SELF-SUPERVISED REPRESENTATION LEARNING AS CONDITIONAL GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-supervised representation learning (SSL) typically learns from view consistency or masked modeling by matching, reconstructing, or predicting specified targets. However, the relationship between an observation and its target can inherently be *one-to-many*: multiple semantically valid representations or visual contents may be plausible given the same observation, yet existing SSL objectives do not explicitly model this conditional uncertainty. We introduce *CondGen*, a new formulation of SSL as conditional generation in representation space, which models the conditional distribution $p_\theta(z_i \mid z_j)$ between representations of different views using diffusion models. By explicitly modeling this distribution, *CondGen* captures the one-to-many relationship across views rather than enforcing consistency with a specified target. To further incorporate global semantic structure, we introduce *Decoupled Sinkhorn Clustering*, which learns balanced semantic prototypes independently of encoder optimization and uses them to refine target representations for conditional generation. Extensive experiments show that *CondGen* learns robust, semantically structured, and spatially coherent representations across diverse tasks, including shape/structural robustness, image generation, video object segmentation, and 3D correspondence estimation. In particular, *CondGen* even significantly surpasses DINOv3 on the unsupervised object discovery task, despite its substantially larger-scale pretraining, demonstrating the potential of conditional representation modeling as a new paradigm for scalable self-supervised learning.

1 INTRODUCTION

Self-supervised representation learning (SSL) is a fundamental pathway toward artificial general intelligence (AGI), as it enables AI systems to learn structured, predictive abstractions of the world from unlabeled observations. Such representations are central to building world models in two complementary ways: they can serve as semantic and structural priors that guide generative models to synthesize coherent and physically plausible images and videos (Zheng et al., 2026; Yu et al., 2024; Singh et al., 2025; Wu et al., 2026), enabling counterfactual reasoning, mental simulation, and planning; or they can directly encode the underlying factors and dynamics of the world in a compact latent space (Assran et al., 2023; Destrade et al., 2025; LeCun et al., 2022; Bardes et al., 2021), providing a more efficient substrate for prediction, reasoning, and decision-making. By discovering the regularities of the world without relying on costly human annotations or exhaustive interaction, self-supervised learning offers a scalable foundation for acquiring the representations, dynamics, and knowledge required for increasingly general and capable AI systems.

SSL has evolved along two dominant paradigms: *view consistency* and *masked modeling*. As shown in Figure 1(a), view-consistency methods, including contrastive learning (Oord et al., 2018; Chen et al., 2020; He et al., 2020; Chen et al., 2021), feature matching (Grill et al., 2020; Chen & He, 2021), and self-distillation (Caron et al., 2021; Oquab et al., 2023; Siméoni et al., 2025), learn representations by enforcing consistency between different augmented views of the same image. Masked modeling methods demonstrated in Figure 1(b), including BEiT (Bao et al., 2021), MAE (He et al., 2022), iBOT (Zhou et al., 2021), and JEPA (Bao et al., 2021; He et al., 2022; Zhou et al., 2021; Assran et al., 2023), instead exploit partial observations to reconstruct missing visual content or predict latent representations, and can be naturally combined with view-consistency objectives, as exemplified by iBOT (Zhou et al., 2021).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

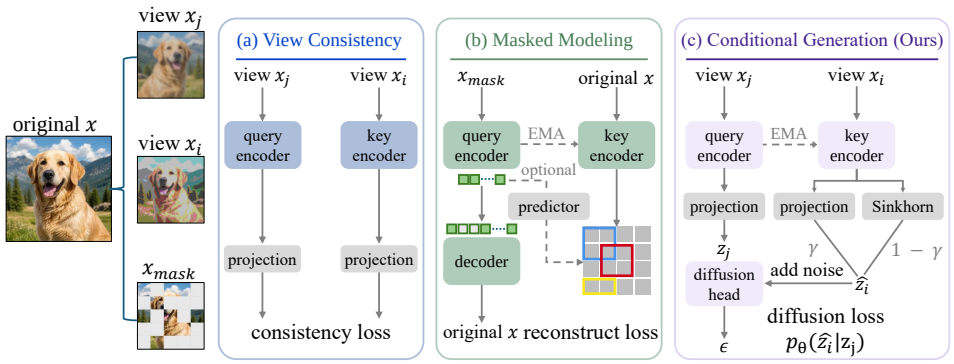


Figure 1: Comparison with existing SSL paradigms. (a) *View consistency* learns invariant representations by enforcing consistency across augmented views, using objectives such as InfoNCE (SimCLR w/o EMA, MoCo w/ EMA), feature matching (BYOL w/ additional predictor), or prototype-level KL divergence (DINO). (b) *Masked modeling* predicts missing visual information, either in pixel space (MAE) or representation space (JEPA). (c) Our *CondGen* formulates SSL as conditional generation with the diffusion mechanism, explicitly modeling the one-to-many uncertainty between representations of different views.

One-to-Many Uncertainty in SSL. Despite these paradigms differing in their supervisory signals, their objectives typically formulate self-supervision by specifying a target to be matched, reconstructed, or predicted. However, the mapping between an observation and its target can inherently be *one-to-many*: a given view may be compatible with multiple semantically valid representations induced by different augmentations, while a partial observation may admit multiple plausible visual contents or latent representations. Existing SSL objectives generally do not explicitly model this conditional uncertainty as a distribution over plausible targets.

Self-supervised Representation Learning as Conditional Generation. To explicitly model this one-to-many relationship, we formulate SSL as *a conditional generation problem (CondGen)* in Figure 1(c): given an augmented view x_i, we model the conditional distribution of its representation z_i induced by representation z_j of another view x_j, *i.e.*, $p_\theta(z_i \mid z_j)$, where $i \neq j$. We instantiate this formulation with the diffusion mechanism and learn the conditional distribution across multiple pairs of augmented views. This formulation goes beyond target-based consistency by explicitly modeling the *distribution of plausible representations* conditioned on an observation, allowing SSL to capture the uncertainty inherent in cross-view representation learning.

Additionally, to capture global semantic structure beyond individual instances, we introduce a *Decoupled Sinkhorn Clustering* mechanism that learns balanced semantic prototypes from the representations learned by *CondGen*. The clustering objective is decoupled from encoder optimization by stopping its gradients at the representation, while the resulting prototypes are used to refine the target representations for conditional generation. This design provides global semantic guidance while avoiding gradient optimization conflicts between semantic clustering and conditional representation learning during training.

Experimental Results. To validate the effectiveness of *CondGen*, we pretrain ViT models on ImageNet and evaluate the learned representations from several complementary perspectives. First, linear-probing experiments demonstrate that *CondGen* learns robust representations with improved generalization under distribution shifts and substantially stronger shape and structural robustness. Second, unsupervised object discovery reveals that *CondGen* captures strong object-level semantic structure: without task-specific training, it achieves 67.3%, 72.3%, and 57.0% CorLoc on VOC07, VOC12, and COCO, respectively, surpassing even DINOv3 despite its substantially larger-scale pretraining. Third, image generation experiments show that *CondGen* representations provide effective semantic and structural priors for both latent- and pixel-space generation, consistently improving generation quality. Finally, evaluations on video instance segmentation and 3D correspondence demonstrate that *CondGen* preserves spatially structured information beyond image-level semantics, including localized and temporally consistent object representations as well as fine-grained geometric correspondence.

Overall, our contributions are summarized as follows:

Table 1: Comparison with previous methods. Unlike existing methods, *CondGen* explicitly models the one-to-many uncertainty.

