STODA: STRUCTURED TILING VIA ORTHOGONAL DESIGN ADAPTERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern foundation models increasingly rely on functional decomposition, representing complex mappings as mixtures of simpler primitives. Sparse Mixture-of-Experts (MoE) architectures provide an efficient realization of this principle, but often suffer from expert collapse. Existing load-balancing heuristics encourage statistical uniformity, yet do not enforce functional diversity by construction. We propose **STODA**: **S**tructured **T**iling via **O**rthogonal **D**esign **A**dapters, a framework that partitions the global mapping space using algebraic structures derived from Orthogonal Designs. Each expert follows a lightweight *view–core–unview* architecture: a fixed Hurwitz–Radon (HR) view transform, a learnable core module, and an inverse mapping. The anti-commutativity and skew-symmetry of the HR family provide a tractable algebraic mechanism for expert diversification: at the generator level they yield orthogonal token views and non-degenerate mixing, while for linearized cores they identify collapse conditions and balanced cross-term structure under routed mixtures. To validate the framework under strict resource constraints, we study STODA in the Parameter-Efficient Fine-Tuning (PEFT) setting for diffusion models. Empirically, STODA matches or surpasses state-of-the-art methods while requiring dynamic computation and fewer active parameters.

1 INTRODUCTION

Modern foundation models are increasingly built around a simple principle: *scale capacity without scaling compute*. Sparse Mixture-of-Experts (MoE) architectures exemplify this trend, particularly in large language models (Dai et al., 2024; Fedus et al., 2022). Their appeal lies in a favorable quality–compute tradeoff: model capacity scales with parameter count, while only a small subset of experts is activated per token (Shazeer et al., 2017; Jiang et al., 2024). Although MoEs are most prominent in language models, similar scaling strategies are emerging in diffusion models as well (Xue et al., 2023; Cheng et al., 2025). These results suggest a broader trend: as models grow larger, conditional specialization becomes an increasingly effective scaling mechanism.

Is MoE only about a smaller active footprint? One motivation for MoE is the reduced active footprint per token. While true, the deeper principle is *functional decomposition*: representing a complex global mapping as a mixture of simpler, specialized mappings, where only a few experts are adaptively activated. Recent work suggests that such compositions are more than a heuristic, motivating a composition-centric view of MoE in which the model mapping is assembled from a set of specialized experts. Two recent ideas (Garipov et al., 2023; Tian, 2025) make this perspective explicit at complementary levels: *distributions* and *solutions*. At the *distribution level*, Garipov et al. (2023) formulates composition for iterative generative processes. Composite targets are constructed by combining base distributions through binary composition rules, instantiated via classifier guidance. At the *solution/parameter level*, Tian (2025) shows that, for Abelian-group reasoning, Fourier modes diagonalize the objective, while a semiring structure enables clean functional composition through parameter composition. Their insight is that the algebraic structure of the modes prevents *interference* during composition. Together, these works suggest a common principle: *when the primitives possess the right structure, composition becomes stable* (Garipov et al., 2023; Tian, 2025).

What does composition mean for diffusion models? In MoE architectures, experts act as functional primitives that *tile* the token representation space. Each expert models a subset of the global

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.

transformation, while the router learns how to select and combine them (Jacobs et al., 1991; Chen et al., 2022). This perspective extends naturally to fine-tuning, where training from scratch is prohibitively expensive. So, we replace the standard monolithic adapter with an MoE adapter, using token heterogeneity in diffusion models to induce expert specialization. Ideally, each expert captures a distinct transformation regime, and routing becomes a learned lookup over the expert library.

Modularity in adapters for diffusion? Training MoE modules can lead to *expert imbalance* or collapse, where routers overuse a subset of experts. This motivates load-balancing/routing heuristics (Lepikhin et al., 2021; Fedus et al., 2022; Zoph et al., 2022). However, balanced usage does not guarantee *functional diversity*: experts may still learn redundant mappings. Thus, the desired "tiling" of the representation space remains incomplete. The composition perspective suggests an alternative: structure experts so their mixtures have controlled cross-terms (Fig. 1). We try to realize this idea geometrically.

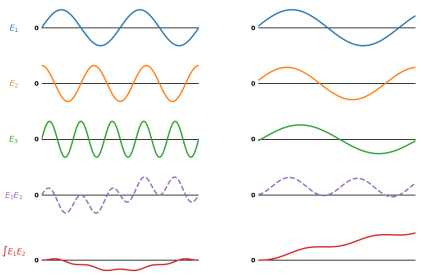


Figure 1: Each panel shows three experts and their sum. The cross-term row measures pointwise correlation between experts. The cumulative integral row measures accumulation. Orthogonal experts: (a) the cross-term oscillates symmetrically (integral bounded near zero). (b) Correlated experts: the cross-term stays positive, integral grows.

What properties would good tiles have? Roughly, we seek expert views that are: (i) *compute efficient*, so routing incurs minimal overhead, (ii) *maximally distinct* across experts, to discourage collapse, (iii) *stable under mixing*, so expert combinations remain well-behaved, and (iv) *principled in scale*, so the number of non-redundant experts is theoretically grounded. Satisfying all four properties simultaneously is nontrivial. In particular, property (iii) is subtle: diversity alone is easy, but *composable diversity* needs mixtures of experts to behave predictably without introducing uncontrolled cross-terms or degeneracies. Put differently, we seek diverse expert views that remain well-conditioned.

This work: STODA. We start with orthogonal designs (Seberry, 2017), particularly the Hurwitz–Radon (HR) family. An HR family consists of skew-symmetric orthogonal matrices that pairwise anti-commute (Shapiro, 2000). In STODA, each expert follows a *view–core–unview* structure: a fixed HR view transform, a lightweight learnable core, and an inverse mapping. Its structure constrains the conjugation of the core, giving expert distinctness and mixing stability. We show that these properties help isolate a structural bias in §4 . Our contributions are: (a) We propose STODA, a compositional PEFT framework for diffusion models that promotes principled expert diversity through orthogonal transforms from the HR family. (b) We show that the HR-based view–core–unview structure provides structural resistance to expert collapse. We further analyze routed mixtures: for top-k routing with $k \geq 2$, HR anti-commutativity yields balanced cross-term structure among active routed experts. (c) We evaluate STODA on standard benchmarks and demonstrate competitive performance against established PEFT methods with a favorable parameter–compute tradeoff. Of independent interest is the analogy to Tian (2025): just as Fourier modes diagonalize objectives in Abelian-group reasoning, HR views diagonalize the expert input space. Both approaches use algebraic structure to suppress interference under composition: parameter space in Tian (2025) and function space in STODA.

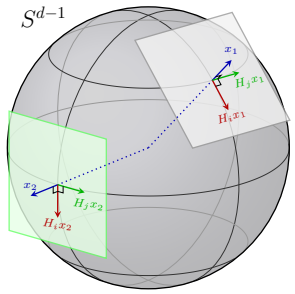


Figure 2: Tangential directions generated by HR matrices on the unit sphere. Although non-trivial HR families exist only in even dimensions (and S^2 admits no nowhere-vanishing vector fields), we use the 3D sphere only for intuition; our construction operates in high-dimensional latent spaces.

2 Preliminaries

Mixture-of-Experts. An MoE layer replaces a dense layer with experts $\{E_i\}_{i=1}^{N}$ and a router that routes the tokens to these experts:

$$\mathrm{MoE}(x) = \sum_{i=1}^{N} g_i(x)\, E_i(x) \qquad \text{s.t.} \qquad g_i(x) \geq 0, \quad \sum_i g_i(x) = 1,$$

In sparse MoE, only the top-k weights are nonzero for each token. We view MoE as *conditional composition*: the router picks a subset of experts whose outputs combine into a final update.

Expert Collapse. A key challenge across these architectures is expert collapse, where the router converges to favoring a few experts (Shazeer et al., 2017), and adapters converge to redundant updates (Zou et al., 2026). Existing MoE-PEFT methods counter this with auxiliary load-balancing losses that enforce uniform utilization (Li et al., 2024; Zadouri et al., 2023; Shazeer et al., 2017), or with frozen random projections that decorrelate experts *in expectation* (Zou et al., 2026). STODA asks instead for distinctness that holds exactly, for every token and before any gradient step. This requires a family of transforms that are mutually orthogonal by construction rather than on average.

Hurwitz–Radon (HR) families. A (real) orthogonal design is a matrix $R \in \mathbb{R}^{n\times n}$ whose entries are drawn from $\{0, \pm x_1, \ldots, \pm x_u\}$. It satisfies the identity $R(x)^\top R(x) = \left(\sum_{\ell=1}^{u} x_\ell^2\right) I_n$. An HR family is a special class of Orthogonal Designs, defined as a set $\{H_1, \ldots, H_s\} \in \mathbb{R}^{n\times n}$ where

$$H_i^\top H_i = I, \quad H_i^\top = -H_i, \quad H_i H_j = -H_j H_i \ \ (i \neq j). \tag{1}$$

Each condition in (1) buys a property STODA needs. Orthogonality makes every view norm-preserving; skew-symmetry makes it *tangential*, $\langle H_i x, x\rangle = 0$; and anti-commutativity makes views from different generators mutually orthogonal, $\langle H_i x, H_j x\rangle = 0$ for every x and every $i \neq j$ (Fig. 3). On S^{n-1} the family therefore traces s everywhere-orthonormal tangent directions (Fig. 2). The same cancellation extends to mixtures: any $M(\alpha) = \alpha_0 I + \sum_i \alpha_i H_i$ is isometric up to scale (Geramita & Pullman, 1974; Lam & Yiu, 1993), so combining views cannot degenerate.

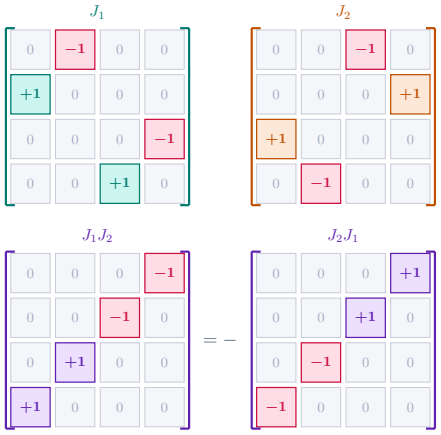


Figure 3: Anti-commutative property of two HR matrices in $\mathbb{R}^4$. Anti-commutativity implies orthogonal/maximally distant transformations.

Two practical consequences come out. The number of generators available in dimension n is $\rho(n) - 1$, where $\rho(n)$ is the Radon–Hurwitz number (Geramita & Pullman, 1974; Lam & Yiu, 1993). This provides a principled ceiling on how many non-redundant experts a layer can host. And they are cheap: HR matrices are built recursively from 2×2 primitives by Kronecker products with entries in $\{0, \pm 1\}$, so applying H_i is a sequence of sign flips and permutations, not a dense product. STODA uses each H_i as a fixed *view generator* $V_i(x) = H_i x$ and lets each expert learn *within* its own view.

