Multimodal Promptable Token Merging for Diffusion Models

Cheng-Yao Hong, Tyng-Luh Liu

Institute of Information Science, Academia Sinica, Taiwan
{sensible, liutyng}@iis.sinica.edu.tw

Abstract

Token compression techniques, such as token merging and pruning, are essential for alleviating the substantial computational burden caused by the proliferation of tokens within attention mechanisms. However, current methods often rely on token-to-token distances or similarity metrics to evaluate token importance, which is inadequate in the context of modern promptable designs and frameworks that are gaining prominence. To address this limitation, we introduce a novel and effective merging strategy called "**M**ultimodal **P**romptable **T**oken **M**erging" (MPTM). The proposed method leverages a multimodal, prompt-centric methodology, assessing the proximity between tokens of each input modality and the multimodal prompt to efficiently eliminate redundant tokens while preserving those rich in information. Extensive experiments demonstrate that MPTM significantly reduces computational costs without compromising essential information in generative image tasks. When integrated into diffusion-based detection architectures, MPTM outperforms existing state-of-the-art methods by 2.3% in object detection tasks. Additionally, when applied to multimodal diffusion models, MPTM maintains high-quality output while achieving a 2.9-fold increase in throughput, highlighting its versatility.

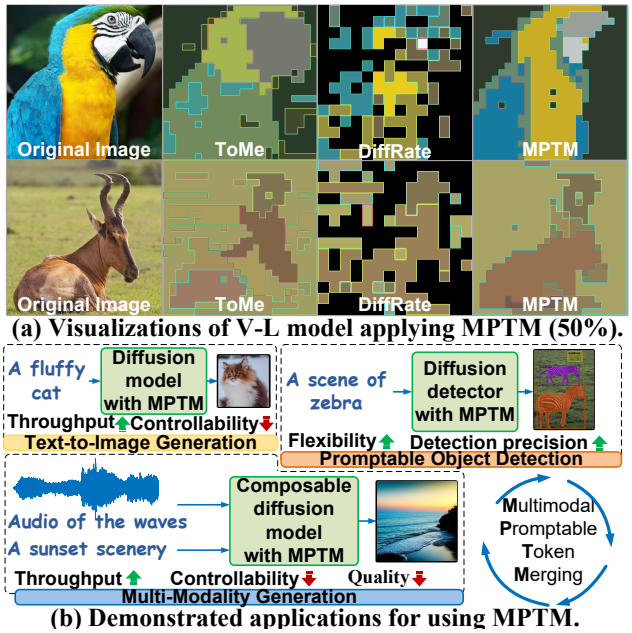

Figure 1: **Motivation perspectives for MPTM**. (a) Compared to residual tokens produced by standard token merging processes, MPTM generates tokens with higher representational quality, outperforming the baseline method, ToMe (Bolya et al. 2023). (b) MPTM demonstrates adaptability and efficiency across diverse architectures, significantly enhancing task performance, particularly in dense prediction tasks, without compromising quality.

1 Introduction

We address the challenge of applying token merging to accelerate transformer-based networks. Unlike existing techniques such as ToMe (Bolya et al. 2023), our method is specifically designed for promptable computer vision applications, introducing a novel strategy for token merging within a multimodal prompt-induced space for each input modality.

Vision Transformers (ViTs) (Dosovitskiy et al. 2021) have emerged as a dominant force, achieving state-of-the-art results in various tasks such as image classification (Liu et al. 2021), object detection (Zhu et al. 2021; Chen et al. 2023b), and semantic segmentation (Xie et al. 2021). The token-based design of ViTs allows for the seamless integration of the masked image modeling (MIM) (He et al. 2022) technique, adapted from the masked language model (MLM) approach in NLP (Devlin et al. 2019). This enables the learning of versatile representations in a self-supervised manner, demonstrating impressive performance across various tasks. Despite their remarkable capabilities, ViTs are often criticized for their substantial computational demands, which limit their practicality.

To address this issue, several model compression techniques have been explored, including weight pruning (Han, Mao, and Dally 2016; Wang et al. 2022), weight quantization (Yuan et al. 2022), knowledge distillation (Pelosin et al. 2022), and neural architecture search (Cai, Zhu, and Han 2019; Gong et al. 2022). Among these, token compression has emerged as a particularly effective strategy for ViTs, given the exponential/quadratic relationship between token

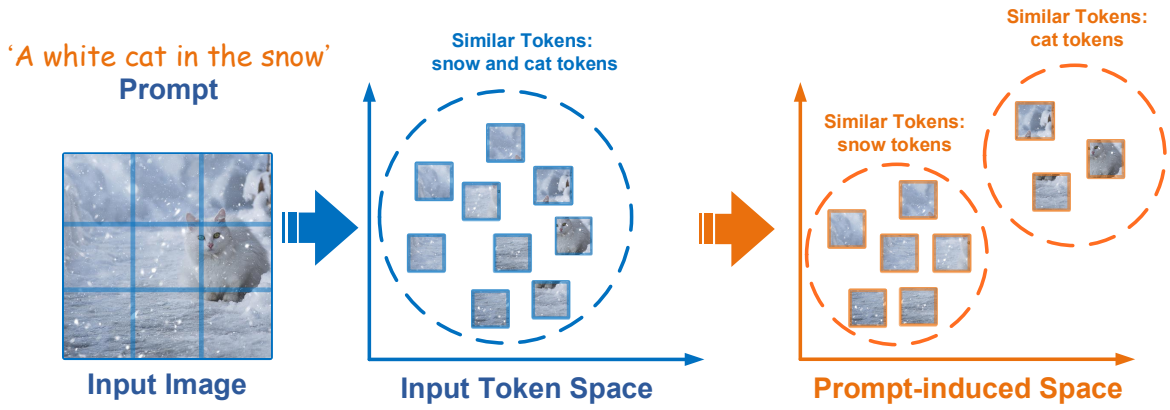

Figure 2: **Example of the potential drawback of the original token merging.** By leveraging semantic information, tokens can be correctly clustered in challenging examples.

length and computational cost. Token compression can be further divided into two categories: token merging (Bolya et al. 2023; Chen et al. 2023a) and pruning (Zeng et al. 2022; Kong et al. 2022; Chen et al. 2023a). Token merging, in particular, has been shown to accelerate both training and inference times. Notably, ToMe (Bolya et al. 2023) introduces bipartite soft matching for token merging, presenting a simple yet effective approach to token reduction. Recently, the latent diffusion model (LDM)(Rombach et al. 2022) has gained considerable attention for its remarkable applications in text-to-image(Nichol et al. 2022; Ge et al. 2023) and text-to-3D conversions (Poole et al. 2022; Lin et al. 2023). LDMs typically consist of a transformer-based U-Net, a latent encoder/decoder, and a prompt encoder. Given that the transformer-based U-Net is central to the model, ToMeSD (Bolya and Hoffman 2023) proposes token merging for stable diffusion, significantly enhancing throughput.

