# Learning Spatial Decay for Vision Transformers

**Yuxin Mao[1], Zhen Qin[2], Jinxing Zhou[3], Bin Fan[1], Jing Zhang[1], Yiran Zhong[3], Yuchao Dai[1]***

[1]School of Electronics and Information, Northwestern Polytechnical University, and Shaanxi Key Laboratory of Information Acquisition and Processing, Xi'an, China,
[2]TapTap,
[3]OpenNLPLab

## Abstract

Vision Transformers (ViTs) have revolutionized computer vision, yet their self-attention mechanism lacks explicit spatial inductive biases, leading to suboptimal performance on spatially-structured tasks. Existing approaches introduce data-independent spatial decay based on fixed distance metrics, applying uniform attention weighting regardless of image content and limiting adaptability to diverse visual scenarios. Inspired by recent advances in large language models where content-aware gating mechanisms (e.g., GLA, HGRN2, FOX) significantly outperform static alternatives, we present the first successful adaptation of data-dependent spatial decay to 2D vision transformers. We introduce **Spatial Decay Transformer (SDT)**, featuring a novel Context-Aware Gating (CAG) mechanism that generates dynamic, data-dependent decay for patch interactions. Our approach learns to modulate spatial attention based on both content relevance and spatial proximity. We address the fundamental challenge of 1D-to-2D adaptation through a unified spatial-content fusion framework that integrates manhattan distance-based spatial priors with learned content representations. Extensive experiments on ImageNet-1K classification and generation tasks demonstrate consistent improvements over strong baselines. Our work establishes data-dependent spatial decay as a new paradigm for enhancing spatial attention in vision transformers.

## Introduction

Vision Transformers (ViTs) (Dosovitskiy et al. 2020) have fundamentally transformed computer vision and multimodal tasks (Zhou et al. 2022, 2024c,a,b, 2025b,a,c; Mao et al. 2024a, 2023, 2024b), achieving state-of-the-art performance across diverse tasks from image classification to generation (Peebles and Xie 2023). The cornerstone of their success lies in the self-attention mechanism (Vaswani et al. 2017), which enables global receptive fields and long-range dependency modeling by treating images as sequences of patches. However, this design choice introduces a critical limitation: the inherent permutation equivariance of self-attention renders it agnostic to the 2D spatial structure of images, treating spatially adjacent patches identically to distant ones.

*Corresponding author (daiyuchao@gmail.com).


This spatial blindness poses significant challenges, as models must learn fundamental spatial relationships purely from data, often requiring extensive training to achieve satisfactory performance. To this end, recent works have sought to inject spatial inductive biases directly into the attention mechanism. Retentive Networks (RMT) (Fan et al. 2024) exemplify this approach by introducing data-independent spatial decay matrices that apply fixed, distance-based attention weighting according to spatial proximity. While providing useful locality bias, this strategy suffers from fundamental rigidity: the same spatial decay pattern is uniformly applied regardless of image content, preventing adaptive focus on semantically relevant regions.

Recent advances in large language models (LLMs) (Li et al. 2025; Chen et al. 2025) offer compelling insights that challenge this static approach. In linear attention mechanisms, data-dependent decay has emerged as a superior paradigm, with models like GLA (Yang et al. 2024), HGRN2 (Qin et al. 2024b; Mao et al. 2025), and Mamba2 (Dao and Gu 2024) demonstrating that content-aware gating significantly outperforms data-independent counterparts. The Forgetting Transformer (Lin et al. 2025) validates this principle within standard quadratic attention, introducing learnable forget gates that dynamically modulate attention based on input content rather than fixed positional relationships. This data-dependent approach enables more nuanced, context-sensitive information flow, leading to substantial improvements in sequence modeling.

Motivated by these successes, we investigate *whether data-dependent decay can be effectively adapted for 2D visual tasks*. This translation from 1D sequential to 2D spatial domains presents unique challenges: unlike sequential data with natural linear temporal relationships, images exhibit complex 2D spatial topologies requiring careful consideration of both horizontal and vertical interactions. Furthermore, extending from 1D positional decay to 2D spatial decay demands novel architectural designs that effectively capture content-dependent spatial relationships while maintaining computational efficiency.

We introduce **Spatial Decay Transformer**, pioneering the application of data-dependent spatial decay to vision transformers through a novel **Context-Aware Gating (CAG)** mechanism. Our approach generates dynamic, content-dependent decay strengths for 2D patch interac-

tions, enabling selective modulation of spatial attention based on both content relevance and spatial proximity. Unlike the fixed manhattan distance-based decay of RMT, our method adapts to the semantic content of each image.

Our technical contribution addresses the 1D-to-2D adaptation challenge through a unified spatial-content fusion framework that integrates manhattan distance-based spatial priors with learned content representations. For computational efficiency in high-resolution stages, we propose a decomposed implementation that maintains content-dependent characteristics while reducing memory complexity.

Extensive experiments on ImageNet-1K classification and generation demonstrate consistent improvements over strong baselines including RMT, with particularly notable gains in tasks requiring fine-grained spatial understanding. Comprehensive ablation studies validate the superiority of data-dependent over data-independent spatial decay, confirming that content-aware gating is crucial for optimal 2D spatial attention.

Our contributions are threefold:

- We systematically identify limitations of data-independent spatial decay in vision transformers and demonstrate the necessity of content-aware spatial attention mechanisms.
- We introduce the first successful adaptation of data-dependent decay from 1D sequential modeling to 2D spatial attention, pioneering a new paradigm for spatial bias injection in vision transformers.
- We propose Spatial Decay Transformer with Context-Aware Gating, achieving significant performance improvements and establishing a strong baseline for spatial attention in vision transformers.

## Related Works

**Vision Transformers and Spatial Biases.** Vision Transformers (ViTs) (Dosovitskiy et al. 2020; Qin et al. 2024a) achieve superior performance over CNNs in image recognition but require massive datasets due to lacking spatial inductive biases. DeiT (Touvron et al. 2021) addresses this limitation through knowledge distillation from CNN teachers, enabling effective training on standard datasets like ImageNet-1K (Krizhevsky, Sutskever, and Hinton 2012). Subsequent research focuses on incorporating spatial information through various position encoding schemes, including absolute (Dosovitskiy et al. 2020), relative (Shaw, Uszkoreit, and Vaswani 2018), rotary (Su et al. 2024), and conditional (Chu et al. 2021) position embeddings. Hierarchical transformers such as PVT (Wang et al. 2022a), Swin Transformer (Liu et al. 2021), and MViT (Li et al. 2022a) build multi-scale feature pyramids to improve dense prediction tasks. Recent hierarchical approaches include FasterViT (Hatamizadeh et al. 2023a) with hierarchical attention for computational efficiency.

