ON COMPLEMENTARITY IN DIFFUSION TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion Transformers (DiTs) have achieved remarkable success in high-quality image generation by stacking multiple homogeneous self-attention layers. However, this homogeneous design may lead to functional redundancy, as different layers and attention heads can learn highly similar transformations. In this work, we investigate complementarity as a structural principle for reducing redundancy in DiTs, encouraging different network components to perform distinct yet cooperative computations. Specifically, we address three key questions: (1) what forms of complementarity can be introduced, (2) where complementarity can be instantiated within a transformer, and (3) how different structural designs affect its effectiveness. First, we explore two representative forms of functional complementarity: rotation complementarity, which introduces distinct rotation patterns across layers, and scale complementarity, which employs different convolutional receptive fields. Second, we implement complementarity from layers to individual attention heads, demonstrating its applicability at different architectural granularities. Third, we systematically examine different symmetry, grouping, and ordering designs, showing that the benefits of complementarity persist across diverse configurations. Experiments demonstrate that incorporating complementarity maintains comparable image generation quality while improving network efficiency. We hope this study can serve as a starting point for exploring underexamined structural properties of DiTs and inspire the development of more efficient and diverse architectures.

1 INTRODUCTION

In recent years, Diffusion Transformers (DiTs)(Peebles & Xie, 2023; Ma et al., 2024a) have emerged as a powerful architecture for high-quality image generation(Rombach et al., 2022; Podell et al., 2023; Yu et al., 2024; Li & He, 2026). By stacking multiple homogeneous self-attention layers, DiT models progressively refine latent representations(Kingma & Welling, 2013) through repeated global context aggregation. This design facilitates iterative denoising(Ho et al., 2020) and feature integration across spatial regions, ledding to impressive performance in large-scale generative modeling(Yao et al., 2025; Chen et al., 2023; Esser et al., 2024; Gao et al., 2023).

However, such uniform layer functions may introduce redundancy, as different layers and attention heads often perform highly similar computations, leading to inefficient use of model capacity (Zhou et al., 2021; Ma et al., 2024b). Specifically, as shown in Fig. 1, the attention functions cross different layers and heads often learn highly similar patterns, which can be inefficient and computationally expensive.

Beyond the homogeneous design, existing DiT architectures suggest that different network components can contribute distinct yet cooperative functions to image generation. PixArt-Σ (Chen et al., 2024) combines attention operations at different spatial granularities through selective K/V compression, while U-DiT (Tian et al., 2024) integrates representations across a hierarchy of spatial resolutions. From a structural perspective, these scale-varying designs can be understood as implicit instances of complementarity: different components process different aspects of the input, and their contributions are integrated within the network. This perspective connects existing architectural choices to a broader question: what forms can complementarity take in DiTs, and how does its realization depend on the organization of network components?

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

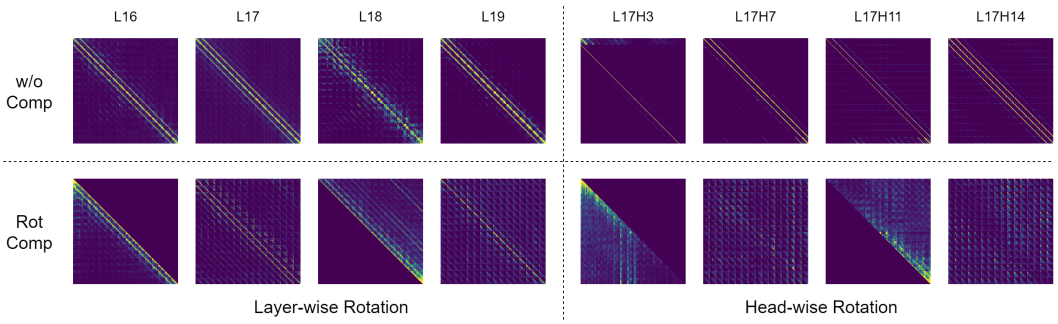

Figure 1: **Visualization of attention patterns in LightningDiT with and without our complementary design.** In the vanilla model, homogeneous global attention across layers and heads tends to produce highly similar attention patterns (top row), indicating functional redundancy. In contrast, our complementary design yields more diverse attention behaviors. A PCA-based diversity analysis can be found in Appendix Section C.2.

In this work, we investigate structural complementarity in DiTs by addressing three core research questions: **RQ1: What** forms of functional complementarity can be introduced into self-attention layers? (Section 2); **RQ2: Where** can complementarity be instantiated within a DiT architecture? (Section 3); and **RQ3: How** do different structural designs affect the effectiveness of complementarity? (Section 4).

To answer the first question, we explore two forms of functional complementarity: rotation complementarity and scale complementarity. Existing DiT models for image generation rely on dense attention instead of causal attention, since diffusion models process fully observed noisy images and require global context rather than sequential token prediction. In rotation complementarity, we aim to leverage the efficiency of causal attention while introducing global context through complementary interactions. Specifically, different causal attention layers operate with distinct rotation patterns, such as $[0°, 90°, 180°, 270°]$, to enhance spatial complementarity across layers and ensure effective global information aggregation. Besides, in scale complementarity, we assign different layers to operate at varying receptive-field scales, enabling the model to capture both fine-grained local details and broader structural patterns. Specifically, we adopt established K/V downsampling operators that inject convolutional layers with different strides before the attention layers to modulate the spatial resolution of token representations, thereby constructing multi-scale features and encouraging diverse contextual modeling.

