

TaskTok: Delving into Task Tokens for Task-driven Image Restoration

Hongjae Lee[✉], Sojung Kang[✉], Jaeseong Yu[✉], and Seung-Won Jung[★][✉]

Korea University, Seoul, Korea

{jimmy9704, topazsco, jsyu624, swjung83}@korea.ac.kr

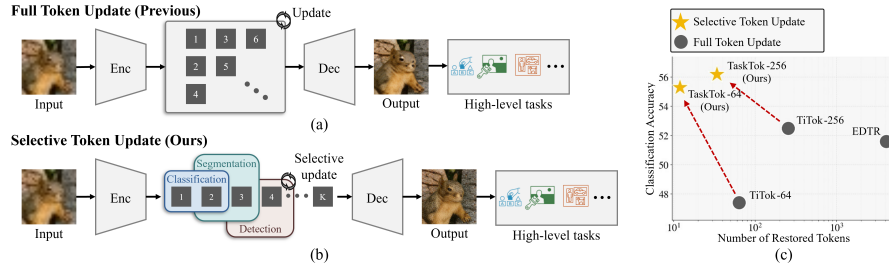


Fig. 1: Overview of **TaskTok**, a task-driven image restoration framework. (a) Existing generative prior-based methods update all latent tokens. (b) In contrast, TaskTok selectively restores only task-relevant tokens, improving performance on high-level tasks. (c) Classification accuracy vs. number of restored tokens. TaskTok achieves higher accuracy while updating substantially fewer tokens than full-token update methods.

Abstract. While traditional image restoration focuses on perceptual quality, Task-Driven Image Restoration (TDIR) aims to maximize the performance of downstream high-level vision tasks. Recent approaches leveraging generative priors have shown promise for TDIR; however, they typically suffer from computational inefficiency and potential semantic alteration by indiscriminately updating all latent tokens. In this paper, we posit that not all visual information is equally important for machine perception. Through an analysis of the latent token space, we observe that task-relevant cues are unevenly distributed across the token sequence, exhibiting index-wise specialization. This suggests that selectively refining a subset of tokens can be sufficient for task-driven objectives. Leveraging this insight, we propose TaskTok, a novel framework that selectively restores only task-relevant tokens via a learnable token switch and a lightweight token refinement module. Extensive experiments across image classification, semantic segmentation, and object detection demonstrate that TaskTok significantly enhances task performance with high computational efficiency. The source code is available at our project page <https://github.com/jimmy9704/TaskTok>.

Keywords: Task-driven image restoration · Image tokenizer · Adaptive token selection

[★] Corresponding author.

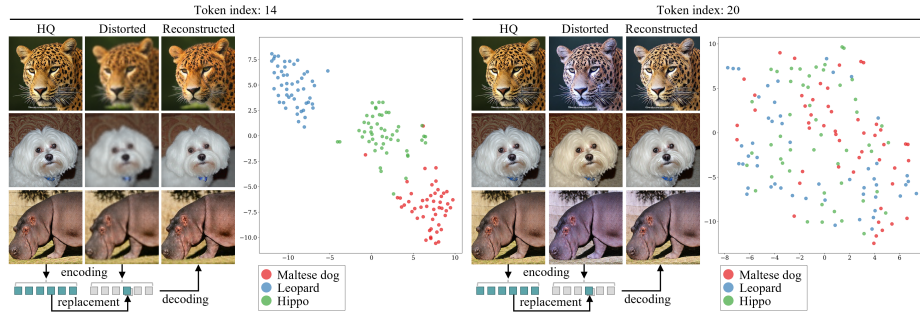


Fig. 2: Correspondence between token-wise visual effects and latent-space properties. Token attribution: Reconstructed images are obtained by replacing tokens at specific indices in a distorted input with those from the corresponding HQ image using the TiTok-64 tokenizer [44]. Token 14 primarily affects fine details (*e.g.*, skin texture), while token 20 influences more global color and tone. **Latent space visualization:** We visualize the t-SNE distribution of latent features extracted from specific token indices of HQ images across three ImageNet classes.

1 Introduction

Real-world images frequently suffer from degradation due to various factors such as transmission loss, sensor limitations, or poor shooting conditions. Under such conditions, essential high-frequency details are often lost, leading to significant performance drops in high-level vision tasks, such as image classification, object detection, and semantic segmentation [5, 16, 29, 41]. While traditional image restoration (IR) methods [22, 23, 42] prioritize perceptual quality, they do not necessarily guarantee machine recognizability. Consequently, task-driven image restoration (TDIR) has emerged to improve downstream model performance by preserving task-relevant semantic cues [3, 11, 16, 17].

Recently, diffusion-based generative priors [23, 33] have gained attention in TDIR due to their ability to robustly recover recognition-critical structures under complex degradations [3, 17]. In these pipelines, a tokenizer [18] encodes the degraded input into a two-dimensional (2D) grid of latent tokens, and a denoising model updates all tokens to produce the restored output (Fig. 1(a)). However, such full-token updates are suboptimal for two reasons: (i) they can unnecessarily alter reliable cues already preserved in the degraded input, causing semantic drift that harms task performance; and (ii) they increase computational cost by updating tokens with limited contribution to the downstream objective. Therefore, selective token updating is desirable in TDIR, as it allows computation to focus on task-relevant tokens while preserving reliable cues in the input.

Nevertheless, selectively updating tokens is structurally challenging in conventional 2D latent grids. In such representations, each token typically corresponds to a local image patch, where multiple factors—such as texture, boundaries, and illumination—are spatially entangled. As a result, isolating and se-



Fig. 3: Task-dependent sensitivity of latent tokens. Token-replaced images and task predictions are obtained by replacing a single token at a specific index in HQ images with a randomly selected token. Replacing a classification-specific token degrades classification performance, whereas replacing a segmentation-specific token degrades segmentation performance. Representative visual effects include token 3 altering material appearance and token 45 modulating background blur.

lecting only task-relevant attributes at the token level becomes difficult, and selective updates are often reduced to coarse, spatially localized modifications.