**Explicit Spatial Priors in Attention.** To address the fundamental lack of spatial inductive biases in standard self-attention, researchers have explored directly incorporating spatial awareness into the attention mechanism. This approach typically involves modifying the attention matrix with explicit, data-independent spatial priors defined by geometric distances between patches. Hybrid architectures like CoAtNet (Dai et al. 2021) and BoTNet (Srinivas et al. 2021) effectively combine convolution and attention. Most relevant to our work, Retentive Vision Transformer (RMT) (Fan et al. 2024) introduces fixed spatial decay based on Manhattan distance, extending the 1d temporal decay mechanism of RetNet (Sun et al. 2023) to spatial domains through explicit spatial decay matrices.

**Data-Dependent Decay in Sequence Modeling.** Recent breakthroughs in natural language processing and sequence modeling demonstrate the superiority of data-dependent state dynamics over static positional information. This paradigm shift is exemplified in the evolution of linear attention mechanisms (Choromanski et al. 2020; Katharopoulos et al. 2020) and state-space models (SSMs) (Gu and Dao 2023). Mamba (Dao and Gu 2024) and related architectures utilize input-dependent state transitions, enabling selective information retention and forgetting based on current token content. This principle is generalized across architectures: Gated Linear Attention (GLA) (Yang et al. 2024) and HGRN2 (Qin et al. 2024b) explicitly incorporate content-aware gating into linear recurrent models, achieving significant performance improvements over data-independent counterparts. Particularly relevant is the Forgetting Transformer (Lin et al. 2025), which integrates learnable forget gates into self-attention, demonstrating that content-based modulation outperforms fixed positional relationships for 1D sequences.

**Uniqueness of Our Solutions.** Building upon these advances, our work pioneers the adaptation of data-dependent mechanisms to 2D spatial domains. While previous approaches focus on temporal decay patterns, we address the unique challenges of 2D topology, including bidirectional spatial dependencies and the need for principled spatial distance metrics that respect image geometry. To our knowledge, we present the first content-aware, dynamic spatial decay mechanism for vision transformers, bridging static spatial priors and dynamic data-dependent mechanisms.

## Method

### Preliminaries

**Vision Transformer Foundation.** Vision Transformers process images by partitioning them into non-overlapping patches, treating each patch as a token in a sequence. Given input features $\mathbf{X} \in \mathbb{R}^{L \times D}$ where $L = H \times W$ represents spatial resolution and $D$ denotes feature dimension, standard self-attention computes:

$$\mathbf{O} = \text{softmax}\left(\mathbf{Q}\mathbf{K}^T/\sqrt{d_k}\right)\mathbf{V} \tag{1}$$

where $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{L \times d_k}$ are query, key, and value projections with linear projection matrices $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v$ respectively. While this mechanism enables global receptive fields, it lacks inherent spatial awareness, treating all patch interactions uniformly regardless of their geometric relationships.

**Data-Dependent Attention Modulation.** Recent advances in large language models have demonstrated the superiority of content-aware gating over fixed positional biases. In

the autoregressive setting, this mechanism applies content-dependent decay where the output at position $i$ accumulates information with learned decay strengths:

$$\mathbf{o}_i = \sum_{j=1}^{i} \alpha_{ij}\mathbf{v}_j, \alpha_{ij} \propto \exp\left(\mathbf{q}_i^T\mathbf{k}_j + \sum_{l=j}^{i-1}\mathbf{g}_l\right) \quad (2)$$

where $\mathbf{g}_l$ represents learned gating values controlling information decay. This recurrent formulation naturally captures the cumulative nature of content-dependent attention modulation.

For parallel computation, this mechanism can be expressed in matrix form. More generally, given input $\mathbf{X}$, we compute content-dependent modulation logits and apply:

$$\mathbf{O} = \text{Softmax}\left(\mathbf{S} + \mathbf{B}_{\text{mod}}\right) \quad (3)$$

where $\mathbf{S} = \mathbf{Q}\mathbf{K}^T/\sqrt{d_k}$ represents the standard attention scores and $\mathbf{B}_{\text{mod}}$ is the content-dependent bias matrix derived from modulation logits. This framework provides a general representation for content-aware attention weighting, where different constructions of $\mathbf{B}_{\text{mod}}$ can encode various attention preferences from temporal dependencies in sequences to spatial relationships in images.

## Context-Aware Gating

**Design Motivation.** Traditional spatial attention mechanisms employ fixed positional encodings (Dosovitskiy et al. 2020; Shaw, Uszkoreit, and Vaswani 2018) or distance-based decay (Fan et al. 2024) that apply uniform spatial biases regardless of image content. This approach fundamentally ignores semantic relationships: semantically related regions should maintain strong attention connections regardless of spatial distance, while irrelevant regions should be suppressed even when spatially adjacent. We propose Context-Aware Gating (CAG) to address this limitation through dynamic, content-dependent spatial attention modulation.

**Content-Dependent Gate Generation.** Given input features $\mathbf{X} \in \mathbb{R}^{B\times H\times W\times D}$, we generate head-specific decay logits through learnable projection:

$$\mathbf{F} = \mathbf{X}\mathbf{W}_g \in \mathbb{R}^{B\times H\times W\times N}, \quad (4)$$

where $\mathbf{W}_g \in \mathbb{R}^{D\times N}$ is a learnable parameter matrix and $N$ is the number of attention heads. This projection enables the model to learn head-specific spatial attention patterns, allowing different heads to focus on different types of spatial relationships (e.g., local texture patterns vs. global object structures).

**Bounded Decay Computation.** The decay logits are transformed into bounded decay strengths through a log-sigmoid activation:

$$\mathbf{G} = \log\sigma(\mathbf{F}) \in \mathbb{R}^{B\times L\times N}, \quad (5)$$

where $L = H \times W$ and spatial dimensions are flattened for computational efficiency. The log-sigmoid transformation ensures decay strengths are bounded in $(-\infty, 0]$, providing stable gradient flow during training.

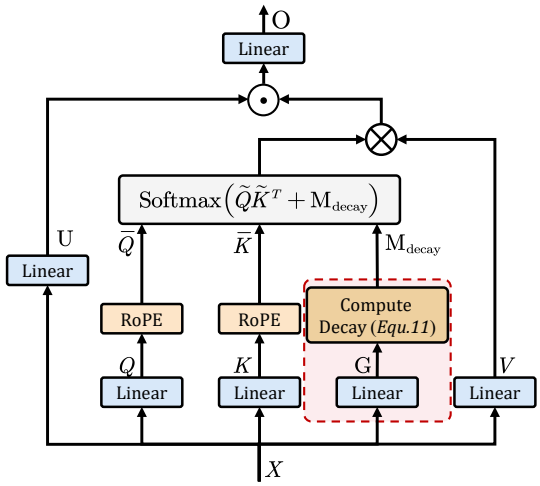


Figure 1: The network structure of the Spatial Decay Layer. The attention weights are modulated by a learned decay map $\mathbf{M}_{\text{decay}}$ computed from $\mathbf{G}$, enabling spatially adaptive attention.