Despite these advancements, existing merging approaches primarily focus on the inherent feature space of tokens, overlooking the cross-modal interactions prevalent in diverse applications. This gap leads us to an intriguing question: **What if we consider the semantic information of multimodal prompts in the merging process?** To explore this, we conducted token merging experiments using a pretrained vision-language model (V-L model) with a simple sentence prompt: "*The scene of the [category name].*" The V-L model includes a modality encoder (for images) and a prompt encoder (for text). When applying our promptable method with a 50% reduction ratio, as illustrated in Figure 1(a), the results closely resemble the original image, outperforming existing token compression techniques. This suggests that incorporating semantic information from the prompt into the merging process can mitigate information loss. To further illustrate the advantage of utilizing semantic information in token merging, consider the example of "A white cat in the snow." As depicted in Figure 2, the original token merging process tends to cluster tokens containing both the white cat and the snow, leading to information loss. In contrast, by evaluating similarity in the prompt-induced space, our method appropriately distinguishes cat tokens from snow tokens, merging the redundant tokens within each cluster.

Motivated by these findings, we propose **M**ultimodal **P**romptable **T**oken **M**erging (MPTM) for diffusion models. Unlike existing token reduction approaches, which are typically applied to transformer-based models in tasks that do not require dense prediction—such as image-text retrieval, visual question answering (VQA), and image captioning—"**the MPTM approach is specifically tailored for attention-based multimodal models that excel in tasks requiring dense prediction across a wide range of modalities**", including images, bounding boxes, audio, and video. The method leverages multimodal prompts for token merging, thereby enriching application scenarios (see Figure 1(b)). When applied to generative models, diffusion-based detectors, and cross-modality models, MPTM not only enhances throughput but also preserves performance. The main contributions can be summarized as follows:

- We present the MPTM framework, which introduces a promptable multimodal conditioning perspective for token reduction, designed to seamlessly integrate with diffusion models that use U-Nets across various input modalities.
- We demonstrate significant performance improvements over state-of-the-art methods, particularly in tasks requiring dense prediction, such as image generation and object detection.
- We showcase the computational efficiency and application versatility of our proposed multimodal generalizations of existing token merging techniques.

2 Related Work

2.1 Transformer-based Tasks

Vision Transformer (ViT) models (Dosovitskiy et al. 2021) have achieved state-of-the-art performance across a variety of downstream tasks, including image classification (Dosovitskiy et al. 2021; Marin et al. 2023), object detection (Zhu et al. 2021; Chen et al. 2023b), semantic segmentation (Xie et al. 2021), and image generation (Bolya and Hoffman 2023; Ristea et al. 2023; Zhang et al. 2022; Rombach et al. 2022). However, ViTs are computationally intensive due to the quadratic

increase in complexity with the number of tokens processed in the stacked attention layers. Encouragingly, (Naseer et al. 2021) show that ViTs are robust to patch-dropping, suggesting that token compression can be used to discard less informative tokens. .

Image Generation. Recent diffusion-based image generators (Bolya and Hoffman 2023; Rombach et al. 2022; Tang et al. 2023a) use multiple diffusion steps to iteratively denoise the initially added noise. Most modern diffusion models rely on a U-Net (Ronneberger, Fischer, and Brox 2015) composed of Transformer blocks, with each block containing self-attention, cross-attention, and multi-layer perception modules. These modules are the precise points where MPTM can be integrated. For image generation via stable diffusion, (Bolya and Hoffman 2023) demonstrate how Token Merging (ToMe) (Bolya et al. 2023) can accelerate the image generation process.

Object Detection. Query-based detectors, such as DETR (Carion et al. 2020) and Deformable DETR (Zhu et al. 2021), achieve promising performance by employing a Transformer design, diverging from traditional methods that rely on classification and box regression on object priors like anchors or proposals. DiffusionDet (Chen et al. 2023b) utilizes diffusion models for object detection by framing the task as generation over the position-size space representing bounding boxes.

Multimodal Generation. The Composable Diffusion (CoDi) model (Tang et al. 2023b) is developed for multimodal generation tasks, enabling the creation of content across different modalities, including videos, images, audio, and text. Our proposed MPTM leverages CoDi to incorporate multimodal prompt conditioning, enabling the underlying diffusion models to achieve more effective multimodal generation.

2.2 Token Compression

Token compression (Haurum et al. 2023) addresses the issue of redundancy in ViTs. Recent studies show that this approach can speed up transformers by eliminating redundant tokens with only a moderate trade-off in accuracy. It can be seamlessly integrated into existing ViT-based models or used as an additional component alongside other network compression methods (Chen et al. 2023a).

Token Pruning. Token pruning aims to discard less informative tokens by measuring per-token importance according to a specific metric. DynamicViT (Rao et al. 2021) adds an extra token selection network, enabling a trained ViT to focus on a subset of tokens. Both DynamicViT and SPVit (Kong et al. 2022) maintain a per-image token-level mask vector, along with manually defined token compression rates, ensuring that all images retain the same number of tokens. EViT (Liang et al. 2022) introduces inattentive token fusion, consolidating information from less informative tokens to form new tokens. ATS (Fayyaz et al. 2022) employs inverse transform sampling to select tokens for pruning.

