieves the best performance.

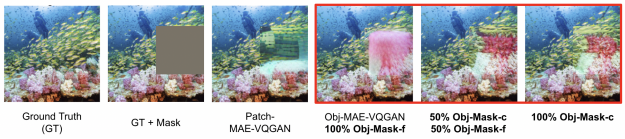

Figure 8. **Failure Cases:** (4): Failure case of reconstruction with fine-grained object masking (Obj-Mask-f). (5)-(6): Remedy by using coarse object masking (Obj-Mask-c)

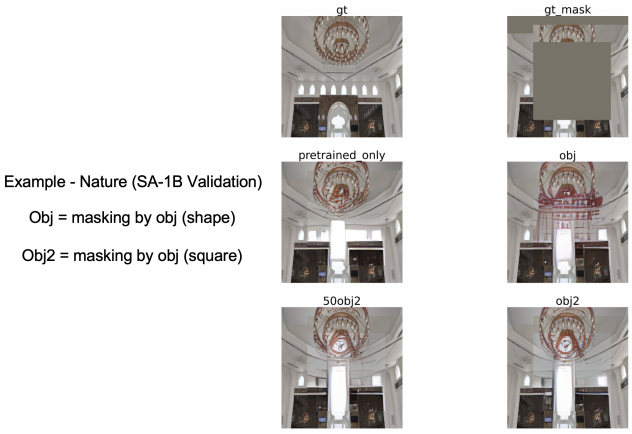

Figure 10. **Ablation Study of Masking Strategies (B)**

Model Variant	VQA (v2.0) Acc. (%)
MIM (w. Obj Rep)	53.02
+ $\mathbb{L}_{MIM}$ only	55.44
+ $\mathbb{L}_{\text{obj}}$ only	52.48
+ $\mathbb{L}_{MIM}$ + $\mathbb{L}_{\text{obj}}$ (Eq. 7)	**56.89**

Table 12. Effect of adding different loss terms in Eq. 7 on VQA (v2.0). Combining both $\mathbb{L}_{MIM}$ and $\mathbb{L}_{\text{obj}}$ achieves the best performance.

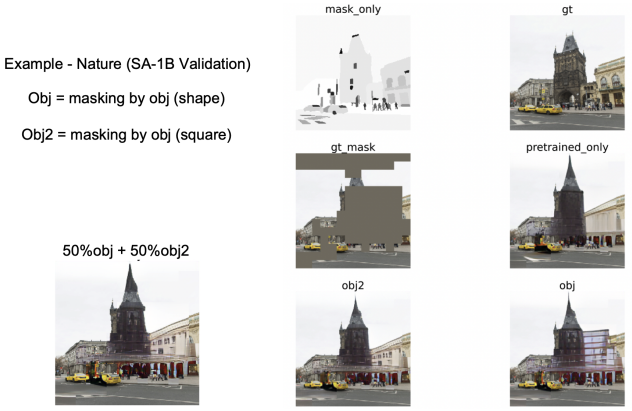

Figure 9. **Ablation Study of Masking Strategies (A)**

D. Additional Discussions.

Model size. Here we show LP results on Cifar-100 classification with ViT-B and ViT-L. Table 11 indicates that our approach is scalable with respect to increasing model sizes.

Additional motivation for using object-level representation. Besides computer vision research, neuroscience studies

Figure 11. **GAN loss** can further help with better details.

have also found that the human brain uses an object-centric approach for visual recognition [6, 7, 38]. Within computer vision research, object segmentations have also been found to be helpful for tasks such as instance segmentation [19] and weakly supervised learning [55]. Hence, we conjecture "object" as a plausible candidate and explore it as the masking unit in MAE by simply masking out random objects and inpainting them instead of random patches.

Generalizability of object-centric objective. The surprising result is that while Patch-MAE severely degrades downstream fine-tuning performance, Obj-MIM can recover such gap in a short GPU-hour, demonstrating that object-centric learning objective enables the learning of highly semantic and generalizable features where the original Patch-MIM cannot, especially given the underlying semantic difference (domain gap) between the datasets.

Furthur enhancing visual details with Gan loss. Generative adversarial networks (GAN) [20] learn representation through the competition of a generator and a discriminator. Recent studies show that adding GAN losses can enhance visual details [17, 24, 37, 47]. Following this intuition, we add an auxiliary GAN loss to our objective in Equation 7:

$$\mathbb{L}_{OBJ-MAE} = \mathbb{L}_{MAE} + \lambda_1 \cdot \mathbb{L}_{\text{obj}} + \lambda_2 \cdot \mathbb{L}_{GAN} \quad (8)$$

This can be achieved by adding a simple discriminator and using the original network as the generator; details can be found in the Appendix. Results in (Figure 11) confirm that GAN loss can help produce more detailed images.