To answer the second question, we further investigate which components of DiTs can benefit from complementary design and observe that complementarity can be realized not only across layers but also at the attention head level. Specifically, we also assign distinct rotation patterns to different attention heads within the same layer, enabling each head to capture complementary directional dependencies. Such head-wise diversification promotes functional specialization, increases representational diversity, and facilitates more effective global context aggregation without introducing additional computational overhead.

Our experiments demonstrate that incorporating these complementary strategies results in DiT models that maintain comparable or slightly improved image generation performance relative to a homogeneous DiT baseline (Yao et al., 2025), while significantly enhancing network efficiency. These findings highlight complementarity as a powerful yet underexplored design principle for improving the architectural efficiency of DiTs. Furthermore, we conduct extensive experiments under different structural design choices, including symmetry configurations, grouping strategies, and ordering schemes, to investigate how complementarity can be effectively induced in practice. Specifically, we compare symmetric versus asymmetric rotation distributions across layers, structured versus random grouping of rotation patterns, and fixed versus random ordering schemes (both inter-group and intra-group). Experimentally, we find that the rotation complementarity is robust for most designs, and scale complementarity requires suitable inter-group orders. These findings provide practical insights into designing complementary DiT architectures with improved efficiency and robustness.

The main contributions of this work are as follows:

- We introduce complementarity as a structural principle for DiTs to mitigate redundancy arising from homogeneous layer functions.
- We design architectures based on rotation complementarity and scale complementarity. The resulting models achieve comparable image generation performance while enhancing computational efficiency.
- We systematically analyze design choices for inducing complementarity, including symmetry, grouping, and ordering strategies, providing practical guidance for constructing functionally diverse yet cooperative DiTs.

2 What to Complement: Identifying Complementary Functions

We begin our methodology by addressing the first research question: *"Which functions can be designed to be complementary?"* We first provide a conceptual definition of complementarity, clarifying the distinction between homogeneous and complementary functional designs, and then outlining three conditions that a complementary function set should satisfy. Building on this foundation, we introduce rotation complementarity and scale complementarity, detailing their motivations, functional mechanisms, and justifications.

2.1 Conceptual Definition of Complementarity

Homogeneous Baseline. Standard DiTs follow a homogeneous design paradigm, which we formulate as follows. Given a collection of N computational units (e.g., blocks or attention heads) $\mathcal{H}_{homo} = \{\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_N\}$, in this paradigm, we instantiate each of them parameterized by a universal function ψ_{base} and learnable parameters θ_i: $\mathcal{F}_i = f(\cdot; \psi_{base}, \theta_i)$, for all $i \in \{1, \ldots, N\}$. While the parameters θ_i vary across units, the structural invariance of ψ_{base} constrains the model to a singular inductive bias, leading to functional redundancy in the latent representation space.

Complementary Functions. In contrast, we propose to replace this uniformity with a set of heterogeneous functions Ψ, where each $\psi \in \Psi$ is designed with a distinct inductive bias. A complementary architecture $\mathcal{H}_{comp} = \{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_N\}$ is then constructed by selecting and parameterizing these functions:

$$\mathcal{G}_i = f(\cdot; \psi_i, \theta_i), \quad \psi_i \in \Psi, \tag{1}$$

where $\psi : \{1, \ldots, N\} \to \Psi$ is an assignment function, and $\psi_i := \psi(i)$ denotes the function type selected for the i-th unit. It indicates that each unit $\mathcal{G}_i$ is specialized by a complementary function $\psi_i \in \Psi$ and learnable parameters θ_i. In the following sections, unless otherwise specified, we use $f(\psi)$ to denote a unit instantiated with function type ψ for simplicity.

To ensure that $\mathcal{H}_{comp}$ serves as an efficient and effective alternative to the baseline, it must satisfy three core conditions. Here we introduce a cost measurement function $\text{Cost}(\cdot)$ and a general aggregation operator $\bigoplus$ to formally introduce these conditions: **(i) Functional diversity.** For any $\psi_i \neq \psi_j \in \Psi$, the units $f(\psi_i)$ and $f(\psi_j)$ must exhibit different inductive biases (e.g., distinct directional sensitivities or receptive field scales), ensuring that they capture diverse feature subspaces with minimal redundancy. **(ii) Additive approximation.** Let $\bigoplus$ denote the aggregation mechanism, such as functional composition ($\mathcal{G}_N \circ \cdots \circ \mathcal{G}_1$) for sequential layers or concatenation ($[\mathcal{G}_1; \ldots; \mathcal{G}_N]$) for parallel attention heads. While an individual function $f(\psi)$ may have a more specialized (narrower) focus than $f(\psi_{base})$, the collective representation of the sequence must approximate the original capacity: $\left(\bigoplus_{i=1}^{N} \mathcal{G}_i\right) \approx \left(\bigoplus_{i=1}^{N} \mathcal{F}_i\right)$. This ensures the "synergy of specialized parts" maintains generative fidelity. **(iii) Computational efficiency.** To ensure efficiency gains, the complexity of each function is constrained such that its execution cost (e.g., sampling speed) is lower than or equal to the baseline: $\text{Cost}(f(\psi)) \leq \text{Cost}(f(\psi_{base})), \quad \forall\, \psi \in \Psi$.

