ETHER: Efficient Finetuning of Large-Scale Models with Hyperplane Reflections

Massimo Bini [1 2] **Karsten Roth** [3 2] **Zeynep Akata** [2 4 5] **Anna Khoreva** [6]

Abstract

Parameter-efficient finetuning (PEFT) has become ubiquitous to adapt foundation models to downstream task requirements while retaining their generalization ability. However, the amount of additionally introduced parameters and compute for successful adaptation and hyperparameter searches can explode quickly, especially when deployed at scale to serve numerous individual requests. To ensure effective, parameter-efficient, and hyperparameter-robust adaptation, we propose the *ETHER* transformation family, which performs *E*fficient fine*T*uning via *H*yp*E*rplane *R*eflections. By design, *ETHER* transformations require *a minimal number of parameters*, are *less likely to deteriorate model performance*, and exhibit *robustness to hyperparameter and learning rate choices*. In particular, we introduce *ETHER* and its relaxation *ETHER+*, which match or outperform existing PEFT methods with significantly fewer parameters (~10-100 times lower than LoRA or OFT) across multiple image synthesis and natural language tasks without *exhaustive hyperparameter tuning*. Finally, we investigate the recent emphasis on Hyperspherical Energy retention for adaptation and raise questions on its practical utility. The code is available at `https://github.com/mwbini/ether`.

1. Introduction

Recently, large-scale foundation models (Bommasani et al., 2021) have demonstrated impressive general-purpose capabilities across both generative and discriminative tasks (Rombach et al., 2022; Touvron et al., 2023a; OpenAI, 2023; Kirillov et al., 2023), showing extensive flexibility and strong performance when further adapted to different, more specialized tasks such as instruction following or controlled image synthesis (Zhang & Agrawala, 2023; Ruiz et al., 2022; Taori et al., 2023; Chiang et al., 2023).

While impressive, these capabilities come with parameter counts increasing into the billions (OpenAI, 2023; Podell et al., 2023a; Touvron et al., 2023b). To allow for affordable and scalable model adaptation that can serve large and diverse client bases, various techniques have been introduced in the literature. They range from full finetuning (Zhao et al., 2024; Zhang et al., 2023a; Stojanovski et al., 2022) to just a few layers of the pretrained model (Kornblith et al., 2019), concatenating additional learning modules (Houlsby et al., 2019; Pfeiffer et al., 2020; Mou et al., 2023), and more recently to adapters on the network weights with lightweight learnable transformations (Qiu et al., 2023; Hu et al., 2022; Kopiczko et al., 2023; Valipour et al., 2023). The latter have proven particularly effective, introducing no inference latency, fewer adaptation parameters, and strong performance.

Conceptually, these methods finetune on smaller datasets to adapt to downstream task and data requirements, without (1) compromising too much on the costly pretraining and (2) incurring concept and semantic drifts by catastrophically overwriting pretrained weights (Kirkpatrick et al., 2017; Lee et al., 2019; Lu et al., 2020; Mehta et al., 2022; Ruiz et al., 2023; Ke et al., 2023; Roth et al., 2024; Garg et al., 2024; Ibrahim et al., 2024). Treading the line for a suitable trade-off between adaptation and retention of the foundational model capabilities thus presents itself as a difficult task to tackle, often requiring costly tuning of hyperparameters such as learning rates. This problem is acknowledged explicitly in Li et al. (2018); Chen et al. (2023a); Gouk et al. (2021) aiming to preserve Euclidean weight distances between pretrained and finetuned models, and implicitly with approaches opting for both lower learning rates (at the cost of more tuning iterations) and inclusion of tuning parameters via summation (Qiu et al., 2023).

In particular, Qiu et al. (2023) hints that a Euclidean distance measure likely fails to fully capture the preservation of the network's ability, suggesting instead Hyperspherical Energy (HE) as an alternative measure. The resulting objective uses orthogonal transformations (OFT) for multiplicative weight changes that control HE. Still, even OFT requires

[1]Bosch IoC Lab, University of Tübingen [2]Helmholtz Munich [3]Tübingen AI Center, University of Tübingen [4]Technical University of Munich [5]Munich Center for Machine Learning [6]Bosch Center for Artificial Intelligence. Correspondence to: Massimo Bini <massimo.bini@uni-tuebingen.de>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

specific and restricted hyperparameter choices such as small learning rates and initialization from identity matrices to ensure sufficient knowledge preservation. In addition, while more robust and stable for finetuning in controllable generation settings compared to LoRA (Qiu et al., 2023), OFT comes with a high computational overhead due to matrix multiplication and a large number of tuning parameters.

In this work, we propose **E**fficient fine**T**uning via **H**yp**E**rplane **R**eflections (*ETHER*) - a new family of weight transformations, efficient in parameter count while preserving model abilities and being robust in convergence and learning rate choices. By default, *ETHER* transformations frame the tuning process as a search for suitable hyperplanes, along which weight vectors can be reflected based on the orthogonal Householder transformation (Householder, 1958). This keeps the distance to the transformation neutral element - the identity matrix - constant by construction and improves training stability while reducing the chance of deteriorating model performance. In addition, being built from single vectors, Householder transformations allow for efficient block-parallel matrix multiplication with minimal performance trade-offs.

However, situations may arise where the hard distance restriction of *ETHER* can prove suboptimal (such as for subject-driven image generation, where finegrained subject-specific semantics need to be retained). As such, we augment the *ETHER* family with *ETHER+* - a relaxation on the default *ETHER* method. More precisely, *ETHER+* derives from the Householder transformation, but breaks the orthogonality and constant distance constraints, introducing multiple hyperplanes that can interact with a weight vector. As a result, *ETHER+* allows for more controlled and finegrained adaptation, while still having a bounded distance to the transformation neutral element, and retaining the *ETHER* benefits of high parameter-efficiency, training stability, and hyperparameter robustness.

Indeed, across subject-driven image generation, controlled image synthesis, natural language understanding and instruction tuning tasks, we find that *ETHER* and especially *ETHER+* match and outperform existing methods using only a few additional tuning parameters (e.g. $100\times$ less than OFT when finetuning Stable Diffusion for controlled image synthesis) - all while presenting stronger learning rate robustness compared to other methods and consequently requiring minimal hyperparameter tuning to achieve strong performance (c.f. Sec. 4). Finally, we also utilize our experimental benchmark findings to further investigate and question the recent emphasis on transformation orthogonality and hyperspherical energy (HE) retention (e.g. Qiu et al. (2023)), showing how non-orthogonal *ETHER+* can achieve strong performance while displaying increased HE.