Token Merging. ToMe (Bolya et al. 2023) progressively combines r tokens within each Transformer block by separating tokens into source and destination groups, then pairing each source token with its most similar destination token. (Bolya and Hoffman 2023) enhance ToMe with token unmerging, enabling its application to diffusion models such as stable diffusion. TokenLearner (Ryoo et al. 2021) learns a weighted average of the entire feature map with a dynamic attention map to retain a small number of tokens. TokenPooling (Marin et al. 2023) reduces the number of tokens using k-means clustering. Recently, DiffRate (Chen et al. 2023a) integrates token pruning and merging with a differentiable compression rate. CrossGET (Shi et al. 2023) introduces innovative graph-based soft matching and a cross-guided ensemble, enhancing the effectiveness of existing vision-language transformer models in tasks involving two modalities (text and image), such as image-text retrieval, visual question answering (VQA), and image captioning.

3 Method

3.1 Preliminaries

Latent Diffusion and Classifier-free Guidance. Latent diffusion models (Rombach et al. 2022) project the original data $\mathbf{x}$ into a latent space $\mathbf{z}$, using an encoder that varies with the data modality (*e.g.*, VQ-VAE (van den Oord, Vinyals, and Kavukcuoglu 2017; Razavi, van den Oord, and Vinyals 2019) for image data and audio VAE (Liu et al. 2023) for audio data), before applying the diffusion model. When integrated with classifier-free guidance (Ho and Salimans 2022), these models can facilitate a conditional diffusion process. The associated loss for latent diffusion is

$$\mathcal{L}_\theta = \|\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, \mathbf{c}, t) - \boldsymbol{\epsilon}\|^2, \tag{1}$$

where $\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, \mathbf{c}, t)$ incorporates a cross-attention mechanism, enabling the fusion of the conditional embedding $\mathbf{c}$ into the latent $\mathbf{z}$. The conditional score estimation is then given by

$$\hat{\boldsymbol{\epsilon}}_\theta(\mathbf{z}|\mathbf{c}) = (1+\omega)\boldsymbol{\epsilon}_\theta(\mathbf{z}, \mathbf{c}) - \omega\boldsymbol{\epsilon}_\theta(\mathbf{z}, \varnothing), \tag{2}$$

where ω adjusts the strength of the classifier guidance and the symbol $\varnothing$ denotes the unconditional embedding.

Token Reduction. Token reduction mechanisms streamline computational complexity in transformer-based models. With input tokens n_i tokens and prompt tokens n_p of d channels, the computational load per layer is proportional to self-attention and cross-attention demands, as shown below,

$$\text{Complexity} \propto \underbrace{n_i^2 \times d}_{\text{self-attention}} + \underbrace{n_i \times n_p \times d}_{\text{cross-attention}}. \tag{3}$$

These approaches evaluate the importance of each token with several matrices (*i.e.* the weight magnitude of tokens or similarity with other tokens) to discard the unimportant tokens or merge the tokens that are similar. When the reduction ratio is $0 < r < 1$, the complexity is reduced to

$$\text{Complexity} \propto \underbrace{(1-r)^2 n_i^2 \times d}_{\text{self-attention}} + \underbrace{(1-r) n_i \times n_p \times d}_{\text{cross-attention}}. \tag{4}$$

Diffusion-based Object Detection. DiffusionDet (Chen et al. 2023b) introduces a novel application of diffusion processes to object detection, conceptualizing it as a "noise-to-box" transformation. The approach represents a bounding box as a 4-tuple to specify its location as $b = (x, y, w, h)$.

Here, x and y denote the center coordinates of the box, while w, and h represent its width and height, respectively. DiffusionDet then utilizes a modified dynamic instance interactive head, derived from Sparse R-CNN (Sun et al. 2021), to enable accurate box prediction:

$$\text{prediction box} = \text{decoder}(\text{box}_{\text{noisy}}, t, \text{feature}). \quad (5)$$

This equation indicates how the decoder function uses noisy box data, time-step t, and relevant features to yield the predicted bounding box location and dimensions.

3.2 Multimodal Promptable Token Merging

Token reduction techniques such as ToMe (Bolya et al. 2023) are developed to lower computational costs concerning attention mechanisms for transformer-based neural network models. A majority of such approaches are formulated by establishing a similarity measure solely based on the token features to facilitate the subsequent operations, including token filtering, grouping or matching. However, the use of "prompt" has emerged as another source of input to the transformer-based models and is shown to enhance their flexibility and effectiveness in several contemporary computer vision-related applications, *e.g.*, text-to-image (Nichol et al. 2022; Ge et al. 2023), video editing (Qi et al. 2023; Chai et al. 2023) and promptable detection and segmentation tasks (Kirillov et al. 2023). Since the inclusion of multimodal prompts may significantly affect the importance of each token, and *redefines* the similarity relations between tokens, we are thus motivated to propose a new token merging technique that primarily takes account of the use of the promptable mechanisms in transformer-based models.

Multimodal Prompt Fusion. We consider a general setting of token merging that both the prompts and the inputs can be multimodal. Assume that there are K modalities of prompts, denoted as $P^{(k)}, k = 1, \ldots, K$ and L modalities of input type. Analogously to the use of embedding alignment in (Tang et al. 2023b), we regulate the K prompt encoders, $\{E_{P^{(k)}}\}_{k=1}^{K}$, to project to the same d-dimensional space, also ensuring that the encoder of each modality yields the same number of prompt tokens. (See Figure 3.)

Given a prompt $P^{(k)}$ of modality k, our method applies the respective prompt encoder $E_{P^{(k)}}$ to obtain n prompt tokens. We express the prompt embedding and tokenization as

$$P^{(k)} \xmapsto{E_{P^{(k)}}} \{\mathbf{p}_1^{(k)}, \mathbf{p}_2^{(k)}, \ldots, \mathbf{p}_n^{(k)}\}, \quad (6)$$

where $\mathbf{p}_i^{(k)} \in \mathbb{R}^d$ is a prompt token of modality k, and d is the number of output channels by $E_{P^{(k)}}$. The embedding alignment among the K prompt encoders naturally leads to the fusion of the resulting multimodal prompt tokens by

$$\mathbf{p}_i = \sum\nolimits_{k=1}^{K} \alpha_k \, \mathbf{p}_i^{(k)}, \quad i = 1, \ldots, n, \quad (7)$$

where $\mathbf{p}_i$ is the resulting ith token after fusion, $\sum \alpha_k = 1$ and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)^\intercal$ is the predefined fusion vector to weigh the influence of each modality in the multimodal conditioning scenario. Notice that when text prompt is present among the K modalities, its corresponding fusion wight α_k would be set to a larger value owing to its pivotal role in the alignment of prompt encoders.