3 STODA: Setting up the Architecture

STODA gives each routed expert a *structured view* of the token, chosen so that **(i)** different experts see different views and **(ii)** these views are well-behaved under router mixing. These two properties (diversity+composability) are not independent: in §4, we will see that the algebraic structure of HR family gives both via the *conjugation* of each expert's learnable core by its view transform.

Let $W \in \mathbb{R}^{d\times d}$ be a frozen pretrained linear map in a diffusion model and let $x \in \mathbb{R}^d$ be the input token. STODA is inserted simply as an additive residual module:

$$y = Wx + \text{STODA}(x). \tag{2}$$

This preserves the pretrained setup while allowing task-specific adaptation through a lightweight routed module (Fig. 4a). Following (Dai et al., 2024; Yang et al., 2024), we use a shared expert E_0 that processes all tokens and a pool of N routed experts $\{E_i\}_{i=1}^N$. The overall output is

$$\text{STODA}(x) = E_0(x) + \sum_{i\in\mathcal{S}} g_i(x)\, E_i(x), \qquad \text{where } \mathcal{S} = \textsc{TopK}(\hat{g}, k) \tag{3}$$

where $g_i(x) \geq 0$ are router weights with sparse top-k support, $\hat{g}(x)$ are the router logits. E_0 anchors the update in the pretrained coordinate frame, while routed experts provide token-dependent specialization (Fig. 4b). Unlike standard MoEs, where N is arbitrary, the Radon-Hurwitz number bounds STODA's maximum number of functionally diverse experts ($\rho(d) - 1$).

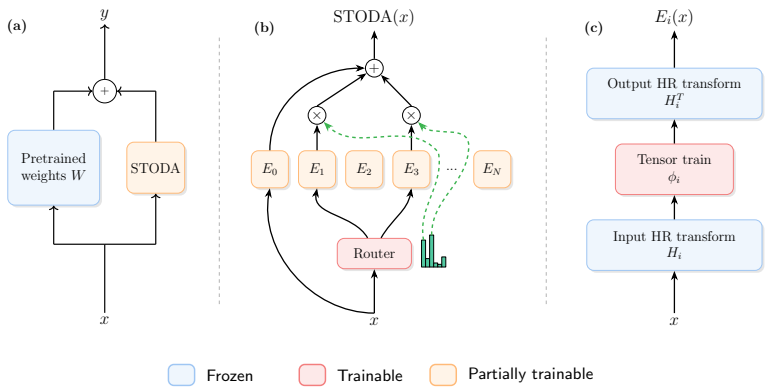


Figure 4: **(a)** Addition of STODA module parallel to the pretrained weights W. **(b)** STODA architecture: shared expert is always selected while the router activates the remaining experts per token. The routed experts are weighted by the gate scores before combining. **(c)** Expert architecture: the input undergoes an HR transformation followed by the tensor train function and then an inverse HR transformation.

3.1 The View–Core–Unview Experts

One way to prevent expert collapse is to ensure that experts see different inputs. In STODA, each expert applies a fixed HR view transform H_i, a lightweight core ϕ_i, and then maps back with $H_i^\top$:

$$E_i(x) = H_i^\top \, \phi_i(H_i x). \tag{4}$$

We parameterize ϕ_i with Tensor Trains to achieve high expressivity with low parameter cost and enable cross-core information flow. Importantly, H_i is *not* the expert itself, but a generator of orthogonal views, and ϕ_i is the learnable function defined on that view (Fig. 4c). **Compute benefits.** Dense random orthogonal rotations cost $\mathcal{O}(d^2)$ per token, which is not ideal. For HR matrices, due to the recursive Kronecker constructions with entries in $\{0, \pm 1\}$, we get multiply-free operations and fast application in $\mathcal{O}(d \log d)$ time— essentially "free" (relative to adapter core compute).

Tiling behavior. The router adds expert outputs, so composition in an MoE is literally addition in function space. The interesting question is *not* whether experts can be added, but what structure on the expert family makes this sum behave (at least in an idealized case) like a "union of tiles" instead of interference? To check this, let us unpack the expert form. Writing $V_i(x) = H_i x$ and $U_i(z) = H_i^\top z$, we express each expert as $E_i = U_i \circ \phi_i \circ V_i$, The router mixes a structured family of *conjugated cores* (these cores are not unrelated functions):

$$F(x) = \sum_{i \in \mathcal{S}(x)} g_i(x) \, H_i^\top \, \phi_i(H_i x). \tag{5}$$

In a generic MoE, nothing prevents experts from learning redundant functions even under this addition. In STODA, due to the view–core–unview structure, we get a structural non-interference bias: the conjugation $H_i^\top(\cdot)\, H_i$ constrains both what the learned core can be for collapse to occur and how expert outputs interact under mixing (discussed in §4).

3.2 The Shared Expert (the anchor) and the Router

In fine-tuning, the backbone has learned a useful coordinate frame. So, the shared expert can operate in the native frame by setting $H_0 = I$ in (4). This gives a stable anchor, while routed experts operate through HR-generated views that are tangential/orthogonal in the sense analyzed in §4. In STODA, diversity comes from the view generators $\{H_i\}$, not from routing schemes. So, to keep routing lightweight, we use a linear router $G(x) = W_g x$ with $W_g \in \mathbb{R}^{N \times d}$, select the top-$k$ logits for each token, and apply a softmax over the selected set to obtain sparse weights $\{g_i(x)\}$, see Alg. 1.

4 Composable Diversity: Why the Experts Mix Stably

We now formalize what the structure in (4) buys us: first, that the HR views themselves are diverse and mix without rank collapse (§4.1). Then, we see that the conjugation of the learned core by the view provides structural resistance to expert collapse and controls cross-terms under routing (§4.2).

4.1 VIEW GEOMETRY AND GENERATOR-LEVEL MIXING

We first notice that different HR generators send each token to orthogonal views.

Proposition 4.1 (Orthogonal tangent frame). *Let $\{H_1, \dots, H_s\}$ be an HR family in $\mathbb{R}^d$. For any $x \in \mathbb{R}^d$: (i) $\langle H_i x,\, x\rangle = 0$ for all i, (ii) $\langle H_i x,\, H_j x\rangle = 0$ for all $i \neq j$, (iii) $\|H_i x\| = \|x\|$ for all i. Therefore, $\{x,\, H_1 x,\, \dots,\, H_s x\}$ is a set of $s{+}1$ pairwise orthogonal vectors, each of norm $\|x\|$.*

On the unit sphere S^{d-1}, each H_i defines a smooth, nowhere-vanishing tangent vector field $x \mapsto H_i x$. The HR family gives s such fields that are everywhere orthonormal. From Adams (1962), we know that the maximum number of linearly independent continuous tangent vector fields on S^{d-1} is exactly $\rho(d) - 1$, where $\rho(d)$ is the Radon–Hurwitz number. The HR family achieves this bound, so $\rho(d) - 1$ is the maximum number of distinct view generators for the geometry. Prop. 4.1 tells us that expert *inputs* are diverse. Next, we see that *mixing* these views does not introduce degeneracy.

Proposition 4.2 (Mixing of HR generators). *Let $\{H_1, \dots, H_N\}$ be an HR family in $\mathbb{R}^d$ and $\alpha_0, \dots, \alpha_N \in \mathbb{R}$. Let $M = \alpha_0 I + \sum_{i=1}^{N} \alpha_i H_i$. Then $M^\top M = \Big(\sum_{i=0}^{N} \alpha_i^2\Big) I$.*

So any linear combination of HR generators is full-rank and isometric up to a scalar factor. This is the "mixing is well-conditioned" statement for view-generators. These two results (orthogonal inputs and non-degenerate mixing) hold for the *views* alone. We now turn to the *expert functions* $E_i = H_i^\top \phi_i(H_i \cdot)$, where the learned core ϕ_i enters through conjugation.

4.2 EXPERT-LEVEL BEHAVIOR

Structural resistance to expert collapse. Prop. 4.1–4.2 do not say anything about the learned cores. Our next result shows that the conjugation $H_i^\top(\cdot)\, H_i$ severely constrains *which* cores can lead to expert collapse. All proofs are in supplement.

Theorem 4.3 (Anti-collapse). *Let $\{H_1, \dots, H_s\}$ be an HR family in $\mathbb{R}^d$. Set $H_0 = I$ for shared expert. Consider experts with a shared linear core $A \in \mathbb{R}^{d\times d}$:*

$$E_i(x) = H_i^\top A\, H_i\, x, \qquad i = 0, 1, \dots, s. \tag{6}$$

Then $E_i = E_j$ for all $i, j \in \{0, \dots, s\}$ if and only if A lies in the commutant of the algebra generated by $\{H_1, \dots, H_s\}$. This commutant has real dimension at most $(d/d_{\text{irr}})^2 \cdot \dim_{\mathbb{R}}(D)$, where d_{irr} is the dimension of the irreducible representation of the Clifford algebra $\mathrm{Cl}(0, s)$ generated by the H_i and $D \in \{\mathbb{R}, \mathbb{C}, \mathbb{H}\}$ is its Schur division algebra ($\dim_{\mathbb{R}}(D) \leq 4$).

Will random orthogonal views not work? For generic orthogonal matrices $\{Q_1, \dots, Q_s\}$, the collapse condition needs A to commute with all products $Q_j Q_i^\top$. These products are generic rotations whose commutant structure depends on their specific eigenspace decomposition. We do not know how to bound this commutant: it can be large if the Q_i share eigenstructure. But the HR family gives a dimension-independent bound ($\dim_{\mathbb{R}}(D)/d_{\text{irr}}^2$ depends only on s, not on d).

From linear cores to nonlinear cores. Thm. 4.3 says that expert collapse forces the core into a tiny algebraic subspace. Since STODA uses nonlinear tensor-train cores, the theorem applies through local linearization: collapse imposes a global equivariance condition $H_i^\top J_\phi(H_i x)\, H_i = J_\phi(x)$, coupling Jacobians at orthogonal points in the latent space. At the origin, this reduces to the Clifford commutant condition. Globally, it creates gradient diversity (not automatically offered by generic linear maps). While this is a linear regime, it isolates the structural bias in the nonlinear model.

Apart from collapse, we also need to understand if *mixing* distinct experts preserves conditioning.

> *When the router forms a token-dependent mixture of expert outputs, what object is being mixed, and where does HR cancellation still matter?*

The answer is that the HR structure shows up most cleanly in the *local linearization* of the routed module. Fix a token x and consider the routed adapter (omitting the shared expert for clarity) with $\mathcal{S}(x)$ being the top-k expert set:

$$F(x) = \sum_{i \in \mathcal{S}(x)} g_i(x)\, H_i^\top\, \phi_i(H_i x), \tag{7}$$

Each expert's Jacobian is a conjugated core Jacobian, $J_{E_i}(x) = H_i^\top J_{\phi_i}(H_i x)\, H_i$. This is the *same* conjugation structure Thm. 4.3 uses for collapse, now used in mixing. Differentiating F gives

$$J_F(x) = \sum_{i\in\mathcal{S}} g_i\, H_i^\top J_{\phi_i}(H_i x) H_i + \underbrace{\sum_{i\in\mathcal{S}} E_i(x)\,(\nabla g_i(x))^\top}_{\text{router-derivative terms}}. \tag{8}$$

Here, (8) separates the mixed Jacobian into a structural view-mixing term and a router-derivative term. The first term captures the stable geometric mixing of the expert cores, which is our main focus. The router term contributes additional local sensitivity, bounded in scale by the router logits (e.g., by $\|W_g\|_2$ for a linear router), and can be controlled by temperature or logit regularization.