2. Related Work

Parameter-Efficient Finetuning (PEFT). PEFT of pre-trained models has seen different strategies evolve in the past years - starting from finetuning protocols and concatenation of learnable modules (Houlsby et al., 2019; Lester et al., 2021; Li & Liang, 2021; Pfeiffer et al., 2020; Guo et al., 2021) to more recently reparametrization of network weights with efficient transformations (Qiu et al., 2023; Hu et al., 2022; Kopiczko et al., 2023; Valipour et al., 2023; Zhang et al., 2023c). The latter have shown convincing trade-offs between adaptation quality, additional parameters, and inference latency. LoRA (Hu et al., 2022) transforms network weights by adding the result of a learnable, low-rank matrix product. On top of LoRA, multiple variations have been proposed, s.a. QLora (Dettmers et al., 2023) with quantized weights, AdaLoRA (Zhang et al., 2023c) with dynamic rank adjustment, and VeRA (Kopiczko et al., 2023) with low-rank frozen random projections and trainable vectors to reduce parameter counts. OFT (Qiu et al., 2023) instead learns matrix multiplier with orthogonality constraints to retain hyperspherical energy. In our work, we use the same paradigm but introduce hyperplane reflections for better parameter efficiency and learning rate robustness.

Controlling Diffusion Generative Models. Diffusion-based generative models show strong compositional generation (Rombach et al., 2022; Mukhopadhyay et al., 2023; Podell et al., 2023b; Karthik et al., 2023; Saharia et al., 2022). Among these, Gal et al. (2022); Ruiz et al. (2023) popularized personalized generation - teaching models to generate variations of user-provided samples. Based on DreamBooth (Ruiz et al., 2023), other works (Liu et al., 2023b; Richardson et al., 2023; Zhang et al., 2023e) followed. ControlNet (Zhang et al., 2023b) shows model controllability through external signals s.a. semantic and depth maps or face landmarks via extra layers at the cost of higher inference latency. Qiu et al. (2023) show controllability through direct finetuning with learnable matrix-multiplication transformations. Our work suggests an alternative, more robust and parameter-efficient approach through hyperplane reflections.

Instruction Tuning Language Models. Large Language Models (LLMs) have shown striking generalization across a wide range of tasks (Zhao et al., 2023; Zhang et al., 2023d; OpenAI, 2023; Touvron et al., 2023a). However, the default training objective often does not exactly match downstream task requirements and intentions. To address this mismatch, Instruction Tuning (Wang et al., 2023; Zhang et al., 2023d; Longpre et al.; Taori et al., 2023) finetunes LLMs using additional `(Instruction, Output)` pairs to explicitly align the model with human preferences. This enhances capabilities and controllability while avoiding costly retraining (Köpf et al., 2023). Recently, methods based on LoRA (Hu

et al., 2022) have been proposed to efficiently achieve this control (Dettmers et al., 2023; Xu et al., 2023; Chen et al., 2023b; Valipour et al., 2023; Kopiczko et al., 2023). This work proposes a strong alternative with further parameter-efficiency and high learning rate robustness.

3. Method

We first discuss adapter-based PEFT in §3.1, before describing and motivating the use of hyperplane reflections in *ETHER* (§3.2). To encourage flexibility in trainable control and adaptation, we propose a simple, yet effective relaxation *ETHER+* in §3.3. Finally, §3.4 describes block-diagonal *ETHER* for improved computational efficiency.

3.1. Preliminaries

Parameter-Efficient Finetuning with Adapters. The most commonly deployed form of PEFT with an adapter is *Low-rank Adaptation* (*LoRA*, Hu et al. (2022)). LoRA parametrizes a change of pretrained weights W as

$$(W + BA)^\intercal x + b$$

where BA is the matrix product of two low-rank matrices, i.e. for $W \in \mathbb{R}^{d\times f}$, $A \in \mathbb{R}^{d\times r}$ and $B \in \mathbb{R}^{r\times f}$. When rank $r << \min(d, f)$, this can bring down required tuning parameters significantly compared to full finetuning. In addition, BA can be absorbed into W during inference to avoid additional latency.

Orthogonal Finetuning (OFT). However, finetuning with LoRA can incur significant, potentially catastrophic weight changes. To ensure better preservation of pretrained model weights, Qiu et al. (2023) propose Orthogonal Finetuning (OFT). Based on the hypothesis that Hyperspherical Energy (HE) needs to be kept unaltered to preserve the original model abilities, OFT proposes the usage of multiplicative orthogonal transformations on the model weights. By retaining pairwise weight angles, HE can remain unaffected. However, to work in practice, Qiu et al. (2023) require the construction of the orthogonal matrix Q via a Cayley parametrization $Q = (I + S)(I - S)^{-1}$, where S is skew-symmetric. Notice that by using this parametrization, they limit the range of possible orthogonal matrices to those with determinant 1, missing orthogonal matrices with determinant equal to -1. As we show, this is relevant, as it excludes reflections, which motivate *ETHER*. To make OFT more parameter efficient, the orthogonal matrix $Q \in \mathbb{R}^{d\times d}$ is built in a block-diagonal fashion, made up of n smaller blocks Q^b of size $\frac{d}{n} \times \frac{d}{n}$. The final OFT transformation on the forward pass can then be described as

$$(Q^B W)^\intercal x + b$$

with block-diagonal Q^B. The trainable parameters are the n matrices $Q^b \in \mathbb{R}^{\frac{d}{n}\times\frac{d}{n}}$ that compose Q^B - more specifically the matrices R^b that build the skew-symmetric matrices $S^b = \frac{1}{2}(R^b - (R^b)^\intercal)$ for Q^b. For finetuning, the R^b are initialized as zero, such that $Q^B|_0 = I$ and consequently $Q^B|_0 W = W$ at the beginning of finetuning.