Prompt-induced Similarity. Besides the multimodal conditioning imposed by prompts, the inputs can also comprise multiple modalities, where we use $\mathbf{z}^{(\ell)}$ to represent an input of data modality $\ell \in \{1, \ldots, L\}$. To handle this aspect of complexity, we consider a network structure of multiple streams, each of which deals with a specific input modality. (See Figure 3(a).) Since our formulation for establishing the prompt-induced similarity is the same for each input modality ℓ, for the sake of simplicity, we hereafter omit the modality index ℓ from the notation. As such, we simply consider an arbitrary transformer layer from one of the multiple network streams, and an input $\mathbf{z}$, yielding m input tokens $\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_m$ and $\mathbf{q}_i \in \mathbb{R}^d$. Let A be the affinity measure that returns $A(\mathbf{q}_i, \mathbf{q}_j)$ as the similarity value between two input tokens $\mathbf{q}_i$ and $\mathbf{q}_j$. Notice that the similarity measure A is constructed without considering the prompt effect, which has been used in existing token reduction methods.

The set of n multimodal prompt tokens in (7) implicitly defines a "multimodal prompt space" which yields a new feature representation $\tilde{\mathbf{q}}$ for each input token $\mathbf{q}$ where

$$\tilde{\mathbf{q}} = (\mathbf{q}^\intercal \mathbf{p}_1, \ldots, \mathbf{q}^\intercal \mathbf{p}_n)^\intercal \in \mathbb{R}^n. \quad (8)$$

With (8), we can define the proposed *prompt-induced* similarity measure $\tilde{A}$ as follows:

$$\tilde{A}(\mathbf{q}_i, \mathbf{q}_j) \coloneqq A(\tilde{\mathbf{q}}_i, \tilde{\mathbf{q}}_j), \quad (9)$$

where we implement A with the cosine similarity. However, obtaining the new feature representation by projecting to the multimodal prompt space introduces extra computation cost. We instead consider first deriving a representative prompt token $\tilde{\mathbf{p}}$ from $\{\mathbf{p}_i\}$ and then approximating (8) by

$$\tilde{\mathbf{q}} \approx (\mathbf{q}^\intercal \tilde{\mathbf{p}}, \ldots, \mathbf{q}^\intercal \tilde{\mathbf{p}})^\intercal \in \mathbb{R}^n, \quad (10)$$

where $\tilde{\mathbf{p}}$ is obtained by either mean-pooling or max-pooling over $\{\mathbf{p}_i\}$. The related ablation experiment regarding the projection approximation can be found in Table 6d in the appendix.

Token Merging. We can now carry out token merging with the prompt-induced similarity measure in (9). We adopt the bipartite soft matching in (Bolya et al. 2023) to achieve this process, which comprises the following three essential steps: **1.** Divide the input tokens into two sets, denoted Set $\mathbb{A}$ and Set $\mathbb{B}$, via alternating selection. **2.** Connect each token in $\mathbb{A}$ to its most similar counterpart in $\mathbb{B}$. **3.** Preserve $r\%$ most similar connections, merge the linked tokens and concatenate the remaining ones in $\mathbb{A}$ and $\mathbb{B}$.

Token Unmerging. As pointed out in token merging for stable diffusion (ToMeSD) (Bolya and Hoffman 2023), diffusion generative models require knowledge of noise removal for each token, necessitating an unmerging process. ToMeSD adopts a straightforward scheme: assigning the merged feature uniformly back to the unmerged tokens by

$$\begin{aligned} &\text{Merge:} \quad \mathbf{q}_{\text{merge}} = \text{mean}(\mathbf{q}_i + \mathbf{q}_j), \\ &\text{Unmerge:} \quad \mathbf{q}_i^{out} = \mathbf{q}_j^{out} = \mathbf{q}_{\text{merge}}, \end{aligned} \quad (11)$$

where $\mathbf{q}_*^{out}$ denotes the output of the attention module. In our Multimodal Promptable Token Merging (MPTM), a τ-threshold strategy is formulated to minimize the redundancy

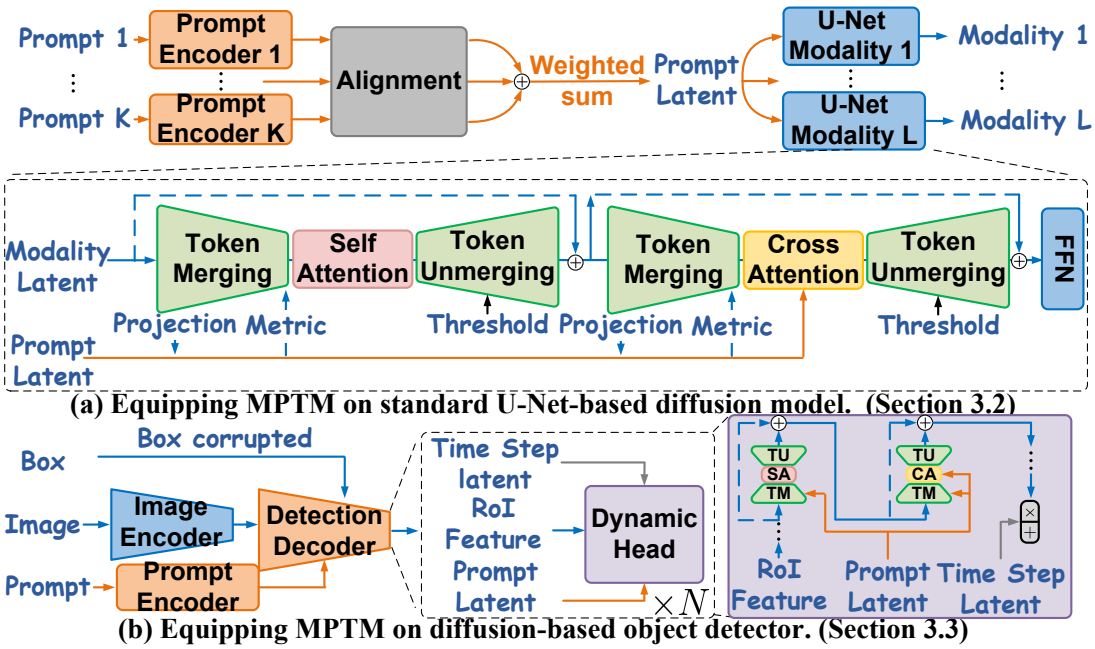

Figure 3: **Two variants of MPTM-enhanced model configurations**. (a) For a standard U-Net-based diffusion model, additional token merging (TM), token unmerging (TU), and prompt are plugged into the basic Transformer block. (b) For the diffusion-based object detector, additional self/cross-attention layers besides the TM, TU, and prompt are inserted in the dynamic head.