2.2 Rotation Complementarity

Motivation. Dense attention in DiTs provides global spatial interaction but is costly and applies similar feature processing across layers. Causal attention is cheaper but lacks global context. We therefore rotate feature maps before applying causal masks, creating function variants that capture complementary directional dependencies. Although each variant models only partial interactions, their aggregation across layers recovers broad spatial connectivity at lower cost.

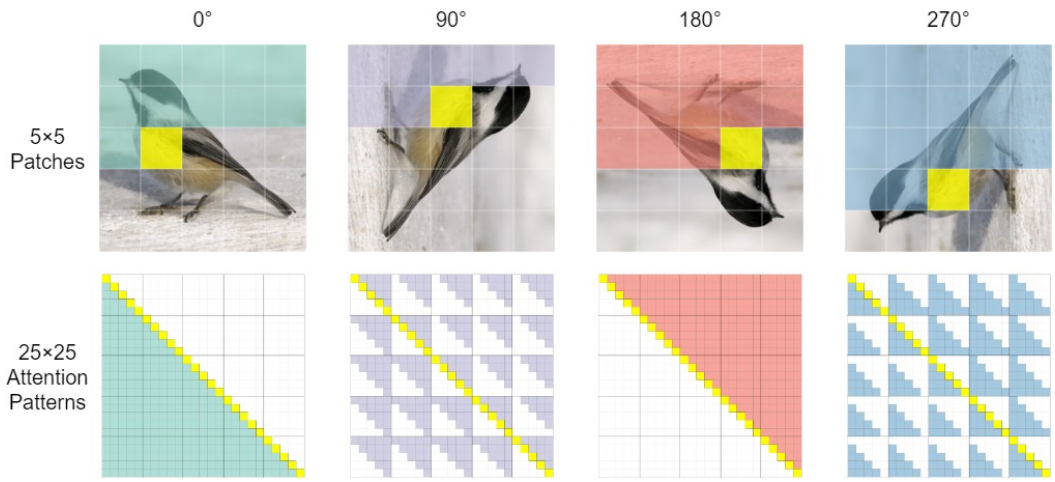

Figure 2: **Illustration of four rotation-based functions.** As shown in the top row, under causal attention, the yellow query token always attends to its prefix tokens, which vary according to the rotation angle, ensuring that information from different spatial directions is fully aggregated. The bottom row shows the attention pattern of the unrotated feature equivalent to the corresponding rotation function, further demonstrating the structural complementarity of rotation design.

Functional Mechanism. To capture diverse directional dependencies while maintaining spatial alignment and global context, we define a set of four rotation-based functions $\Psi_{rot} = \{0°, 90°, 180°, 270°\}$ corresponding to the four cardinal rotation angles as shown in Fig. 2. For example, $90° \in \Psi_{\text{rot}}$ denotes a function instantiated with a $90°$ spatial rotation. To implement these functions, we propose a "Rotate-Attend-Restore" pipeline which allows each function to focus on a unique directional scan (e.g., scanning from bottom-right to top-left at $180°$) while ensuring that the resulting features are perfectly aligned for subsequent layer integration. Specifically, given a feature map $\mathbf{X} \in \mathbb{R}^{H\times W\times d}$, where $H \times W$ denotes the spatial resolution and d is the channel dimension, the rotation-based function $f^{\text{rot}}(\psi)$ operates as follows.

First, given the rotation function $\psi \in \Psi_{\text{rot}}$, the input $\mathbf{X}$ is spatially re-oriented in a clock-wise manner via $\mathbf{X}_{rot} = \text{Rot}(\mathbf{X}, \psi)$ and flattened into a sequence $\mathbf{X}_{flat} \in \mathbb{R}^{L\times d}$ with $L = H \times W$. The attention is then computed with a causal mask $\mathbf{M} \in \mathbb{R}^{L\times L}$ to enforce a specific scanning direction:

$$\mathbf{X}_{attn} = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}} + \mathbf{M}\right)\mathbf{V}, \quad \text{where } \mathbf{M}[i,j] = \begin{cases} 0, & i \geq j \\ -\infty, & i < j \end{cases}. \tag{2}$$

Here, $\{\mathbf{Q}, \mathbf{K}, \mathbf{V}\} = \text{Linear}(\mathbf{X}_{flat}) \in \mathbb{R}^{L\times d}$ are the query, key, and value projections. Then, the ordinary residual connection is applied as $\mathbf{X}_{out} = \mathbf{X}_{attn} + \mathbf{X}_{flat}$. Finally, to ensure that the output feature map $\mathbf{X}' \in \mathbb{R}^{H\times W\times C}$ remains spatially consistent with the original input $\mathbf{X}$, we apply an inverse rotation after reshaping the sequence $\mathbf{X}_{out} \in \mathbb{R}^{L\times C}$ back into a 2D grid: $\mathbf{X}' = \text{Rot}^{-1}(\text{Reshape}(\mathbf{X}_{out}), \psi)$, where $\text{Rot}^{-1}(\cdot, \psi)$ restores the feature map to its canonical orientation.