Each H_i is orthogonal, so conjugation preserves spectral norms ($\|H_i^\top A H_i\|_2 = \|A\|_2$), and the mixture is a sum of *norm-preserving* conjugations rather than unrelated Jacobians. To see what HR structure controls, we decompose each core Jacobian into an isotropic and an anisotropic part:

$$J_{\phi_i}(H_i x) = \beta_i(x)\, I + \Delta_i(x), \qquad \beta_i(x) = \tfrac{1}{d}\,\mathrm{tr}\big(J_{\phi_i}(H_i x)\big), \tag{9}$$

where $\Delta_i(x)$ is trace-free. The structural term of the mixture then becomes

$$\sum_{i\in\mathcal{S}} g_i\, H_i^\top J_{\phi_i}(H_i x) H_i = \underbrace{\bar\beta\, I}_{\text{isotropic}} + \underbrace{\sum_{i\in\mathcal{S}} g_i\, H_i^\top \Delta_i H_i}_{D\,=\,\text{anisotropic residual}}, \tag{10}$$

where $\bar\beta = \sum_i g_i \beta_i$. The isotropic component is perfectly conditioned. The anisotropic residual D satisfies $\|D\|_2 \le \sum_i g_i \|\Delta_i\|_2$ by norm invariance and triangle inequality, but this bound holds for *any* orthogonal views. The question that determines mixing quality is whether the anisotropic residuals from different experts interfere constructively (inflating $\|D\|$) or cancel out?

Cross-term structure under HR mixing. This is where anti-commutativity enters. Expanding $\|D\|_F^2$, the answer depends on cross-terms $X_{ij} = \langle H_i^\top \Delta_i H_i,\, H_j^\top \Delta_j H_j\rangle_F$ between distinct experts: $\|D\|_F^2 = \sum_{i\in\mathcal{S}} g_i^2 \|\Delta_i\|_F^2 + 2\sum_{\substack{i<j\\ i,j\in\mathcal{S}}} g_i\, g_j\, X_{ij}$. Each cross-term reduces to a conjugation by the HR product $P_{ij} = H_j H_i$: $X_{ij} = \langle \Delta_i,\ P_{ij}^\top \Delta_j\, P_{ij}\rangle_F$. Anti-commutativity forces P_{ij} to be skew-symmetric, which means $A \mapsto P_{ij}^\top A\, P_{ij}$ has balanced ± 1 eigenspaces of dimension $d^2/2$ each. For a generic perturbation, the projections onto these eigenspaces roughly cancel, driving X_{ij} toward zero, see Prop. A.1 in the Appendix.

Remark. Existing exact compositionality results rely on constrained algebraic settings. Tian (2025) obtains a global composition result for Abelian-group reasoning because that task admits a Fourier decomposition which diagonalizes the objective. Large-scale diffusion models admit no such decomposition. We instead give algebraic statements through Prop. 4.1, Prop. 4.2, Thm. 4.3 and identify the local constraint that the nonlinear architecture inherits from these results. Finally, note that the Jacobians above measure local sensitivity across the token space at a fixed point in parameter space: they describe function space, not evolution over training steps. This geometric structure prevents expert redundancy prior to any gradient update step. This separates STODA from diversity induced by a training heuristic such as an auxiliary load-balancing loss.

Table 1: Subject-driven generation on DreamBooth (SD 1.5). * = metrics from prior work. STODA performs on par with, or exceeds all baselines at ∼6–15× fewer parameters.

Method	#params	DINO↑	CLIP-I↑	CLIP-T↑	LPIPS↑
Real Images	-	0.703	0.864	-	0.695
DreamBooth*	859M	0.614	0.778	0.239	0.737
LoRA*	25.2M	0.613	0.765	0.237	0.744
OFT*	23.3M	0.632	0.785	0.237	0.746
STODA	1.6M/4.1M	**0.634**	**0.801**	**0.242**	0.743

5 Experiments

We evaluate STODA on three tasks: subject-driven generation (§5.1), style fine-tuning (§5.2), and human preference alignment (§5.3). For our experiments, STODA uses 1 shared and 7 routed experts with random HR assignments. While our mixing theory generally applies to top-k routing, we evaluate under stricter top-1 routing. We treat this as a conservative stress test: eliminating the



Figure 6: Subject-driven generation (SD 1.5). STODA preserves subject identity but follows the prompt (at a fraction of parameter cost). Models trained once per subject.

capacity benefits of multi-expert blending strictly isolates and demonstrates the inherent efficacy of the individual HR orthogonal views. We use two tensor-train cores per expert (bond dimension 2), and HR view assignments drawn at random. For the experimental dimensions, $\rho(320) = 12$ (SD 1.5, max. 11 generators) and $\rho(1152) = 16$ (PixArt-Σ, max. 15 generators), so our 7 routed experts are comfortably within both bounds. Additional experiments/ablations/datasets are in the appendix.

5.1 Preserve Subject Identity under Strict Parameter Constraints

We fine-tune attention blocks of Stable Diffusion v1.5 (Rombach et al., 2022) on the DreamBooth dataset (Ruiz et al., 2022) using the reconstruction and prior preservation loss as described in Ruiz et al. (2022). We train for 1000 iterations with $\lambda = 1$ ($\text{lr} = 3 \times 10^{-3}$). We use the evaluation setup from Qiu et al. (2024) and report the baseline numbers from their work. Table 1 presents the subject fidelity (DINO, CLIP-I), prompt fidelity (CLIP-T), and diversity (LPIPS) of the generated images. STODA performs on par or better with 1.6M active / 4.1M total parameters: around $15\times$ fewer active parameters than LoRA and OFT.

Fig. 6 shows some representative results. We see that while DreamBooth overfits to subject appearance, LoRA loses subject identity (backpack, bowl) and OFT distorts subjects (dog, cat). STODA preserves both subject detail and prompt fidelity at a lower cost per token.



Figure 5: DPO-aligned generation (PixArt-Σ). STODA gives better subject staging, contrast, and prompt binding compared to LoRA and OFT.

5.2 Style Fine-tuning/Adaptation

We fine-tune three architectures: PixArt-Σ (Chen et al., 2024), SANA (Xie et al., 2024), and Nitro-T (Haridas et al., 2025) on five style datasets from CIVITAI (mnemic, 2024) (BarbieCore, Cyberpunk, ElementFire, Expedition, Hornify). All attention matrices are adapted. We provide all the hyperparameters used for this experiment in §M.2. PixArt-Σ uses a diffusion loss. SANA and Nitro-T use flow matching. Table 2 shows that STODA achieves the best average FID across all three architectures with a competitive CLIP score, at the same total parameter count as LoRA but $\sim$50% lower active parameters per token.

Table 2: Style fine-tuning. STODA achieves the best average FID with a competitive CLIP score.

	Method	Active params	Total params	Barbie		Cyberpunk		Element		Expedition		Hornify		Average	
				FID↓	CLIP↑	FID↓	CLIP↑	FID↓	CLIP↑	FID↓	CLIP↑	FID↓	CLIP↑	FID↓	CLIP↑
PixArt	LoRA	10.3M	10.3M	180.4	**30.2**	118.8	28.6	149.5	26.7	134.9	26.4	169.9	26.6	150.7	27.7
	OFT	12.3M	12.3M	170.5	30.1	123.3	29.0	165.7	**27.6**	132.6	**26.8**	164.9	**26.9**	151.4	**28.0**
	MixLoRA	6.2M	10.3M	175.9	30.1	120.2	28.6	166.3	26.6	137.4	26.6	174.5	26.5	154.8	27.6
	FlyLoRA	5.7M	10.3M	173.1	29.8	121.0	28.8	155.9	26.8	133.5	26.3	169.9	26.7	150.6	27.7
	STODA	5.7M	10.3M	**170.3**	29.6	**111.9**	**29.1**	**131.6**	27.0	**127.2**	26.5	**151.3**	26.5	**138.5**	27.8
SANA	LoRA	10.3M	10.3M	189.5	31.9	127.8	**29.6**	129.5	26.6	145.5	27.6	177.1	27.9	153.9	28.7
	OFT	12.3M	12.3M	184.3	31.7	124.7	29.4	133.4	**27.7**	140.8	27.9	164.3	27.7	149.5	**28.9**
	MixLoRA	6.2M	10.3M	186.8	**32.0**	127.3	29.5	130.9	27.1	145.7	27.8	176.5	28	153.4	28.8
	FlyLoRA	5.7M	10.3M	186.3	31.5	124.2	29.3	131.7	27.1	138.2	27.3	164.2	27.6	148.9	28.5
	STODA	5.7M	10.3M	**183.4**	**32.0**	**124.1**	29.5	**128.0**	26.6	**139.1**	**28.3**	**163.2**	**28.1**	**147.6**	**28.9**
NitroT	LoRA	10.3M	10.3M	255.2	26.1	200.8	25.7	245.2	24.3	239.8	23.5	213.6	25.5	230.9	25.0
	OFT	12.3M	12.3M	256.6	26.6	198.2	25.9	242.5	24.8	229.1	24.3	214.7	25.9	228.2	25.5
	MixLoRA	6.2M	10.3M	255.9	26.2	203.4	25.7	245.9	**25.0**	240.3	23.6	213.5	25.3	231.8	25.2
	FlyLoRA	5.7M	10.3M	254.5	26.3	199.5	25.8	247.8	24.6	**224.4**	24.2	214.5	25.6	228.1	25.3
	STODA	5.7M	10.3M	**253.1**	**27.0**	**197.8**	**26.0**	**240.9**	24.3	229.5	**24.5**	**212.6**	**26.0**	**226.8**	**25.6**

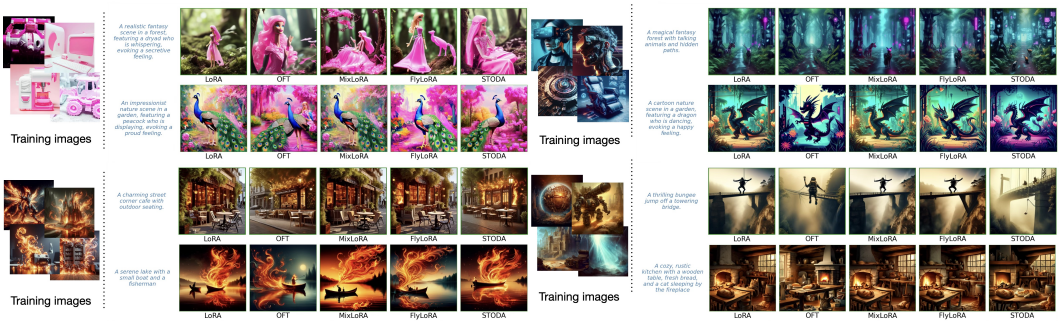


Figure 8: Style fine-tuning with different methods. STODA shows stronger style alignment and prompt adherence than baselines. Best viewed in color, zoomed in.