To overcome this limitation, we turn to recently proposed one-dimensional (1D) tokenizers [12, 15, 43, 44]. As shown in Fig. 1(b), a 1D tokenizer compresses an image into a fixed-length sequence of K tokens. Prior work [1] suggests that these tokens can exhibit index-wise specialization, where different indices show different sensitivities to visual attributes. Here, a token index refers to its sequence position. This index-structured representation provides a natural handle for selecting and refining only task-relevant tokens in TDIR. Building on TiTok [44], we further probe token indices and their associated visual attributes (Fig. 2). For instance, replacing token 14 primarily affects fine details such as skin texture, whereas token 20 tends to influence more global appearance statistics such as color and tone. Consistent with these observations, the t-SNE visualization in Fig. 2 indicates that token 20 is less class-structured and appears to encode relatively generic information. In contrast, token 14 exhibits more separated clusters, suggesting stronger alignment with content-specific variations and fine-grained detail cues such as texture. Therefore, for TDIR, it is reasonable to prioritize refining tokens that encode these semantically specific, fine-detail cues, as they are more likely to directly contribute to task performance, which in turn motivates our selective token refinement strategy.

Furthermore, the set of task-relevant tokens varies across downstream tasks. As illustrated in Fig. 3, perturbing tokens that primarily modulate material appearance tends to have a larger impact on classification performance, whereas perturbing tokens that modulate background blur tends to affect semantic segmentation more strongly. In other words, determining which tokens should be restored is inherently task-dependent, motivating a strategy that selectively updates task-relevant token subsets.

Based on these observations and empirical analyses, we propose TaskTok, a novel framework that selectively restores only task-relevant tokens (Fig. 1(b)). Unlike existing large-scale generative models, TaskTok is distinguished by two key components: (i) a token switch mechanism that learns to select which tokens to restore during training, and (ii) a lightweight refinement module designed

to refine the selected tokens. We evaluate TaskTok on three high-level vision tasks—image classification, object detection, and semantic segmentation—under complex degradation scenarios. As shown in Fig. 1(c), TaskTok achieves highly competitive task performance while restoring significantly fewer tokens than existing methods [17, 44]. Our main contributions are summarized as follows:

- We propose TaskTok, a TDIR framework that selectively refines task-relevant tokens via a learnable token switch and a lightweight token refinement module.
- We present a text-guided token attribution analysis that maps token-wise variations to natural-language attributes for interpreting task-selected tokens.
- We demonstrate that TaskTok significantly enhances downstream task performance with high computational efficiency.

2 Related Work

Task-driven image restoration. TDIR aims to recover recognition-critical cues under degradations, rather than optimizing only pixel fidelity or perceptual quality [11, 26, 36, 41]. Since attaching a generic restorer as a simple pre-processing step often yields limited downstream gains [5, 19, 29], recent approaches couple restoration with task networks, training them with task losses or feature-level guidance [14, 20, 24, 35]. For instance, SR4IR [16] leverages a perceptual loss within the task network’s feature space to recover task-relevant details and stabilizes training with tailored strategies. More recently, EDTR [17] introduces diffusion priors [33] into TDIR for complex degradations. Despite this progress, existing methods still update entire latent representations (tokens or grids). This is computationally expensive and can unnecessarily overwrite reliable cues already preserved in the degraded input. In contrast, our approach explicitly estimates task-dependent token importance and selectively refines only a subset of task-relevant tokens, rather than uniformly updating all latent representations.

1D tokenization. Many image generation methods [9, 23, 33, 38, 42] rely on tokenization to compress images into a low-dimensional latent space for efficient processing. Conventional image tokenizers [18, 31, 39] rely on 2D latent grids that preserve spatial layouts. Since this naturally ties each token to a specific local region, it becomes difficult to disentangle and selectively manipulate task-relevant attributes. In contrast, TiTok [44] introduces a 1D tokenizer that removes the fixed correspondence between tokens and 2D patches, learning a compact latent sequence without an explicit spatial arrangement and significantly reducing the number of tokens for efficient processing. Prior work [1] further observes that TiTok-style 1D token sequences exhibit position-dependent semantics, where different token indices tend to specialize in different visual attributes. Building on this line of research, 1D tokenization has been extended to text-conditioned image generation [12, 15] and to compact video tokenization [21, 37, 43] for efficient synthesis. Motivated by the index-wise specialization property of 1D tokens, we operate in a compact 1D token space [44] and selectively refine only a subset of task-relevant tokens for TDIR.

3 Proposed Method

3.1 Preliminary

TiTok [44] is a transformer-based model that transforms an image into a compact 1D latent representation, distinct from traditional 2D grid-based tokenizers. This approach removes spatial redundancies inherent in images, enabling more efficient compression. At its core, TiTok compresses an image into a fixed number K of 1D latent tokens, where the i -th token is denoted as $\mathbf{z}_i \in \mathbb{R}^D$ and $\mathbf{Z} = [\mathbf{z}_1; \dots; \mathbf{z}_K] \in \mathbb{R}^{K \times D}$. In the tokenization phase, the input image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ is processed by a Vision Transformer (ViT) [7] encoder to produce the compact latent embeddings $\mathbf{Z}$. Unlike standard methods that preserve the 2D spatial grid, the encoder aggregates image information directly into these 1D tokens.

In the de-tokenization phase, this latent representation $\mathbf{Z}$ is quantized using a codebook before being fed into the ViT decoder. The decoder then predicts the entire pixel space based on the semantic information compressed within these tokens to reconstruct the image $\hat{\mathbf{I}}$. The overall process of encoding the image into latent tokens and subsequently reconstructing it can be formally expressed as:

$$\mathbf{Z} = \text{Enc}(\mathbf{I}), \quad (1)$$

$$\hat{\mathbf{I}} = \text{Dec}(\text{Quant}(\mathbf{Z})). \quad (2)$$

Here, Enc and Dec denote the ViT encoder and decoder, respectively, and Quant represents the vector quantization operation that maps $\mathbf{Z}$ to the nearest entries in a learnable codebook.

3.2 Motivation: A Subset of Tokens Dominates Task Performance

To identify task-relevant tokens in TDIR, we conduct a token-level contribution analysis. We measure ImageNet [6] Top-1 accuracy using TiTok-64 [44], which encodes each image into a sequence of 64 tokens. For our analysis, we sample 300 images from the ImageNet training set and extract two sets of token sequences for each image: a high-quality (HQ) sequence, $\mathbf{Z}^{HQ}$, from the clean input, and a low-quality (LQ) sequence, $\mathbf{Z}^{LQ}$, from its degraded counterpart.