Justification. We justify that Ψ_{rot} satisfies the three core conditions required for structural complementarity: **(i) Diversity via anisotropic priors.** By varying the rotation angle $\psi \in \{0°, 90°, 180°, 270°\}$, we generate four heterogeneous anisotropic priors that result in different attention patterns. **(ii) Additive approximation and structural completeness.** While a single causal function ψ restricts the receptive field, the collective ensemble ensures structural completeness. In our formulation, each function ψ with causal attention equals applying a specific mask $\mathbf{M}_\psi$ to the attention computation without rotation, as illustrated in Fig. 2. We define the attendable index set for $\mathbf{M}_\psi$ as $\Omega_\psi = \{(i,j) \mid \mathbf{M}_\psi[i,j] = 0\}$. For any pair $\psi_i, \psi_j \in \Psi_{\text{rot}}$ representing diametrically opposed rotations (i.e., $|\psi_i - \psi_j| = 180°$), they form a spatial complement satisfying: $\Omega_{\psi_i} \cup \Omega_{\psi_j} = \mathcal{I} \times \mathcal{I} = \Omega_{\text{base}}$, where $\mathcal{I}$ is the set of token indices and Ω_{base} denotes the attendable index set of dense attention. By applying four functions together, each token pair can be covered twice on both horizontal and vertical trajectories. This structural property ensures that the integrated representation can approximate the full spatial reachability of the homogeneous baseline. By implementing this design in different components of the transformer (Vaswani et al., 2017) architecture

in Section 3, we experimentally demonstrate that such an additive approximation can be practically achieved beyond the theoretical justification. **(iii) Efficiency through causal truncation.** Compared to dense attention in ψ_{base}, the causal mask $\mathbf{M}$ in $\psi \in \Psi_{\text{rot}}$ achieves approximately 50% sparsity, halving the theoretical computational cost of the attention operation.

2.3 Scale Complementarity

Motivation. Multi-scale hierarchical designs have achieved remarkable success in computer vision. In this section, we explore scale-wise complementarity for DiT-based image generation. To this end, we use convolutional downsampling for keys and values with varying strides as an established operator, which enables the model to capture diverse yet complementary representations across multiple scale levels.

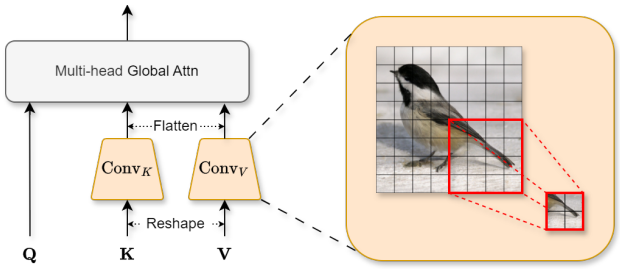

Figure 3: **Scale-based functions.** Unit-specific ratios ψ control separate convolutions that downsample $\mathbf{K}$ and $\mathbf{V}$.

Functional Mechanism. To address the redundancy in spatial resolution, we define a set of scale-based functions $\Psi_{scale} = \{1, ..., K\}$, which indicates the functions with a list of distinct downsampling ratios $\psi \in \Psi_{scale}$. For a feature map $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$ and its flattened sequence $\mathbf{X}_{flat} \in \mathbb{R}^{L \times C}$ where $L = H \times W$, the function ψ executes the following pipeline:

Similarly, we compute $\{\mathbf{Q}, \mathbf{K}, \mathbf{V}\} = \text{Linear}(\mathbf{X}_{flat})$ to obtain the query, key, and value. We then spatially downsample $\mathbf{K}$ and $\mathbf{V}$ with depth-wise 2D convolutions: $\tilde{\mathbf{K}} = \text{Conv}_K(\mathbf{K}, \psi)$ and $\tilde{\mathbf{V}} = \text{Conv}_V(\mathbf{V}, \psi)$, where $\tilde{\mathbf{K}}, \tilde{\mathbf{V}} \in \mathbb{R}^{\tilde{L} \times d}$ and $\tilde{L} = \frac{L}{\psi^2}$. Here, we omit the reshape operation before and after convolution for simplicity. We do not downsample $\mathbf{Q}$, so the attention output $\mathbf{X}_{attn}$ retains the input resolution. Attention is computed between the full-resolution query and the condensed key-value pairs as $\mathbf{X}_{attn} = \text{Softmax}\left(\frac{\mathbf{Q}\tilde{\mathbf{K}}^\top}{\sqrt{d}}\right)\tilde{\mathbf{V}}$.

The final output $\mathbf{X}'$ is computed by reshaping $\mathbf{X}_{out} = \mathbf{X}_{attn} + \mathbf{X}_{flat}$ back into a 2D grid. When ψ equals 1, we remove the convolution operation and simply use standard units. This "Conv-Atten" mechanism forces each function to specialize in a unique spatial granularity.