Objective	Existing work	One-to-many Uncertainty
Contrastive Learning	CPC, CPCv2, SimCLR, MoCo, MoCov2, MoCov3, InfoMin	✗
Feature Matching	BYOL, SimSiam	✗
Self-distillation& Online Clustering	SwAV, DINO, iBOT, DINOv2, DINOv3	✗
Masked Modeling	BEiT, MAE, iBOT, JEPA	✗
Conditional Generation	**CondGen (Ours)**	✓

- We propose *CondGen*, a new formulation of self-supervised representation learning as *conditional generation*, which explicitly models the one-to-many relationship between representations induced by different views.
- We instantiate *CondGen* with diffusion models to learn conditional distributions over multi-view representations, and introduce *Decoupled Sinkhorn Clustering* to incorporate global semantic structure while avoiding direct optimization conflicts with conditional generation.
- We conduct extensive experiments on ImageNet and diverse downstream tasks, demonstrating that *CondGen* learns robust and semantically structured representations, achieves strong object-level understanding in unsupervised object discovery, provides effective priors for visual generation, and preserves transferable spatial structure for video segmentation and geometric correspondence.

2 Related Work

In this section, we first review the evolution of self-supervised learning regarding its objectives and then discuss their applications for world models. Table 1 shows a summary of existing methods.

Self-Supervised Learning via View Consistency. The development of self-supervised learning (SSL) has been largely driven by the evolution of its learning objectives. Early *contrastive learning* methods formulate SSL as instance discrimination with the InfoNCE objective (Oord et al., 2018), encouraging representations of augmented views of the same image to agree while separating those of different images. SimCLR (Chen et al., 2020) relies on large batches to provide sufficient negative samples, while MoCo (He et al., 2020) introduces a momentum-updated memory queue to maintain a large and consistent set of negatives. *Feature matching* methods subsequently eliminate explicit negative samples and shift the objective from contrastive discrimination to feature matching: BYOL (Grill et al., 2020) and SimSiam (Chen & He, 2021) directly align representations of different augmented views using regression-based objectives. Then SwAV (Caron et al., 2020) proposes the online clustering paradigm and DINOs (Caron et al., 2021; Siméoni et al., 2025; Oquab et al., 2023) further casts SSL as self-distillation, where a momentum teacher produces semantic prototype assignments, and the student is optimized to match them through a KL-divergence objective.

Self-Supervised Learning via Masked Modeling. Another line of work draws inspiration from masked language modeling in BERT (Devlin et al., 2019) and introduces *masked image modeling* into SSL. BEiT (Bao et al., 2021), MAE (He et al., 2022), and iBOT (Zhou et al., 2021) learn representations by predicting or reconstructing masked image tokens or pixels from their visible context. More recently, JEPA (Assran et al., 2023) moves beyond reconstructing observable content by predicting the latent representations of masked patches from their surrounding context, encouraging the encoder to capture higher-level and more predictable structure while discarding low-level, view-specific details.

Self-Supervised Foundation Models for World Modeling. The scalability of SSL has enabled powerful visual foundation models such as DINOv2 (Oquab et al., 2023) and DINOv3 (Siméoni et al., 2025), which are pretrained on large and diverse datasets and exhibit strong transferability across a broad range of vision tasks. Importantly, self-supervised representations also provide a critical foundation for visual generation, where incorporating DINOv2/DINOv3 features can improve the semantic fidelity and perceptual quality of generated images and videos (Zheng et al., 2026; Yu et al., 2024; Singh et al., 2025; Wu et al., 2026). By providing structured representations of the visual world, SSL can facilitate generative models to capture plausible visual states and their temporal dynamics, which are essential for building generative world models capable of counterfactual reasoning, mental simulation, and planning. Alternatively, SSL can directly encode the underlying

factors and regularities of the world in compact latent spaces (Assran et al., 2023; Destrade et al., 2025; LeCun et al., 2022; Bardes et al., 2021), providing an efficient substrate for prediction, reasoning, and decision-making. These complementary roles highlight self-supervised representation learning as a fundamental building block toward AGI.

3 METHOD

3.1 MOTIVATION

Existing SSL methods learn representations through view consistency or masked modeling, using objectives that match, reconstruct, or predict a specified target. However, the relationship between an observation and its target can inherently be *one-to-many*: different augmentations may induce multiple semantically valid representations, and partial observations may admit multiple plausible targets. Existing objectives do not explicitly model this conditional uncertainty as a distribution over plausible representations.

To address this limitation, we formulate SSL as *conditional generation in representation space*. Given representations z_j and z_i from two views of an image, we learn the conditional distribution:

$$p_\theta(z_i \mid z_j), \tag{1}$$

rather than enforcing consistency with a single target. We instantiate this formulation with diffusion models, enabling the model to capture the one-to-many relationship between representations across views and learn representations that are less dependent on view-specific variations.

3.2 SELF-SUPERVISED LEARNING AS CONDITIONAL GENERATION (CONDGEN)

Network Architecture. Following MoCo (He et al., 2020) and DINO (Caron et al., 2021), *CondGen* consists of three components: a backbone encoder f_θ, an MLP projection head g_θ, and a diffusion head h_θ. The *query* network is parameterized by θ, while the *key* network, parameterized by ξ, shares the same backbone and the MLP head. Particularly, the *key* network parameters are updated as an exponential moving average (EMA) of the *query* network parameters:

$$\xi = \tau\xi + (1-\tau)\theta, \;\; where\; \tau \in [0, 1). \tag{2}$$

Conditional Generative Representation Modeling. Given an input x, we sample two stochastic augmentations $x_i, x_j \sim \mathcal{T}(x)$. The *query* and *key* networks produce representations:

$$z_j = g_\theta(f_\theta(x_j)),\; z_i = g_\xi(f_\xi(x_i)). \tag{3}$$

Then, we aim to model the conditional distribution $p_\theta(z_i|z_j)$, which captures the relationship between representations of different views.

To model $p_\theta(z_i \mid z_j)$, we adopt a conditional diffusion framework. We define a forward noising process that gradually perturbs the target representation z_i:

$$q(z_i^m \mid z_i^{m-1}) = \mathcal{N}(z_i^m; \sqrt{1-\beta_m} z_i^{m-1}, \beta_m \mathbf{I}), \tag{4}$$

where $\{\beta_m\}_{m=1}^T$ is a variance schedule, $T = 1000$ is the maximum step, $z_i^1 = z_i$.

This yields the closed-form:

$$q(z_i^m \mid z_i^1) = \mathcal{N}(z_i^m; \sqrt{\bar{\alpha}_m} z_i^1, (1-\bar{\alpha}_m)\mathbf{I}), \;\; where\; \bar{\alpha}_m = \Pi_{m=1}^T (1-\beta_m). \tag{5}$$

The diffusion head h_θ is trained to predict the noise ϵ added to z_i, with the condition of z_j:

$$\mathcal{L}_{\text{CondGen}} = \mathbb{E}_{z_i, z_j, m, \epsilon} ||\epsilon - h_\theta(z_i^m, z_j, m)||^2, \;\; \epsilon \in \mathcal{N}(\mathbf{0}, \mathbf{1}). \tag{6}$$

Decoupled Sinkhorn Clustering. While Eq. 6 learns intra-instance semantic representation by maximizing the conditional likelihood $p_\theta(z_i \mid z_j)$, it does not explicitly enforce inter-instance semantic structure. To address this problem, we introduce a *Decoupled Sinkhorn Clustering* objective

with a learnable clustering head $g_{\text{cluster}}(\cdot)$ and a set of learnable prototypes $\texttt{proto}$. Cluster assignments are computed as:

$$q(x) = \sigma\big(g_{\text{clust}}(f_\xi(x)) \cdot \texttt{proto}/\tau\big), \tag{7}$$

where $\sigma(\cdot)$ denotes the softmax, τ is a temperature hyperparameter. Then, a Sinkhorn balancing step produces pseudo-labels, leading to the clustering loss:

$$\mathcal{L}_{\text{cluster}} = -\frac{1}{2}\big(q^{\text{sinkhorn}}(x_i)\log q(x_j) + q^{\text{sinkhorn}}(x_j)\log q(x_i)\big). \tag{8}$$

Note that *the clustering head $g_{cluster}$ is on top of the EMA key network and we do not backpropagate gradients from $\mathcal{L}_{cluster}$ into the encoder backbone*. We refer to this design as "*Decoupled*" because the clustering objective does not directly optimize the encoder. This separation is important in practice: directly propagating clustering gradients to the encoder introduces optimization interference with $\mathcal{L}_{\text{CondGen}}$, resulting in unstable convergence and degraded representation quality.