Greedy token-importance ordering. To determine the relative importance of each token without exhaustive search, we employ a greedy search strategy. Starting with the all-HQ sequence, we iteratively replace each remaining HQ token with its LQ counterpart. At each step, we identify the token index whose HQ→LQ replacement yields the highest resulting Top-1 accuracy—indicating the least task relevance—and permanently apply the replacement. We repeat this process until all HQ tokens are replaced by LQ tokens. Since the greedy search identifies the least task-relevant tokens first, we reverse the resulting order

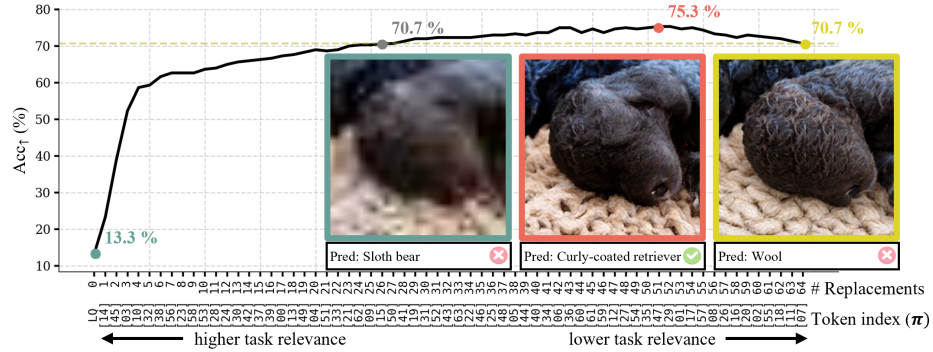


Fig. 4: Impact of token contribution to classification accuracy. We measure Top-1 accuracy on the ImageNet classification task by progressively replacing LQ tokens with HQ counterparts individually. Here, tokens are encoded via TiTok-64 [44], where LQ images are degraded by downsampling, blur, noise, and JPEG artifacts. The token replacement order follows the greedy-derived ranking π (Sec. 3.2). The curve demonstrates that task performance rapidly saturates at an early stage of token replacement.

to obtain the final ranking π (from most to least task-relevant). This procedure is detailed in the supplementary material.

Task-relevant tokens are key to performance recovery. We progressively restore tokens from LQ to HQ following the importance ranking π and track the Top-1 accuracy (Fig. 4). The resulting curve reveals that accuracy rises sharply with the restoration of just a small subset of task-relevant tokens, after which improvements saturate. This demonstrates that classification performance is primarily driven by a few dominant tokens, while the majority contribute only marginally. Interestingly, we observe that accuracy peaks before all tokens are restored. This implies that representations optimized for faithful image reconstruction are not always optimal for downstream tasks, as certain high-frequency details may act as spurious cues. These findings strongly motivate a selective restoration strategy that prioritizes computation for task-relevant tokens over uniform refinement. We observe a similar trend for other downstream tasks, such as object detection and semantic segmentation; results are provided in the supplementary material.

3.3 TaskTok Architecture

Unlike conventional restoration methods [17, 23, 42] that uniformly process all latent tokens, our approach selectively refines only task-relevant tokens within a 1D token space. We leverage TiTok [44] to encode images into a fixed-length 1D latent token sequence. An overview of the proposed TaskTok framework is illustrated in Fig. 5. Since TiTok is trained on clean images, directly tokenizing inputs with complex degradations (*e.g.*, JPEG artifacts, severe blur, noise) yields distorted latent representations. To address this, we apply a lightweight

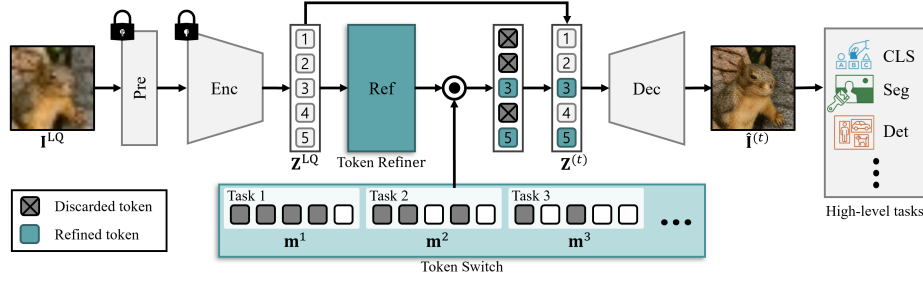


Fig. 5: Overview of TaskTok, consisting of two key modules: a lightweight token refiner module that restores degraded representations in the 1D token space and a token switch module that determines task-specific binary masks for selective token updating. See Sec. 3.3 for details on selective token refinement and architectural configurations.

SwinIR [22]-based pre-restoration module $Pre(\cdot)$ before tokenization and keep it frozen.

Given a degraded input I^{LQ} , the LQ token sequence is obtained as:

$$Z^{LQ} = Enc(Pre(I^{LQ})). \quad (3)$$

Here, the TiTok encoder $Enc(\cdot)$ remains frozen throughout training to preserve stable, index-wise visual attributes of the pre-trained token representations. The token refiner $Ref(\cdot)$ outputs refined token representations from Z^{LQ} , while the token switch $Sw(\cdot)$ determines which token is refined or preserved for every token. **Token refiner.** The token refiner is a lightweight module designed to restore degraded representations in the 1D token space, taking the LQ token sequence Z^{LQ} as input. Specifically, the token refiner is implemented with a small number of transformer layers (*e.g.*, 6), leveraging global token-to-token context to restore token-wise semantic attributions. The refined token sequence $\tilde{Z}$ is given as $\tilde{Z} = Ref(Z^{LQ})$.

Token switch. While the greedy search (Sec. 3.2) provides an importance ordering π , relying solely on a fixed heuristic is potentially suboptimal. We therefore leverage the greedy ranking to initialize the token switch for task-relevant token selection.