Justification. We justify the use of $\mathcal{S}_{\text{scale}}$ by verifying whether it satisfies the three fundamental conditions one by one: **(i) Diversity via multi-resolutions.** By varying the downsampling factor $\psi \in \Psi_{\text{scale}}$, we induce heterogeneous scale-wise priors, encouraging different units to capture diverse frequency components with reduced redundancy. **(ii) Additive approximation for complementarity.** Although each $\psi \in \Psi_{\text{scale}}$ operates at a specific resolution, their combination spans multiple levels of granularity, enabling the model to capture both global context and fine-grained details. **(iii) Efficiency through reduced attention cost.** In a unit with $\psi \neq 1$, the attention complexity is reduced from $O(L^2)$ to $O(\frac{L^2}{\psi^2})$. While depth-wise convolution introduces a linear computational cost, the quadratic savings in attention become dominant as the sequence length L increases, ensuring that integrating multiple scale-wise functions remains computationally advantageous, particularly in high-resolution regimes.

3 Where to Apply: Accessing Different Components

In this section, we address the second research question, *"What components of DiT can exhibit complementarity?"*, by analyzing complementary designs at both the layer and attention-head levels. Through experiments, we observe that both rotation and scale complementarity are suited for layer-wise implementation, and that rotation complementarity can be extended to the head-wise setting.

Experimental Setup. We adopt LightningDiT-XL/1 (Yao et al., 2025) as the baseline model, which consists of $N_l = 28$ homogeneous layers parameterized by ψ_{base} with global self-attention, and $N_h = 16$ attention heads per layer. All LightningDiT variants are trained and evaluated on ImageNet256 (Krizhevsky et al., 2012), which is widely used for image synthesis (Li et al., 2024),

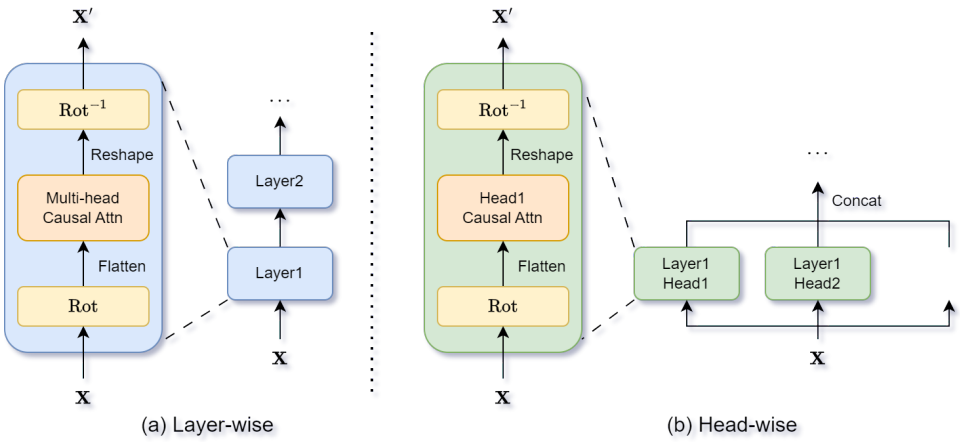

Figure 4: **Layer-wise and head-wise rotation complementarity.** (a) Layer-wise rotations are applied sequentially across transformer layers. (b) Head-wise rotations are applied independently across attention heads and concatenated afterward.

Table 1: **Performance of complementary strategies on LightningDiT-XL.** Models are trained on ImageNet for 64 epochs. Inference latency denotes backbone sampling time per image, excluding VAE decoding. We report the average speed of generating 100 images on a single NVIDIA A100 GPU with a batch size of 4.

Setting	Method	Fidelity				Efficiency
		w/o cfg		w cfg		Inference Latency↓
		FID↓	IS↑	FID↓	IS↑	(sec/img)
Base Model	LightningDiT	5.14	132.83 ± 3.23	2.11	250.15 ± 4.14	2.32
Layer-wise Comp.	Layer-Rot	5.28	131.07 ± 2.70	2.11	246.02 ± 3.62	2.03(12.50%↓)
	Layer-Scale	5.00	136.13 ± 2.48	2.01	252.66 ± 2.92	2.10(9.48%↓)
Head-wise Comp.	Head-Rot	5.35	131.08 ± 2.45	2.10	244.52 ± 5.22	2.27(2.15%↓)

following the original training scheme. Beyond LightningDiT, we further extend complementarity to SiT, assessing its effectiveness across different architectures(Section 3.3). Further implementation detials are provided in Appendix B .

3.1 Layer-wise Complementarity

Implementation Details. For layer-wise complementarity, we treat each layer as a unit. Given $\mathcal{H}_{comp} = \{\mathcal{G}_1^l, \mathcal{G}_2^l, ..., \mathcal{G}_{N_l}^l\}$, the aggregation mechanism $\bigoplus$ is functional composition $(\mathcal{G}_{N_l}^l \circ \cdots \circ \mathcal{G}_1^l)$. The layer- and head-wise implementations are illustrated in Fig. 4.