Prototype-Guided Conditional Generation. The learned prototypes provide a global semantic partition of the representation space and are subsequently used to refine the target representation for conditional generation. Specifically, we construct a semantically-guided target representation as:

$$\hat{z}_i \quad = \quad \gamma z_i + (1-\gamma)\Big[q(x_i)\cdot \texttt{proto}^\top\Big], \tag{9}$$

where $\gamma \in [0,1]$ controls the balance between the original representation and its semantic prototype reconstruction, and $\texttt{proto}^\top$ is the transpose of $\texttt{proto}$. *The refined representation $\hat{z}_t$ is detached from gradient computation and serves as the target representation in the diffusion objective*:

$$\hat{\mathcal{L}}_{\text{CondGen}} = \mathbb{E}_{z_i,z_j,m,\epsilon}||\epsilon - h_\theta\big(\hat{z}_i^m, z_j, m\big)||^2, \quad \epsilon \sim \mathcal{N}(\mathbf{0},\mathbf{1}), \tag{10}$$

thereby encouraging the diffusion model to generate representations that are not only conditionally consistent across views but also aligned with globally coherent semantic prototypes.

3.3 Analysis

We show that directly optimizing conditional generation in equation 6 and clustering in equation 8 objectives through a shared encoder can lead to conflicting gradients, motivating our decoupled optimization strategy. Then, we discuss how the decoupled design enables the two modules to reinforce each other.

Assumption 1 (Semantic Granularity). We assume the representation admits an orthogonal decomposition $z = s_{global} + s_{instance}$, where s_{global} denotes prototype-level semantics shared across semantically similar instances, and $s_{instance}$ denotes instance-level semantics that distinguish different instances within the same semantic category.

Proposition 1 (Gradient Conflict). Suppose conditional generation preserves predictive instance-level semantics to maximize the conditional likelihood $p_\theta(z_i \mid z_j)$, whereas Sinkhorn clustering reduces intra-prototype variation by suppressing instance-level semantic differences. Then the gradients induced by the two objectives are negatively aligned on the instance-level semantic subspace, leading to optimization conflict on the shared encoder.

Proof is provided in Appendix A.1.

Interpretation. Proposition 1 motivates our optimization strategy: instead of jointly updating the encoder with both objectives, we detach the gradients of the Sinkhorn loss from the shared encoder and use the learned semantic prototypes only to refine the diffusion target. This eliminates optimization interference while allowing conditional generation and semantic clustering to complement each other. Moreover, we also empirically confirm the necessity of the decoupled optimization shown in Figure 2: our method with *Decoupled Sinkhorn Clustering* converges faster and achieves much higher performance, while jointly optimizing with Sinkhorn clustering even crashes during training.

Mutual Reinforcement under Decoupled Optimization. Proposition 1 motivates detaching the Sinkhorn gradients from the shared encoder, thereby eliminating direct optimization interference while preserving the interaction between semantic clustering and conditional generation. We hypothesize that the two components mutually reinforce each other: prototype refinement provides semantically informed targets that improve conditional representation learning, while the resulting improved representations facilitate more effective semantic prototype estimation.

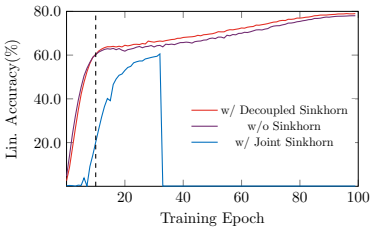


Figure 2: Ablation on the gradient conflicts between conditional generation and Sinkhorn clustering.

Table 2: Ablation on mutual reinforcement between conditional representation learning and the *Decoupled Sinkhorn Clustering*.

Settings	Lin.	Clustering loss (↓)
(I). w/ noise prototype	76.5	-
(II). w/ 50th epoch prototype	77.4	-
(III). w/o prototype	78.0	4.14
(IV). Full	79.0	3.87

We verify these two directions separately. First, removing prototype refinement from the conditional generation target degrades the quality of the learned representations, as indicated by the results of (III) and (IV) in Table 2, demonstrating the benefit of prototype-guided target construction. Second, with the Sinkhorn clustering procedure kept unchanged, representations learned with prototype refinement also reduces the clustering loss from 4.14 to 3.87, indicating that the improved representations provide a better basis for semantic prototype estimation. Together, these results provide empirical evidence of mutual reinforcement: semantic prototypes improve conditional representation learning through target refinement, while the resulting representations facilitate more effective clustering.

4 Experiments

To validate the effectiveness of our method, we pretrain the ViT models on ImageNet using self-supervised learning with *CondGen* and evaluate the resulting representations across a diverse set of downstream tasks.

Training Setup. We follow MAR (Li et al., 2024) and GenDD (Cui et al., 2025) for the implementation of the diffusion head, which serves solely as an optimization mechanism and is discarded after pretraining. Following DINO (Caron et al., 2021), we generate two global views with crop scale $(0.08, 0.4)$ and ten local views with crop scale $(0.4, 1.0)$. We use a total batch size of 512 across eight NVIDIA RTX 6000 Ada GPUs. The learning rate is linearly warmed up from 0 to 2×10^{-3} over 20 epochs and then cosine-decayed to 1×10^{-6}. We optimize the model with AdamW using a weight decay of 0.1. We train ViT-B for 100 epochs and ViT-S for 400 epochs.

4.1 Robust Representation Evaluation

To evaluate the quality and robustness of representations learned by *CondGen*, we adopt linear probing, training a linear classifier on frozen representations using ImageNet training data and evaluating it on diverse test sets. We assess three complementary aspects of representation quality: standard semantic recognition, robustness to distribution shifts, and shape and structural robustness.

Evaluation Benchmarks. For standard semantic recognition, we report top-1 accuracy on the ImageNet validation set. To evaluate robustness to distribution shifts, we use ImageNetV2 (Recht et al., 2019), ImageNet-R (Hendrycks et al., 2021), and ImageNet-C (Hendrycks & Dietterich, 2019), which respectively capture changes in data distribution, visual rendition, and common image corruptions. To assess shape and structural robustness, we use ImageNet-Sketch (Wang et al., 2019) and Stylized-ImageNet (Geirhos et al., 2018), and construct ImageNet-Edge and ImageNet-Silhouette by converting ImageNet validation images into edge maps and silhouettes using OpenCV. These benchmarks progressively suppress or alter appearance cues, providing a complementary assessment of whether representations preserve semantic and structural information beyond texture and color. Visual examples are shown in Figure 3.