3.2. *ETHER*: Finetuning with Hyperplane Reflections

Fundamentally, *ETHER* (**E**fficient fine**T**uning via **H**yp**E**rplane **R**eflections) sets up weight transformations as hyperplane reflections. These reflections can be obtained via the Householder transformation matrix $H \in \mathbb{R}^{d\times d}$ with

$$H = I - 2uu^\intercal \tag{1}$$

with $u \in \mathbb{R}^d$ the hyperplane unit normal vector and the corresponding outer product $uu^\intercal$. The reflection can be easily intuited when applied to a weight vector $w \in \mathbb{R}^d$:

$$Hw = (I - 2uu^\intercal)w = w - 2u(u^\intercal w).$$

Transformation H effectively subtracts twice the component of w projected on u, thereby reflecting it with respect to the hyperplane defined by u (see Fig. 1). By construction, hyperplane reflections are well-suited for the efficient finetuning of pretrained models, as they keep the distance to the transformation neutral element - the identity matrix - constant, which minimizes the risk of divergence from the pretrained model and deterioration of model performance (c.f. Fig. 4). This can be easily shown by computing the Frobenius norm of the difference between the Householder matrix H and the identity matrix I:

$$\|H - I\|_F = \|I - 2uu^\intercal - I\|_F = 2 \cdot \|uu^\intercal\|_F = 2 \tag{2}$$

The above equation leverages the fact that for any matrix M

$$\|M\|_F = \sqrt{\mathrm{Tr}(MM^\intercal)}$$

and that with $M = uu^\intercal$ and u having unit length $u_1^2 + u_2^2 + ... + u_d^2 = 1$, one can simply write (with $(uu^\intercal)^\intercal = uu^\intercal$)

$$\|uu^\intercal\|_F = \sqrt{\textstyle\sum_{i=1}^d u_i^2} = 1.$$

Since the finetuning process simply consists of finding the optimal directions of the reflection hyperplanes with bounded deviations from the transformation neutral element, it allows for (i) a very *low number of extra parameters* corresponding to the unit vectors u, and (ii) the usage of high learning rates, as *the risk of divergence is minimized*. This allows for general *learning rate robustness* and encourages fast convergence by default, as consistently high learning rates can be selected; reducing computational resources required to achieve good performance (e.g. Fig. 6).

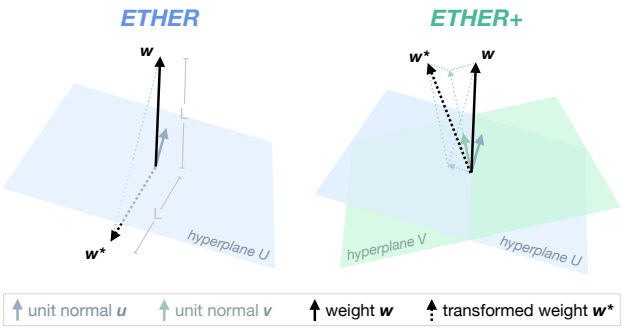

Figure 1. ***ETHER* and *ETHER+* sketches.** We visualize either a single hyperplane reflection for *ETHER* or two interacting hyperplanes for *ETHER+*, parametrized unit normals u (and v). Unlike *ETHER*, the final result of *ETHER+* does not have to retain the original length L, as the need for hard reflections is softened, and orthogonality is no longer guaranteed.

Interestingly, as this transformation is orthogonal ($HH^\intercal = I$), it falls under the umbrella of orthogonal transformations motivated in OFT (Qiu et al., 2023) from the perspective of Hyperspherical Energy control to better preserve model pretraining. However, OFT leverages the Cayley parametrization of orthogonal matrices, which only produces determinant 1 matrices. By construction, this excludes Householder matrices from OFT, which have determinant -1! However, as noted above, it is indeed in this particular setting and through the use of Householder transformations that high parameter efficiency, strong pretraining retention, and learning rate robustness arise.

On top of that, we further investigate the importance of Hyperspherical Energy retention by conducting a control study comparing OFT against its non-orthogonal variant (*Naive*)[1] Our experiments do not show significant differences in terms of control and training stability, suggesting that such properties stem from the multiplicative finetuning approach rather than the underlying HE retention, contrasting insights in Qiu et al. (2023) (c.f. Sec. 5.3). These findings partly motivate the exploration of a relaxed variant of the Householder reflection in the next section 3.3, which demonstrates that loosening the orthogonality constraint not only maintains good performance but can even lead to enhanced results.

3.3. Relaxing Orthogonality in *ETHER*

While finetuning via hyperplane reflections has several promising qualities as highlighted above, there is no free lunch. In particular, situations may arise where the strength of the transformation and inherent deviation from the identity may be too large by default, such as for potentially more nuanced tasks like subject-driven generation (Ruiz et al., 2023). To allow for more nuanced transformations while retaining beneficial properties of *ETHER* - parameter efficiency and learning rate robustness through bounded deviations from the transformation neutral element - we propose the *ETHER+* relaxation

$$H^+ = I - uu^\intercal + vv^\intercal$$

with unit vectors $u, v \in \mathbb{R}^d$. This is a simple variation of the Householder transformation that now allows for interaction between two distinct hyperplanes (see Fig. 1). This helps to control the transformation strength as $uu^\intercal$ and $vv^\intercal$ can weaken or even cancel each other out to return the identity transformation in the limit where $u = v$. In addition, the transformation distance remains bounded, as the relaxed variant H^+ always has $\|H^+ - I\|_F \leq 2$, i.e.

$$\max \left\|H^+ - I\right\|_F \leq \max \|H - I\|_F \,.$$

This follows immediately from the triangle inequality of norms, i.e. $\|vv^\intercal - uu^\intercal\|_F \leq \|vv^\intercal\|_F + \|uu^\intercal\|_F = 2$. Due to the weaker strength of this new transformation, we apply it both on the left (H^+) and right ($\tilde{H}^+$) of the weight matrix W, such that the forward pass becomes

$$\left(H^+ W \tilde{H}^+\right)^\intercal x + b.$$

Consequently, *ETHER+* effectively leverages a sequence of hyperplane interactions that no longer have to retain length to allow for more nuanced weight adjustment while still minimizing the risk of diverging from the pretrained model (as also shown e.g. in Figs. 3, 4, 5 and 6).

Figure 2. **Block-Parallel Computation scheme** between d-dimensional block-diagonal transformation with n blocks and a $d \times f$ -dimensional weight matrix W.

3.4. Efficient *ETHER* through Block-Parallelism

In multiplicative finetuning like OFT or *ETHER*, further computational load is introduced through additional matrix multiplications. To mitigate this issue, we introduce a block-diagonal formulation of *ETHER* similar to block-diagonal OFT described in §3.1. For this, we break down the Householder transformation H (eq. 1) into its corresponding

[1] *Naive* employs an unconstrained block-diagonal transformation matrix N^B made up of n blocks and initialized as an identity matrix, i.e. having the same number of trainable parameters and initialization as OFT's transformation matrix Q^B.