For rotation complementarity, we use Ψ_{rot} as the function set. We instantiate the assignment policy in Eq. (1) as $\psi(i) = 90°\big((i-1) \bmod 4\big)$, cycling rotation patterns with network depth. This formulation indicates that the rotation patterns are applied cyclically as the network depth increases. Experimentally, we observe that applying only two rotations cyclically, such as $\Psi_{\text{rot}} = \{0°, 180°\}$, is already sufficient to achieve strong performance. Since we use `scaled_dot_product_attention` in `torch.nn.functional` (Paszke et al., 2019) to compute attention, we implement the causal attention by setting the parameter `is_causal=True`.

For scale complementarity, we use $\Psi_{scale} = \{1, 2, 4\}$ as function set, where $\psi = 1$ denotes the original unit without scale reduction. By balancing efficiency and effectiveness, we assign scales for $N_l = 28$ using $\{\psi_i\}_{i=1}^{28} = [1^{\times 4}, 2^{\times 8}, 4^{\times 4}, 2^{\times 8}, 1^{\times 4}]$, where each superscript denotes the number of repetitions. For each convolution block in the layers, we set `stride = kernel_size` $= \psi_i$, achieving non-overlapped receptive fields across different convolution steps. The same pair of convolution blocks is shared across heads in the same layer for efficiency.



Figure 5: **Qualitative samples on ImageNet256.** All variants retain high visual fidelity.

Results. Table 1 shows that both layer-wise designs preserve quality after 64 epochs while reducing latency. Layer-Rot uses Interleaved ($G = 2$), matches the baseline FID of 2.11, and reduces latency by 12.50%. Layer-Scale improves FID to 2.01 and IS to 256.55(Table 1), with a 9.48% latency reduction. Visual samples in Fig. 5 (and detailed comparisons in Fig. 9) confirm that fidelity is maintained.

The efficiency gains become more pronounced at higher resolutions (Fig. 6). Attention scales quadratically with sequence length, dominating computation in this regime, whereas rotations and convolutions contribute minor costs. Complementary designs thus save a larger fraction of sampling time as attention grows costlier.

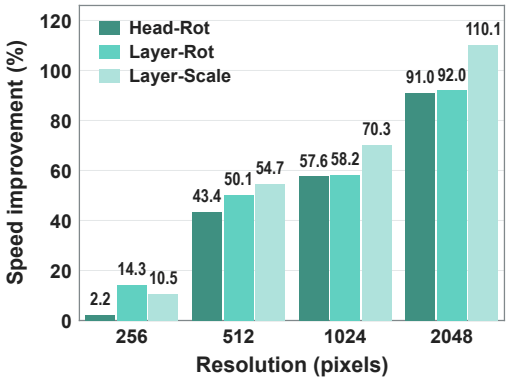

Figure 6: **Sampling speedups.** Backbone sampling time excludes VAE decoding; complementarity yields larger speedups at higher resolutions.

3.2 Head-wise Complementarity

Implementation Details. For head-wise complementarity, the basic units $\mathcal{H}_{comp} = \{\mathcal{G}_i^h\}_{i=1}^{N_h}$ are instantiated as the set of attention heads in a single layer Fig. 4. Thus, the aggregation mechanism $\bigoplus$ is defined by concatenation ($[\mathcal{G}_1^h; ...; \mathcal{G}_{N_h}^h]$). In this configuration, we only implement rotation complementarity and exclude scale complementarity from the head-wise design due to its prohibitive parameter overhead. Unlike the layer-wise design, where convolution kernels can be shared across heads in the same layer, implementing scale-wise functions at the head level would require unique kernels for each downsampling ratio within every individual layer. Furthermore, since $\mathbf{K}$ and $\mathbf{V}$ projections do not share kernels and these kernels are not reused across layers, the cumulative parameter count would scale linearly with both the number of layers and head groups.

For rotation complementarity, we partition the $N_h = 16$ attention heads into four groups of contiguous heads with different rotation angles, assigning them in parallel as $\psi(i) = 90° \left\lfloor \frac{i-1}{4} \right\rfloor$ for $i \in \{1, \ldots, N_h\}$. Similar to the layer-wise implementation, the causal dependency is enforced via `is_causal=True` in the `scaled_dot_product_attention` operator. After the attention scores are computed and restored to their canonical orientation, the outputs of all groups are concatenated to form the final attention output. Pseudocode is provided in the appendix.

Results. As shown in Table 1, head-wise rotation remains competitive. While it still offers a speedup over the baseline by using causal attention, the efficiency gain is less significant than the layer-wise counterpart. This is because head-wise implementation requires more rotations and explicit restore operations to maintain spatial alignment before concatenation, introducing additional overhead.

3.3 Results Using SiT as the Base Model

We implement our designs on SiT-S with 12 layers and train the models on ImageNet-256 for 800 epochs. Table 2 reports FID and IS for all methods with classifier-free guidance(CFG). For Layer-Scale, the schedule of downsampling ratios is $[1^{\times 2}, 2^{\times 3}, 4^{\times 2}, 2^{\times 3}, 1^{\times 2}]$. All variants remain comparable with SiT, while Layer-Scale achieves better FID than the base model.