Results Analysis. Table 3 shows that *CondGen* consistently improves robustness, particularly to changes in visual appearance and structure. For ViT-S/16, *CondGen* achieves a competitive 76.50% linear-probe accuracy and the best results on Edge, Silhouette, and ImageNet-Sketch, yielding the highest overall average of 37.66%. The gains are more pronounced for ViT-B/16: with only 100 epochs, *CondGen* reaches 79.00% linear-probe accuracy and achieves the best performance on all



Edge ImageNet Sketch ImageNet



Silhouette ImageNet Stylized ImageNet

Figure 3: Shape/structural robustness benchmarks.

Table 3: Robust representation evaluation with linear probing top-1 accuracy (%).

Method	Epochs	Lin.	Distribution Shifts Robustness			Shape/Structural Robustness				Avg
			ImageNetV2	ImageNet-R	ImageNet-C	Edge	Styleized	Silhouette	ImageNet-S	
ViT-S/16										
MoCov3	**300**	73.20	60.87	31.23	43.93	10.20	20.45	2.07	20.25	32.78
iBOT	800	**77.90**	**66.53**	**36.74**	**49.01**	13.91	**27.67**	2.62	25.10	37.44
DINO	800	77.00	65.15	34.33	46.25	11.06	24.03	2.22	21.97	35.25
CondGen	400	76.50	65.25	36.07	47.34	**17.14**	27.30	**4.77**	**26.92**	**37.66**
ViT-B/16										
MoCov3	300	76.70	65.08	35.16	49.07	12.13	23.81	3.16	23.64	36.09
MAE	1600	68.00	54.41	25.23	29.41	5.58	9.94	1.63	15.52	26.22
iBOT	400	**79.50**	67.41	39.43	52.56	16.46	31.26	2.90	27.43	39.62
I-JEPA	600	72.90	-	-	-	-	-	-	-	-
DINO	**300**	78.20	66.56	38.04	51.22	14.45	28.00	2.17	25.47	38.01
CondGen	**100**	79.00	**68.66**	**40.43**	**55.45**	**24.58**	**33.76**	**5.92**	**31.77**	**42.45**

evaluated distribution-shift and shape/structural benchmarks, improving over DINO by **10.13%**, **5.76%**, **3.75%**, and **6.30%** on Edge, Stylized-ImageNet, Silhouette, and ImageNet-Sketch, respectively. The particularly large gains on Edge and Silhouette, where appearance cues are heavily suppressed, suggest that *CondGen* better preserves structural and semantic information while remaining robust to appearance variations.

4.2 Image Generation

Recent studies (Zheng et al., 2026; Yu et al., 2024; Singh et al., 2025; Wu et al., 2026) have shown that self-supervised representations provide semantically and spatially structured latent spaces that benefit diffusion-based image and video generation. Motivated by these findings, we investigate whether the representations learned by *CondGen* can improve visual generation. Specifically, we evaluate *CondGen* representations in both latent- and pixel-space diffusion frameworks and compare their generation performance with existing self-supervised representations pretrained on ImageNet.

Experimental Setup. For latent-space image generation, we follow RAEv2 (Singh et al., 2026) and use its open-source implementation. Specifically, each self-supervised model is first used as a frozen image encoder to construct the latent space, followed by training a decoder to reconstruct images from these representations. The diffusion model (DiT^{DH}-XL) is then trained on the resulting latent representations, with the generation architecture and training protocol kept fixed across different pretrained encoders. For pixel-space image generation, we adopt JiT (Li & He, 2025) as the baseline and incorporate iREPA (Singh et al., 2025) to leverage pretrained self-supervised representations during diffusion training. Specifically, iREPA aligns the intermediate representations of the diffusion model from noisy inputs with the self-supervised representations extracted from the corresponding clean images, thereby injecting representation-level supervision without modifying the pixel-space generation architecture. The ViT-B model is used as the backbone. For quantitative evaluation, we report rFID and PSNR for the RAEv2 stage-1 reconstruction, and gFID, IS, and FD$_r^6$ for the final image generation stage of both RAEv2 and JiT.

Table 4: Image generation on ImageNet 256 × 256 with latent and pixel diffusion models.

Method	SSL Model	Pretrained Data	Epoch	Stage-1		Stage-2		
				rFID(↓)	PSNR(↑)	gFID(↓)	IS(↑)	FD_r^6(↓)
SiT-XL/2	None	None	800	-	-	2.12	-	8.44
SiT-XL/2-REPA	DINOv2-B	LVD-142M	800	-	-	1.42	305.70	5.45
RAEv2	DINOv3-L	LVD-1689M	80	0.29	22.57	1.06	255.30	2.17
RAEv2	DINOv1-B	INet(1.2M)	50	0.22	**28.18**	1.46	229.53	5.34
RAEv2	iBOT-B	INet(1.2M)	50	0.22	28.02	1.58	217.70	5.42
RAEv2	CondGen-B	INet(1.2M)	50	**0.21**	27.93	1.32	227.97	5.21
RAEv2	CondGen-B	INet(1.2M)	80	**0.21**	27.93	**1.22**	**238.64**	**4.79**
JiT	None	INet(1.2M)	80	-	-	6.77	218.76	23.10
JiT	DINOv1-B	INet(1.2M)	80	-	-	5.53	228.81	21.88
JiT	CondGen-B	INet(1.2M)	80	-	-	**5.38**	**233.98**	**21.22**

Table 5: DAVIS 2017 video object segmentation. We compare with DINO, iBOT, and MAE in the same setting with an image resolution of 480p.

Method	Data	Arch.	$(\mathcal{J}\&\mathcal{F})_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
Supervised					
ImageNet	INet	ViT-S/8	66.0	63.9	68.1
STM	I/D/Y	RN50	81.8	79.2	84.3
Self-supervised					
DINO	INet	ViT-B/16	62.4	60.8	64.0
iBOT	INet	ViT-B/16	62.7	**61.8**	63.7
MAE	INet	ViT-B/16	51.0	49.4	52.6
CondGen	INet	ViT-B/16	**63.1**	61.6	**64.7**

Table 6: Unsupervised object discovery.

Method	ViT	VOC07	VOC12	COCO
Self-supervised on LVD-1689M				
DINOv3	7B/16	66.1	69.5	55.1
Weakly-supervised				
SigLIPv2	g/16	20.5	24.7	18.6
Self-supervised on ImageNet				
DINO	B/16	60.4	64.7	50.5
iBOT	B/16	43.4	50.5	34.1
MAE	B/16	32.0	36.3	26.6
CondGen	B/16	**67.3**	**72.3**	**57.0**

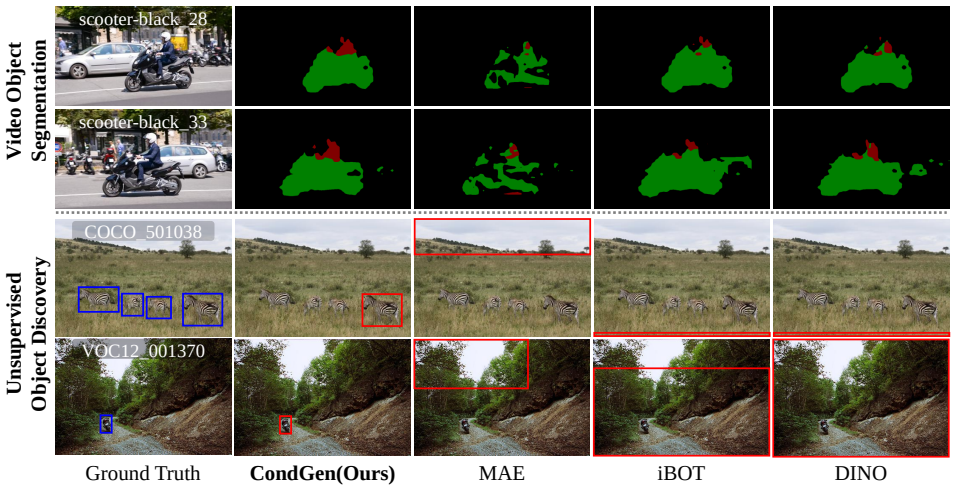


Figure 4: Qualitative comparisons for video object segmentation and unsupervised object discovery.