Also, STODA improves on recent MoE-PEFT baselines, including MixLoRA (Li et al., 2024) and FlyLoRA (Zou et al., 2026), on average. This shows that the diversity induced by HR views provides more stable expert specialization than the statistical or implicit routing heuristics used in prior work. Fig. 8 shows visually that STODA produces better style alignment and color fidelity while adhering closely to the prompt.

We note that the absolute FID scores in this setting are relatively high. This is a natural result of evaluating on the small-scale reference sets (156–440 images) standard in style adaptation literature (Hu et al., 2025). The relative margin between PEFT methods, corroborated by CLIP and FID scores, provides a comparison of the performance of various methods.

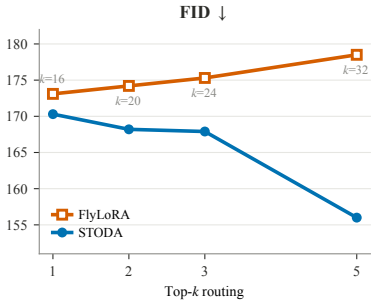


Figure 7: Multi-expert scaling. STODA consistently improves as more experts are activated.

5.3 Human Preference Alignment with DPO

We fine-tune PixArt-Σ with the Diffusion-DPO objective (Wallace et al., 2023) on the Pick-a-Pic dataset (Kirstain et al., 2023) to align generations with human preferences. Adapters are added to all attention matrices while the backbone model remains frozen. Fig. 5 shows that STODA gives better subject staging (centered composition, stronger contrast) and stricter prompt binding (accurately representing complex subjects) compared to LoRA and OFT, but maintains parameter efficiency. Table 5 in §G reports human-preference metrics (ImageReward, Aesthetic-v2, PickScore, HPSv2) on these generations, where STODA ranks first on three of the four (coming in a close second on Aesthetic-v2).

Expert utilization and diversity. Fig. 10 shows the routing distribution across experts. With **no** auxiliary load-balancing loss, STODA achieves near-uniform expert utilization. Beyond utilization, we also measure functional diversity directly: over a batch of 1024 evaluation tokens we sample the local expert Jacobians $J_{E_i}(x)$ and compute their normalized Frobenius inner product $\mathrm{Sim}(J_i, J_j) = \langle J_i, J_j \rangle_F / (\|J_i\|_F \, \|J_j\|_F)$ for $i \neq j$, obtaining $0.0007 \pm 0.0003 \approx 0$. Balanced routing is therefore accompanied by genuinely distinct expert functions, not merely balanced traffic. We show that HR structure plays a crucial role in expert diversity by comparing it to other orthogonal matrices in §D.

Multi-expert mixing. Our main experiments use top-1 routing, which already activates two experts (E_0 plus one routed). To test whether the advantage persists as more experts mix, we change $k = 1, 2, 3,$ and 5 and compare against FlyLoRA at matched active-parameter budgets (Fig. 7). STODA improves monotonically in FID, whereas FlyLoRA degrades over the same range. Added capacity therefore helps only when the additional experts are distinct; without this, mixing more active experts causes destructive interference.

Application to Language Models. To test whether the benefits of STODA extend to Language modeling, we instruction-tune Qwen-3 4B on the Alpaca dataset and evaluated on the LM Evaluation Harness. STODA obtains the best average score on eight benchmarks, against LoRA at roughly 27% fewer total and 63% fewer active parameters. The full table is available in §6.

6 Related Work

Parameter-Efficient Fine-Tuning. PEFT methods adapt large foundational models without full retraining by injecting lightweight, trainable parameters into a frozen backbone. Techniques such as Adapters (Houlsby et al., 2019; Pfeiffer et al., 2020; Karimi Mahabadi et al., 2021), Prefix-Tuning (Li & Liang, 2021; Lester et al., 2021), and Low-Rank Adaptation (Hu et al., 2022) were initially popularized in natural language processing. Recently, these methods have seen widespread use in text-to-image diffusion models. Frameworks like DreamBooth (Ruiz et al., 2022), ControlNet (Zhang et al., 2023), and T2I-Adapter (Mou et al., 2024) leverage PEFT for efficient subject personalization and spatial conditioning of massive U-Net and DiT architectures.

Mixture-of-Experts in PEFT. The MoE paradigm scales capacity by routing tokens to specialized subnetworks (Fedus et al., 2022; Dai et al., 2024), an approach extended to low-rank adaptation (Zadouri et al., 2023; Dou et al., 2024; Li et al., 2024; Zou et al., 2026). A persistent challenge across these architectures is expert collapse, where the router converges to favoring a few experts and the imbalance is self-reinforcing (Shazeer et al., 2017). To prevent collapse, existing MoE-PEFT methods use auxiliary load-balancing losses to enforce uniform utilization (Li et al., 2024; Zadouri et al., 2023; Shazeer et al., 2017) or learned diversity (Zou et al., 2026). In contrast, STODA introduces algebraic diversity. By imposing structural constraints that guarantee the geometric distinctness of expert inputs before any learning occurs, STODA is the first MoE-PEFT method to derive its anti-collapse guarantee directly from a closed algebraic structure (Clifford algebra $\mathrm{Cl}(0, s)$).

Hurwitz-Radon Matrices. The Hurwitz-Radon (HR) family is primarily used in the algebraic problem of composing quadratic forms (Shapiro, 2000). The HR family is linked to Clifford algebra $\mathrm{Cl}(0, s)$, anti-commutation of generators gives the backbone for STODA's expert diversity. This aligns with recent work leveraging geometric algebra in ML, such as the use of Clifford layers (Pence et al., 2025; Brandstetter et al., 2022; Xia et al., 2025) and using "convex lifting" of non-convex models using Clifford algebra (Pilanci, 2024; Wang et al., 2024). There is also interesting work on efficiency under the *Look Ma, No Matrices* framework (De Keninck, 2024) in computer graphics, substituting dense compute with structured algebraic operations.

7 Conclusions

We described STODA, a structured MoE adapter for diffusion model fine-tuning that uses the HR family to provide each expert with an algebraically distinct view of the input token. The view–core–unview design gives provable guarantees: pointwise orthogonal expert inputs (Prop. 4.1), non-degenerate view mixing (Prop. 4.2), structural resistance to expert collapse via Clifford algebra theory (Thm. 4.3), and balanced cross-term suppression under expert mixing (Prop. A.1). Empirically, STODA matches or outperforms LoRA, OFT, etc. across subject-driven generation, style adaptation, and preference alignment at a fraction of the active parameter cost. Deploying STODA on other modalities (language, video) and to pretraining-scale MoE remains an interesting direction.

AI USE STATEMENT

We used AI tools to integrate the core implementation into optimized workflows in order to run large-scale experiments. We also used AI tools to correct grammatical errors in our text. They were not used to generate research ideas, design experiments, or to perform analysis.

ETHICS STATEMENT

All of our experiments use publicly available models and datasets. We collect no new data and do not use human subjects. The goal of our work is to study the application of algebraic objects in image generation models and to make fine-tuning more efficient. So, any risks of general fine-tuning apply here as well. We are not aware of any ethical risks specific to our method.

REPRODUCIBILITY STATEMENT

We have aimed to make the work reproducible by providing the main methodological details, experimental setup, implementation choices, and evaluation protocol in the paper and appendix. We will release code and configuration details upon acceptance.

REFERENCES

J. F. Adams. Vector fields on spheres. *Annals of Mathematics*, 1962. URL http://www.jstor.org/stable/1970213.

Johannes Brandstetter, Rianne van den Berg, Max Welling, and Jayesh K Gupta. Clifford neural layers for pde modeling. *arXiv preprint arXiv:2209.04934*, 2022.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

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 *ECCV*, 2024.

Zixiang Chen, Yihe Deng, Yue Wu, et al. Towards understanding the mixture-of-experts layer in deep learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23049–23062. Curran Associates, Inc., 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 *Proceedings of the 42nd International Conference on Machine Learning*, Proceedings of Machine Learning Research, pp. 10010–10024. PMLR, 2025.

Damai Dai, Chengqi Deng, Chenggang Zhao, et al. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models, 2024. URL https://arxiv.org/abs/2401.06066.

Tri Dao. fast-hadamard-transform: Fast hadamard transform for pytorch, 2024. URL https://github.com/Dao-AILab/fast-hadamard-transform.

Steven De Keninck. Look, ma, no matrices! In *ACM SIGGRAPH 2024 Talks*, SIGGRAPH '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400705151. doi: 10.1145/3641233.3665801. URL https://doi.org/10.1145/3641233.3665801.

Shihan Dou, Enyu Zhou, Yan Liu, Songyang Gao, Jun Zhao, Wei Shen, Yuhao Zhou, Zhiheng Xi, Xiao Wang, Xiaoran Fan, Shiliang Pu, Jiang Zhu, Rui Zheng, Tao Gui, Qi Zhang, and Xuanjing Huang. Loramoe: Alleviate world knowledge forgetting in large language models via moe-style plugin, 2024. URL https://arxiv.org/abs/2312.09979.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity, 2022. URL https://arxiv.org/abs/2101.03961.

Timur Garipov, Sebastiaan De Peuter, Ge Yang, et al. Compositional sculpting of iterative generative processes. In *Advances in Neural Information Processing Systems*, volume 36, pp. 12665–12702. Curran Associates, Inc., 2023.

A. V. Geramita and N. J. Pullman. A theorem of hurwitz and radon and orthogonal projective modules. *Proceedings of the American Mathematical Society*, 42, 1974. ISSN 00029939, 10886826.

Akash Haridas, Tong Shen, Jingai Yu, Dong Zhou, Dong Li, Vikram Appia, and Emad Barsoum. Nitro-t: Training a text-to-image diffusion model from scratch in 1 day, July 2025.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp, 2019. URL https://arxiv.org/abs/1902.00751.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Teng Hu, Jiangning Zhang, Ran Yi, Hongrui Huang, Yabiao Wang, and Lizhuang Ma. Sara: High-efficient diffusion model fine-tuning with progressive sparse low-rank adaptation, 2025. URL https://arxiv.org/abs/2409.06633.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 03 1991.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, et al. Mixtral of experts, 2024. URL https://arxiv.org/abs/2401.04088.

Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. Compacter: Efficient low-rank hypercomplex adapter layers. In *Advances in Neural Information Processing Systems*, volume 34, pp. 1022–1035, 2021.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=G5RwHpBUv0.

Kee Yuen Lam and Paul Yiu. Linear spaces of real matrices of constant rank. *Linear Algebra and its Applications*, 1993. URL https://www.sciencedirect.com/science/article/pii/002437959390257O.

Yann LeCun, Corinna Cortes, and Christopher JC Burges. The mnist database of handwritten digits. *http://yann.lecun.com/exdb/mnist/*, 1998.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, et al. {GS}hard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations*, 2021.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, November 2021.