The token switch optimizes a task-specific learnable probability vector $\mathbf{p}^{(t)} \in [0, 1]^K$ for each task $t \in \{1, \dots, T\}$. To obtain the binary decision mask $\mathbf{m}^{(t)} \in \{0, 1\}^K$, we apply a binarization operation on $\mathbf{p}^{(t)}$ (*e.g.*, thresholding at 0.5). Since this operation is non-differentiable, we employ the straight-through estimator during training. Given the input LQ and refined token sequences, Z^{LQ} and $\tilde{Z}$, the token switch for task t , denoted as $Sw^{(t)}$, determines the final token sequence $Z^{(t)}$ as follows:

$$Z^{(t)} = Sw^{(t)}(Z^{LQ}, \tilde{Z}) = \mathbf{m}^{(t)} \odot \tilde{Z} + (1 - \mathbf{m}^{(t)}) \odot Z^{LQ}. \quad (4)$$

To facilitate stable training, we initialize the switch probabilities using the greedy order $\pi^{(t)}$ for each downstream task t . Specifically, we linearly map the

rank of each token to a probability $p_k^{(t)}$, assigning 1 to the most important token and 0 to the least:

$$p_k^{(t)} = \frac{K - r_k^{(t)}}{K - 1}, \quad r_k^{(t)} \in \{1, \dots, K\}, \quad (5)$$

where $r_k^{(t)}$ denotes the rank of the k -th token in $\boldsymbol{\pi}^{(t)}$.

While hard quantization favors perceptual image quality, it introduces quantization errors that lead to semantic shift. Thus, we bypass codebook mapping and directly feed the selectively refined continuous features into the decoder. Finally, the restored image for task t , $\hat{\mathbf{I}}^{(t)}$, is obtained as:

$$\hat{\mathbf{I}}^{(t)} = \text{Dec}(\mathbf{Z}^{(t)}). \quad (6)$$

3.4 Training Framework

We adopt the two-stage training protocol of EDTR [17], which decouples the optimization of the restoration network and the task network for stable training. We train TaskTok jointly across all target tasks.

Stage 1: training TaskTok. We optimize TaskTok while freezing the downstream task networks. Following EDTR, we use the high-level feature distillation loss $\mathcal{L}_{HLF}$ [17], which minimizes the feature discrepancy between the restored output and its HQ counterpart in the task network’s feature space. In addition, we introduce two losses tailored for selective token restoration. First, we apply token-level supervision only to the selected tokens using the task-specific binary mask $\mathbf{m}^{(t)}$:

$$\mathcal{L}_{tok}^{(t)} = \frac{1}{\|\mathbf{m}^{(t)}\|_1} \left\| \mathbf{m}^{(t)} \odot (\tilde{\mathbf{Z}} - \mathbf{Z}^{HQ}) \right\|_1, \quad \tilde{\mathbf{Z}} = \text{Ref}(\mathbf{Z}^{LQ}). \quad (7)$$

Second, to prevent the model from trivially selecting all tokens and to concentrate computation on task-relevant tokens, we minimize the average selection probability for each task t :

$$\mathcal{L}_{mask}^{(t)} = \frac{1}{K} \sum_{k=1}^K p_k^{(t)}. \quad (8)$$

We jointly optimize all target tasks by minimizing the expected per-task loss; in practice, we sample a task t per iteration and optimize $\mathcal{L}_{stage1}^{(t)}$:

$$\mathcal{L}_{stage1}^{(t)} = \mathcal{L}_{HLF}^{(t)} + \lambda_{tok} \mathcal{L}_{tok}^{(t)} + \lambda_{mask} \mathcal{L}_{mask}^{(t)}, \quad (9)$$

where λ_{tok} and λ_{mask} balance the loss terms.

Stage 2: training the task network. We freeze TaskTok and optimize the downstream task networks. We follow EDTR and train each task network with the standard task loss $\mathcal{L}_{task}$ (e.g., cross-entropy) together with the feature matching loss $\mathcal{L}_{FM}$ [17] to distill knowledge from the HQ-pretrained teacher network.

4 Experiments

We evaluate TaskTok under degraded inputs across three representative high-level vision tasks: image classification (CLS), semantic segmentation (Seg), and object detection (Det). Our goal is twofold: (i) to measure how well TaskTok recovers downstream task performance from LQ inputs, and (ii) to verify that selective, task-relevant token restoration reduces unnecessary token updates and yields a better performance–efficiency trade-off.

Datasets. We use standard benchmarks for each task: ImageNet [6] for classification, and PASCAL VOC2012 [10] for semantic segmentation and object detection. To assess generalization, we additionally report classification results on unseen datasets, CUB200 [40] and Oxford-IIIT Pet [28].

Degradation settings. Following prior works [17, 46], we consider two composite degradation settings:

- Mix-A: downsampling ($s = 8$) + JPEG compression ($q = 75$)
- Mix-B: downsampling ($s \in [1, 16]$) + Gaussian blur ($\sigma \in [0, 8]$) + Gaussian noise ($\sigma \in [0, 10]$) + JPEG compression ($q \in [50, 100]$)

Metrics. We use standard task metrics: Top-1 accuracy (Acc) for classification, mIoU for segmentation, and mAP for detection. For visual quality, we report PSNR and LPIPS [45]. For efficiency, we report the number of restored tokens (#Tok) and throughput, measured on an NVIDIA RTX 4090 GPU using 512×512 inputs.

Downstream architectures. We employ representative task networks for each downstream task: We use ResNet [13] for classification, DeepLabV3 [4] for semantic segmentation, and Faster-RCNN [32] for object detection. Unless stated otherwise, we additionally evaluate classification using unseen backbones (ViT [7] and ConvNeXt [27]) that are not used during TaskTok training.

4.1 Main Results

Tab. 1 and Fig. 6 present quantitative and qualitative results under two composite degradation settings (Mix-A/Mix-B) across image classification, semantic segmentation, and object detection. Here, TiTok- n and TaskTok- n denote models that encode an input image into n tokens, and visual quality is measured on the classification dataset (ImageNet [6]). As shown in Tab. 1, the proposed TaskTok consistently outperforms prior methods across all three tasks.