caused by error propagation when adopting a high reduction ratio r for token merging. Specifically, we have

$$\mathbf{q}_*^{out} = \begin{cases} \mathbf{q}_{\text{merge}}, & \tilde{A}(\mathbf{q}_i, \mathbf{q}_j) \geq \tau, \\ \mathbf{q}_*, & \tilde{A}(\mathbf{q}_i, \mathbf{q}_j) < \tau, \end{cases} \tag{12}$$

where $\tau > 0$ is a specified similarity threshold, allowing the unmerging operation to assign $\mathbf{q}_{\text{merge}}$ to both output tokens if they are sufficiently similar to each other before merging. Furthermore, we also propose enhancing the mechanism by substituting the fixed-value threshold with a learnable threshold for achieving optimal performance during training. The results of our unmerging scheme are detailed in the corresponding ablation study, as reported in Table 7c in the Appendix.

3.3 MPTM for Object Detection

DiffusionDet (Chen et al. 2023b) introduces the innovative "noise-to-box" concept in object detection tasks. During its training phase, the detection decoder predicts box features from a mix of corrupted boxes, image features, and the degree of corruption. In the sampling process, random noise boxes are processed through the detection decoder and subsequently refined in the DDIM step to produce the final outcomes. However, the simplicity of this approach and the disparity between its training and sampling phases might curtail its potential. DiffusionDet incorporates a dynamic head for detection decoding, as proposed in Sparse R-CNN (Sun et al. 2021). Integrating extra self-attention and cross-attention modules into the dynamic head aims to amalgamate prompt features, a staple in latent diffusion models, which is theoretically conducive to learning enhanced features for promptable object detection. However, this enhancement introduces considerable computational demands to the sizeable object detection models. To manage this challenge efficiently, a token merging mechanism is utilized. As shown in Figure 3(b), token merging blocks are adeptly integrated into the detection decoder. This modification enables the resulting model to achieve generalized object detection in a more flexible setting. For the experiment in Section 4.2 on applying MPTM to DiffusionDet, we have used prompts such as"*The scene of the [categories in the ground truth]*" and negative prompts like "*The scene of the [categories not in the ground truth].*" to significantly diversify the application scenarios of object detection.

4 Experiments

To reiterate, the primary focus of MTPM is on diffusion models, especially for tasks that require dense prediction, where few token compression approaches, such as ToMeSD (Bolya and Hoffman 2023), are applicable. Accordingly, our experimental evaluations have primarily concentrated on these tasks. Although this is outside our main research scope, we have also included comparisons with existing approaches on traditional discriminative models for tasks like image classification. These comparisons can be found in Appendix D.1. Table 1 details the experimental setup used to evaluate token merging methods across various dense prediction tasks involving different data modalities. These tasks include text-to-image generation, prompt-based object detection, and multimodal generation using multimodal prompts. The results demonstrate that MTPM effectively prunes less informative tokens while preserving critical ones, resulting in enhanced overall performance. To ensure statistical robustness, all re-

Dense Prediction Task	Baseline Architecture	Training	Input Modalities → Output Modalities
Text-to-Image Generation	Stable Diffusion v1.5	✗	Text → Image
Promptable Object Detection	Diffusion Detector	✓	Text → Box
Multimodal Generation	Composable Diffusion	✗	(Text, Video, Audio, Image) → (Text, Video, Audio, Image)

Table 1: **Task settings.** MPTM is a training-free add-on for most pretrained baseline models. However, for the object detection task, both MPTM and ToMe need model retraining to work properly with DiffusionDet. In the other two tasks, τ is predefined.

Method	r%	ImangNet-1k CLIP-R (↑)	FID (↓)	s/image (↓)	GB/image (↓)
Image Generative Architecture (Rombach et al. 2022)					
SD v1.5	-	62.11	33.14	3.11	3.41
Token Merging for Stable Diffusion (Bolya and Hoffman 2023)					
+ToMe	10	60.11±0.12	32.98±0.04	2.58	2.99
	20	58.23±0.26	33.03±0.04	2.31	2.17
	30	55.17±0.81	33.05±0.06	2.08	1.71
	40	51.56±1.44	33.08±0.06	1.87	1.26
	50	49.33±1.78	33.15±0.07	1.69	0.89
	60	47.05±2.16	33.43±0.07	1.54	0.60
Promptable Token Merging for Stable Diffusion					
+ **MPTM**	10	62.33±0.04	32.91±0.03	2.65	3.02
	20	61.65±0.16	32.90±0.03	2.45	2.22
	30	60.52±0.37	32.92±0.05	2.18	1.75
	40	59.42±0.54	32.96±0.05	1.93	1.27
	50	58.51±0.72	33.07±0.06	1.82	0.91
	60	57.46±1.23	33.14±0.07	1.61	0.65

Table 2: **Performance comparison of text-to-image generation on ImageNet-1k dataset.** MPTM achieves better image generation outcomes (FID, CLIP-R) with minor tradeoffs in throughput (s/image) and memory (GB/image) compared to ToMeSD. The blue highlights indicate that MPTM achieves competitive quality and improved consistency with higher throughput, utilizing a more challenging reduction ratio (60%, MPTM compared to 50%, ToMe) than the existing approach.

sults are reported as the mean of three independent runs. Due to space constraints, the detailed training methodology is provided in Appendix B.1, and the explicit ablation studies of the method's components are presented in Appendix C. Additionally, comprehensive information about the training datasets is available in Appendix B.2, and the architecture specifics of the multimodal diffusion models are elaborated in Appendix B.4.