Results Analysis. Table 4 shows that *CondGen* provides effective representations for both latent- and pixel-space generation. With RAEv2, *CondGen-B* achieves comparable stage-1 reconstruction quality to DINOv1-B and iBOT-B, while obtaining the best generation performance among ImageNet-pretrained encoders. After 50 epochs, it reaches a gFID of 1.32 and FD_r^6 of 5.21, which further improve to 1.22 and 4.79 after 80 epochs. Notably, the gFID of 1.22 surpasses SiT-XL/2-REPA with DINOv2-B (1.42), despite the latter being trained for 800 epochs. While RAEv2 with DINOv3-L achieves a lower gFID of 1.06, it benefits from substantially larger-scale pretraining on LVD-1689M, leaving the scaling of *CondGen* to larger models and datasets for future work.

For pixel-space generation with JiT, *CondGen-B* consistently outperforms DINOv1-B across all metrics, improving gFID from 5.53 to 5.38, IS from 228.81 to 233.98, and FD_r^6 from 21.88 to 21.22. These consistent gains across two generation paradigms demonstrate that *CondGen* representations provide effective semantic and structural priors for visual generation.

Table 7: Correspondence results on NAVI and ScanNet using frozen ViT-B/16 backbones. We aggregate features from layers 2, 5, 8, and 11 for all methods. We report recall under different 3D and 2D localization error thresholds, with higher values being better.

Method	NAVI (3D Recall, %)			NAVI (2D Recall, %)			ScanNet (2D Recall, %)		
	@0.01m	@0.02m	@0.05m	@5px	@25px	@50px	@5px	@10px	@20px
DINO	26.40	43.25	75.83	4.41	23.61	38.24	23.08	31.59	**41.94**
iBOT	26.58	**44.35**	**77.49**	4.22	23.91	**39.79**	19.33	27.69	38.32
MAE	20.43	35.94	72.57	3.56	18.42	31.33	8.39	12.42	20.29
CondGen	**27.43**	44.22	76.35	**4.63**	**24.86**	39.60	**23.50**	**31.70**	41.37

4.3 Unsupervised Object Discovery

We evaluate whether *CondGen* representations support unsupervised object discovery without task-specific training. Following the protocol of TokenCut (Wang et al., 2023), we apply its non-parametric graph-based segmentation directly to frozen patch features and report CorLoc on VOC 2007, VOC 2012 (Everingham et al., 2015), and COCO-20k (Lin et al., 2014). As shown in Table 11, *CondGen* achieves 67.3%, 72.3%, and 57.0% CorLoc on the three datasets, substantially outperforming DINO, iBOT, and MAE pretrained on ImageNet. Notably, *CondGen* also surpasses DINOv3, despite DINOv3 being pretrained on the much larger LVD-1689M dataset. These results demonstrate that *CondGen* learns representations with strong object-level structure that can be directly exploited for unsupervised object discovery. Refer to Figure 4 for qualitative comparisons.

4.4 Video Object Segmentation

We evaluate the spatial and temporal properties of the learned representations on DAVIS-2017 (Pont-Tuset et al., 2017) using nearest-neighbor matching between patch tokens of consecutive frames, following work (Jabri et al., 2020), without task-specific training. As shown in Table 5, *CondGen* achieves the best overall $(\mathcal{J}\&\mathcal{F})_m$ of 63.1 and the best contour accuracy $\mathcal{F}_m$ of 64.7 among the evaluated self-supervised methods, outperforming DINO, iBOT, and MAE. These results indicate that *CondGen* preserves spatially localized and temporally consistent information that transfers effectively to video object segmentation. Refer to Figure 4 for qualitative comparisons.

4.5 3D Correspondence Estimation

To evaluate whether *CondGen* preserves geometric information across views, we assess its multi-view correspondence capability following the protocol of Probe3D (El Banani et al., 2024). Specifically, given frozen ViT-B/16 features, we measure whether patches corresponding to the same 3D keypoint can be reliably matched across different views. We aggregate features from layers 2, 5, 8, and 11 and report recall under different 3D and 2D localization error thresholds on NAVI and ScanNet. As shown in Table 7, *CondGen* achieves consistently strong correspondence performance across both datasets and localization settings. Particularly, CondGen achieves the best performance with strict localization thresholds, where accurate correspondence requires preserving precise spatial information. These results indicate that *CondGen* representations retain fine-grained spatial correspondence across views, despite being pretrained solely on 2D image data.

5 Conclusion

We presented *CondGen*, a conditional generative paradigm for self-supervised representation learning that explicitly models the one-to-many uncertainty between representations of different views using diffusion. By combining conditional generation with *Decoupled Sinkhorn Clustering*, *CondGen* captures both instance-level and global semantic structure without introducing optimization conflicts. Extensive experiments show that *CondGen* learns robust, semantically structured, and spatially coherent representations, with consistent gains in distribution-shift and shape/structural robustness, both latent- and pixel-space image generation, video object segmentation, and 3D correspondence estimation. Surprisingly, *Condgen* model significantly surpasses DINOv3 on the unsupervised object discovery task, despite its larger scale pretraining. These results demonstrate the potential of conditional representation modeling as a new direction for scalable self-supervised learning that can support more general AI systems.

AI USE STATEMENT

We used generative AI tools to polish the writing of the manuscript. The authors reviewed the AI-assisted edits and take full responsibility for the final content of this work.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629. IEEE, 2023.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9630–9640. IEEE, 2021.

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

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15745–15753. IEEE, 2021.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9620–9629. IEEE, 2021.

Jiequan Cui, Beier Zhu, Qingshan Xu, Xiaogang Xu, Pengguang Chen, Xiaojuan Qi, Bei Yu, Hanwang Zhang, and Richang Hong. Generative distribution distillation. *arXiv preprint arXiv:2507.14503*, 2025.

Matthieu Destrade, Oumayma Bounou, Quentin Le Lidec, Jean Ponce, and Yann LeCun. Value-guided action planning with jepa world models. *arXiv preprint arXiv:2601.00844*, 2025.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3d awareness of visual foundation models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21795–21806. IEEE, 2024.

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.

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. *arXiv preprint arXiv:1811.12231*, 2018.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 9726–9735. IEEE, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15979–15988. IEEE, 2022.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8320–8329. IEEE, 2021.

Allan Jabri, Andrew Owens, and Alexei Efros. Space-time correspondence as a contrastive random walk. *Advances in neural information processing systems*, 33:19545–19560, 2020.

Yann LeCun et al. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. *arXiv preprint arXiv:2511.13720*, 2025.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems*, 37: 56424–56445, 2024.

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*, pp. 740–755. Springer, 2014.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pp. 5389–5400. PMLR, 2019.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Jaskirat Singh, Xingjian Leng, Zongze Wu, Liang Zheng, Richard Zhang, Eli Shechtman, and Saining Xie. What matters for representation alignment: Global information or spatial structure? In *The Fourteenth International Conference on Learning Representations*, 2025.

Jaskirat Singh, Boyang Zheng, Zongze Wu, Richard Zhang, Eli Shechtman, and Saining Xie. Improved baselines with representation autoencoders. *arXiv preprint arXiv:2605.18324*, 2026.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in neural information processing systems*, 32, 2019.