## Spatial Decay Extension: From 1D to 2D

**Theoretical Challenges.** The extension of data-dependent decay from 1D sequential modeling to 2D spatial attention constitutes a non-trivial theoretical challenge. In 1D causal attention (Yang et al. 2024; Qin et al. 2024b), temporal ordering enables efficient cumulative computation:

$$\mathbf{M}_{\text{1D}}[i,j] = \sum_{k=j}^{i-1}\mathbf{g}_k, \quad \text{for } i > j. \quad (6)$$

However, 2D spatial grids lack inherent ordering, presenting three challenges: (1) *bidirectional dependencies* where each position interacts with neighbors in all directions, (2) *non-causal relationships* where spatial proximity lacks temporal precedence, and (3) *topological complexity* requiring sophisticated distance metrics beyond sequential indexing.

**Spatial Distance Formulation.** We establish a mathematical framework for spatial distance computation in discrete 2D grids. Given spatial positions $i$ and $j$ with coordinates $\mathbf{p}_i = (h_i, w_i)$ and $\mathbf{p}_j = (h_j, w_j)$ respectively, we define the manhattan distance metric:

$$d_{\text{M}}(\mathbf{p}_i, \mathbf{p}_j) = |h_i - h_j| + |w_i - w_j|. \quad (7)$$

This choice is theoretically motivated by its natural alignment with discrete grid topology, computational efficiency, and proven effectiveness in spatial modeling tasks. The manhattan distance preserves the grid-aligned structure inherent in vision transformers while providing a principled measure of spatial proximity.

**Content-Dependent Spatial Fusion Framework.** We propose a unified spatial-content fusion mechanism that bridges fixed 2D image geometric priors and adaptive content representations.

For notational clarity in the subsequent formulation, we omit the batch dimension $B$. Thus, we consider the gate for

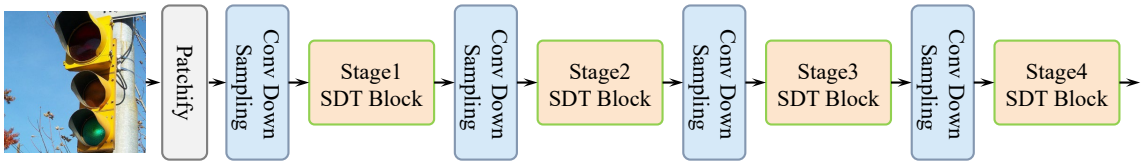


Figure 2: Overall architecture of the proposed Learnable Spatial Decay based Vision Transformer. The model consists of four stages of Spatial Decay Transformer (SDT) Blocks, and each stage consists of several Spatial Decay Layers as shown in Fig. 1.

a single batch item as $\mathbf{G} \in \mathbb{R}^{L\times N}$. The combined decay strength $\mathbf{M}_{\text{combined}}[i,j]$ (which is an $N$-dimensional vector) between any two spatial positions $i,j \in [1,\dots,L]$ is then defined as:

$$\mathbf{M}_{\text{combined}}[i,j] = \frac{1}{2}(\mathbf{G}[i,:] + \mathbf{G}[j,:]) \cdot d_{\text{M}}(\mathbf{p}_i, \mathbf{p}_j) \cdot \alpha, \quad (8)$$

where $\mathbf{G}[i,:] \in \mathbb{R}^N$ and $\mathbf{G}[j,:] \in \mathbb{R}^N$ are the content-dependent gating vectors for spatial positions $i$ and $j$, respectively. This clarifies the ambiguity raised by the reviewer: we are combining the $N$-dimensional gating vectors corresponding to the $i$-th and $j$-th spatial locations. The rationale for using their average, $\frac{1}{2}(\mathbf{G}[i,:] + \mathbf{G}[j,:])$, is to establish a symmetric and balanced measure of mutual influence. This operation ensures that the content-dependent modulation between position $i$ and position $j$ is reciprocal and jointly determined by the content at both locations. This averaged gating value then modulates the fixed scalar distance $d_{\text{M}}(\mathbf{p}_i, \mathbf{p}_j)$ (which is broadcast across all feature dimensions), and $\alpha \in \mathbb{R}^+$ is a scaling factor.

**Final Spatial Decay Mask and Integration.** The complete spatial decay mask is computed as:

$$\mathbf{M}_{\text{decay}}[i,j] = -|\mathbf{M}_{\text{combined}}[i,j]|. \quad (9)$$

The negative absolute value operation serves dual purposes: (1) it maps all values to the non-positive domain, ensuring attention score reduction rather than amplification, and (2) it maintains gradient flow stability by preventing unbounded positive values during backpropagation. Based on Equ. (3), the final attention computation becomes:

$$\mathbf{O} = \text{Softmax}\left(\mathbf{Q}\mathbf{K}^T/\sqrt{d_k} + \mathbf{M}_{\text{decay}}\right)\mathbf{V}. \quad (10)$$

**Efficient Decomposed Implementation.** In hierarchical vision architectures, early stages with high spatial resolution face computational challenges when computing the full $L \times L$ spatial decay mask $\mathbf{M}_{\text{decay}}$, leading to excessive memory consumption. To address this challenge, we propose a decomposed implementation that is selectively applied only to the high-resolution stages.

The decomposed approach separately computes attention scores for horizontal and vertical directions, applying one-dimensional data-dependent decay to each direction:

$$\begin{aligned}
\mathbf{F}_H = \mathbf{X}\mathbf{W}_{g,H}, \mathbf{F}_W = \mathbf{X}\mathbf{W}_{g,W}, \\
\mathbf{M}_{\text{decay},H}[i,j] = -\sum_{k=\min(i_h,j_h)}^{\max(i_h,j_h)-1} |\log\sigma(\mathbf{F}_H[k])|, \\
\mathbf{M}_{\text{decay},W}[i,j] = -\sum_{k=\min(i_w,j_w)}^{\max(i_w,j_w)-1} |\log\sigma(\mathbf{F}_W[k])|.
\end{aligned} \quad (11)$$

The decomposed attention is then computed as:

$$\mathbf{O} = \text{Attn}_H(\text{Attn}_W\mathbf{V})^T, \quad (12)$$

where $\text{Attn}_H$ and $\text{Attn}_W$ apply the respective decay masks. This decomposition reduces complexity from $\mathcal{O}(L^2)$ to $\mathcal{O}(H^2+W^2)$ while maintaining the data-dependent characteristics of our gating mechanism. For the latter two stages with reduced spatial resolution, we employ the full 2D spatial decay mask for optimal performance.