4.1 Text-to-Image Generation

We perform text-to-image generation on the ImageNet-1k dataset as ToMeSD (Bolya and Hoffman 2023). The stable diffusion version 1.5 (SD v1.5) (Rombach et al. 2022), involving 50 diffusion steps via PLMS, is used to generate two samples per category on ImageNet-1k, resulting in a total of 2,000 images of resolution 512×512. The classifier-free guidance scale is set at 7.5, and the prompt we used for this task is "*The image of* [*category name*]." The setting details can be referred to in Appendix B.3.

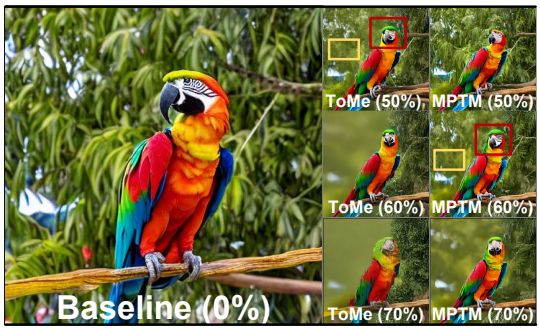

Figure 4: **Visualization of text-to-image generation.** Compared to ToMe on the stable diffusion-based image generation, MPTM effectively preserves the image content fidelity even at large merging ratios. The yellow and red boxes indicate missing neck feathers, blurred tree leaves, and changed head feathers, respectively.

Performance Comparison. Table 2 shows the quantitative results obtained from comparing our MPTM to ToMe. The results demonstrate that MPTM consistently achieves higher image generation quality based on the FID score while also maintaining better image-text alignment according to the CLIP-R metric. Despite the improvements, there is only a minor decrease in throughput and an increase in memory usage. Figure 4 visualizes the qualitative results, which indicate that MPTM successfully preserves the image content fidelity as the token merging ratio increases.

4.2 Promptable Object Detection

We employ the DiffusionDet (Chen et al. 2023b), a diffusion-based object detection method, for promptable object detection on the Large Vocabulary Instance Segmentation (LVIS) (Gupta, Dollár, and Girshick 2019) and COCO (Lin et al. 2014) datasets. The object detection is carried out using ResNet-50, ResNet-101 (He et al. 2016), and Swin Transformer (Liu et al. 2021) as the backbones for encoding images. Additionally, a modified dynamic head is used as the detection decoder detailed in Section 3.3. Furthermore, we have leveraged the prompt encoder, which is a text encoder from CLIP (Radford et al. 2021), to extract the latent prompt. The prompt we used for this task is "*The scene of the* [*category name 1, ..., category name N*].", in which the categories are derived from the ground truth. The data augmentation, dataset contents and full comparison results of the LVIS and COCO are available in Appendix B and D.3.

Method	AP (↑)	AP_{50} (↑)	AP_{75} (↑)	AP_r (↑)	AP_c (↑)	AP_f (↑)	FPS (↓)
Diffusion-based with ResNet50 (He et al. 2016)							
DiffusionDet (Chen et al. 2023c)	29.4	40.4	31.0	22.7	27.2	34.7	19
DiffusionDet w/ ToMe (50%)	25.1 ± 0.3	35.9 ± 0.4	27.2 ± 0.3	18.3 ± 0.3	23.1 ± 0.3	30.5 ± 0.5	17
DiffusionDet w/ MPTM (50%)	29.5 ± 0.2	40.7 ± 0.3	31.2 ± 0.2	22.9 ± 0.2	27.3 ± 0.3	35.1 ± 0.4	17
DiffusionDet w/ MPTM (50%) + NP	**30.3** ± **0.3**	**41.1** ± **0.4**	**32.1** ± **0.3**	**23.8** ± **0.4**	**28.3** ± **0.3**	**35.2** ± **0.4**	17
Diffusion-based with Swin-B Transformer (Liu et al. 2021)							
DiffusionDet (Chen et al. 2023c)	39.5	52.3	42.0	33.0	38.5	43.5	11
DiffusionDet w/ ToMe (50%)	34.2 ± 0.8	47.2 ± 0.9	37.1 ± 0.7	28.4 ± 0.7	33.4 ± 0.6	38.3 ± 0.7	9
DiffusionDet w/ MPTM (50%)	39.9 ± 0.6	52.5 ± 0.8	42.2 ± 0.5	34.3 ± 0.6	38.6 ± 0.7	43.9 ± 0.6	9
DiffusionDet w/ MPTM (50%) + NP	**41.8** ± **0.7**	**55.6** ± **0.7**	**44.8** ± **0.6**	**36.6** ± **0.7**	**40.6** ± **0.7**	**47.2** ± **0.7**	9

Table 3: **Performance comparison of promptable object detection on LVIS V1.0 `val` dataset.** All diffusion detectors in this table employ one iteration step and are evaluated with 300 boxes.

Method	AudioCaps			
	FID_t (↓)	**FID_a (↓)**	**FID_{t+a} (↓)**	**s/image (↓)**
Image Generative Architecture (Tang et al. 2023b)				
CoDi†	16.23	16.21	16.77	4.72
Token Merging for Stable Diffusion (Bolya and Hoffman 2023)				
+ToMe(50%)	16.57±0.31	16.69±0.42	16.92±0.52	1.22
Promptable Token Merging for Stable Diffusion				
+ **MPTM (60%)**	16.32±0.24	16.36±0.33	16.79±0.47	1.19

Table 4: **Performance comparison of multi-modality image generation on AudioCaps `val` dataset.** The button markers for FID indicate the input modality (i.e., t = text and a = audio). † For calculating FID, this experiment was conducted using a subset of the AudioCaps validation dataset.

Method	AudioCaps	
	SIM-VA (↑)	**s/sample (↓)**
Multimodal Generative Architecture (Tang et al. 2023b)		
CoDi†	0.247	32.14
Token Merging for Stable Diffusion (Bolya and Hoffman 2023)		
+ToMe(50%)	0.194±0.035	12.65
Promptable Token Merging for Stable Diffusion		
+ **MPTM (60%)**	0.241±0.022	10.23

Table 5: **Performance Comparison of Multi-inputs-outputs ((Text + Image → Video + Audio) on AudioCaps `val` dataset.** SIM-VA denotes the cosine similarity between generated embeddings of the modalities.