Table 1. **Better computational efficiency through block-diagonality** on **Phi1.5**-1.3B and **Llama-2**-7B, with internal dimensions of 2048 and 4096 respectively. As the number of blocks n increases, so does the computational efficiency, quantified by the decrease in TFLOPs required for a single backward pass (using a sample with longest sequence length). The larger the model's internal dimension, the larger the efficiency gain.

	Phi1.5-1.3B		**Llama-2-7B**	
	TFLOPs	rel. drop	TFLOPs	rel. drop
$\text{LoRA}_{r=8}$	6.04	-	6.85	-
$\text{OFT}_{n=256}$	9.13	-	25.26	-
$\textit{ETHER}_{n=1}$	9.13	-	25.26	-
$\textit{ETHER}_{n=4}$	7.07	-23%	12.07	-52%
$\textit{ETHER}_{n=32}$	6.71	-27%	8.22	-68%
$\textit{ETHER+}_{n=1}$	10.78	-	51.65	-
$\textit{ETHER+}_{n=4}$	7.69	-29%	18.66	-64%
$\textit{ETHER+}_{n=32}$	6.79	-37%	9.04	-83%

block-diagonal variant H^B:

$$\text{diag}(H^1 \cdots H^n) = I - 2 \begin{pmatrix} \hat{u}_1 \hat{u}_1^\intercal & & \\ & \ddots & \\ & & \hat{u}_n \hat{u}_n^\intercal \end{pmatrix}$$

with each i-th block-plane parameterized by $\hat{u}_i \in \mathbb{R}^{\frac{d}{n}}$. Of course, one can do the same for H^+. In both cases, such a block-diagonal formulation reduces the cost of computing H. More importantly, each i-th block now only affects the corresponding i-th block-row in the weight matrix W. This means we can split W into n sub-blocks $W^i \in \mathbb{R}^{\frac{d}{n} \times f}$, each of which is uniquely altered by its corresponding H^i counterpart. As a result, the full weight transformation can now be separated into smaller block-specific operations, reducing the overall number of computations. Furthermore, these operations can now be fully block-parallelized, significantly increasing training speed! In terms of computations, for each full-matrix-multiplication between H and W of sizes $d \times d$ and $d \times f$ respectively, $d(df)$ multiplications and $(d-1)df$ additions are necessary, accounting for $\mathcal{O}(d^2 f)$ operations. With our block-parallel scheme, we reduce these to n block-specific $\frac{d}{n}(\frac{d}{n} f)$ multiplications and $\frac{d-1}{n}(\frac{d}{n} f)$ additions, resulting in $\mathcal{O}(\frac{d^2 f}{n})$ operations (see Tab. 1).

Furthermore, with each block being built from a single vector of dimension $\frac{d}{n}$, *ETHER* transformations' construction ensures that the total number of trainable parameters remains constant for any n number of blocks. This stands in contrast to block-diagonal OFT, where the use of higher block counts was introduced to minimize the number of parameters while introducing noticeable decreases in adaptation performance! Instead, for block-diagonal *ETHER*, we find performance to be consistent over increasing block counts (see App. D), allowing for an improved computational footprint with negligible performance decrease.

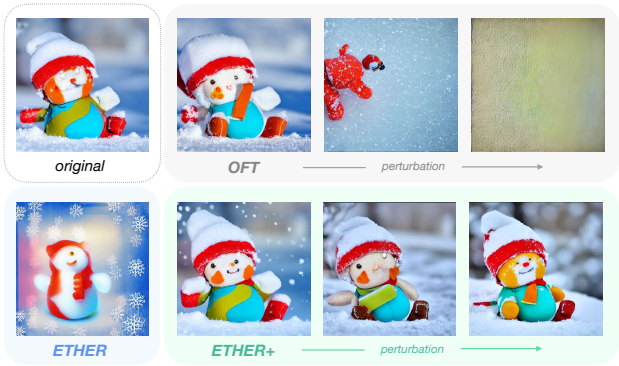

Figure 3. **Change in model behavior as a function of perturbation strength**, i.e. distance between weight transformation and identity matrix. As *ETHER* and *ETHER+* are upper-bounded in perturbation by construction, catastrophic deterioration of model performances is rarely encountered, and weight transformations remain controllable even for maximal deviations. For standard approaches, s.a. OFT, larger deviations from the identity matrix may occur during training and result in substantial divergence from the pretrained model. Notice also that by breaking orthogonality constraints in *ETHER+*, both smaller and stronger semantic variants can be learned.

4. Intriguing Properties of ETHER

This section investigates and highlights the bounded distance and non-deteriorating nature of *ETHER*/*ETHER+* in more detail while providing insights into its favorable learning rate robustness and the reliable use of high learning rates for fast convergence. For completeness, we also report here comparisons with the unconstrained *Naive* method, to better show the impact of orthogonality as proposed by Qiu et al. (2023), and how our method provides much stronger robustness. Finally, we include a discussion on the parameter efficiency. For all experiments in this section, please see §5.1 for relevant implementation details.

Non-Deteriorating Nature. Because both *ETHER* and *ETHER+* are upper-bounded in their possible perturbation over the pretrained weight matrices (as measured for example by the distance to the transformation neutral element, the identity matrix), finetuning with both methods will guarantee suitable results for most hyperparameter choices. This is easily visualized in Fig. 3 by looking at generation samples after perturbing Stable Diffusion with randomly sampled transformations for each approach - OFT, *ETHER* and *ETHER+* - respectively. While *ETHER* uses a fixed-distance transformation (c.f. Eq. 2) that introduces a noticeable change (but still retaining semantics), *ETHER+* can obtain both finegrained visual control as well as stronger semantic changes. Conversely, unbounded methods like OFT catastrophically deteriorate a model's generative abilities as the perturbation strength increases.

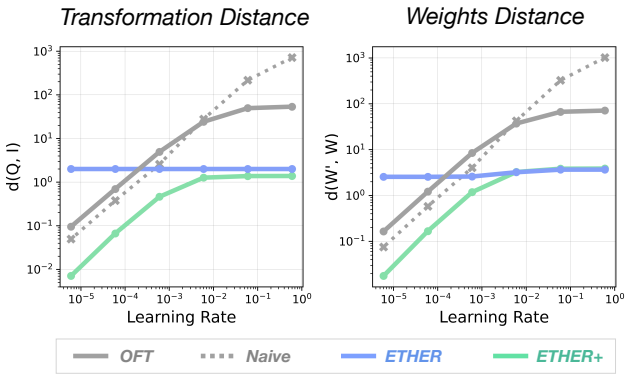

Figure 4. **Distances as a function of learning rates** between transformation and identity matrix (*Transformation Distance*), and finetuned and pretrained weights (*Weights Distance*). Distances obtained for subject-driven generation finetuning at convergence (1200 iterations). Results show distances magnitudes higher and unbounded for non-*ETHER* methods in both cases as learning rates increase.