## Overall Architecture

**Spatial Decay Attention Layer.** The Spatial Decay Attention Layer is composed of a Spatial Decay Attention (SDA) and a Feed-Forward Network (FFN). Each SDA block integrates: (1) Multi-head Context-Aware Spatial Gating that generates head-specific decay logits, (2) Rotary Position Embedding (Su et al. 2024) for enhanced positional awareness, (3) Local Position Encoding (Chu et al. 2021) through depthwise convolutions, and (4) A low-rank output gate for comprehensive attention control, as shown in Fig. 1. The architecture seamlessly combines hierarchical feature learning with adaptive, content-aware spatial attention for superior performance across diverse vision tasks.

$$\mathbf{O} = \mathbf{U} \odot \left[\text{Softmax}\left(\frac{\tilde{\mathbf{Q}}\tilde{\mathbf{K}}^T}{\sqrt{d_k}} + \mathbf{M}_{\text{decay}}\right)\mathbf{V} + \text{LPE}(\mathbf{V})\right], \quad (13)$$

where $\tilde{\mathbf{Q}} = R(\mathbf{Q})$ and $\tilde{\mathbf{K}} = R(\mathbf{K})$ are RoPE-enhanced queries and keys. $\mathbf{Q},\mathbf{K},\mathbf{V}$ is projected via linear projection matrices $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v$. $\mathbf{U} = \sigma(\mathbf{X}\mathbf{W}_{u_1})\mathbf{W}_{u_2}$ with $\mathbf{W}_{u_1}, \mathbf{W}_{u_2}$ is low-rank projection matrices for output gating, and $\text{LPE}(\cdot)$ denotes local position encoding via depthwise convolutions.

**Hierarchical Design.** Our Spatial Decay Transformer (SDT) can be implemented in both hierarchical architecture and plain structure configurations. When employing the hierarchical architecture (SDT-H), the model adopts a multi-stage design using the downsampling strategy (Liu et al.

| Model | Parmas (M) | FLOPs (G) | Acc (%) |
|---|---|---|---|
| PVTv2-b1 (Wang et al. 2022a) | 13 | 2.1 | 78.7 |
| QuadTree-B-b1 (Tang et al. 2022) | 14 | 2.3 | 80.0 |
| RegionViT-T (Chen, Panda, and Fan 2022) | 14 | 2.4 | 80.4 |
| MPViT-XS (Lee et al. 2022) | 11 | 2.9 | 80.9 |
| tiny-MOAT-2 (Yang et al. 2023) | 10 | 2.3 | 81.0 |
| VAN-B1 (Guo et al. 2022b) | 14 | 2.5 | 81.1 |
| BiFormer-T (Zhu et al. 2023) | 13 | 2.2 | 81.4 |
| Conv2Former-N (Hou et al. 2024) | 15 | 2.2 | 81.5 |
| CrossFormer-T (Wang et al. 2022b) | 28 | 2.9 | 81.5 |
| NAT-M (Hassani et al. 2023) | 20 | 2.7 | 81.8 |
| QnA-T (Arar, Shamir, and Bermano 2022) | 16 | 2.5 | 82.0 |
| GC-ViT-XT (Hatamizadeh et al. 2023b) | 20 | 2.6 | 82.0 |
| SMT-T (Lin et al. 2023) | 12 | 2.4 | 82.2 |
| RMT-T (Fan et al. 2024) | 14 | 2.5 | 82.4 |
| SDT-H-T (Ours) | 14 | 2.7 | **82.7** |
| DeiT-S (Touvron et al. 2021) | 22 | 4.6 | 79.9 |
| Swin-T (Liu et al. 2021) | 29 | 4.5 | 81.3 |
| ConvNeXt-T (Liu et al. 2022) | 29 | 4.5 | 82.1 |
| Focal-T (Yang et al. 2021) | 29 | 4.9 | 82.2 |
| FocalNet-T (Yang et al. 2022a) | 29 | 4.5 | 82.3 |
| RegionViT-S (Chen, Panda, and Fan 2022) | 31 | 5.3 | 82.6 |
| CSWin-T (Dong et al. 2022) | 23 | 4.3 | 82.7 |
| MPViT-S (Lee et al. 2022) | 23 | 4.7 | 83.0 |
| ScalableViT-S (Yang et al. 2022b) | 32 | 4.2 | 83.1 |
| MOAT-0 (Yang et al. 2023) | 28 | 5.7 | 83.3 |
| Ortho-S (Huang, Zhou, and He 2022) | 24 | 4.5 | 83.4 |
| CMT-S (Guo et al. 2022a) | 25 | 4.0 | 83.5 |
| MaxViT-T (Tu et al. 2022) | 31 | 5.6 | 83.6 |
| SMT-S (Lin et al. 2023) | 20 | 4.8 | 83.7 |
| BiFormer-S (Zhu et al. 2023) | 26 | 4.5 | 83.8 |
| RMT-S (Fan et al. 2024) | 27 | 4.5 | 84.1 |
| SDT-H-S (Ours) | 27 | 4.8 | **84.2** |
| Swin-S (Liu et al. 2021) | 50 | 8.7 | 83.0 |
| ConvNeXt-S (Liu et al. 2022) | 50 | 8.7 | 83.1 |
| CrossFormer-B (Wang et al. 2022b) | 52 | 9.2 | 83.4 |
| NAT-S (Hassani et al. 2023) | 51 | 7.8 | 83.7 |
| Quadtree-B-b4 (Tang et al. 2022) | 64 | 11.5 | 84.0 |
| Ortho-B (Huang, Zhou, and He 2022) | 50 | 8.6 | 84.0 |
| ScaleViT-B (Yang et al. 2022b) | 81 | 8.6 | 84.1 |
| MOAT-1 (Yang et al. 2023) | 42 | 9.1 | 84.2 |
| InternImage-S (Wang et al. 2023) | 50 | 8.0 | 84.2 |
| DaViT-S (Ding et al. 2022) | 50 | 8.8 | 84.2 |
| GC-ViT-S (Hatamizadeh et al. 2023b) | 51 | 8.5 | 84.3 |
| BiFormer-B (Zhu et al. 2023) | 57 | 9.8 | 84.3 |
| MViTv2-B (Li et al. 2022b) | 52 | 10.2 | 84.4 |
| iFormer-B (Si et al. 2022) | 48 | 9.4 | 84.6 |
| RMT-B (Fan et al. 2024) | 54 | 9.7 | 85.0 |
| SDT-H-B (Ours) | 54 | 10.8 | **85.1** |

Table 1: Performance comparison for image classification task on ImageNet-1K.