Yangtao Wang, Xi Shen, Yuan Yuan, Yuming Du, Maomao Li, Shell Xu Hu, James L Crowley, and Dominique Vaufreydaz. Tokencut: Segmenting objects in images and videos with self-supervised transformer and normalized cut. *IEEE transactions on pattern analysis and machine intelligence*, 45(12):15790–15801, 2023.

Ge Wu, Shen Zhang, Ruijing Shi, Shanghua Gao, Zhenyuan Chen, Lei Wang, Zhaowei Chen, Hongcheng Gao, Yao Tang, Ming-Ming Cheng, et al. Representation entanglement for generation: Training diffusion transformers is much easier than you think. *Advances in Neural Information Processing Systems*, 38:7714–7743, 2026.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *International Conference on Learning Representations*, volume 2026, pp. 35791–35820, 2026.

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.

A Additional Proofs

A.1 Proof of Proposition 1

Let

$$h_j = f_\theta(x_j), \qquad z_j = g_\theta(h_j), \tag{11}$$

where f_θ and g_θ denote the shared encoder and projection head, respectively.

By the chain rule, the gradients propagated to the encoder are

$$\frac{\partial \mathcal{L}_{\text{CondGen}}}{\partial h_j} = J_g^\top \frac{\partial \mathcal{L}_{\text{CondGen}}}{\partial z_j}, \qquad \frac{\partial \mathcal{L}_{\text{cluster}}}{\partial h_j} = J_g^\top \frac{\partial \mathcal{L}_{\text{cluster}}}{\partial z_j}, \tag{12}$$

where

$$J_g = \frac{\partial z_j}{\partial h_j}$$

is the Jacobian of the projection head.

Let $P_{instance}$ denote the orthogonal projection onto the instance-level semantic subspace associated with $s_{instance}$. By Assumption 1, conditional generation maximizes the conditional likelihood $p(z_i \mid z_j)$ and therefore preserves all predictive information, including the instance-level semantic component. Consequently,

$$P_{instance} \frac{\partial \mathcal{L}_{\text{CondGen}}}{\partial z_j} = -\alpha u, \qquad \alpha > 0, \tag{13}$$

where u is a unit vector in the instance-level semantic subspace.

In contrast, Sinkhorn clustering minimizes intra-prototype variation by suppressing instance-level semantic differences. Hence,

$$P_{instance} \frac{\partial \mathcal{L}_{\text{cluster}}}{\partial z_j} = \beta u, \qquad \beta > 0. \tag{14}$$

Therefore,

$$\begin{aligned} &\left\langle P_{instance} \frac{\partial \mathcal{L}_{\text{CondGen}}}{\partial z_j}, P_{instance} \frac{\partial \mathcal{L}_{\text{cluster}}}{\partial z_j} \right\rangle \\ =&(-\alpha u)^\top (\beta u) = -\alpha\beta \|u\|^2 < 0. \end{aligned} \tag{15}$$

Since

$$\frac{\partial \mathcal{L}}{\partial h_j} = J_g^\top \frac{\partial \mathcal{L}}{\partial z_j},$$

the negative alignment is preserved after back-propagation through the projection head when J_g is locally well-conditioned. Thus, the two objectives impose conflicting optimization directions on the shared encoder, completing the proof. □

A.2 Training Efficiency

We compare the training efficiency of *CondGen* with representative view-consistency methods under the same setting, using a batch size of 512 images distributed across 8 RTX 6000 Ada GPUs. As shown in Table 8, *CondGen* uses 37.0 GB of GPU memory and achieves 1.61 s/iteration, slightly improving over DINO in both memory usage (37.2 GB) and training speed (1.65 s/iteration). Compared with iBOT, it reduces memory consumption from 42.9 GB to 37.0 GB and training time from 1.76 to 1.61 s/iteration. This efficiency is facilitated by the lightweight design of *CondGen*: conditional generation is performed only on the global CLS token with a lightweight diffusion head, without additional patch-level supervision. In contrast, DINO uses a high-dimensional projection head for prototype-based KL-divergence supervision, while iBOT additionally supervises patch tokens. Overall, *CondGen* achieves comparable training efficiency to DINO while requiring substantially less memory and computation than iBOT.

Table 8: Training efficiency comparisons. ViT-B is used with a batch of 512 images on 8 RTX 6000 Ada GPUs.

Method	Memory	Training Speed
iBOT	42932M	1.77s/iter
DINO	37234M	1.65s/iter
CondGen	37040M	1.61s/iter

Method	SSL Model	Pretrained Data	Epoch	CFG scale	Sampler	gFID(↓)	IS(↑)	FD_r^6(↓)
JiT	None	None	80	2.9	Heun50	10.47	138.02	31.74
				4.3		7.073	195.47	25.04
				4.4		7.021	198.07	24.79
				4.5		6.935	202.22	24.43
				4.6		6.867	206.02	24.11
				4.7		6.842	208.59	23.90
				4.8		6.801	211.91	23.63
				4.9		6.781	214.65	23.41
				5.0		6.774	216.75	23.25
				5.1		**6.768**	218.76	23.10
JiT + DINOv1	DINOv1-B	ImageNet(1.2M)	80	2.1	Heun50	12.71	110.24	32.29
				2.3		10.68	125.50	30.21
				2.7		8.199	151.19	27.24
				2.9		7.366	163.72	26.04
				3.1		6.862	172.50	25.23
				4.3		5.542	222.80	22.08
				4.4		5.529	225.36	21.96
				4.5		5.533	228.81	21.88
				4.6		5.538	231.94	21.82
				4.7		5.543	234.19	21.75
JiT + CondGen	CondGen-B	ImageNet(1.2M)	80	2.1	Heun50	12.49	112.31	31.75
				2.3		10.52	127.36	29.67
				2.7		8.053	153.90	26.67
				2.9		7.220	166.43	25.48
				3.1		6.718	175.60	24.66
				3.2		6.461	181.09	24.22
				3.3		6.234	186.74	23.80
				3.5		5.898	195.96	23.10
				3.7		5.667	205.31	22.54
				3.8		5.575	210.02	22.30
				4.0		5.461	217.52	21.90
				4.3		5.378	227.77	21.43
				4.4		5.375	230.52	21.31
				4.5		5.376	233.98	21.22
				4.6		5.387	237.22	21.15
				4.7		5.403	239.58	21.06
				5.0		5.469	247.17	20.92

Table 9: Effect of CFG scale on ImageNet 256 $\times$ 256 generation with JiT-B/16. All generators are trained for 80 epochs and evaluated on 50,000 samples using Heun with 50 steps, guidance interval $[0.1, 1.0]$, seed 0, and EMA decay 0.9999. gFID values are rounded to four significant figures. Gray rows represent the optimal CFG scale for balancing the selection of all metrics.

A.3 Comprehensive results on Image Generation

For fair comparisons, we select the best classifier-free guidance scale for each setting in Table 9. Here, we also report the detailed results of FD_r^6 in Tabel 10.

A.4 More Visual Examples for Unsupervised Object Discovery

We show more visual examples for the unsupervised object discovery task in Figure 5 on VOC07, Figure 6 on VOC12, and Figure 7 on COCO.

Method	SSL Model	Epoch	Representation Fréchet Distance (FD_r) ↓						FD_r^6 ↓
			Incep.	ConvNeXt	DINOv2	MAE	SigLIP	CLIP	
RAEv2	DINOv1-B	50	0.85	1.44	4.93	3.62	10.60	10.59	5.34
RAEv2	iBOT-B	50	0.97	1.58	4.84	3.47	11.21	10.42	5.42
RAEv2	CondGen-B	50	0.79	1.42	4.53	3.59	9.88	11.02	5.21
RAEv2	CondGen-B	80	**0.74**	**1.24**	**4.09**	**3.42**	**8.89**	**10.37**	**4.79**
JiT	None	80	4.03	7.05	20.03	18.05	46.23	43.19	23.10
JiT	DINOv1-B	80	3.30	6.61	15.94	18.17	45.60	41.68	21.88
JiT	CondGen-B	80	**3.22**	**6.47**	**15.53**	**17.64**	**43.79**	**40.64**	**21.22**

Table 10: Image generation on ImageNet 256 × 256 with latent and pixel diffusion models.