This results in a much more controlled generation setting for *ETHER* and *ETHER+* finetuning. This is also depicted quantitatively in Fig. 4, which shows distances between the learned transformation and the transformed weights (at convergence) to the identity matrix and the pretrained weights, respectively, as a function of the learning rate. As can be seen, larger learning rate values for OFT and *Naive* finetuning (OFT without orthogonality constraints) result in distances that are orders of magnitude higher than those of *ETHER* and *ETHER+*, leading to catastrophic deterioration and model collapse (see Fig. 8 in App.).

Learning Rate and Hyperparameter Robustness. Practically, the non-deteriorating nature of *ETHER* and *ETHER+* manifests in learning rate robustness during finetuning. As the risks of divergence and collapse are minimized, training stability becomes much less dependent on the choice of learning rate. This is seen when evaluating performance (e.g. mIoU for controllable image synthesis in Fig. 5) and model convergence (Fig. 6) against learning rates. For non-*ETHER* methods, Fig. 5 shows significant performance drops for high learning rates, while Fig. 6 reveals fast convergence speeds for *ETHER+* with learning rates covering multiple magnitudes, much more general than e.g. OFT.

This means that not only can good performance be guaranteed for most learning rate choices, but fast convergence as well, with competitive results already after the first epoch. Since *ETHER* also only introduces a single hyperparameter, the number of diagonal blocks, which marginally impacts performance (c.f. §3.4), *ETHER* methods become very attractive for practical usage, as the need for grid-search and cautious low learning rate training for good performance (c.f. §1) is reduced.

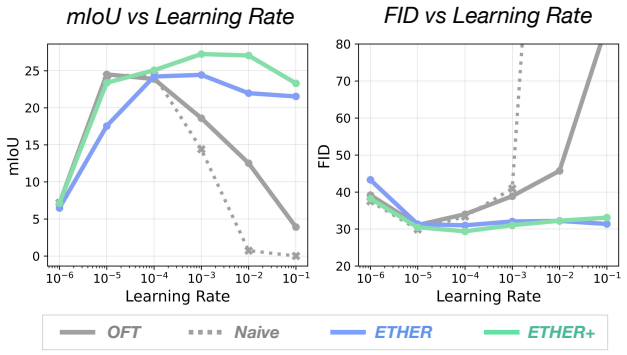

Figure 5. **mIoU and FID performances as a function of learning rates.** Results are obtained for controllable generation S2I finetuning on Stable Diffusion, and reveal a much stronger learning rate robustness of *ETHER*-based methods; retaining strong performance across entire learning rate magnitudes.

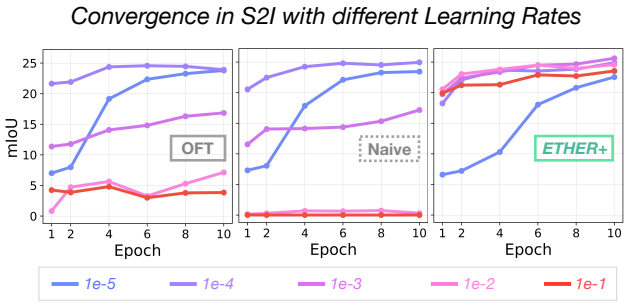

Figure 6. **Achieved controllability (mIoU) per epoch for different finetuning methods.** This figure extends Fig. 5 and highlights in detail how only a learning rate of 10^{-4} allows for optimal convergence in OFT and Naive, while for *ETHER+* fastest convergence speeds are stably achieved across magnitudes.

Parameter Efficiency. Finally, we provide a more detailed exploration on the parameter efficiency of *ETHER*-based methods. Let L be the number of finetuned layers, d and f the respective weight dimensions for $W \in \mathbb{R}^{d \times f}$. Then the parameter complexity for OFT can be written as $\mathcal{O}(\frac{Ld^2}{n})$ (Qiu et al., 2023) with n number of diagonal blocks[2]. Similarly, for LoRA we get $\mathcal{O}(Lr(d+f))$, while for *ETHER* and *ETHER+* we only have $\mathcal{O}(Ld)$ and $\mathcal{O}(L(d+f))$ respectively. With respect to both LoRA and OFT, this omits at the very least the rank multiplier r, or a potentially quadratic scaling. As already motivated in Sec. 3, this results in incredibly efficient finetuning while achieving comparable or stronger performances. For example, when finetuning Stable Diffusion as done above, *ETHER* and *ETHER+* use 120 times and 30 times fewer parameters than OFT respectively.

[2]Qiu et al. (2023) note a possible $\mathcal{O}(Ld)$ if $n = \alpha d$. However, in practice, equally scaling n with d disproportionally reduces adaptation parameters for large weight matrices. As OFT is fairly dependent on the parameter count, we omit this estimate.

5. Benchmark Experiments

We first investigate generative model adaptation in Sec. 5.1, with a focus on subject-driven image synthesis (§5.1.1) and controllable image synthesis (§5.1.2) following recent works (Qiu et al., 2023; Liu et al., 2023a). Sec. 5.2 then correspondingly investigates language model adaptation, looking at both natural language understanding (§5.2.1) and instruction tuning (§5.2.2). Finally, we study the importance of orthogonality and hyperspherical energy on finetuning performance in Sec. 5.3.

5.1. *ETHER* for Image-generative Model Adaptation

For our experiments on diffusion-based generative models, we apply the finetuning methods on the pretrained Stable Diffusion-v1.5 (Rombach et al., 2022), following the setting from OFT (Qiu et al., 2023). Our experiments follow best practices and hyperparameter choices for each method. For implementation details, please refer to App. C.

5.1.1. Subject-driven Generation

We first deploy *ETHER* and *ETHER+* on subject-driven generation following Ruiz et al. (2023); Qiu et al. (2023); finetuning the generative model for each of the 30 subjects and 25 prompts. For each combination, we generate four images, and measure image quality via a DINO (Caron et al., 2021) and a CLIP image encoder (Radford et al., 2021), text-prompt fidelity via a CLIP text encoder, and image diversity using LPIPS (Zhang et al., 2018).