Performance Comparison.

The results of using MPTM with a 50% token merging ratio in DiffusionDet over three backbones, ResNet-50, ResNet-101, and Swin Transformer, are summarized in Table 3 and Appendix D.3. Two important findings emerged from the analysis. First, plugging MPTM in DiffusionDet as "DiffusionDet w/ MPTM", without Negative Prompt (NP), consistently resulted in higher AP scores compared to "DiffusionDet w/ ToMe" across all three backbones. The degradation of "DiffusionDet w/ ToMe" can be mainly attributed to the absence of semantic consideration in its merging process. In contrast, "DiffusionDet w/ MPTM" successfully retains the performance of DiffusionDet by incorporating semantic information via prompt information. Second, in comparison to "DiffusionDet w/ MPTM", the addition of negative prompts in "DiffusionDet w/ MPTM + NP", with the Swin Transformer backbone, led to notable performance gains of 1.9% AP on LVIS. The results show the effectiveness of negative prompts, a notable advantage of the MPTM framework.

4.3 Multimodal Generation

The CoDi model (Tang et al. 2023b) is capable of creating content across different modalities, such as videos, images, audio, and text. We integrate MPTMs within CoDi's inlaid U-Nets, as described in Appendix B.4. Detailed settings can be found in Appendix B.3. Table 4 demonstrates that MPTM not only yields superior image quality but also achieves enhanced throughput compared to ToMe, regardless of the originating modality, be it text, audio, or a combination thereof. Table 5 addresses scenarios involving multiple input and output modalities. The findings corroborate MPTM's consistent performance, further improving throughput. These experiments underscore the robustness and adaptability of our MPTM in facilitating broad generative tasks, marking a significant advancement in the realm of multimodal generation.

5 Conclusions

We have introduced Multimodal Promptable Token Merging (MPTM), an innovative approach to token reduction that leverages prompt-induced semantic relationships during the merging process to optimize both quality and efficiency. MPTM is designed for seamless integration into diffusion models utilizing U-Nets, requiring no additional training, which underscores its practicality. In addition to its application in generative models, MPTM has demonstrated significant potential in detection tasks, enhancing detector flexibility through prompt-based input adaptation. The versatility of MPTM extends to generative models handling multi-modality inputs, highlighting its broad applicability.

Acknowledgments

This work was supported in part by NSTC grants 111-2221-E-001-015-MY3, 111-2221-E-001-011-MY2, 113-2221-E-001-010-MY3 and 113-2634-F-007-002 of Taiwan. We thank National Center for High-performance Computing for providing computing resources.

References

Bolya, D.; Fu, C.; Dai, X.; Zhang, P.; Feichtenhofer, C.; and Hoffman, J. 2023. Token Merging: Your ViT But Faster. In *Int. Conf. Learn. Represent. (ICLR)*.

Bolya, D.; and Hoffman, J. 2023. Token Merging for Fast Stable Diffusion. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh. (CVPRW)*, 4599–4603.

Cai, H.; Zhu, L.; and Han, S. 2019. ProxylessNAS: Direct Neural Architecture Search on Target Task and Hardware. In *Int. Conf. Learn. Represent. (ICLR)*.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-End Object Detection with Transformers. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J., eds., *Eur. Conf. Comput. Vis. (ECCV)*, volume 12346, 213–229.

Chai, W.; Guo, X.; Wang, G.; and Lu, Y. 2023. StableVideo: Text-driven Consistency-aware Diffusion Video Editing. In *Int. Conf. Comput. Vis. (ICCV)*.

Chen, M.; Shao, W.; Xu, P.; Lin, M.; Zhang, K.; Chao, F.; Ji, R.; Qiao, Y.; and Luo, P. 2023a. DiffRate : Differentiable Compression Rate for Efficient Vision Transformers. In *Int. Conf. Comput. Vis. (ICCV)*.

Chen, S.; Sun, P.; Song, Y.; and Luo, P. 2023b. DiffusionDet: Diffusion Model for Object Detection. In *Int. Conf. Comput. Vis. (ICCV)*.

Chen, S.; Sun, P.; Song, Y.; and Luo, P. 2023c. Diffusiondet: Diffusion model for object detection. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 19830–19843.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*, 4171–4186.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Int. Conf. Learn. Represent. (ICLR)*.

Fayyaz, M.; Koohpayegani, S. A.; Jafari, F. R.; Sengupta, S.; Joze, H. R. V.; Sommerlade, E.; Pirsiavash, H.; and Gall, J. 2022. Adaptive Token Sampling for Efficient Vision Transformers. In *Eur. Conf. Comput. Vis. (ECCV)*, 396–414.

Ge, S.; Park, T.; Zhu, J.; and Huang, J. 2023. Expressive Text-to-Image Generation with Rich Text. In *Int. Conf. Comput. Vis. (ICCV)*.

Gong, C.; Wang, D.; Li, M.; Chen, X.; Yan, Z.; Tian, Y.; Liu, Q.; and Chandra, V. 2022. NASViT: Neural Architecture Search for Efficient Vision Transformers with Gradient Conflict aware Supernet Training. In *Int. Conf. Learn. Represent. (ICLR)*.

Gupta, A.; Dollár, P.; and Girshick, R. B. 2019. LVIS: A Dataset for Large Vocabulary Instance Segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 5356–5364.

Han, S.; Mao, H.; and Dally, W. J. 2016. Deep Compression: Compressing Deep Neural Network with Pruning, Trained Quantization and Huffman Coding. In Bengio, Y.; and LeCun, Y., eds., *Int. Conf. Learn. Represent. (ICLR)*.

Haurum, J. B.; Escalera, S.; Taylor, G. W.; and Moeslund, T. B. 2023. Which Tokens to Use? Investigating Token Reduction in Vision Transformers. In *Int. Conf. Comput. Vis. Worksh. (ICCVW)*.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. B. 2022. Masked Autoencoders Are Scalable Vision Learners. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 15979–15988.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 770–778.