2021), with structural design aligned with that of RMT. As shown in Fig. 2, SDT-H comprises four stages, each progressively reducing spatial resolution while increasing feature dimension. This hierarchical design enables effective multi-scale feature learning, which is crucial for vision tasks. We employ a decomposed implementation in the first two stages and a global implementation in the final two stages. Alternatively, SDT can be configured as a plain structure (SDT-P) without hierarchical downsampling for scenarios requiring consistent spatial resolution throughout the network.

## Experiments

We evaluate the performance and scalability of our proposed Spatial Decay Transformer (SDT) as a substitute for existing models on image classification (using SDT-H) and image generation (using SDT-P) tasks.

### Settings

**Image Classification.** We train all models from scratch on the ImageNet-1K dataset (Krizhevsky, Sutskever, and Hinton 2012). For fair comparison, we adopt the same training protocol as in (Fan et al. 2024), relying solely on classification loss as supervision. The stochastic depth rates for different size Tiny/Small/Base are set to 0.1, 0.15, and 0.4, respectively. We employ the AdamW (Loshchilov and Hutter 2017) optimizer in conjunction with a cosine learning rate schedule. The initial learning rate is set to 0.001, with a weight decay of 0.05 and a batch size of 1024. For data augmentation and regularization, we follow the strong strategy used in (Fan et al. 2024), including RandAugment (Cubuk et al. 2020), Mixup (Zhang et al. 2018), CutMix (Yun et al. 2019), and Random Erasing (Zhong et al. 2020).

**Image Generation.** We build our model upon the Diffusion Transformer (DiT) (Peebles and Xie 2023), employing our proposed SDT as the denoising network (Note that we are using a plain structure instead of a hierarchical structure). We scale the model across various configurations (S, B, L, XL) with a fixed patch size of 2, consistent with DiT. We conduct experiments on the ImageNet dataset (Krizhevsky, Sutskever, and Hinton 2012) at $256 \times 256$ resolution. We compare performance against representative image generation methods, including ADM (Dhariwal and Nichol 2021), CDM (Ho et al. 2022), LDM (Rombach et al. 2022), and DiT (Peebles and Xie 2023). All models are trained for 400K steps with a batch size of 256 to evaluate scaling capabilities. For the largest model variant, we extend training to 0.8M steps with a batch size of 512 (compared to 7M steps in DiT) to enhance generative performance.

### Main Results

**Image Classification.** We evaluate our models against state-of-the-art vision transformers on the ImageNet-1K validation set. As shown in Table 1, our approach demonstrates competitive performance across various model scales.

Our method consistently outperforms the data-independent spatial decay method, RMT. This comparison provides strong evidence for the effectiveness of

| Model | FID↓ | sFID↓ | IS↑ | Pre.↑ | Rec.↑ | Params |
|---|---|---|---|---|---|---|
| ADM | 10.94 | 6.02 | 100.98 | 0.69 | 0.63 | - |
| ADM-U | 7.49 | 5.13 | 127.49 | 0.72 | 0.63 | - |
| ADM-G | 4.59 | 5.25 | 186.70 | 0.82 | 0.52 | - |
| CDM | 4.88 | - | 158.71 | - | - | - |
| LDM-8 | 15.51 | - | 79.03 | 0.65 | 0.63 | 395M |
| LDM-8-G | 7.76 | - | 209.52 | 0.84 | 0.35 | 506M |
| LDM-4 | 10.56 | - | 103.49 | 0.71 | 0.62 | 400M |
| LDM-4-G | 3.60 | - | 247.67 | **0.87** | 0.48 | 400M |
| DiT-S | 68.40 | - | - | - | - | 32M |
| DiT-B | 43.47 | - | - | - | - | 130M |
| DiT-L | 23.33 | - | - | - | - | 459M |
| DiT-XL | 19.47 | - | - | - | - | 675M |
| DiT-XL | 9.62 | 6.85 | 121.50 | 0.67 | 0.67 | 675M |
| DiT-XL-G | 2.27 | 4.60 | 278.24 | 0.83 | 0.57 | 675M |
| SDT-P-S | 60.70 | 10.42 | 23.15 | 0.40 | 0.60 | 33M |
| SDT-P-B | 37.47 | 6.82 | 60.49 | 0.53 | 0.60 | 132M |
| SDT-P-L | 21.28 | 5.99 | 63.79 | 0.62 | 0.64 | 462M |
| SDT-P-XL | 18.20 | 5.74 | 71.82 | 0.64 | 0.63 | 679M |
| SDT-P-XL | 6.82 | 6.21 | 158.34 | 0.71 | 0.64 | 679M |
| SDT-P-XL-G | **2.25** | **4.59** | **279.85** | 0.83 | **0.58** | 679M |

Table 2: Performance comparison for image generation task on ImageNet-1K. SDT-P-XL achieves state-of-the-art FID with or without classifier-free guidance (-G). "Pre." and "Rec." represents the "Precision" and "Recall". The best result is highlighted with **bold**.

data-dependent, context-aware gating over static, content-agnostic decay mechanisms. Additionally, our models achieve competitive results with other leading architectures. Notably, SDT-S surpasses established models including ConvNeXt-S, BiFormer-S, and NAT-S, demonstrating strong performance within its complexity class. The consistent improvements across different scales validate the robustness and scalability of our spatial decay mechanism.

Fig. 3 presents training dynamics comparing our SDT-H-T with RMT-T. The results show that our model achieves faster convergence and superior final accuracy, further supporting the effectiveness of the proposed context-aware spatial decay approach.

**Image Generation.** We evaluate this variant, termed SDT-P, on class-conditional image generation using ImageNet at $256 \times 256$ resolution, as shown in Table 2. We can observe that SDT-P-XL-G achieves competitive performance with a FID of 2.25, marginally outperforming DiT-XL-G (2.27 FID) while maintaining comparable parameter count. These results suggest that the enhanced spatial reasoning capabilities of our model translate effectively to generative tasks. Specifically, the ability to dynamically attend to relevant contextual information appears beneficial for image generation. The improved performance indicates that our Context-Aware Gating (CAG) mechanism provides beneficial guidance during the denoising process, contributing to more coherent output generation.