Quantitative Results. Results are shown in Tab. 2. On subject-driven generation, we find competitive performance for both image quality, text-prompt fidelity and image diversity, particularly for *ETHER+* (e.g. DINO and CLIP-I scores of 0.666 vs 0.652 and 0.8 vs 0.794 for OFT, respectively). Most importantly, we achieve this performance while only utilizing a fraction of tuning parameters; with *ETHER+* only introducing $0.4M$ as compared to $11.6M$ by OFT. As hypothesized in Sec. 3, for nuanced finetuning, *ETHER*'s transformation strength seems to be too high to retain key semantic concepts in subject-driven generation, falling short in image quality with respect to other methods (e.g. also qualitatively depicted in Fig 3), despite achieving strong image diversity and text-prompt fidelity.

5.1.2. Controllable Image Generation

This section applies *ETHER* for controllability of Stable Diffusion following Qiu et al. (2023) for the Semantic Map to Image (S2I) task on ADE20K (Zhou et al., 2018). We use the trainable encoder from ControlNet (Zhang et al., 2023b) for the control signal and perform finetuning on the Stable Diffusion weights only. We report a baseline with just the control signal encoder to highlight relative gains through finetuning. Evaluations are performed on 2000 images generated from the validation set using mean Intersection-over-Union (mIoU) and accuracy of semantic maps over generated images using UperNet-101 (Xiao et al., 2018) pretrained on ADE20K. Finally, we measure the similarity between generated and original images via FID (Heusel et al., 2018). For OFT, we also test magnitude re-fitting (Qiu et al., 2023) for an additional epoch. We note that LoRA did not provide good results even for larger hyperparameter grids and was therefore omitted from Tab. 3 (more details in App. F).

Quantitative Results. Results are depicted in Tab. 3, and clearly demonstrate competitive control with both *ETHER* and *ETHER+*. Unlike subject-driven image generation, we find that *ETHER* performs on the same level as OFT multiplicative finetuning while using over $100\times$ fewer parameters (e.g. 24.6 versus 24.5 mIoU of OFT with $0.1M$ versus $13.2M$ parameters). Introducing magnitude re-fitting to OFT yields only limited gains while adding $0.2M$ parameters. Similar to Tab. 2 for subject-driven image generation, we find that for controllable image synthesis, the *ETHER+* relaxation provides additional performance gains (e.g. 27.3 vs 24.5 mIoU and 68.1 vs 62.8 Acc against OFT). Taking into account the more robust (Fig. 5) and faster convergence (Fig. 6), this presents *ETHER+* as a practically attractive finetuning alternative.

Table 2. **Subject-driven Generation Results.** We use r to denote rank, and n the number of diagonal blocks. We measure image quality (DINO, CLIP-I), text-prompt fidelity (CLIP-T) and image diversity (LPIPS). *ETHER+* addresses finegrained adaptation shortcomings of *ETHER* (c.f. Sec. 3.3) and achieves strong performance with only few adaptation parameters.

	#params	DINO ↑	CLIP-I ↑	CLIP-T ↑	LPIPS ↑
Real Images	-	0.703	0.864	-	0.695
DreamBooth	859.5M	0.644	0.793	0.236	0.709
LoRA$_{r=4}$	0.8M	0.660	0.796	0.231	0.714
OFT$_{n=4}$	11.6M	0.652	0.794	0.241	0.725
ETHER	0.1M	0.567	0.746	**0.256**	**0.766**
ETHER+	0.4M	**0.666**	**0.800**	0.240	0.729

Table 3. **Semantic Map to Image Results.** We use n to denote the number of diagonal blocks. *ETHER* and particularly *ETHER+* achieve strong synthesis control (mIoU, Acc) with few parameters while retaining good image alignment (FID). We indicate with (+ magn. r.f.) the OFT version with magnitude re-fitting.

	#params	mIoU ↑	Acc ↑	FID ↓
Encoder-only	0	8.2	38.0	41.2
OFT$_{n=4}$	13.2M	24.5	62.8	31.1
OFT$_{n=4}$ (+ magn. r.f.)	13.4M	24.6	63.3	**30.8**
ETHER	0.1M	24.6	63.3	32.0
ETHER+	0.4M	**27.3**	**68.1**	31.0

Table 4. **GLUE benchmark.** Comparisons of different methods finetuning DeBERTaV3-base. Results of all baselines are taken from (Liu et al., 2023a). We use r to denote rank, and n the number of diagonal blocks. As can be seen, *ETHER* and *ETHER+* achieve competitive performances across metrics while utilizing fewer parameters (up to a magnitude in the case of *ETHER*) while also retaining all practical benefits such as learning rate robustness depicted e.g. in Sec. 4.

	#params	MNLI↑	SST-2↑	CoLA↑	QQP↑	QNLI↑	RTE↑	MRPC↑	STS-B↑	**Avg↑**
Full Finet.	184M	89.90	95.63	69.19	**92.40**	94.03	83.75	89.46	91.60	88.25
BitFit	0.10M	89.37	94.84	66.96	88.41	92.24	78.70	87.75	91.35	86.20
H-Adapter	1.22M	90.13	95.53	68.64	91.91	94.11	84.48	89.95	91.48	88.28
P-Adapter	1.18M	90.33	95.61	68.77	92.04	94.29	85.20	89.46	91.54	88.41
LoRA$_{r=8}$	1.33M	90.65	94.95	69.82	91.99	93.87	85.20	89.95	91.60	88.50
AdaLoRA	1.27M	**90.76**	96.10	71.45	92.23	**94.55**	88.09	90.69	91.84	89.46
OFT$_{n=16}$	0.79M	90.33	96.33	**73.91**	92.10	94.07	87.36	92.16	91.91	89.77
BOFT$^{m=2}_{n=8}$	0.75M	90.25	**96.44**	72.95	92.10	94.23	88.81	92.40	91.92	89.89
ETHER	0.09M	90.23	96.10	71.31	91.42	94.31	**89.53**	**93.68**	92.30	89.86
ETHER+	0.33M	90.52	96.33	72.64	92.22	94.33	**89.53**	92.89	**92.35**	**90.10**

5.2. *ETHER* for Language Models Adaptation

To understand the applicability of the *ETHER* transformation family in the language domain, we follow Liu et al. (2023a)'s and Hu et al. (2022)'s experimental setup. For fair comparisons, we run grid searches over the most relevant hyperparameters in common value ranges. For additional implementation details, please refer to App. C.