Classification. In terms of classification, TaskTok achieves higher accuracy while restoring substantially fewer tokens than EDTR and TiTok. For instance, under the Mix-B setting, TaskTok-256 restores only 34 tokens compared to EDTR and TiTok-256, improving Top-1 accuracy by +4.6% and +3.7%, respectively. This suggests that selectively restoring and refining task-relevant tokens, rather than uniformly updating all latent representations, aligns more directly with the downstream objective. Notably, although TaskTok yields stronger task performance, it does not always achieve the best PSNR/LPIPS, indicating that

Table 1: Performance across high-level vision tasks, visual quality, and efficiency under composite degradations. #Tok denotes the number of restored tokens; for pixel/feature-space restorers (*e.g.*, SwinIR, SR4IR) that update dense representations, #Tok is not applicable. Best is in **bold** and second best is underlined

IR Methods	CLS		Seg		Det		Visual Quality		Throughput (img/s) $\uparrow$
	Acc $\uparrow$	#Tok	mIoU $\uparrow$	#Tok	mAP $\uparrow$	#Tok	PSNR $\uparrow$	LPIPS $\downarrow$	
Oracle (Clean)	66.2	—	67.0	—	36.9	—	—	—	—
Mix-A	No restoration	45.9	—	42.9	—	18.0	—	22.17	0.6383
	SwinIR [22]	50.4	—	53.8	—	22.1	—	23.28	0.4095
	SR4IR [16]	51.4	—	54.5	—	24.1	—	22.04	0.2163
	TiTok-64 [44]	52.2	64	56.8	64	23.6	64	18.02	0.2684
	TiTok-256 [44]	56.6	256	62.1	256	26.5	256	19.67	0.2257
	EDTR [17]	55.7	4096	62.2	4096	<u>29.2</u>	4096	<u>22.20</u>	0.2099
	TaskTok-64 (Ours)	<u>58.1</u>	12	<u>63.3</u>	22	26.8	23	19.23	0.2427
	TaskTok-256 (Ours)	59.9	34	63.9	79	30.1	78	20.53	<u>0.2123</u>
Mix-B	No restoration	42.9	—	40.2	—	16.7	—	21.79	0.6354
	SwinIR [22]	46.4	—	49.9	—	20.1	—	23.46	0.4997
	SR4IR [16]	46.8	—	51.7	—	23.1	—	<u>22.45</u>	0.2482
	TiTok-64 [44]	47.4	64	55.5	64	22.3	64	18.12	0.2893
	TiTok-256 [44]	52.5	256	60.4	256	25.7	256	19.51	0.2470
	EDTR [17]	51.6	4096	60.9	4096	<u>28.1</u>	4096	22.28	0.2310
	TaskTok-64 (Ours)	<u>55.3</u>	12	<u>61.9</u>	22	25.9	23	19.12	0.2595
	TaskTok-256 (Ours)	56.2	34	62.7	79	28.6	78	20.45	<u>0.2338</u>

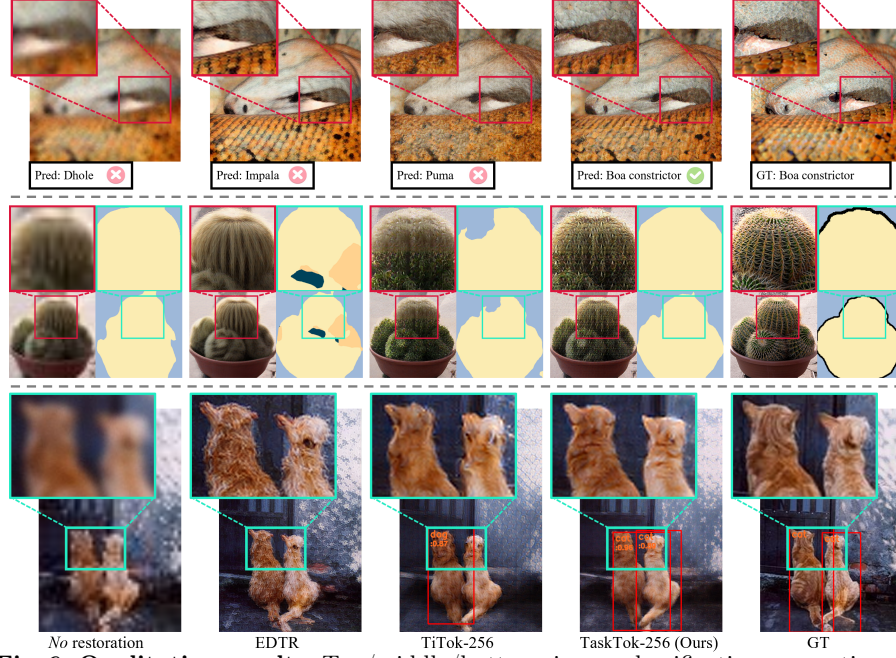


Fig. 6: Qualitative results. Top/middle/bottom: image classification, semantic segmentation, and object detection.

Table 2: Generalization to unseen datasets. We evaluate Top-1 accuracy on two datasets not used during TaskTok training.

IR Method	Unseen Datasets	
	CUB200	Oxford-IIIT Pet
<i>No restoration</i>	10.2	16.4
TiTok-64 [44]	34.1	70.2
TiTok-256 [44]	48.7	74.0
EDTR [17]	50.3	73.7
TaskTok-64	43.4	75.9
TaskTok-256	54.5	79.6

Table 3: Generalization to unseen classifier architectures. We evaluate Top-1 accuracy with classification backbones not used during TaskTok training.

IR Method	Unseen Backbones	
	ViT-B/16	ConvNeXt-B
<i>No restoration</i>	38.7	46.1
TiTok-64 [44]	49.6	50.7
TiTok-256 [44]	55.8	60.3
EDTR [17]	55.3	62.6
TaskTok-64	55.5	60.9
TaskTok-256	58.4	66.5

optimizing conventional fidelity/perceptual metrics does not necessarily translate into better recognition. Instead, TaskTok prioritizes restoring task-relevant semantic cues while preserving reliable information already present in the input. **Segmentation and detection.** A similar trend is observed for semantic segmentation and object detection. Under Mix-A, for semantic segmentation, TaskTok-256 improves mIoU by +1.7% over EDTR while restoring only 79 tokens. For object detection, TaskTok-256 improves mAP by +0.9% over EDTR with the restored token budget limited to 78.

Efficiency. In terms of efficiency, TaskTok restores a very small number of tokens (#Tok), enabling computation to focus on task-relevant representations and yielding a favorable performance–efficiency trade-off in throughput. In particular, TaskTok-256 runs $8.3\times$ faster than the one-step diffusion model, EDTR, demonstrating that selective token restoration can simultaneously improve performance and reduce computation.

Qualitative results. Visual comparisons in Fig. 6 further support these findings. Without restoration, critical high-frequency details and structural cues are lost, leading to noticeable performance degradation. While EDTR and TiTok may generate visually sharper details, such details can be task-irrelevant or even misleading. In contrast, TaskTok focuses on recovering cues that matter for the downstream tasks, resulting in more consistent and reliable predictions.