Method	ViT	τ	VOC07	VOC12	COCO
Agglomerative backbones					
AM-RADIOv2.5	g/14	–	55.0	59.7	45.9
PEspatial	G/14	–	51.2	56.0	43.9
Weakly-supervised backbones					
SigLIPv2	g/16	–	20.5	24.7	18.6
PEcore	G/14	–	14.2	18.2	13.5
Self-supervised backbones					
DINO	S/16	–	61.1	66.0	48.7
DINO	B/16	–	60.1	64.4	50.5
DINOv2	g/14	–	55.6	60.4	45.4
Web-DINO	7B/14	–	26.1	29.7	20.9
DINOv3	**7B/16**	–	**66.1**	**69.5**	**55.1**
Our ViT-B/16 encoder evaluation					
		0.10	44.64	50.94	33.52
		0.15	52.07	57.86	39.69
		0.20	57.83	62.96	46.12
DINO	B/16	0.25	**60.35**	**64.68**	50.07
		0.30	60.19	64.49	**50.51**
		0.35	57.55	61.97	48.45
		0.40	50.57	55.75	44.08
		0.10	41.95	48.99	32.37
		0.15	**43.44**	**50.53**	33.83
		0.20	43.17	50.39	**34.07**
iBOT	B/16	0.25	42.01	49.06	33.89
		0.30	41.27	48.00	33.60
		0.35	40.43	46.79	33.03
		0.40	39.41	45.64	32.64
		0.10	31.51	**36.27**	25.28
		0.15	31.41	35.99	25.83
		0.20	31.55	35.50	26.22
MAE	B/16	0.25	31.15	35.50	26.54
		0.30	31.47	35.23	26.56
		0.35	**31.97**	35.11	**26.60**
		0.40	31.69	35.17	26.59
		0.10	25.16	30.40	18.85
		0.15	32.25	38.28	23.93
		0.20	38.42	44.29	27.57
		0.25	47.48	53.36	33.20
CondGen	B/16	0.30	57.41	63.27	41.57
		0.35	63.60	69.48	-50.29
		0.40	66.83	71.84	55.54
		0.45	**67.33**	**72.27**	**57.01**
		0.50	66.45	71.39	56.74

Table 11: Unsupervised object discovery results. CorLoc (%, ↑) on VOC07/VOC12 and COCO20K. "–" denotes that τ is not applicable; reported values use the corresponding default or best setting. For frozen ViT-B/16 encoders, we follow TokenCut (Wang et al., 2023) for graph partitioning and scoring at full image resolution, using final-layer patch features after the final LayerNorm and excluding the CLS token. We evaluate $\tau \in 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50$ and report each setting separately. We retain the binary graph with $\varepsilon = 10^{-5}$, include difficult and truncated VOC objects, and count a localization as correct when the predicted box has IoU ≥ 0.5 with any ground-truth box.

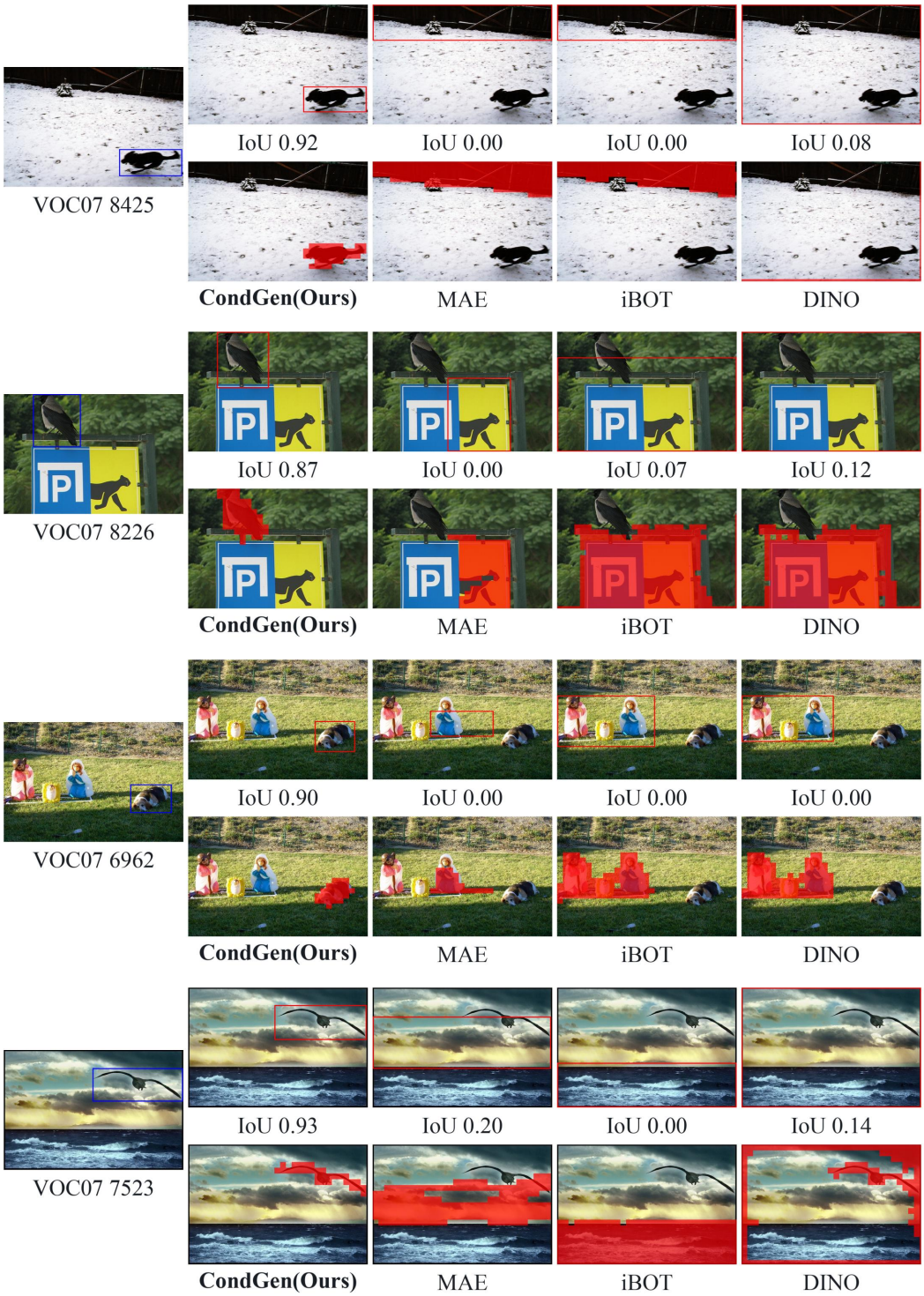


Figure 5: More visualization results of different methods on unsupervised object discovery tasks on the VOC07 dataset.