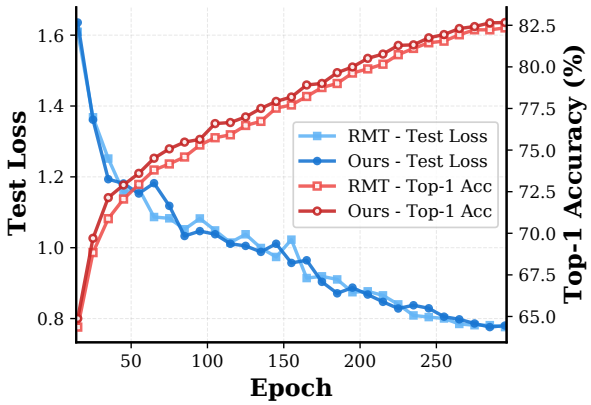


Figure 3: Training loss comparison between our proposed SDT-H-T and RMT-T. The blue curve represents our SDT-H-T, and the orange curve represents RMT-T.

We further analyze the scaling properties of our approach across four model sizes on ImageNet. As illustrated in Figure 4, performance consistently improves from S to XL scale, demonstrating favorable scaling characteristics. This trend suggests the potential applicability of our method to larger-scale diffusion models, where spatial attention mechanisms become increasingly important for maintaining generation quality at higher resolutions.

## Ablation Studies

To dissect our model and validate the contributions of its key components, we conduct a series of ablation studies. We extend our analysis beyond image classification and image generation to demonstrate the robustness and general applicability of our design principles. All experiments use the Tiny (Ours-T) configuration for classification (reporting Top-1 Acc) and a comparable Base (Ours-B) model for generation (reporting FID).

**Effectiveness of the Context-Aware Gating mechanism.** We propose a Context-Aware Gating mechanism that enables data-dependent decay computation to learn the importance of correlations between image patches. Unlike traditional fixed decay patterns, this mechanism adaptively determines decay values based on the actual content and context of the input data, providing more flexible and effective attention modeling. To thoroughly evaluate the effectiveness of our Context-Aware Gating mechanism, we conduct ablation studies with the following variants: (1) No Decay: We remove all decay mechanisms and employ only the original attention mechanism without any positional or content-based decay modulation. This serves as our baseline to demonstrate the necessity of decay mechanisms. (2) Spatial Decay: We implement a data-independent decay approach that computes decay values solely based on the relative positional relationships between image patches. This variant captures spatial proximity but ignores content-dependent correlations. (3) Context-Aware Gating (Ours): Our proposed data-dependent decay mechanism that learns to adaptively weight patch correlations based on both spatial relationships and content similarity.

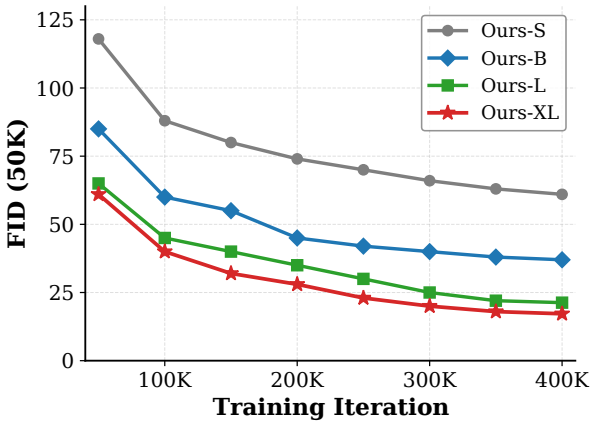


Figure 4: Scaling up the SDT-P enhances the FID during every iterations of training. We present the FID-50K across training iterations for four SDT-P models. Enhancing the SDT-P backbone results in improved generative models for all sizes of models.

| Decay Mechanism | Acc (%)↑ | FID↓ |
|---|---|---|
| No decay | 81.9 | 43.22 |
| Spatial Decay | 82.4 | 40.29 |
| **Ours** | **82.7** | **37.47** |

Table 3: Ablation study on Context-Aware Gating variants on image classification (Acc) and generation (FID).

The quantitative comparison is presented in Table 3. The results demonstrate the superiority of our Context-Aware Gating mechanism. The baseline without decay shows limited performance, highlighting the importance of incorporating decay mechanisms. The Spatial Decay variant provides improvements by considering positional relationships, but its data-independent nature limits its adaptability. In contrast, our Context-Aware Gating achieves the best performance by dynamically adjusting decay patterns based on input content, effectively capturing both spatial and semantic correlations between patches.

**Effectiveness of the Content-Dependent Spatial Fusion.** In extending the 1D decay mechanism to 2D scenarios, we propose a Content-Dependent Spatial Fusion Framework (CDSF) that integrates spatial relationships with learnable decay patterns. This framework combines the manhattan distance between image patches with computed learnable decay values for spatially-aware attention. To validate the effectiveness of our spatial fusion mechanism, we design several variants: (1) 1D Decay: We treat the 2D image as a flattened 1D sequence and directly apply the decay computation from Equ. (6). (2) Bidirectional Decay: Based on 1D decay, we adopt a bidirectional scanning strategy inspired by Vision-Mamba (Zhu et al. 2024), averaging the forward and backward decay computations. (3) Decomposed 2D Decay: We decompose the image along height and width dimensions, computing decay values independently for each dimension.

| Decay Mechanism | Acc (%)↑ | FID↓ |
|---|---|---|
| 1D Decay | 82.2 | 42.25 |
| Bidirectional Decay | 82.3 | 41.32 |
| Decomposed 2D Decay | 82.5 | 39.11 |
| **Ours** | **82.7** | **37.47** |

Table 4: Ablation study on Content-Dependent Spatial Fusion variants on image classification (Acc) and generation (FID).

| Decay Mechanism | Acc (%)↑ | FID↓ |
|---|---|---|
| $\alpha = 0.05$ | 82.5 | 40.57 |
| $\alpha = \mathbf{0.1}$ | **82.7** | **37.47** |
| $\alpha = 0.15$ | 82.6 | 39.45 |
| $\alpha = 0.2$ | 82.4 | 39.79 |

Table 5: **Ablation study on the $\alpha$ factor** in Equ. (8) on image classification (Acc) and generation (FID).

(4) Ours Content-Dependent Spatial Fusion: Our proposed framework that integrates both spatial distance and content-dependent decay computation.

As shown in Table 4, our proposed CDSF consistently outperforms all variants. The 1D decay shows limited performance due to ignoring 2D spatial structure. Bidirectional Decay provides modest improvements through bidirectional processing. The decomposed approach captures dimension-specific patterns but lacks cross-dimensional interaction. In contrast, our CDSF achieves superior performance by effectively combining spatial proximity with content-dependent decay, validating our design.

**Impact of $\alpha$.** As shown in Equ. 8, alpha controls the strength of decay, larger values indicate stronger decay. We conduct experiments to verify the impact of alpha. The experimental results, shown in Table (5), show that an alpha of 0.1 achieves optimal decay strength.