5.2.1. Natural Language Understanding

We begin by deploying *ETHER* and *ETHER+* on the widely utilized (Devlin et al., 2019; Liu et al., 2019; He et al., 2023; Kopiczko et al., 2023) GLUE benchmark (Wang et al., 2018), finetuning a pretrained DeBERTaV3-base model (He et al., 2023) following Liu et al. (2023a), from which we report the baselines' results. GLUE comprises various English sentence understanding tasks, such as inference tasks (MNLI, QNLI, RTE), classification of sentiment (SST-2) or correct English grammatical structures (CoLA), and semantic similarity and equivalence prediction (MRPC, QQP, STS-B). CoLA scores report the Matthews correlation coefficient, MNLI matched accuracy, and STS-B average correlation. All other tasks are evaluated on accuracy.

Quantitative Results. Results in Tab. 4 show that *ETHER* and *ETHER+* match and even outperform previous methods with significantly fewer parameters. For example, *ETHER* outperforms the second-best BOFT on the RTE inference task (89.53 vs 88.81) or equivalence prediction on MRPC (93.68 vs 92.40) while using just one-ninth of the parameters ($0.085M$ compared to $0.75M$). *ETHER+* sets both the best performance on STS-B and particularly the highest overall score (90.10) using less than half of the parameters of BOFT. These results provide additional support for the practical viability of *ETHER* transformations, now for natural language adaptation - being a strong, but much more parameter-efficient competitor.

Table 5. **Instruction Tuning.** We use r to denote rank, and n the number of diagonal blocks. Both *ETHER* and *ETHER+* outperform LoRA/OFT which use up to a magnitude more parameters, and beat VeRA with similar parameter counts.

	#params	MMLU↑	ARC↑	Tru-1↑	Tru-2↑
Llama-2-7B	-	41.81	42.92	25.21	38.95
VeRA$_{r=64}$	0.27M	42.30	45.13	27.41	41.04
VeRA$_{r=256}$	1.05M	42.21	43.85	25.33	39.02
LoRA$_{r=1}$	0.52M	42.40	44.62	27.05	41.94
LoRA$_{r=8}$	4.19M	43.61	46.16	28.76	42.21
OFT$_{n=256}$	2.09M	42.92	44.88	27.42	41.11
ETHER$_{n=32}$	0.26M	44.57	45.14	27.91	41.83
ETHER+$_{n=32}$	1.04M	**44.87**	**46.50**	**29.38**	**43.51**

5.2.2. Instruction Tuning

Our instruction tuning experiments make use of Llama-2-7B (Touvron et al., 2023b) as pretrained model, finetuning it on the Alpaca dataset (Taori et al., 2023) for one epoch. To operate on a consumer GPU, we truncate the maximum sequence length to 256 and use bfloat16 precision (Kalamkar et al., 2019). We evaluate 0-shot performance of our instruction-tuned model on (i) Massive Multitask Language Understanding (MMLU) (Hendrycks et al., 2021) with 57 different tasks in four different subjects (STEM, Humanities, Social Sciences, Others); (ii) the AI2 Reasoning Challenge (ARC) (Clark et al., 2018), a common-sense reasoning dataset of questions from science grade exams; (iii) TruthfulQA (Lin et al., 2022) comprising 817 questions spanning 38 categories testing how much the model (wrongly) relies on imitation of human text to answer.

Quantitative Results. Results in Tab. 5 show that both *ETHER* and *ETHER+* outperform comparable finetuning approaches while utilizing fewer parameters. Across all metrics, the Llama-2-7B baseline is consistently surpassed by significant margins (e.g. 44.87 MMLU for *ETHER+* vs the 41.81 baseline, or 46.50 vs 42.92 ARC score). Despite being the most parameter-efficient method, *ETHER* outperforms all baselines with comparable number of parameters, such as the recently introduced VeRA (Kopiczko et al., 2023) with rank $r = 64$, and LoRA rank 1. Surprisingly, increasing the rank of VeRA to 256 leads to a decrease in performance, while LoRA rank 8 shows better results but is still outperformed on MMLU despite having $16\times$ more parameters. On the other hand, *ETHER+* surpasses all other methods across all benchmarks, while having $4\times$ fewer parameters than LoRA rank 8.

5.3. Hyperspherical Energy for Effective PEFT

Qiu et al. (2023) link finetuning stability and performance obtained by transforming the weights via matrix-multiplication to the orthogonality of the transformations, and a consequently unaltered hyperspherical energy (HE). To test this assumption, we have included an OFT control

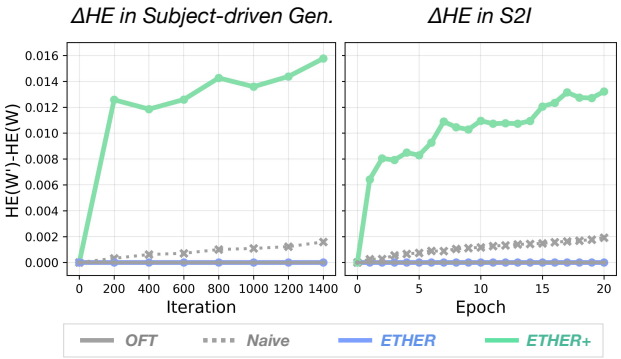

Figure 7. **Difference in HE** between finetuned/pretrained models for Subject-driven Generation and S2I. Notice that by removing the orthogonality constraint, both *ETHER+* and Naive alter the HE of the pretrained model, while OFT and *ETHER* do not.

Table 6. **OFT vs Naive.** OFT performance-test against its non-orthogonal counterpart Naive. We show that results don't differ significantly, questioning the relevance of HE retaining for finetuning performance.

	Subject-driven Generation				S2I		
	DINO	CLIP-I	CLIP-T	LPIPS	mIoU	Acc	FID
$\text{OFT}_{n=4}$	0.652	0.794	0.241	0.725	24.5	62.8	31.1
$\text{Naive}_{n=4}$	0.648	0.793	0.245	0.730	24.3	62.9	29.9

baseline (*Naive*), which does not utilize orthogonality constraints, on the same finetuning settings in which OFT was proposed. Results at convergence, as reported in Tab. 6, do not show significant differences, while actually introducing the overhead of computing the Cayley parametrizations (which also involve computing the inverse of a matrix). We also included the *Naive* baseline in the learning rate robustness studies in Fig. 4 and Fig. 5, showcasing that while differences are present for high learning rates, the optimal working range remains unaltered. Finally, we validate that the HE indeed varies during training, as reported in Fig. 7.

In contrast, on these same evaluations, our newly proposed *ETHER* transformation family, by introducing a boundary on the Euclidean distance on the transformation side, achieves stronger performance and greater robustness. This is especially true for the non-orthogonal *ETHER+*, which alters the overall HE even more than *Naive* (Fig. 7). This evidence diminishes the role of the HE and instead emphasizes the greater importance of the Euclidean distance, establishing the *ETHER* family as a favorable option in multiplicative finetuning settings.