4.2 Generalizability of TaskTok

Cross-dataset generalization. Tab. 2 reports Top-1 classification accuracy on unseen datasets, CUB200 and Oxford-IIIT Pet, to evaluate dataset-level generalization. TaskTok demonstrates consistent robustness on unseen data. Notably, TaskTok-256 achieves the best performance on CUB200 (54.5%), outperforming EDTR (50.3%) by +4.2% and TiTok-256 (48.7%) by +5.8%. A similar trend is observed on Oxford-IIIT Pet, indicating that TaskTok effectively generalizes to diverse data distributions without overfitting.

Cross-architecture generalization. Tab. 3 further evaluates whether the gains persist on unseen classifier architectures that were not used during TaskTok

Table 4: Ablation study on TaskTok components. Pre: pre-restoration module; Ref: token refiner; Sw: token switch; π : importance order initialization; VQ: codebook quantization.

Variant	Ablation toggles							CLS		Seg		Det	
	Pre	Ref	$\mathcal{L}_{tok}$	Sw	$\mathcal{L}_{mask}$	π	VQ	Acc $\uparrow$	#Tok	mIoU $\uparrow$	#Tok	mAP $\uparrow$	#Tok
TiTok-64 [44]	$\times$	$\times$	$\times$	$\times$	$\times$	$\times$	$\checkmark$	47.4	64	55.5	64	22.3	64
+ Pre-restoration	$\checkmark$	$\times$	$\times$	$\times$	$\times$	$\times$	$\checkmark$	48.5	64	56.9	64	23.1	64
+ Token refiner	$\checkmark$	$\checkmark$	$\times$	$\times$	$\times$	$\times$	$\checkmark$	46.3	64	56.0	64	21.8	64
+ $\mathcal{L}_{tok}$	$\checkmark$	$\checkmark$	$\checkmark$	$\times$	$\times$	$\times$	$\checkmark$	46.7	64	56.2	64	22.2	64
+ Token switch	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\times$	$\times$	$\checkmark$	46.9	31	57.1	33	22.1	35
+ $\mathcal{L}_{mask}$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\times$	$\checkmark$	50.9	14	58.3	26	22.7	25
+ π initialization	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	53.0	10	60.4	24	24.3	22
TaskTok-64	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\times$	55.3	12	61.9	22	25.9	23

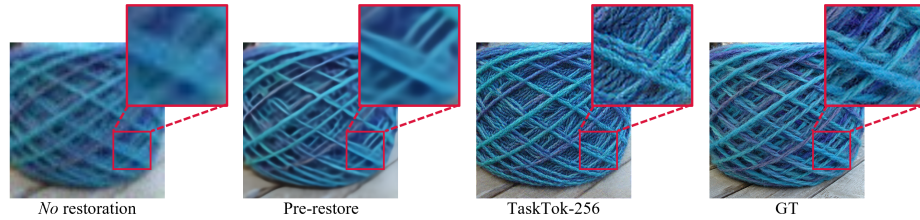


Fig. 7: Qualitative examples of the pre-restoration outputs and the TaskTok results.

training. We measure Top-1 accuracy using classification backbones pretrained on HQ images (*e.g.*, ViT-B/16 and ConvNeXt-B). TaskTok maintains stable improvements across these unseen models; for instance, TaskTok-256 achieves 66.5% Top-1 accuracy with ConvNeXt-B. These findings confirm that TaskTok learns robust, generalizable restoration representations independent of specific classifier designs. All generalization results are reported under Mix-B.

4.3 Ablation Studies

Tab. 4 summarizes the ablation results in the Mix-B setting, analyzing the contribution of each component step by step. We first apply a pre-restoration module before TiTok to suppress degradation artifacts. This mitigates the risk that artifacts act as misleading generation cues, and consequently improves performance across all three tasks. Fig. 7 presents qualitative examples of the pre-restoration outputs and the final TaskTok results. While pre-restoration mainly suppresses degradation artifacts before tokenization, TaskTok restores task-relevant cues through selective token refinement. In contrast, adding the token refiner to restore all tokens degrades performance, and incorporating $\mathcal{L}_{tok}$ yields only marginal gains. This suggests that (i) uniformly restoring every token is not always beneficial for downstream tasks, and (ii) the lightweight refiner is insufficient for fully restoring the entire token sequence. Introducing the

Table 5: Cross-task token transfer. Using tokens selected by a task (CLS/Seg/Det) for other tasks.

Token Set	CLS (Acc↑)	Seg (mIoU↑)	Det (mAP↑)
CLS-selected	55.3	59.9	24.2
Seg-selected	53.9	61.9	25.3
Det-selected	53.5	61.3	25.9

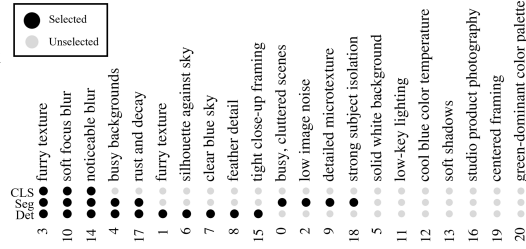


Fig. 8: Task-wise token attribute selection (TaskTok-64). First 21 latent tokens (0–20).

token switch leads to a clear improvement by enabling selective updates of task-relevant tokens. In particular, with the initialization of the token order π and the mask regularizer $\mathcal{L}_{mask}$ to discourage selecting unnecessary tokens, the model learns an optimal token subset and achieves gains of 5.6%, 4.9%, and 2.0% on CLS/Seg/Det over TiTok-64. Finally, removing quantization alleviates the semantic shift induced by quantization, yielding the best performance of the final TaskTok model.

4.4 Overhead of the Token Refiner

TaskTok employs a selective update strategy, refining only task-relevant tokens instead of uniformly processing all latent representations. Since its objective is to focus on refining the small subset selected by the token switch, the token refiner can be designed as a highly lightweight transformer. Consequently, for TaskTok-64 (total latency: 13.26 ms/img), the refiner introduces a negligible overhead of just 0.08 ms/img (0.60%) and contributes only 19.0 M parameters (4.6%) to the 413.6 M total. These results confirm that selective restoration incurs minimal cost while significantly enhancing downstream task performance.