## Conclusion

We have introduced Spatial Decay Transformer, which successfully adapts data-dependent decay mechanisms from language models to 2D spatial attention in vision transformers. Our Context-Aware Gating (CAG) mechanism generates dynamic, content-dependent spatial decay that adapts to image content, overcoming the rigidity of fixed distance-based approaches like RMT. Through a unified spatial-content fusion framework, we address the fundamental challenge of extending 1D sequential decay to 2D spatial domains. Extensive experiments on ImageNet-1K classification and generation demonstrate consistent improvements over strong baselines, validating that content-aware spatial gating is crucial for optimal 2D attention. Our work establishes data-dependent spatial decay as a new paradigm for vision transformers, opening promising directions for other spatial-structured vision tasks or spatial-temporal vision tasks.

## Acknowledgments

This research was supported in part by the National Natural Science Foundation of China (62271410, 12150007). Yuxin Mao is sponsored by the Innovation Foundation for Doctor Dissertation of NWPU (CX2024014).

## References

Arar, M.; Shamir, A.; and Bermano, A. H. 2022. Learned Queries for Efficient Local Attention. In *CVPR*.

Chen, A.; Li, A.; Gong, B.; Jiang, B.; Fei, B.; Yang, B.; Shan, B.; Yu, C.; Wang, C.; Zhu, C.; et al. 2025. MiniMax-M1: Scaling Test-Time Compute Efficiently with Lightning Attention. *arXiv preprint arXiv:2506.13585*.

Chen, C.-F. R.; Panda, R.; and Fan, Q. 2022. RegionViT: Regional-to-Local Attention for Vision Transformers. In *ICLR*.

Choromanski, K. M.; Likhosherstov, V.; Dohan, D.; Song, X.; Gane, A.; Sarlos, T.; Hawkins, P.; Davis, J. Q.; Mohiuddin, A.; Kaiser, L.; et al. 2020. Rethinking Attention with Performers. In *ICLR*.

Chu, X.; Tian, Z.; Zhang, B.; Wang, X.; and Shen, C. 2021. Conditional Positional Encodings for Vision Transformers. In *ICLR*.

Cubuk, E. D.; Zoph, B.; Shlens, J.; and Le, Q. V. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPRW*, 702–703.

Dai, Z.; Liu, H.; Le, Q. V.; and Tan, M. 2021. Coatnet: Marrying convolution and attention for all data sizes. *NeurIPS*, 34: 3965–3977.

Dao, T.; and Gu, A. 2024. Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality. In *ICLR*.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *NeurIPS*, 34: 8780–8794.

Ding, M.; Xiao, B.; Codella, N.; et al. 2022. DaViT: Dual Attention Vision Transformers. In *ECCV*.

Dong, X.; Bao, J.; Chen, D.; et al. 2022. CSWin Transformer: A General Vision Transformer Backbone with Cross-Shaped Windows. In *CVPR*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.

Fan, Q.; Huang, H.; Chen, M.; Liu, H.; and He, R. 2024. Rmt: Retentive networks meet vision transformers. In *CVPR*, 5641–5651.

Gu, A.; and Dao, T. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Guo, J.; Han, K.; Wu, H.; Xu, C.; Tang, Y.; Xu, C.; and Wang, Y. 2022a. CMT: Convolutional neural networks meet vision transformers. In *CVPR*.

Guo, M.-H.; Lu, C.-Z.; Liu, Z.-N.; Cheng, M.-M.; and Hu, S.-M. 2022b. Visual Attention Network. *arXiv preprint arXiv:2202.09741*.

Hassani, A.; Walton, S.; Li, J.; Li, S.; and Shi, H. 2023. Neighborhood Attention Transformer. In *CVPR*.

Hatamizadeh, A.; Heinrich, G.; Yin, H.; Tao, A.; Alvarez, J. M.; Kautz, J.; and Molchanov, P. 2023a. FasterViT: Fast Vision Transformers with Hierarchical Attention. In *ICLR*.

Hatamizadeh, A.; Yin, H.; Heinrich, G.; Kautz, J.; and Molchanov, P. 2023b. Global context vision transformers. In *ICML*.

Ho, J.; Saharia, C.; Chan, W.; Fleet, D. J.; Norouzi, M.; and Salimans, T. 2022. Cascaded diffusion models for high fidelity image generation. *JMLR*, 23(47): 1–33.

Hou, Q.; Lu, C.-Z.; Cheng, M.-M.; and Feng, J. 2024. Conv2former: A simple transformer-style convnet for visual recognition. *IEEE TPAMI*, 46(12): 8274–8283.

Huang, H.; Zhou, X.; and He, R. 2022. Orthogonal Transformer: An Efficient Vision Transformer Backbone with Token Orthogonalization. In *NeurIPS*.

Katharopoulos, A.; Vyas, A.; Pappas, N.; and Fleuret, F. 2020. Transformers are rnns: Fast autoregressive transformers with linear attention. In *ICML*, 5156–5165. PMLR.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *NeurIPS*, 25.

Lee, Y.; Kim, J.; Willette, J.; and Hwang, S. J. 2022. MPViT: Multi-Path Vision Transformer for Dense Prediction. In *CVPR*.

Li, A.; Gong, B.; Yang, B.; Shan, B.; Liu, C.; Zhu, C.; Zhang, C.; Guo, C.; Chen, D.; Li, D.; et al. 2025. Minimax-01: Scaling foundation models with lightning attention. *arXiv preprint arXiv:2501.08313*.

Li, Y.; Wu, C.-Y.; Fan, H.; Mangalam, K.; Xiong, B.; Malik, J.; and Feichtenhofer, C. 2022a. Mvitv2: Improved multiscale vision transformers for classification and detection. In *CVPR*, 4804–4814.

Li, Y.; Wu, C.-Y.; Fan, H.; Mangalam, K.; Xiong, B.; Malik, J.; and Feichtenhofer, C. 2022b. MViTv2: Improved multiscale vision transformers for classification and detection. In *CVPR*.

Lin, W.; Wu, Z.; Chen, J.; Huang, J.; and Jin, L. 2023. Scale-Aware Modulation Meet Transformer. In *ICCV*.

Lin, Z.; Nikishin, E.; He, X.; and Courville, A. 2025. Forgetting Transformer: Softmax Attention with a Forget Gate. In *ICLR*.

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 *ICCV*.

Liu, Z.; Mao, H.; Wu, C.-Y.; et al. 2022. A convnet for the 2020s. In *CVPR*.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Mao, Y.; Qin, Z.; Zhou, J.; Deng, H.; Shen, X.; Fan, B.; Zhang, J.; Zhong, Y.; and Dai, Y. 2025. Autoregressive Image Generation with Linear Complexity: A Spatial-Aware Decay Perspective. *arXiv preprint arXiv:2507.01652*.