Dengchun Li, Yingzi Ma, Naizheng Wang, and et al. Mixlora: Enhancing large language models fine-tuning with lora-based mixture of experts, 2024. URL https://arxiv.org/abs/2404.15159.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Association for Computational Linguistics*. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.353. URL https://aclanthology.org/2021.acl-long.353/.

mnemic. Lora datasets / training data list + civitai dataset guide, mar 2024. URL https://civitai.com/articles/2138/lora-datasets-training-data-list-civitai-dataset-guide. Accessed: 2026-03-03.

Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2I-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023. URL https://arxiv.org/abs/2212.09748.

Travis Pence, Daisuke Yamada, and Vikas Singh. Composing linear layers from irreducibles. In *Neural Information Processing Systems*, 2025.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, et al. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, October 2020. doi: 10.18653/v1/2020.emnlp-demos.7.

Mert Pilanci. From complexity to clarity: Analytical expressions of deep neural network weights via clifford algebra and convexity. *Transactions on machine learning research*, 2024.

Zeju Qiu, Weiyang Liu, Haiwen Feng, et al. Controlling text-to-image diffusion by orthogonal finetuning, 2024. URL https://arxiv.org/abs/2306.07280.

Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arxiv:2208.12242*, 2022.

Christoph Schuhmann. LAION-aesthetics predictor v2. https://github.com/christophschuhmann/improved-aesthetic-predictor, 2022.

Jennifer Seberry. *Orthogonal Designs*. Springer International Publishing, 2017.

Daniel B. Shapiro. *Compositions of Quadratic Forms*. De Gruyter, Berlin, New York, 2000. ISBN 9783110824834. doi: doi:10.1515/9783110824834.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer, 2017. URL https://arxiv.org/abs/1701.06538.

Yuandong Tian. Composing global solutions to reasoning tasks via algebraic objects in neural nets. In *Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=tD7MLq0dbZ.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization, 2023. URL https://arxiv.org/abs/2311.12908.

Yifei Wang, Sungyoon Kim, Paul Chu, Indu Subramaniam, and Mert Pilanci. Randomized geometric algebra methods for convex neural networks. *arXiv preprint arXiv:2406.02806*, 2024.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

Tianxiang Xia, Max Neuwinger, and Lin Xiao. Fast clifford neural layers. *arXiv preprint arXiv:2507.01040*, 2025.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. Sana: Efficient high-resolution image synthesis with linear diffusion transformers, 2024. URL https://arxiv.org/abs/2410.10629.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Zeyue Xue, Guanglu Song, Qiushan Guo, et al. Raphael: Text-to-image generation via large mixture of diffusion paths. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

An Yang, Baosong Yang, Binyuan Hui, et al. Qwen2 technical report, 2024. URL https://arxiv.org/abs/2407.10671.

Ted Zadouri, Ahmet Üstün, Arash Ahmadian, et al. Pushing mixture of experts to the limit: Extremely parameter efficient moe for instruction tuning, 2023. URL https://arxiv.org/abs/2309.05444.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3836–3847, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. St-moe: Designing stable and transferable sparse expert models, 2022. URL https://arxiv.org/abs/2202.08906.

Heming Zou, Yunliang Zang, Wutong Xu, Yao Zhu, and Xiangyang Ji. FlyloRA: Boosting task decoupling and parameter efficiency via implicit rank-wise mixture-of-experts. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=nGQLYn13Xf.

APPENDIX

In this Appendix, we provide the proofs for the theory and additional details related to the experiments reported in the main paper. In Section A, we derive the proofs for the propositions and theorem stated in the main paper. Section B provides the algorithms for the STODA Adapter module and the experts, showing the steps where the view-core-unview operations are performed. In Section C, we describe how STODA handles different dimensions in modern architectures. In Section D, we replace the HR views with different orthogonal matrices and observe that HR gives the lowest expert overlap and the best FID. Section E demonstrates the stability of training with STODA using a few examples from subject-based generation. In Section F, we show that the idea of HR based expert design is not just restricted to fine-tuning, but applies to pre-training as well. Section G reports human preference metrics for the DPO-aligned models, where STODA ranks first on three of the four metrics and a close second on the Aesthetic-v2 score. In Section H, we instruction-tune a language model with STODA and show that the benefits of HR based experts are not limited to image generation. Section I compares a single STODA expert against LoRA at equal trainable parameter counts. We observe that the margin widens as the budget grows. In Section J, we vary the random seed, the number of routed experts, and the router initialization, and observe that STODA is not sensitive to the HR view assignment. Section K analyzes the performance of STODA as a function of the bonding dimension. We observe that the performance improves in general as we increase the bonding dimension. In Section L, we compare the computational performance of HR matrices and random orthogonal matrices and observe that HR matrices perform much better than random orthogonal matrices. Section M provides additional details and generated images that were omitted in the main text. In Section N, we discuss the limitations of our theory and experiments, including the gap between the linear core analysis and the nonlinear model we implement.

A PROOFS FOR THE PROPOSITIONS AND THEOREM 1

Proof of Proposition 4.1. For any $x \in \mathbb{R}^d$ and H_i in the HR family, we have

(i)

$$\begin{aligned}\langle H_i x, x\rangle &= x^\top H_i^\top x \\ &= -x^\top H_i x\end{aligned}$$

Since H_i is skew-symmetric, $x^\top H_i x = 0$ for all x.

(ii)

$$\begin{aligned}\langle H_i x, H_j x\rangle &= x^\top H_i^\top H_j\, x \\ &= -x^\top (H_i H_j)\, x\end{aligned}$$

Now $(H_i H_j)^\top = H_j^\top H_i^\top = (-H_j)(-H_i) = H_j H_i = -H_i H_j$, so $H_i H_j$ is skew-symmetric and the quadratic form vanishes.

(iii)

$$\begin{aligned}\|H_i x\|^2 &= x^\top H_i^\top H_i\, x \\ &= x^\top x && \text{since } H_i^\top H_i = I. \\ &= \|x\|^2\end{aligned}$$

□

Proof of Proposition 4.2. By definition of an HR family, we have $H_i^\top H_i = I$, $H_i^\top = -H_i$, and $H_i H_j = -H_j H_i$ for all $i \neq j$. This implies $H_i^2 = -I$. Now, we expand $M^\top M$ using $H_i^\top = -H_i$. Distributing the terms, the linear cross-terms cancel exactly, giving us

$$\begin{aligned}M^\top M &= \alpha_0^2 I + \alpha_0 \sum_{j=1}^{N} \alpha_j H_j - \alpha_0 \sum_{i=1}^{N} \alpha_i H_i - \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j H_i H_j \\ &= \alpha_0^2 I - \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j H_i H_j\end{aligned}$$

Table 3: Summary of theoretical guarantees. The first three results hold for all routing regimes; cross-term control becomes active under top-k with $k \geq 2$.

Result	Guarantee	Regime
Prop. 4.1	Pointwise orthogonal expert inputs	All
Prop. 4.2	Scaled isometry under view mixing	All
Thm. 4.3	Collapse $\Rightarrow$ core in $\sim$1.6% of param. space	All
Prop. A.1	Cross-term suppression via balanced eigenspaces	≥ 2 experts (including E_0)

We split the remaining sum into diagonal ($i = j$) and off-diagonal ($i \neq j$) components:

$$M^\top M = \alpha_0^2 I - \sum_{i=1}^{N} \alpha_i^2 H_i^2 - \sum_{i<j} \alpha_i \alpha_j (H_i H_j + H_j H_i)$$

Applying $H_i^2 = -I$ and the anti-commutativity property $H_i H_j + H_j H_i = 0$, the off-diagonal terms vanish:

$$M^\top M = \alpha_0^2 I + \sum_{i=1}^{N} \alpha_i^2 I = \Big(\sum_{i=0}^{N} \alpha_i^2 \Big) I$$

This completes the proof. □

Proof of Theorem 4.3. Since $E_0(x) = Ax$, the condition $E_0 = E_i$ requires $H_i^\top A\, H_i = A$ for all i. Using $H_i^\top = -H_i$, this means $(-H_i)\, A\, H_i = A$. We left-multiply by $H_i^\top = -H_i$:

$$\begin{aligned} (-H_i)(-H_i)\, A\, H_i &= (-H_i)\, A \\ \Longrightarrow \qquad H_i^2\, A\, H_i &= -H_i\, A \\ \Longrightarrow \qquad -A\, H_i &= -H_i\, A, \end{aligned}$$

using $H_i^2 = -I$. Therefore $A\, H_i = H_i\, A$ for all i: the matrix A commutes with every generator. The set of such matrices forms the commutant of the representation $\{H_1, \ldots, H_s\}$ on $\mathbb{R}^d$. By the Artin–Wedderburn theorem, this commutant has dimension at most $(d/d_{\text{irr}})^2 \cdot \dim_{\mathbb{R}}(D)$, where $D \in \{\mathbb{R}, \mathbb{C}, \mathbb{H}\}$ is the Schur division algebra determined by $s \bmod 8$. For $s = 7$ (used in experiments): $d_{\text{irr}} = 8$ and $D = \mathbb{R}$, giving a commutant of dimension at most $(d/8)^2$, which is $\frac{1}{64} \approx 1.6\%$ of the full parameter space d^2. □

Cross-term control for top-k HR mixing. While the structural views in STODA are pointwise orthogonal, stable mixing further requires that the anisotropic residuals of different experts do not interfere constructively. This is guaranteed by the anti-commutativity of the HR family, which ensures that the superoperator $\mathcal{C}_{P_{ij}}$ acts as an isometric involution with balanced eigenspaces of equal dimension. Consequently, cross-terms are suppressed toward zero, maintaining a well-conditioned mixed Jacobian and stable gradient flow even in a top-1 routing regime where the shared expert E_0 and a routed expert E_i interact. More precisely

Proposition A.1 (Cross-term control under top-k HR mixing)**.** *Let $\{H_1, \ldots, H_s\}$ be an HR family. For each pair $i \neq j$, let $P_{ij} = H_j H_i$. Then:*

(i) *P_{ij} is skew-symmetric and orthogonal, with $P_{ij}^2 = -I$.*

(ii) *The superoperator $\mathcal{C}_{P_{ij}}$ is an isometric involution on $\mathbb{R}^{d\times d}$ with eigenvalues ± 1. The $+1$ and -1 eigenspaces have equal dimension $d^2/2$.*

(iii) *For identical core perturbations $\Delta_i = \Delta_j = \Delta$, the cross-term decomposes as*

$$X_{ij} \;=\; \|\Delta_+^{ij}\|_F^2 \;-\; \|\Delta_-^{ij}\|_F^2, \tag{11}$$

where $\Delta = \Delta_+^{ij} + \Delta_-^{ij}$ is the decomposition into the $+1$ and -1 eigenspaces of $\mathcal{C}_{P_{ij}}$.