Table 2: **Results of complementarity on SiT-S.** Models are trained on ImageNet for 800 epochs (with CFG).

Method	SiT	Layer-Rot	Layer-Scale	Head-Rot
FID↓	11.02	11.17	10.76	11.25
IS↑	262.11 ± 3.66	260.19 ± 3.75	275.00 ± 5.17	261.19 ± 3.00

4 How Design Matters: Sensitivity to Architectural Assignments

In this section, we address the third research question, *"How do different structural designs influence the effectiveness of complementarity?"*, by evaluating complementary mechanisms under various design choices, including symmetry configurations, grouping strategies, and ordering schemes. Our results show that the proposed complementary designs remain effective across diverse architectural permutations, highlighting their robustness as a general structural principle.

4.1 Effect of Complementarity

To isolate the effect of complementarity, we evaluate different rotation angle assignments across layers (28 layers total). Without comprehensive angle coverage, identical angles (all $0°$) yields non-complementary designs. In contrast, alternating opposing angles ($0°$ and $180°$) or cycling through four orientations ($[0°, 90°, 180°, 270°] \times 7$) provides full angle coverage and thus achieves complementarity. As shown in Table 3, complementary assignments achieve superior performance compared to non-complementary ones, validating the effectiveness of complementarity.

Table 3: **Effect of complementarity.** FID for layer-wise rotation with different angle assignments.

Rotation Assignment	FID↓
$[0°] \times 28$	114.81
$[0°, 90°, 180°, 270°] \times 7$	2.15
$[0°, 180°] \times 14$	2.11

4.2 Symmetry Configurations

Symmetry in our context refers to whether different functions appear with equal frequency across the architecture. To evaluate our model's sensitivity to this distribution, we compare three layer-wise configurations: **Symmetric**: The 4-angle cyclic pattern ($[0°, 90°, 180°, 270°] \times 7$) introduced above, where all orientations are uniformly distributed. **Asymmetric (Imbalanced)**: Angles appear with unequal frequencies (10, 8, 6, and 4, respectively) in a randomly shuffled order. **Asymmetric (Incomplete)**: One cardinal direction ($180°$) is excluded, using only three angles with a distribution of $[0°, 90°, 270°] \times 9 + [0°]$. As shown in Table 4, FID differences across these variants are marginal, suggesting that directional diversity matters more than a balanced or exhaustive distribution.

4.3 Grouping Strategies

We further investigate how the spatial grouping of functions affects the synergy of complementarity across two dimensions: rotation-wise and scale-wise groupings.

For rotation complementarity, we contrast four configurations: **(1) Interleaved** ($G = 4$): $[0°, 90°, 180°, 270°] \times 7$; **(2) Interleaved** ($G = 2$): $[0°, 180°] \times 14$; **(3) Homogeneous Blocks**: 7 layers of each angle clustered sequentially (e.g., $7 \times 0°$ followed by $7 \times 90°$, ...); **(4) Randomized**: Angles appearing 7 times each in a shuffled order. All variants yield similar competitive results except Homogeneous Blocks(Table 4).This indicates that local homogeneity harms rotation complementarity, as long sequences of identical directional priors restrict the model's ability to maintain a multi-perspective representation space in intermediate layers.

For scale complementarity, we evaluate the following arrangements of downsampling ratios $r \in \{1, 2, 4\}$: **(1) Intra-group Homogeneous Blocks**: The symmetric sandwich schedule defined above, with identical functions grouped into contiguous clusters; **(2) Cyclic Micro-groups**: $[1, 2, 2, 4, 2, 2, 1] \times 4$; **(3) Randomized**: Shuffled ratios maintaining the same total proportion.

The results in Table 4 show that Intra-group Homogeneous Blocks 1 achieve the best performance, whereas the heterogeneous 2 and randomized 3 settings are slightly less effective. Interestingly,

Table 4: **Effect of symmetry and grouping designs.** FID under rotation symmetry, rotation grouping, and scale grouping choices.The implementation is based on LightningDiT-XL.

	Rotation symmetry			Rotation grouping				Scale grouping		
Setting	Sym.	Imbal.	Incomp.	Interl.(4)	Interl.(2)	Homo.	Rand.	Intra-Homo.	Cyclic	Rand.
FID↓	2.15	2.17	2.11	2.15	2.11	2.40	2.14	2.01	2.15	2.16

Table 5: **Effect of Ordering Schemes.** FID(cfg) under different intra-group rotation sequences and inter-group scale schedules. The implementation is based on LightningDiT.

	Rotation Ordering (Intra-group)			Scale Ordering (Inter-group)		
Setting	Clockwise	Counter-cw	Random	Small-Large-Small	Random Shuffle	Monotonic Inc.
FID↓	2.15	2.17	2.10	2.01	2.10	2.48

while local homogeneity harms rotation settings, it appears beneficial for scale complementarity. This suggests that structural continuity in resolution helps stabilize feature aggregation, whereas rotation requires more frequent "perspective switching" to thrive. Nonetheless, the small performance gap across all scale settings confirms the robustness of the complementarity principle.