Ho, J.; and Salimans, T. 2022. Classifier-Free Diffusion Guidance. *CoRR*, abs/2207.12598.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Int. Conf. Comput. Vis. (ICCV)*.

Kong, Z.; Dong, P.; Ma, X.; Meng, X.; Niu, W.; Sun, M.; Shen, X.; Yuan, G.; Ren, B.; Tang, H.; Qin, M.; and Wang, Y. 2022. SPViT: Enabling Faster Vision Transformers via Latency-Aware Soft Token Pruning. In *Eur. Conf. Comput. Vis. (ECCV)*, 620–640.

Liang, Y.; Ge, C.; Tong, Z.; Song, Y.; Wang, J.; and Xie, P. 2022. Not All Patches are What You Need: Expediting Vision Transformers via Token Reorganizations. In *Int. Conf. Learn. Represent. (ICLR)*.

Lin, C.; Gao, J.; Tang, L.; Takikawa, T.; Zeng, X.; Huang, X.; Kreis, K.; Fidler, S.; Liu, M.; and Lin, T. 2023. Magic3D: High-Resolution Text-to-3D Content Creation. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 300–309.

Lin, T.; Maire, M.; Belongie, S. J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In Fleet, D. J.; Pajdla, T.; Schiele, B.; and Tuytelaars, T., eds., *Eur. Conf. Comput. Vis. (ECCV)*, 740–755.

Liu, H.; Chen, Z.; Yuan, Y.; Mei, X.; Liu, X.; Mandic, D. P.; Wang, W.; and Plumbley, M. D. 2023. AudioLDM: Text-to-Audio Generation with Latent Diffusion Models. In *Int. Conf. Machine Learning (ICML)*, 21450–21474.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *Int. Conf. Comput. Vis. (ICCV)*, 9992–10002.

Marin, D.; Chang, J. R.; Ranjan, A.; Prabhu, A.; Rastegari, M.; and Tuzel, O. 2023. Token Pooling in Vision Transformers for Image Classification. In *IEEE Winter Conf. App. Comput. Vis. (WAVC)*, 12–21.

Naseer, M.; Ranasinghe, K.; Khan, S.; Hayat, M.; Khan, F. S.; and Yang, M. 2021. Intriguing Properties of Vision Transformers. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 23296–23308.

Nichol, A. Q.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvári, C.; Niu, G.; and Sabato, S., eds., *Int. Conf. Machine Learning (ICML)*.

Pelosin, F.; Jha, S.; Torsello, A.; Raducanu, B.; and van de Weijer, J. 2022. Towards Exemplar-Free Continual Learning in Vision Transformers: an Account of Attention, Functional and Weight Regularization. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 3819–3828.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2022. DreamFusion: Text-to-3D using 2D Diffusion. In *Int. Conf. Learn. Represent. (ICLR)*.

Qi, C.; Cun, X.; Zhang, Y.; Lei, C.; Wang, X.; Shan, Y.; and Chen, Q. 2023. FateZero: Fusing Attentions for Zero-shot Text-based Video Editing Supplemental Material. In *Int. Conf. Comput. Vis. (ICCV)*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Int. Conf. Machine Learning (ICML)*, 8748–8763.

Rao, Y.; Zhao, W.; Liu, B.; Lu, J.; Zhou, J.; and Hsieh, C. 2021. DynamicViT: Efficient Vision Transformers with Dynamic Token Sparsification. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 13937–13949.

Razavi, A.; van den Oord, A.; and Vinyals, O. 2019. Generating Diverse High-Fidelity Images with VQ-VAE-2. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 14837–14847.

Ristea, N.; Miron, A.; Savencu, O.; Georgescu, M.; Verga, N.; Khan, F. S.; and Ionescu, R. T. 2023. CyTran: A cycle-consistent transformer with multi-level consistency for non-contrast to contrast CT translation. *Neurocomputing*, 538: 126211.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 10674–10685.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *MICCAI*, 234–241.

Ryoo, M. S.; Piergiovanni, A. J.; Arnab, A.; Dehghani, M.; and Angelova, A. 2021. TokenLearner: Adaptive Space-Time Tokenization for Videos. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 12786–12797.

Shi, D.; Tao, C.; Rao, A.; Yang, Z.; Yuan, C.; and Wang, J. 2023. Crossget: Cross-guided ensemble of tokens for accelerating vision-language transformers. *CoRR*, abs/2305.17455.

Sun, P.; Zhang, R.; Jiang, Y.; Kong, T.; Xu, C.; Zhan, W.; Tomizuka, M.; Li, L.; Yuan, Z.; Wang, C.; and Luo, P. 2021. Sparse R-CNN: End-to-End Object Detection With Learnable Proposals. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 14454–14463.

Tang, Z.; Yang, Z.; Zhu, C.; Zeng, M.; and Bansal, M. 2023a. Any-to-Any Generation via Composable Diffusion. *CoRR*, abs/2305.11846.

Tang, Z.; Yang, Z.; Zhu, C.; Zeng, M.; and Bansal, M. 2023b. Any-to-Any Generation via Composable Diffusion. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*.

van den Oord, A.; Vinyals, O.; and Kavukcuoglu, K. 2017. Neural Discrete Representation Learning. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 6306–6315.

Wang, Z.; Luo, H.; Wang, P.; Ding, F.; Wang, F.; and Li, H. 2022. VTC-LFC: Vision Transformer Compression with Low-Frequency Components. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Álvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 12077–12090.

Yuan, Z.; Xue, C.; Chen, Y.; Wu, Q.; and Sun, G. 2022. PTQ4ViT: Post-training Quantization for Vision Transformers with Twin Uniform Quantization. In *Eur. Conf. Comput. Vis. (ECCV)*, 191–207.

Zeng, W.; Jin, S.; Liu, W.; Qian, C.; Luo, P.; Ouyang, W.; and Wang, X. 2022. Not All Tokens Are Equal: Human-centric Visual Analysis via Token Clustering Transformer. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 11091–11101.

Zhang, B.; Gu, S.; Zhang, B.; Bao, J.; Chen, D.; Wen, F.; Wang, Y.; and Guo, B. 2022. StyleSwin: Transformer-based GAN for High-resolution Image Generation. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 11294–11304.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2021. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *Int. Conf. Learn. Represent. (ICLR)*.