Proof. (i) Skew-symmetry:

$$\begin{aligned} P_{ij}^\top = (H_j H_i)^\top &= H_i^\top H_j^\top \\ &= (-H_i)(-H_j) \\ &= H_i H_j \\ &= -H_j H_i = -P_{ij} \end{aligned}$$

, using anti-commutativity. Orthogonality: $P_{ij}^\top P_{ij} = (-P_{ij})P_{ij} = -P_{ij}^2$. And

$$\begin{aligned} P_{ij}^2 &= H_j H_i H_j H_i \\ &= -H_j H_j H_i H_i \\ &= -(-I)(-I) = -I \end{aligned}$$

, again using anti-commutativity in the first step. So $P_{ij}^\top P_{ij} = I$.

(ii) Since P_{ij} is orthogonal, $\mathcal{C}_{P_{ij}}$ is an isometry on $(\mathbb{R}^{d\times d}, \|\cdot\|_F)$. The involution property $\mathcal{C}_{P_{ij}}^2 = \mathrm{id}$ follows from $(P_{ij}^\top)^2 = P_{ij}^2 = -I$; $\mathcal{C}_{P_{ij}}(\mathcal{C}_{P_{ij}}(A)) = (P_{ij}^\top)^2\, A\, P_{ij}^2 = (-I)A(-I) = A$. So eigenvalues are ± 1. For the dimension count: in vectorized form, $\mathcal{C}_{P_{ij}}$ acts as $P_{ij}^\top \otimes P_{ij}^\top$, whose trace is $(\operatorname{tr} P_{ij}^\top)^2 = (\operatorname{tr} P_{ij})^2$. Since P_{ij} is skew-symmetric, $\operatorname{tr}(P_{ij}) = 0$.

Thus $\operatorname{tr}(\mathcal{C}_{P_{ij}}) = \dim(V_+) - \dim(V_-) = 0$, giving $\dim(V_+) = \dim(V_-) = d^2/2$.

(iii) Let $\Delta = \Delta_+ + \Delta_-$ and $\mathcal{C}_{P_{ij}}(\Delta) = \Delta_+ - \Delta_-$. We have

$$\begin{aligned} \langle \Delta,\, \mathcal{C}_{P_{ij}}(\Delta)\rangle_F &= \langle \Delta_+ + \Delta_-,\, \Delta_+ - \Delta_-\rangle_F \\ &= \|\Delta_+\|_F^2 - \|\Delta_-\|_F^2, \end{aligned}$$

since $V_+ \perp V_-$ (eigenspaces of a self-adjoint operator). □

The balanced eigenspace property $\dim(V_+) = \dim(V_-)$ is the HR-specific mechanism: because the eigenspaces are of equal dimension, a generic traceless Δ projects roughly equally onto both, driving X_{ij} toward zero. For random orthogonal views $\{Q_i\}$, the products $Q_j Q_i^\top$ are not skew-symmetric, so the eigenspaces can be unbalanced and cross-terms systematically positive. This causes constructive interference in the anisotropic residual. When the isotropic part dominates ($|\bar{\beta}| > \|D\|_2$), the structural term of the mixed Jacobian has condition number $\kappa(\bar{\beta}\, I + D) \leq (|\bar{\beta}| + \|D\|_2)/(|\bar{\beta}| - \|D\|_2)$. So, the HR cross-term suppression keeps κ close to 1, ensuring stable gradient flow through the mixture. Table 3 summarizes our theoretical results and the regimes they apply to.

B Algorithms for STODA Adapter and Expert

The algorithms for the full STODA adapter module and the individual experts are presented in Alg. 1 and Alg. 2, respectively.

C Hardware-Aligned Dimensionality and Arbitrary Dimensions

As noted in Section 3, the Radon-Hurwitz number $\rho(d)$ dictates the maximum size of an orthogonal tangent frame on S^{d-1}. While $\rho(d)$ theoretically restricts the number of orthogonal generators for odd or prime dimensions, this constraint aligns well with modern hardware properties. Foundation models universally employ hidden dimensions that are highly composite—typically multiples of 64 or 128—to maximize Tensor Core utilization and ensure memory coalescing during dense GEMM operations. Because these dimensions are inherently divisible by large powers of 2, they natively yield high Radon-Hurwitz numbers without any architectural modifications.

For the exceedingly rare case of an arbitrary or odd dimension, STODA applies trivial zero-padding to reach the nearest hardware-optimal power of 2 before the view transform and then truncating the output afterward. This technique is commonly used in efficient sequence transformations (Dao, 2024) and ensures the algebraic structure is preserved with negligible computational overhead.

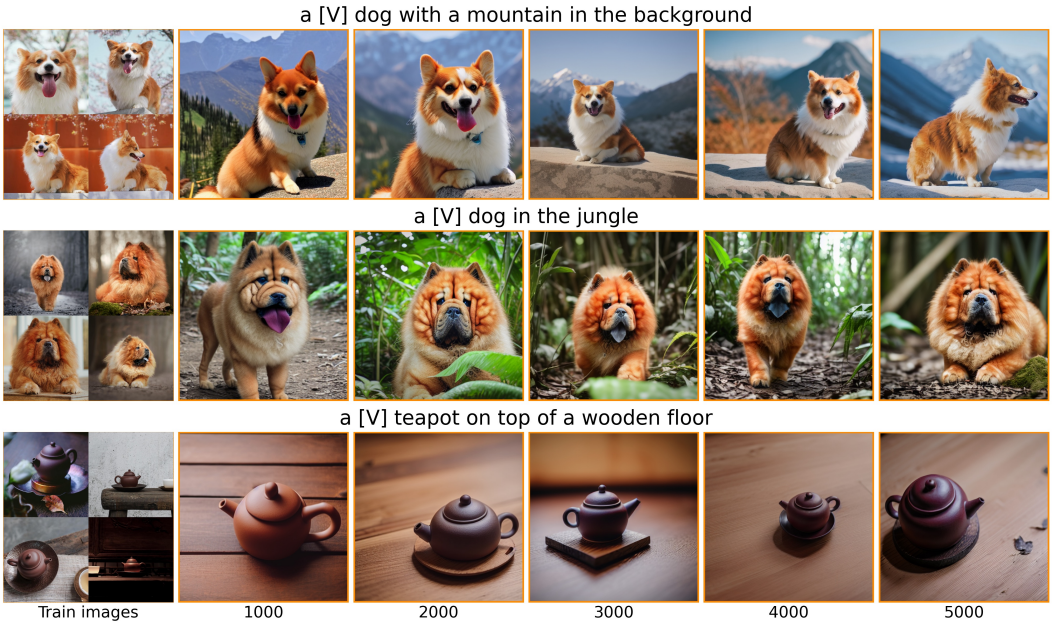


Figure 9: Training stability of STODA. STODA maintains consistent subject features and prompt adherence over extended training horizons without learning rate annealing.

Algorithm 1 STODA

Require: Input x, principal expert E_0, residual experts $\{E_i\}_{i=1}^{N}$, router G, pretrained weights W, sparsity k

1: Pretrained layer output $y_0 = Wx$
2: Principal expert output $e_0 = E_0(x)$
3: Gate logits $\hat{g} = G(x)$
4: Top-k selection $\mathcal{S} = \textsc{TopK}(\hat{g},\, k)$
5: Normalized gate weights $g_i = \dfrac{\exp(\hat{g}_i)}{\sum_{j\in\mathcal{S}} \exp(\hat{g}_j)}, \quad \forall i \in \mathcal{S}$
6: Sparse residual output $y_\Delta = e_0 + \sum_{i\in\mathcal{S}} g_i\, E_i(x)$
7: Layer output $y = y_0 + y_\Delta$
return y

D Does the HR structure matter?

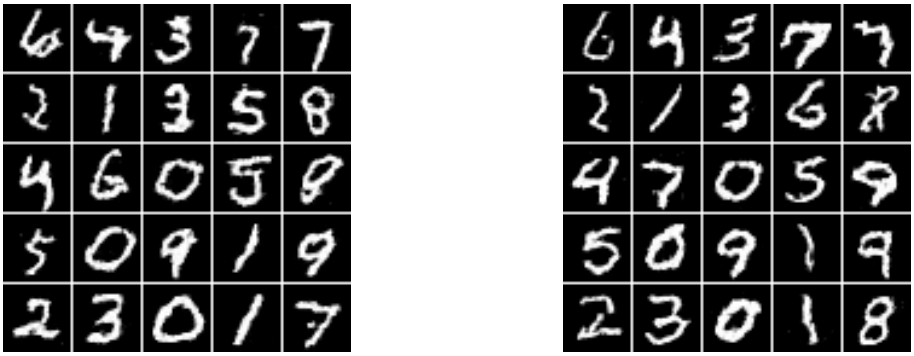
To isolate the HR family's contribution from the rest of the architecture, we fine-tune PixArt-Σ on the BarbieCore dataset, replacing the HR view transforms with dense random orthogonal matrices, random sign flips, and the identity map (an unconstrained MoE), while holding every other component fixed. To measure functional specialization alongside generation quality, we compute the mean off-diagonal cosine similarity between expert outputs. Table 4 shows that HR views yield both the lowest expert overlap and the best FID. Unconstrained sign flips and random rotations lack guaranteed cross-term cancellation, whereas the skew-symmetry and anti-commutativity of the HR generators enforce geometric distinctness by construction. The small CLIP difference is the standard style-dominance effect in diffusion fine-tuning, caused by the adapter's increased capacity.

Table 4: Ablation on the expert view transform (PixArt-Σ, BarbieCore). HR shows better FID and expert diversity.

Transformation	FID↓	CLIP↑	Avg. Sim↓
Random orthogonal	171.3	29.7	0.00891
Random sign flips	173.5	30.0	0.00564
Identity	173.8	**30.1**	0.00594
HR (STODA)	**170.3**	29.6	**0.00315**

Algorithm 2 STODA Expert

Require: Input x, HR matrix H, tensor trains $\{tt_i\}_{i=1}^{T}$
1: Rotate input, $h = Hx$
2: For tt_i in $\{tt_i\}_{i=1}^{T}$:
$h = tt_i(h)$
3: Undo the rotation, $y = H^T h$
return y



(a) Generated images by the HR-MoE model

(b) Generated images by the dense DiT model

Figure 11: Images generated by the DiT models trained on MNIST. (a) the HR-MoE model uses an MoE architecture with HR based experts, (b) the dense model uses the same depth and hidden dimensions as the HR-MoE model, but with a dense MLP.

E Training Stability

In this experiment, we evaluate the stability and convergence behavior of STODA during fine-tuning. For this, we use a constant learning rate of 1e-4 throughout training, without annealing, and train for longer. We choose a few subjects from the DreamBooth dataset and observe the generated images as training progresses. Figure 9 shows these generated images at 1000-step intervals from 1000 to 5000 steps. STODA demonstrates rapid capture of the subject's identity within 1000 steps and remains consistent throughout training. The structural features of the subject and the semantic integration of the background information from the prompt remain well preserved. The anti-commutativity and skew-symmetry of the HR generators provide controlled cross-term suppression during expert mixing. This ensures a well-conditioned optimization landscape, preventing catastrophic overfitting and texture burning.