Mao, Y.; Shen, X.; Zhang, J.; Qin, Z.; Zhou, J.; Xiang, M.; Zhong, Y.; and Dai, Y. 2024a. Tavgbench: Benchmarking text to audible-video generation. In *ACM MM*, 6607–6616.

Mao, Y.; Zhang, J.; Wan, Z.; Tian, X.; Li, A.; Lv, Y.; and Dai, Y. 2024b. Generative transformer for accurate and reliable salient object detection. *IEEE TCSVT*.

Mao, Y.; Zhang, J.; Xiang, M.; Zhong, Y.; and Dai, Y. 2023. Multimodal variational auto-encoder based audio-visual segmentation. In *ICCV*, 954–965.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *ICCV*, 4195–4205.

Qin, Z.; Mao, Y.; Shen, X.; Li, D.; Zhang, J.; Dai, Y.; and Zhong, Y. 2024a. You only scan once: Efficient multi-dimension sequential modeling with lightnet. *arXiv preprint arXiv:2405.21022*.

Qin, Z.; Yang, S.; Sun, W.; Shen, X.; Li, D.; Sun, W.; and Zhong, Y. 2024b. HGRN2: Gated Linear RNNs with State Expansion. In *First Conference on Language Modeling*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*, 10684–10695.

Shaw, P.; Uszkoreit, J.; and Vaswani, A. 2018. Self-Attention with Relative Position Representations. In *NAACL*, 464–468.

Si, C.; Yu, W.; Zhou, P.; Zhou, Y.; Wang, X.; and YAN, S. 2022. Inception Transformer. In *NeurIPS*.

Srinivas, A.; Lin, T.-Y.; Parmar, N.; Shlens, J.; Abbeel, P.; and Vaswani, A. 2021. Bottleneck transformers for visual recognition. In *CVPR*, 16519–16529.

Su, J.; Ahmed, M.; Lu, Y.; Pan, S.; Bo, W.; and Liu, Y. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568: 127063.

Sun, Y.; Dong, L.; Huang, S.; Ma, S.; Xia, Y.; Xue, J.; Wang, J.; and Wei, F. 2023. Retentive network: A successor to transformer for large language models. *arXiv preprint arXiv:2307.08621*.

Tang, S.; Zhang, J.; Zhu, S.; et al. 2022. Quadtree attention for vision transformers. In *ICLR*.

Touvron, H.; Cord, M.; Douze, M.; et al. 2021. Training data-efficient image transformers & distillation through attention. In *ICML*.

Tu, Z.; Talebi, H.; Zhang, H.; Yang, F.; Milanfar, P.; Bovik, A.; and Li, Y. 2022. MaxViT: Multi-Axis Vision Transformer. In *ECCV*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *NeurIPS*, 30.

Wang, W.; Dai, J.; Chen, Z.; Huang, Z.; Li, Z.; Zhu, X.; Hu, X.; Lu, T.; Lu, L.; Li, H.; et al. 2023. InternImage: Exploring Large-Scale Vision Foundation Models with Deformable Convolutions. In *CVPR*.

Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2022a. Pvtv2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3): 1–10.

Wang, W.; Yao, L.; Chen, L.; Lin, B.; Cai, D.; He, X.; and Liu, W. 2022b. CrossFormer: A Versatile Vision Transformer Hinging on Cross-scale Attention. In *ICLR*.

Yang, C.; Qiao, S.; Yu, Q.; et al. 2023. MOAT: Alternating mobile convolution and attention brings strong vision models. In *ICLR*.

Yang, J.; Li, C.; Dai, X.; and Gao, J. 2022a. Focal Modulation Networks. In *NeurIPS*.

Yang, J.; Li, C.; Zhang, P.; Dai, X.; Xiao, B.; Yuan, L.; and Gao, J. 2021. Focal Self-Attention for Local-Global Interactions in Vision Transformers. In *NeurIPS*.

Yang, R.; Ma, H.; Wu, J.; Tang, Y.; Xiao, X.; Zheng, M.; and Li, X. 2022b. ScalableViT: Rethinking the context-oriented generalization of vision transformer. In *ECCV*.

Yang, S.; Wang, B.; Shen, Y.; Panda, R.; and Kim, Y. 2024. Gated Linear Attention Transformers with Hardware-Efficient Training. In *ICLR*.

Yun, S.; Han, D.; Oh, S. J.; Chun, S.; Choe, J.; and Yoo, Y. 2019. CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features. In *ICCV*.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018. mixup: Beyond Empirical Risk Minimization. In *ICLR*.

Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; and Yang, Y. 2020. Random erasing data augmentation. In *AAAI*, 13001–13008.

Zhou, J.; Guo, D.; Guo, R.; Mao, Y.; Hu, J.; Zhong, Y.; Chang, X.; and Wang, M. 2025a. Towards open-vocabulary audio-visual event localization. In *CVPR*, 8362–8371.

Zhou, J.; Guo, D.; Mao, Y.; Zhong, Y.; Chang, X.; and Wang, M. 2024a. Label-anticipated event disentanglement for audio-visual video parsing. In *ECCV*, 35–51. Springer.

Zhou, J.; Guo, D.; Zhong, Y.; and Wang, M. 2024b. Advancing weakly-supervised audio-visual video parsing via segment-wise Pseudo Labeling. *IJCV*, 1–22.

Zhou, J.; Li, Z.; Yu, Y.; Zhou, Y.; Guo, R.; Li, G.; Mao, Y.; Han, M.; Chang, X.; and Wang, M. 2025b. Mettle: Meta-Token Learning for Memory-Efficient Audio-Visual Adaptation. *arXiv preprint arXiv:2506.23271*.

Zhou, J.; Shen, X.; Wang, J.; Zhang, J.; Sun, W.; Zhang, J.; Birchfield, S.; Guo, D.; Kong, L.; Wang, M.; and Zhong, Y. 2024c. Audio-visual segmentation with semantics. *IJCV*.

Zhou, J.; Wang, J.; Zhang, J.; Sun, W.; Zhang, J.; Birchfield, S.; Guo, D.; Kong, L.; Wang, M.; and Zhong, Y. 2022. Audio–visual segmentation. In *ECCV*, 386–403.

Zhou, J.; Zhou, Z.; Zhou, Y.; Mao, Y.; Duan, Z.; and Guo, D. 2025c. Clasp: Cross-modal salient anchor-based semantic propagation for weakly-supervised dense audio-visual event localization. *arXiv preprint arXiv:2508.04566*.

Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; and Wang, X. 2024. Vision mamba: efficient visual representation learning with bidirectional state space model. In *ICML*, 62429–62442.

Zhu, L.; Wang, X.; Ke, Z.; Zhang, W.; and Lau, R. 2023. BiFormer: Vision Transformer with Bi-Level Routing Attention. In *CVPR*.