4.5 Cross-Task Transferability

Tab. 5 compares the performance when a token subset selected for one task is directly applied to another. Using task-specific tokens yields the best performance for each respective task. For instance, evaluating CLS with tokens selected by Seg/Det drops the accuracy by 1.4% / 1.8% from 55.3%. This clearly indicates that performance relies on whether the restored tokens are truly task-critical, rather than merely the total number of restored tokens. Interestingly, the Seg $\leftrightarrow$ Det transfer exhibits much smaller performance drops (0.6% / 0.6%), implying that segmentation and detection share underlying structural cues. Overall, these results demonstrate that our model successfully identifies optimal, task-specific subsets, which is crucial for maximizing downstream performance. All results are reported under Mix-B.

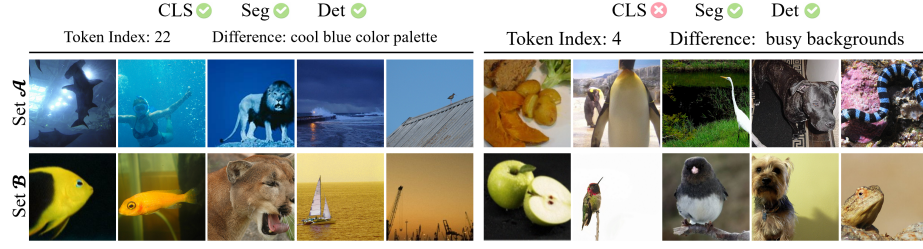


Fig. 9: Qualitative examples of text-based attribute analysis. Images from two distinct clusters (top and bottom rows) for specific token indices. Difference indicates the core visual variation between the two sets, summarized by an image-set difference captioning method [8].

Algorithm 1: Text-Guided Token Attribution Analysis

Input: Image set $\mathcal{D}$, frozen tokenizer encoder $\text{Enc}(\cdot)$, token index k , clustering algorithm $\text{Cluster}(\cdot)$, image-set difference captioner $\text{DiffCap}(\cdot, \cdot)$
Output: Proxy attribute label c_k for token index k
Initialize token set $\mathcal{Z}_k \leftarrow []$;
foreach $\mathbf{I} \in \mathcal{D}$ **do**
 Extract $\mathbf{z}_k \leftarrow \text{Enc}(\mathbf{I})[k]$;
 Append $\mathbf{z}_k$ to $\mathcal{Z}_k$;
Obtain clusters $\{\mathcal{C}_1, \dots, \mathcal{C}_M\} \leftarrow \text{Cluster}(\mathcal{Z}_k)$;
Compute the center $\boldsymbol{\mu}_m$ for each cluster $\mathcal{C}_m$;
Select $(a, b) \leftarrow \arg \max_{i \neq j} d(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j)$, $1 \leq i, j \leq M$;
Define image sets $\mathcal{A}, \mathcal{B}$ from the images assigned to $\mathcal{C}_a$ and $\mathcal{C}_b$;
Generate caption $c_k \leftarrow \text{DiffCap}(\mathcal{A}, \mathcal{B})$;
return c_k ;

4.6 Task-Specific Token Attribution Analysis

The visual semantics encoded by latent tokens are highly entangled in a high-dimensional space, making them difficult to define individually. To interpret TaskTok’s selective update mechanism from a human-understandable semantic perspective, we conduct a text-guided attribution analysis by mapping the axes of visual variation into natural language descriptions.

Text-based attribute analysis. We summarize the attribution pipeline in Algorithm 1. We use 30K clean and degraded images from the ImageNet [6] validation set. For each token index, we collect its representations across images and cluster them using HDBSCAN [2]. A visual attribute can appear as multiple coherent clusters; for example, color may split into several hue-specific clusters. We therefore select the two most distant clusters and define their corresponding image sets as $\mathcal{A}$ and $\mathcal{B}$. We then follow VisDiff [8] to describe their dominant visual difference: LLaVA-1.5 [25] first generates captions for both sets, GPT-5 [34] proposes candidate difference descriptions, and CLIP [30] ranks them. Fig. 9 provides qualitative examples demonstrating that the clustering results

for each token index capture a dominant axis of visual variation, which the generated difference captions consistently summarize. This supports using the difference captions as proxy labels for the dominant visual attribute modulated by each token index.

Task-specific token selection trends. Fig. 8 summarizes the attribute analysis for the first 21 TaskTok-64 token indices alongside their selection indicators across three downstream tasks.

Our analysis shows that universally selected tokens are primarily associated with fine-grained details and morphological sharpness (*e.g.*, furry texture, soft-focus blur, and noticeable blur), whereas consistently excluded tokens tend to reflect global appearance factors such as color (*e.g.*, cool-blue color temperature, green-dominant color palette) or lighting. This suggests that such global attributes are either already sufficiently preserved in the degraded inputs or do not decisively impact the final performance of these high-level tasks.

Meanwhile, both Seg and Det tend to select tokens related to background clutter and scene complexity, consistent with their need for object localization and boundary delineation. Notably, requiring dense pixel-level precision, Seg exhibits a heightened sensitivity to localized detail variations, such as image noise and microtexture. Overall, this analysis offers an interpretable perspective on the visual attributes prioritized by TaskTok’s selective refinement mechanism. The complete token attribution results, along with additional qualitative examples of the image-set difference analysis, are provided in the supplementary material.

5 Conclusion

This paper proposes TaskTok, a novel TDIR framework that selectively refines task-relevant representations in a 1D latent token space. We observe that task-specific tokens play a critical role in downstream recognition, and that indiscriminate restoration may distort semantic cues essential for machine perception. To address this, we propose a set of task-aware components, including a learnable token switch and a lightweight token refiner, to selectively restore only the token subsets that are most relevant to each downstream objective. Experimental results demonstrate that TaskTok significantly improves the performance of various high-level vision tasks, such as image classification, semantic segmentation, and object detection, under complex image degradation scenarios. We anticipate that our approach will provide new insights into task-aware and computationally efficient token processing for TDIR.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2025-16068196).