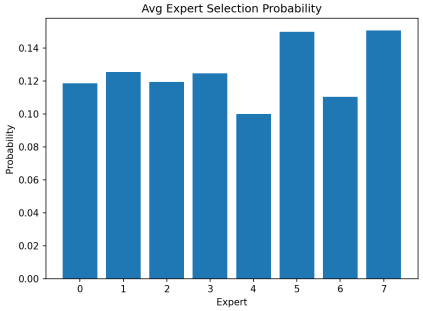


Figure 10: Expert selection probability across routed experts, showing near-uniform utilization (no auxiliary balancing losses!).

F Applying HR-Experts to DiT Pre-Training

In this experiment, we investigate whether the HR-based expert design is effective for pre-training Diffusion Transformers (Peebles & Xie, 2023). To evaluate this, we pre-train a model to generate images using the MNIST dataset (LeCun et al., 1998). The architecture consists of a four-layer Diffusion Transformer with a hidden size of 64 and a patch size of 4. We train the model for 20,000 steps using a batch size of 32 and a learning rate of 1×10^{-3}.

To integrate our method, we replace the standard dense MLP layers with HR-based MoE layers, configuring each expert with a hidden dimension of 64. This setup results in a total parameter count of 384K, of which 284K are active during each forward pass.

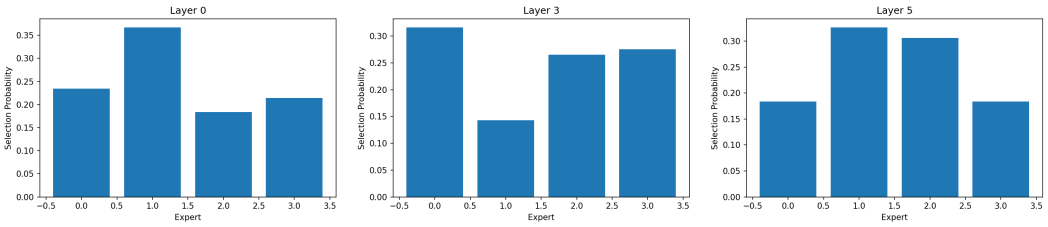


Figure 12: The expert selection probabilities at three different layers in a DiT trained on MNIST. We can see that each expert is selected close to a uniform distribution, and there is no expert collapse.

During training, we use a total of four experts and apply a top-2 routing strategy for each token. Figure 12 illustrates the routing probabilities for each expert. We can observe that the expert selection probabilities across three randomly selected layers demonstrate that the experts are routed close to a uniform distribution without any expert collapse. Finally, Figure 11a displays the generated images at the end of the pre-training phase, confirming the viability of our approach for pre-training Diffusion Transformers. For comparison, we show the images generated by a dense DiT model with the same depth and the hidden dimension in figure 11b. The dense DiT model has a total parameter count of 483K parameters. Figure 11 shows that the generated quality of the HR-MoE model is quite similar to the dense model even at a lower parameter count.

G Quantitative Evaluation of DPO Alignment

Here we score the PixArt-Σ generations with four human-preference metrics: ImageReward (Xu et al., 2023), the LAION aesthetic predictor v2 (Schuhmann, 2022), PickScore (Kirstain et al., 2023) and HPSv2 (Wu et al., 2023). Table 5 reports the results. STODA ranks first on ImageReward, PickScore and HPSv2, and is within 0.007 of the best aesthetic score. The clearest margin is on ImageReward ($+0.018$ over LoRA, $+0.010$ over OFT); the PickScore and HPSv2 gaps are narrower. We used the same parameter ratios as the style transfer experiment. STODA obtains these scores with fewer active parameters than either baseline.

Table 5: Automatic human-preference metrics for DPO-aligned generation (PixArt-Σ, Pick-a-Pic). Higher is better for all four metrics.

Method	**ImageReward↑**	**Aesthetic-v2↑**	**PickScore↑**	**HPSv2↑**
LoRA	0.9734	**5.7972**	22.4030	0.2934
OFT	0.9820	5.7349	22.3496	0.2932
STODA	**0.9915**	5.7905	**22.4102**	**0.2950**

H Application to Language Models

Table 6 reports the full benchmark breakdown for the instruction-tuning experiment described in the main text. STODA obtains the best average score and the best result on seven of the eight benchmarks.

Table 6: Instruction tuning Qwen-3 4B on Alpaca, evaluated with the LM Evaluation Harness. STODA uses one shared and four routed experts with top-1 routing.

Method	**Active / Total**	**ARC-c**	**ARC-e**	**BoolQ**	**HellaSwag**	**OBQA**	**PIQA**	**RTE**	**WinoG.**	**Avg**
Pretrained	–	54.01	78.58	**85.11**	68.42	40.00	74.92	75.81	66.06	67.86
LoRA (r=16)	5.9/5.9M	54.35	80.05	84.89	68.37	40.40	76.22	76.53	66.61	68.43
OFT (block 64)	5.8/5.8M	54.18	79.76	85.05	68.68	40.60	76.33	76.17	66.69	68.43
STODA	2.2/4.3M	**55.80**	**82.20**	83.82	**70.92**	**41.60**	**76.88**	**77.98**	**67.32**	**69.57**

I MATCHED-PARAMETER COMPARISON WITH LORA

A rank-r LoRA update satisfies $\mathrm{rank}(\Delta W) \leq r$, whereas a Tensor-Train parameterization constrains tensor-network ranks rather than the matrix rank of the represented operator, so a compact TT core need not correspond to a low-matrix-rank update. To test whether this matters at equal budget, we fine-tune PixArt-Σ on BarbieCore with LoRA at ranks 20, 32 and 64, and STODA at bond dimensions chosen to match the same trainable-parameter counts. STODA is better at all three budgets, and the margin widens from 9.1 to 16.9 FID as the budget grows, i.e. the advantage is present at the level of the individual expert parameterization and is not solely a product of the multi-expert construction.

Table 7: Matched trainable-parameter comparison (PixArt-Σ, BarbieCore).

Trainable params	Rank / bond dim	LoRA FID↓	LoRA CLIP↑	STODA FID↓	STODA CLIP↑	FID gap
10.3M	20	180.4	**30.2**	171.3	29.9	9.1
16.5M	32	178.1	**30.2**	164.6	29.8	13.5
33.0M	64	177.1	30.1	**160.2**	29.8	**16.9**

J SENSITIVITY TO SEEDS, EXPERT COUNT AND ROUTER INITIALIZATION

All runs in this section fine-tune PixArt-Σ on BarbieCore with the settings of the main paper. Across five random seeds, which reshuffle both the HR view assignment and the router initialization, STODA attains CLIP 29.95 ± 0.21 and FID 169.29 ± 1.32; performance is therefore not sensitive to which HR generator a given expert receives. Table 8 varies the number of routed experts: FID improves monotonically as experts are added, at the cost of a larger stored adapter. Table 9 varies the router initialization, where a Gaussian scheme with a small scale is slightly better than the default. Sensitivity to the tensor-train bond dimension is reported separately in §K. The non-monotonic behavior of the CLIP score is a direct result of the style dominance effect caused by the increased capacity of the adapter.

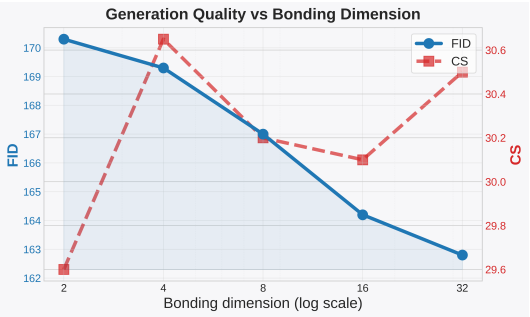


Figure 13: Performance of STODA as a function of bonding dimension. We see a consistent improvement in FID score as the bonding dimension increases and an overall improvement in CLIP score compared to the bonding dimension of 2.

Table 8: Number of routed experts.

Routed experts	CLIP↑	FID↓	Adapter size
7	**29.6**	170.3	46 MB
12	29.2	164.5	67 MB
15	29.3	**153.4**	83 MB

Table 9: Router initialization.

Initialization	CLIP↑	FID↓
Default (baseline)	29.6	170.3
Gaussian s=0.1	**30.0**	**168.5**
Gaussian s=0.6	29.7	168.6
Orthogonal s=0.1	29.8	169.1

K PERFORMANCE VS BONDING DIMENSION

In this experiment, we study the impact of bonding dimension on the performance of STODA. Bonding dimension differentiates Tensor Trains from the Kronecker product by allowing information flow between the tensor cores. To evaluate the effect, we fine-tune the PixArt-Σ model on the BarbieCore style and measure the FID and the CLIP score of the generated images.

Figure 13 shows the performance of STODA as a function of bonding dimension ranging from 2 to 32. We see that the FID score consistently improves as we increase the bonding dimension,

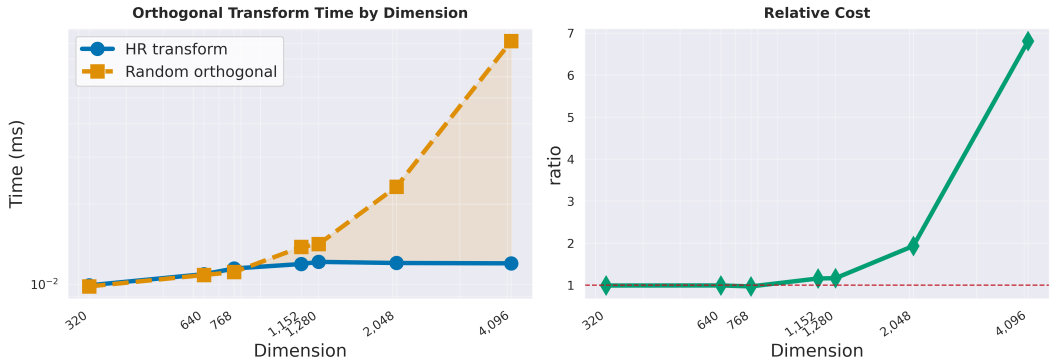


Figure 14: Time taken for HR vs random orthogonal transformation on the left and their relative cost on the right.

showing that the model can use the additional expressivity to align the generated image distribution more closely with the target style. While the CLIP score exhibits some fluctuation, it shows a clear overall improvement compared to the baseline score of 29.60 at a dimension of 2. This demonstrates that increasing the bonding dimension enhances prompt-image alignment alongside the gains in distributional fidelity.

L HR vs Random Orthogonal Transform Time

In this experiment, we compare HR vs Random Orthogonal transform in terms of computational cost. We measure the time for the common attention dimensions found in Stable Diffusion v1.5 and DiT models. All the times are measured on an NVIDIA-A100 machine. For a random orthogonal transform, we need $\mathcal{O}(d^2)$ computations while HR only needs $\mathcal{O}(d \log d)$ owing to the efficient Kronecker structure of HR matrices. Figure 14 shows the time taken by the HR and random orthogonal transform and their relative cost (ratio of the two times). We can see that at very low dimensions, both the methods perform similarly, but as the dimension increases, HR performs much better than random orthogonal transform, almost $7\times$ faster, at a dimension of 4096.

M Experiment Details

In this section, we provide additional details about the experiments that we omitted in the main text due to space constraints. We also provide additional images generated by the models. Note that we do not retrain the models again. We use the same models that were used in the main text and provide additional generated images.