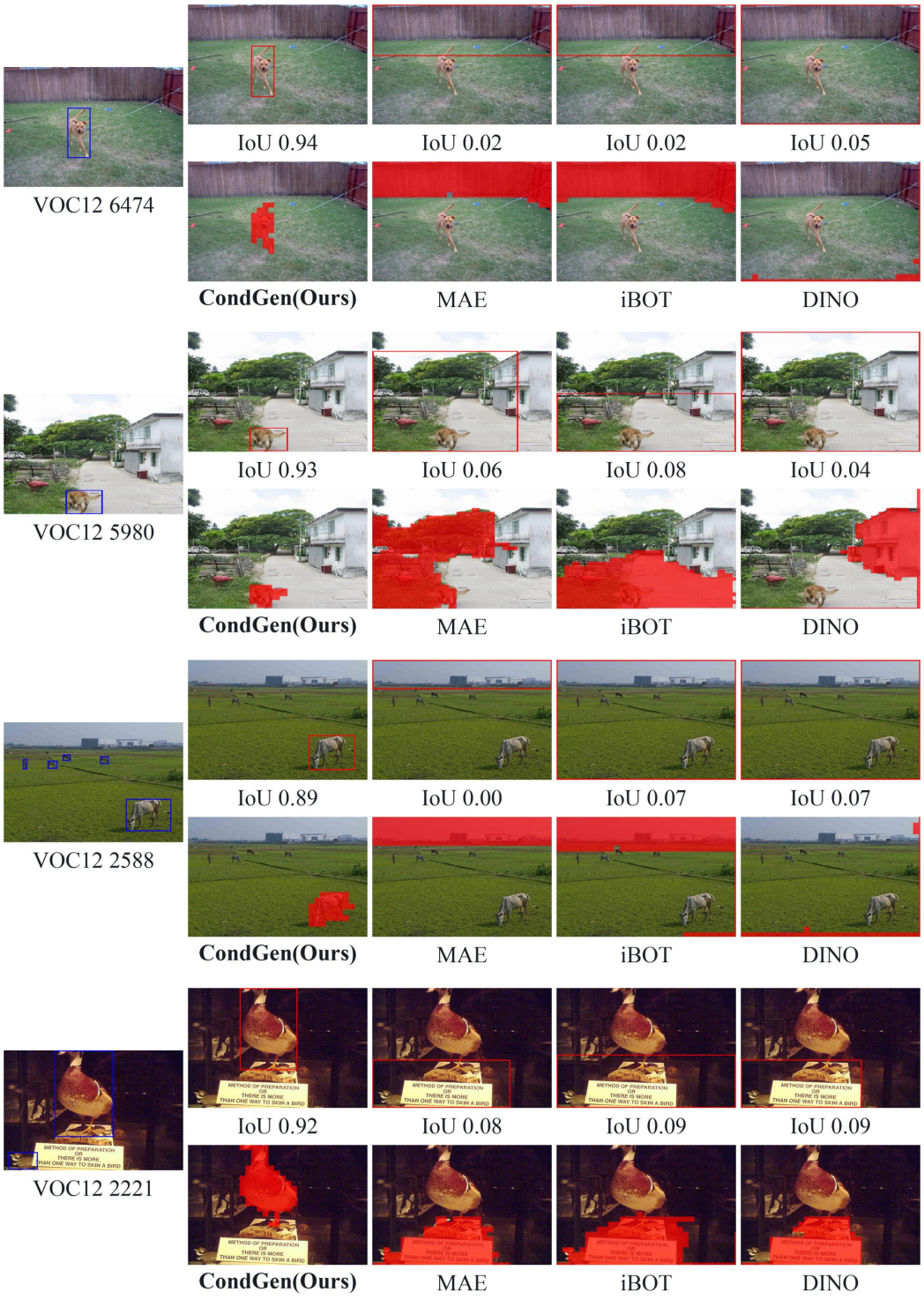


Figure 6: More visualization results of different methods on unsupervised object discovery tasks on the VOC12 (Everingham et al., 2015) dataset.

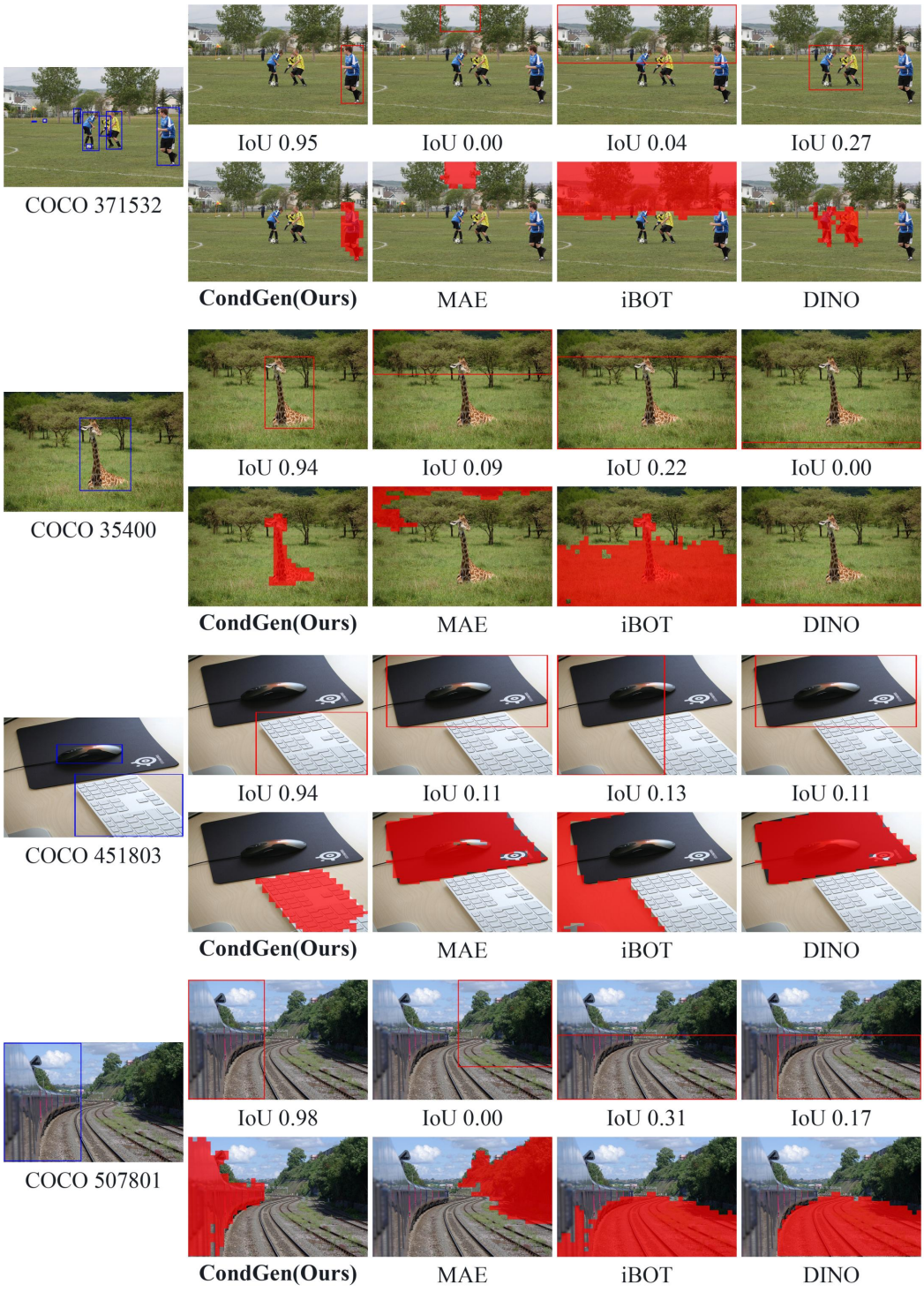


Figure 7: More visualization results of different methods on unsupervised object discovery tasks on the COCO-20k (Lin et al., 2014) dataset.