References

1. Beyer, L.L., Li, T., Chen, X., Karaman, S., He, K.: Highly compressed tokenizer can generate without training. In: ICML (2025)
2. Campello, R.J., Moulavi, D., Sander, J.: Density-based clustering based on hierarchical density estimates. In: *Advances in Knowledge Discovery and Data Mining*. pp. 160–172 (2013)
3. Chen, I., Chen, W.T., Liu, Y.W., Chiang, Y.C., Kuo, S.Y., Yang, M.H., et al.: UniRestore: Unified perceptual and task-oriented image restoration model using diffusion prior. In: CVPR. pp. 17969–17979 (2025)
4. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE TPAMI* **40**(4), 834–848 (2017)
5. Dai, D., Wang, Y., Chen, Y., Van Gool, L.: Is image super-resolution helpful for other vision tasks? In: WACV. pp. 1–9 (2016)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: CVPR. pp. 248–255 (2009)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2020)
8. Dunlap, L., Zhang, Y., Wang, X., Zhong, R., Darrell, T., Steinhardt, J., Gonzalez, J.E., Yeung-Levy, S.: Describing differences in image sets with natural language. In: CVPR. pp. 24199–24208 (2024)
9. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: CVPR. pp. 12873–12883 (2021)
10. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The Pascal visual object classes (VOC) challenge. *IJCV* **88**, 303–338 (2010)
11. Haris, M., Shakhnarovich, G., Ukita, N.: Task-driven super resolution: Object detection in low-resolution images. In: *International Conference on Neural Information Processing*. pp. 387–395 (2021)
12. He, J., Yu, Q., Liu, Q., Chen, L.C.: FlowTok: Flowing seamlessly across text and image tokens. In: ICCV. pp. 16629–16640 (2025)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
14. Huang, S.C., Le, T.H., Jaw, D.W.: DSNet: Joint semantic learning for object detection in inclement weather conditions. *IEEE TPAMI* **43**(8), 2623–2633 (2020)
15. Kim, D., He, J., Yu, Q., Yang, C., Shen, X., Kwak, S., Chen, L.C.: Democratizing text-to-image masked generative models with compact text-aware one-dimensional tokens. In: ICCV. pp. 18442–18452 (2025)
16. Kim, J., Oh, J., Lee, K.M.: Beyond image super-resolution for image recognition with task-driven perceptual loss. In: CVPR. pp. 2651–2661 (2024)
17. Kim, J., Oh, J., Lee, K.M.: Exploiting diffusion prior for task-driven image restoration. In: ICCV. pp. 10151–10161 (2025)
18. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: ICLR (2014)
19. Li, B., Peng, X., Wang, Z., Xu, J., Feng, D.: AOD-Net: All-in-one dehazing network. In: ICCV. pp. 4770–4778 (2017)
20. Li, C., Zhou, H., Liu, Y., Yang, C., Xie, Y., Li, Z., Zhu, L.: Detection-friendly dehazing: Object detection in real-world hazy scenes. *IEEE TPAMI* **45**(7), 8284–8295 (2023)

21. Li, Y., Tian, C., Xia, R., Liao, N., Guo, W., Yan, J., Li, H., Dai, J., Li, H., Yang, X.: Learning adaptive and temporally causal video tokenization in a 1D latent space. arXiv preprint arXiv:2505.17011 (2025)
22. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: SwinIR: Image restoration using swin transformer. In: ICCV. pp. 1833–1844 (2021)
23. Lin, X., He, J., Chen, Z., Lyu, Z., Dai, B., Yu, F., Qiao, Y., Ouyang, W., Dong, C.: DiffBIR: Toward blind image restoration with generative diffusion prior. In: ECCV. pp. 430–448 (2024)
24. Liu, D., Wen, B., Jiao, J., Liu, X., Wang, Z., Huang, T.S.: Connecting image denoising and high-level vision tasks via deep learning. IEEE TIP **29**, 3695–3706 (2020)
25. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023)
26. Liu, W., Ren, G., Yu, R., Guo, S., Zhu, J., Zhang, L.: Image-adaptive YOLO for object detection in adverse weather conditions. In: AAAI. pp. 1792–1800 (2022)
27. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: CVPR. pp. 11976–11986 (2022)
28. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.: Cats and dogs. In: CVPR. pp. 3498–3505 (2012)
29. Pei, Y., Huang, Y., Zou, Q., Lu, Y., Wang, S.: Does haze removal help CNN-based image classification? In: ECCV. pp. 682–697 (2018)
30. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML (2021)
31. Razavi, A., Van den Oord, A., Vinyals, O.: Generating diverse high-fidelity images with VQ-VAE-2. In: NeurIPS (2019)
32. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: NeurIPS (2015)
33. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
34. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: Openai GPT-5 system card. arXiv preprint arXiv:2601.03267 (2025)
35. Son, T., Kang, J., Kim, N., Cho, S., Kwak, S.: URIE: Universal image enhancement for visual recognition in the wild. In: ECCV. pp. 749–765 (2020)
36. Sun, S., Ren, W., Wang, T., Cao, X.: Rethinking image restoration for object detection. In: NeurIPS. pp. 4461–4474 (2022)
37. Tan, Z., Xue, B., Jia, J., Wang, J., Ye, W., Shi, S., Sun, M., Wu, W., Chen, Q., Jiang, P.: SweetTok: Semantic-aware spatial-temporal tokenizer for compact video discretization. In: ICCV. pp. 23541–23550 (2025)
38. Tian, K., Jiang, Y., Yuan, Z., Peng, B., Wang, L.: Visual autoregressive modeling: Scalable image generation via next-scale prediction. In: NeurIPS. pp. 84839–84865 (2024)
39. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. In: NeurIPS (2017)
40. Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., Perona, P.: Caltech-ucsd birds 200. Technical Report California Institute of Technology. CNS-TR-2010-001. (2010)
41. Wu, J., Jin, Z.: Unsupervised variational translator for bridging image restoration and high-level vision tasks. In: ECCV. pp. 214–231 (2024)
42. Wu, R., Yang, T., Sun, L., Zhang, Z., Li, S., Zhang, L.: SeeSR: Towards semantics-aware real-world image super-resolution. In: CVPR. pp. 25456–25467 (2024)

43. Yin, A., Tan, X., Shen, K., Leng, Y., Zhou, X., Li, J., Tang, S.: The best of both worlds: Integrating language models and diffusion models for video generation. In: ICCV. pp. 15604–15615 (2025)
44. Yu, Q., Weber, M., Deng, X., Shen, X., Cremers, D., Chen, L.C.: An image is worth 32 tokens for reconstruction and generation. In: NeurIPS. pp. 128940–128966 (2024)
45. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR. pp. 586–595 (2018)
46. Zhou, S., Chan, K., Li, C., Loy, C.C.: Towards robust blind face restoration with codebook lookup transformer. In: NeurIPS. pp. 30599–30611 (2022)