M.1 Subject-driven Generation

The goal of the experiment is to learn new objects from only a few examples of the object shown to the model. To this end, we learn a unique identifier to anchor the object in the model's text space and fine-tune the attention matrices that feed this information into the generation pipeline. We use Stable Diffusion v1.5 (Rombach et al., 2022) as the pretrained network and evaluate STODA and other methods such as DreamBooth (Ruiz et al., 2022), LoRA (Hu et al., 2022), and OFT (Qiu et al., 2024).

Table 10: Dataset sizes of different style datasets.

Dataset	#images
BarbieCore	316
Cyberpunk	440
ElementFire	156
Expedition	396
Hornify	236

Dataset We use the DreamBooth dataset (Ruiz et al., 2022), which contains a collection of 30 subjects that include unique objects and pets such as animal-themed backpacks, toys, animals, cartoons, etc. For evaluation, they have 25 prompts, which are divided as 20 recontextualization prompts (placing the object in different scenarios) and 5 property modification prompts (changing color, texture, etc.) for objects, and 10 recon-

textualization, 10 accessorization (changing outfit, accessories, etc.), and 5 property modification prompts for live subjects.

Objective We follow the loss function in Dreambooth (Ruiz et al., 2022) — diffusion reconstruction loss and the prior preservation loss.

$$\mathbb{E}_{x,c,\epsilon,\epsilon',t}[w_t||\hat{x}_\theta(\alpha_t x + \sigma_t \epsilon, c)||_2^2 + \lambda w_{t'}||\hat{x}_\theta(\alpha_{t'} x_{pr} + \sigma_{t'}\epsilon', c_{pr}) - x_{pr}||_2^2] \tag{12}$$

Metrics We evaluate subject fidelity - which measures how well the fine-tuned model preserves the details in the subject, prompt fidelity - which describes how well the generated images match the description in the text prompt and generation diversity, which measures the variance in the generated images. For image fidelity, we use CLIP-I (Radford et al., 2021), DINO (Caron et al., 2021) scores which measure the average pairwise cosine similarity of CLIP and ViT S/16 DINO embeddings of the generated and real images, respectively. For text fidelity, we measure the CLIP-T (Radford et al., 2021) score, which is the average cosine similarity of the clip embeddings of the text prompt and the generated images. To evaluate the diversity in the generated images, we use LPIPS (Zhang et al., 2018), which calculates the cosine similarity between the generated images of a given subject for a given prompt.

M.2 Fine-Tuning for Different Styles

Dataset In this experiment, we fine-tune Text-to-Image models on five widely used datasets for fine-tuning on CIVITAI (mnemic, 2024), namely BarbieCore, Cyberpunk, ElementFire, Expedition, and Hornify. Table 10 shows the size of these datasets. Here, the goal is to learn new styles from the downstream dataset. We use models that are available on the Hugging Face hub that are designed for consumer devices, specifically, Pixart-Σ (Chen et al., 2024), SANA (Xie et al., 2024), and Nitro-T (Haridas et al., 2025).

Objective To adapt the models to these specific aesthetic domains, we optimize the adapters by minimizing the discrepancy between the model predictions and ground-truth target. Pixart-Σ is trained with the diffusion objective, so we minimize the loss function

$$\mathcal{L}_{style} = \mathbb{E}_{x,c,\epsilon,t}\left[|\epsilon - \epsilon_\theta(z_t, t, c)|_2^2\right] \tag{13}$$

where z_t is the style-conditioned latent at timestep t, ϵ is the Gaussian noise added to the clean latent x, c represents the style-specific text embedding, and ϵ_θ denotes the predicted model. Conversely, SANA (Xie et al., 2024) and Nitro-T (Haridas et al., 2025) were trained with the flow matching objective, so, we optimize the adapters for these models by minimizing the error between predicted and ground-truth vector field (velocity):

$$\mathcal{L}_{style} = \mathbb{E}_{x,c,\epsilon,t}\left[\|(x-\epsilon) - v_\theta(z_t, t, c)\|_2^2\right] \tag{14}$$

where v_θ is the velocity predicted by the flow model.

Hyperparameters Across all three backbones reported in the paper, we adapt all the attention projections. We use the AdamW optimizer with a weight decay of $1e-2$ and a constant learning rate scheduler. We pick the learning rate for STODA between $\{1e-4, 1e-3\}$, and the baselines get the best learning rate between $5e-5$ and $3e-4$. We use a batch size of 16 for all the experiments. We train for 2000 optimizer steps with gradient clipping applied with an ℓ_2 norm of 1.

Metrics We fine-tune all the attention matrices with different methods and measure the CLIP score and FID of the generated images and the images in the training set.

Figure 15, Figure 16 show additional generated images with different styles.

M.3 Direct Preference Optimization

In this experiment, we fine-tune the Pixart-Σ model (Chen et al., 2024) to align with human preferences using Diffusion-DPO (Wallace et al., 2023).

Dataset We use the Pick-a-Pic dataset (Kirstain et al., 2023), which contains text-to-image prompts paired with human preference judgments. This dataset provides bias towards vivid aesthetics, clear compositions, and prompt adherence, providing guidance for human preference alignment.

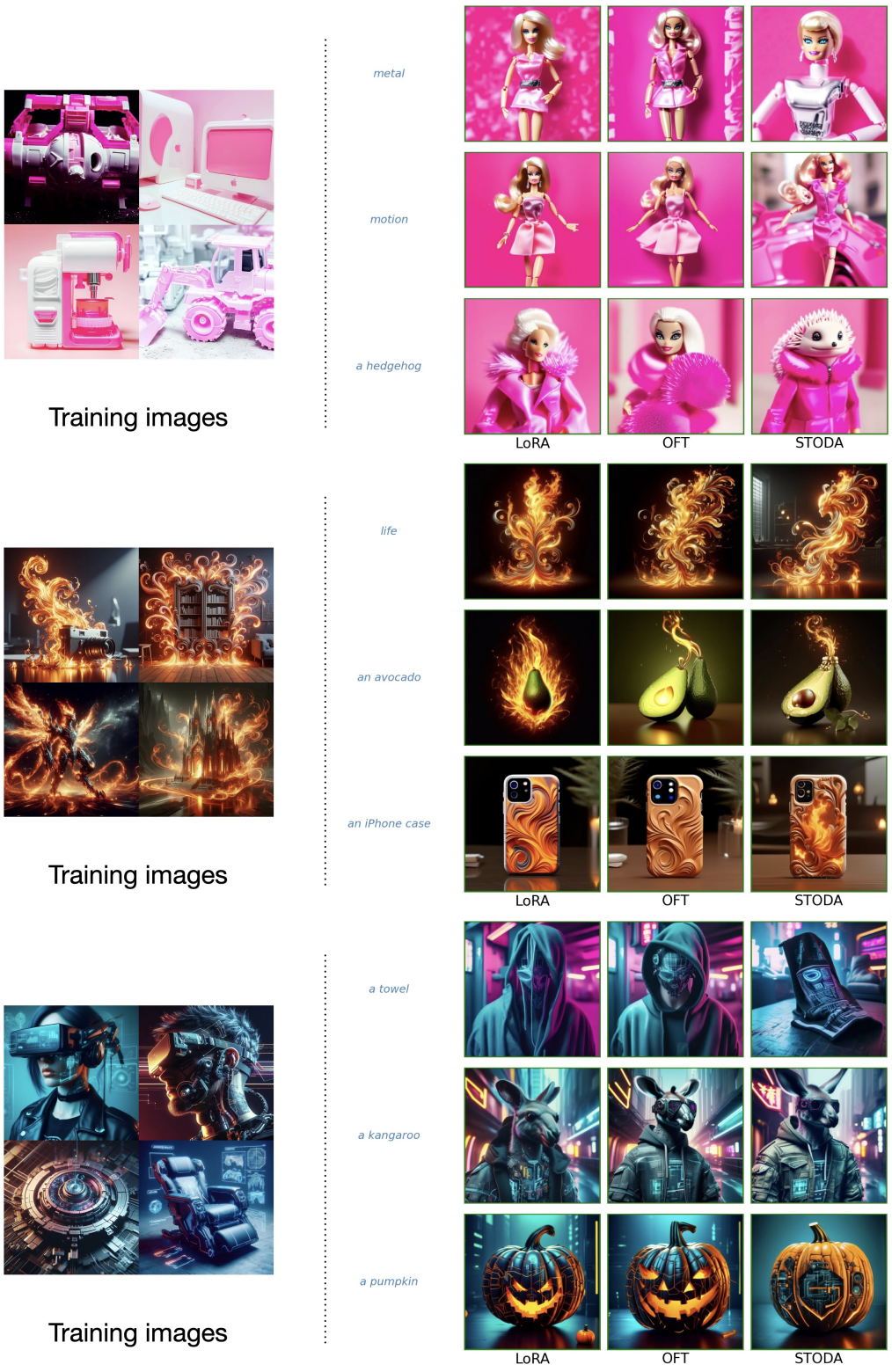


Figure 15: Example images generated by different methods for style-based generation.



Figure 16: Additional images for the style-based generation experiment.

Objective We use the Diffusion-DPO objective (Wallace et al., 2023) to align the model with the Pick-a-Pic dataset. Diffusion-DPO reformulates the standard DPO objective for continuous diffusion models, directly updating the policy to maximize the likelihood of the chosen image x_w and minimize the likelihood of the rejected image x_l.

$$\mathcal{L}_{DPO} = \mathbb{E}_{t,\epsilon,(x_w,x_l),c}\left[-\log\sigma\left(\beta\left(\|\epsilon-\epsilon_\theta(z_t^{(l)},t,c)\|_2^2-\|\epsilon-\epsilon_\theta(z_t^{(w)},t,c)\|_2^2\right)\right)\right] \tag{15}$$

where $z_t^{(w)}$ and $z_t^{(l)}$ are the noisy latents for the chosen and rejected images respectively, and β controls the strength of the reference model divergence penalty. We freeze the backbone model and add adapters to all the attention matrices and train them with the Diffusion-DPO objective. We use similar hyperparameters as the style transfer experiment.

N Limitations

The anti-collapse theorem is stated for linear cores. Thm. 4.3 characterizes collapse for a shared linear core A. STODA uses nonlinear tensor-train cores, and the theorem transfers only through local linearization: it identifies the algebraic obstruction the nonlinear model inherits at a point, not a global characterization of the learned function.

Cross-term suppression is exact only in a special case. Prop. A.1(iii) assumes identical core perturbations $\Delta_i = \Delta_j$. For distinct perturbations, we argue that a generic Δ projects roughly equally onto the two eigenspaces of $\mathcal{C}_{P_{ij}}$, which is a genericity argument rather than a bound.

Most results are single runs. We report seed variation only for PixArt-Σ on BarbieCore (§J). Due to our limited compute, we cannot do multiple runs of all the experiments.

The expert count is bounded by the dimension. $\rho(d) - 1$ caps the number of non-redundant views a layer can host. We use this as a design principle, but it also means the pool is limited for small or odd d, where we fall back to zero-padding (§C).