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 $blr = 1.5\times10^{-4}$, 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 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.

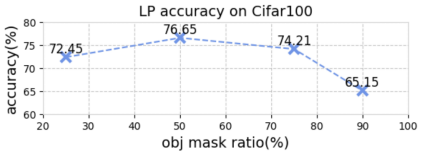

Figure 6. **Effect of object mask ratio:** The number of objects masked out during masked image modeling.

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.

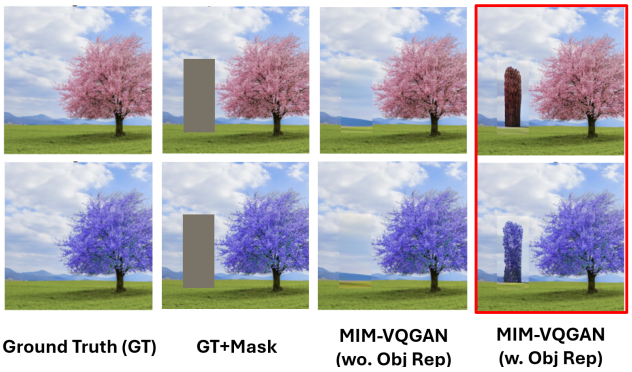

Figure 7. **Extend of color learning example**

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.

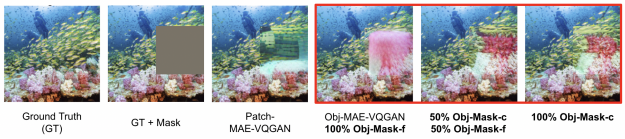

Figure 8. **Failure Cases:** (4): Failure case of reconstruction with fine-grained object masking (Obj-Mask-f). (5)-(6): Remedy by using coarse object masking (Obj-Mask-c)

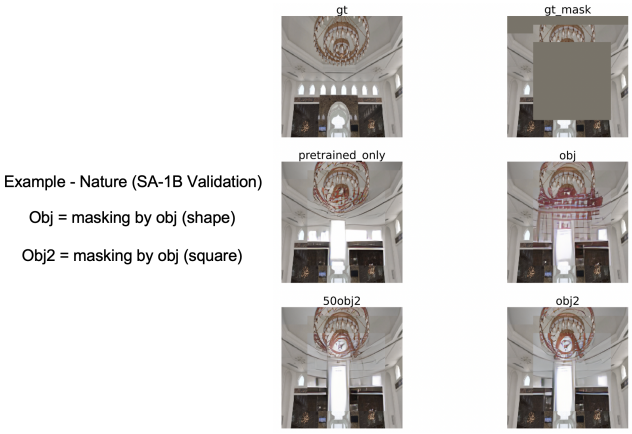

Figure 10. **Ablation Study of Masking Strategies (B)**

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.

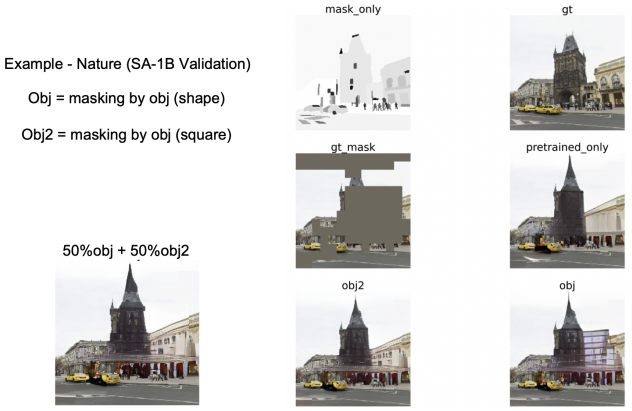

Figure 9. **Ablation Study of Masking Strategies (A)**

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 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.

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

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.