4.4 Ordering Schemes

We finally investigate the sequential order of functions. We compare various ordering patterns to determine whether the model is sensitive to the "trajectory" of transformations across layers.

For rotation complementarity, we evaluate intra-group ordering schemes: **(1) Clockwise**: $[0°, 90°, 180°, 270°] \times 7$. **(2) Counter-clockwise**: $[270°, 180°, 90°, 0°] \times 7$. **(3) Random**: Layers are divided into groups of four, where each group contains all four distinct angles from $\mathcal{S}_{rot}$, but their internal sequence is randomized. For the tested schedules, FID ranges from 2.10 to 2.17(Table 5). This relatively narrow range suggests limited sensitivity to the evaluated orders, demonstrating that the deep transformer architecture can integrate diverse directional priors regardless of their specific sequential arrangement.

For scale complementarity, we investigate whether the sequential arrangement of resolution scales impacts generative quality. Here, we maintain the homogeneous block clustering strategy—where layers with the same downsampling ratio are grouped together—and specifically compare the impact of different inter-group ordering schemes: **(1) Small-Large-Small**: The symmetric sandwich schedule defined above. **(2) Random Shuffle**: In our random seed, it is $[1..., 2..., 4..., 2..., 2..., 1..., 2...]$. **(3) Monotonic Increasing**: A gradual transition from high to low resolution $[1\ldots, 1\ldots, 2\ldots, 2\ldots, 2\ldots, 2\ldots, 4\ldots]$. The system maintains its integrity except a slight performance drop in the Monotonic scheme(Table 5), which is likely due to the model ending in a low-resolution $r = 4$ that limits precision of the final reconstruction stage.

5 Conclusion

In this paper, we introduce complementarity as a structural principle for DiT design. By replacing homogeneous self-attention layers with heterogeneous yet cooperative functions, we design two practical instantiations, rotation complementarity and scale complementarity, and apply them at both layer-wise and head-wise levels. Extensive experiments demonstrate that complementary designs maintain comparable generative performance while can benefit efficiency. Moreover, analyses of symmetry, grouping, and ordering strategies show that complementarity is robust across various architectural configurations. **Limitations:** Our findings characterize preliminary properties of complementarity within the studied configurations. How these insights generalize to other tasks and a broader architectural space, and how complementarity can guide more efficient structural designs, remain open questions for future work. We hope that the analysis presented in this paper can serve as a strong starting point and practical reference for the community to further explore structural complementarity in generative transformers.

AI use statement. *We used AI-assisted tools to support manuscript writing and revision, and to modify experimental code. All AI-assisted outputs, code changes, and experimental results were reviewed and verified by the authors.*

REFERENCES

Srinadh Bhojanapalli, Ayan Chakrabarti, Andreas Veit, Michal Lukasik, Himanshu Jain, Frederick Liu, Yin-Wen Chang, and Sanjiv Kumar. Leveraging redundancy in attention with reuse transformers. *arXiv preprint arXiv:2110.06821*, 2021.

Daniel Bolya and Judy Hoffman. Token merging for fast stable diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4599–4603, 2023.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023.

Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *European Conference on Computer Vision*, pp. 74–91. Springer, 2024.

Tianlong Chen, Zhenyu Zhang, Yu Cheng, Ahmed Awadallah, and Zhangyang Wang. The principle of diversity: Training stronger vision transformers calls for reducing all levels of redundancy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12020–12030, 2022.

Kun Cheng, Xiao He, Lei Yu, Zhijun Tu, Mingrui Zhu, Nannan Wang, Xinbo Gao, and Jie Hu. Diff-moe: Diffusion transformer with time-aware and space-adaptive experts. In *Forty-second International Conference on Machine Learning*, 2025.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Shanghua Gao, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. Masked diffusion transformer is a strong image synthesizer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 23164–23173, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36115–36125, 2026.

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

Junqing Lin, Xingyu Zheng, Pei Cheng, Bin Fu, Jingwei Sun, and Guangzhong Sun. Token pruning for in-context generation in diffusion transformers. *arXiv preprint arXiv:2602.01609*, 2026.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pp. 23–40. Springer, 2024a.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. Deepcache: Accelerating diffusion models for free. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15762–15772, 2024b.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Ethan Smith, Nayan Saxena, and Aninda Saha. Todo: Token downsampling for efficient generation of high-resolution images. *arXiv preprint arXiv:2402.13573*, 2024.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. 2023.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

Yuchuan Tian, Zhijun Tu, Hanting Chen, Jie Hu, Chao Xu, and Yunhe Wang. U-dits: Downsample tokens in u-shaped diffusion transformers. *Advances in Neural Information Processing Systems*, 37:51994–52013, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15703–15712, 2025.

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

Daquan Zhou, Bingyi Kang, Xiaojie Jin, Linjie Yang, Xiaochen Lian, Zihang Jiang, Qibin Hou, and Jiashi Feng. Deepvit: Towards deeper vision transformer. *arXiv preprint arXiv:2103.11886*, 2021.

Junhan Zhu, Hesong Wang, Mingluo Su, Zefang Wang, and Huan Wang. Obs-diff: Accurate pruning for diffusion models in one-shot. *arXiv preprint arXiv:2510.06751*, 2025.