6. Conclusions

Our paper introduces the *ETHER* family of transformations for parameter-efficient finetuning. Based on the Householder formulation of hyperplane reflections, *ETHER* methods frame finetuning as a search for unit normal vectors that define hyperplanes along which weight vectors are reflected. In doing so, *ETHER* (and its relaxation *ETHER+* for more finegrained adaptation) fix (or upper bound) the distance of learned transformations from the identity matrix (the transformation neutral element), thereby minimizing the risk of finetuning divergence. Put together, *ETHER* methods operate more parameter-efficiently than other PEFT methods (e.g., around 10-100 times less than LoRA or OFT), have higher learning rate robustness and encourage fast convergence. Consequently, *ETHER* transformations require less expansive hyperparameter searches to achieve good performance, making them very attractive for practical deployment.

Limitations. Of course, there is no free lunch. While both *ETHER* and its relaxation *ETHER+* show strong results with few parameters across a broad range of tasks, increasing the expressive power of the transformation is not as straightforward as in other methods, such as LoRA, where one can adjust the rank parameter to more closely approximate full finetuning. Moreover, multiplicative methods introduce a computational overhead during training compared to additive methods. Thanks to our block-parallel scheme, we make significant progress towards closing the gap between multiplicative and additive approaches; however, multiplicative methods still lag behind. This introduces a trade-off between parameter efficiency and computational overhead when achieving similar performance levels.

Impact Statement

This paper presents work that looks into better and more efficient finetuning of foundation models. By bringing down the need for compute-expensive hyperparameter grid searches and encouraging fast convergence, both the cost and environmental footprint of serving individually adapted models at scale can be brought down. Of course, with most advancement in the field of Machine Learning, there is potential for misuse and societal consequences, however, none of which we feel are specific to our proposed method and which need to be highlighted explicitly.

Acknowledgements

Massimo Bini was supported by Bosch Industry on Campus Lab at University of Tübingen. Karsten Roth thanks the European Laboratory for Learning and Intelligent Systems (ELLIS) PhD program and the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for support. Zeynep Akata and Karsten Roth were supported by DFG project number 276693517, by BMBF FKZ: 01IS18039A, by the ERC (853489 - DEXIM), by EXC number 2064/1 – project number 390727645.

References

AI, L. Litgpt. https://github.com/Lightning-AI/litgpt, 2023.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N. S., Chen, A. S., Creel, K. A., Davis, J., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L. E., Goel, K., Goodman, N. D., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D. E., Hong, J., Hsu, K., Huang, J., Icard, T. F., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P. W., Krass, M. S., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X. L., Li, X., Ma, T., Malik, A., Manning, C. D., Mirchandani, S. P., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J. C., Nilforoshan, H., Nyarko, J. F., Ogut, G., Orr, L., Papadimitriou, I., Park, J. S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y. H., Ruiz, C., Ryan, J., R'e, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K. P., Tamkin, A., Taori, R., Thomas, A. W., Tramèr, F., Wang, R. E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S. M., Yasunaga, M., You, J., Zaharia, M. A., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., and Liang, P. On the opportunities and risks of foundation models. *ArXiv*, 2021. URL https://crfm.stanford.edu/assets/report.pdf.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging Properties in Self-Supervised Vision Transformers, May 2021. URL http://arxiv.org/abs/2104.14294. arXiv:2104.14294 [cs].

Chen, J., Zhang, A., Shi, X., Li, M., Smola, A., and Yang, D. Parameter-Efficient Fine-Tuning Design Spaces, January 2023a. URL https://arxiv.org/abs/2301.01821v1.

Chen, Y., Qian, S., Tang, H., Lai, X., Liu, Z., Han, S., and Jia, J. LongLoRA: Efficient Fine-tuning of Long-Context Large Language Models, September 2023b. URL http://arxiv.org/abs/2309.12307. arXiv:2309.12307 [cs].

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., and Xing, E. P. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafjord, O. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge, March 2018. URL http://arxiv.org/abs/1803.05457. arXiv:1803.05457 [cs].

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. QLoRA: Efficient Finetuning of Quantized LLMs, May 2023. URL http://arxiv.org/abs/2305.14314. arXiv:2305.14314 [cs].

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T. (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., and Cohen-Or, D. An image is worth one word: Personalizing text-to-image generation using textual inversion, 2022.

Gao, L., Tow, J., Abbasi, B., Biderman, S., Black, S., DiPofi, A., Foster, C., Golding, L., Hsu, J., Le Noac'h, A., Li, H., McDonell, K., Muennighoff, N., Ociepa, C., Phang, J., Reynolds, L., Schoelkopf, H., Skowron, A., Sutawika, L., Tang, E., Thite, A., Wang, B., Wang, K., and Zou, A. A framework for few-shot language model evaluation, 12 2023. URL https://zenodo.org/records/10256836.

Garg, S., Farajtabar, M., Pouransari, H., Vemulapalli, R., Mehta, S., Tuzel, O., Shankar, V., and Faghri, F. TiC-CLIP: Continual training of CLIP models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=TLADT8Wrhn.

Gouk, H., Hospedales, T. M., and Pontil, M. Distance-Based Regularisation of Deep Networks for Fine-Tuning, January 2021. URL http://arxiv.org/abs/2002.08253. arXiv:2002.08253 [cs, stat].

Guo, D., Rush, A. M., and Kim, Y. Parameter-efficient transfer learning with diff pruning, 2021.

He, P., Gao, J., and Chen, W. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=sE7-XhLxHA.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring Massive Multitask Language Understanding, January 2021. URL http://arxiv.org/abs/2009.03300. arXiv:2009.03300 [cs].

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2018.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., Attariyan, M., and Gelly, S. Parameter-Efficient Transfer Learning for NLP, February 2019. URL https://arxiv.org/abs/1902.00751v2.

Householder, A. S. Unitary triangularization of a nonsymmetric matrix. *J. ACM*, 5(4):339–342, oct 1958. ISSN 0004-5411. doi: 10.1145/320941.320947. URL https://doi.org/10.1145/320941.320947.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Ibrahim, A., Thérien, B., Gupta, K., Richter, M. L., Anthony, Q., Lesort, T., Belilovsky, E., and Rish, I. Simple and scalable strategies to continually pre-train large language models, 2024.

Kalamkar, D., Mudigere, D., Mellempudi, N., Das, D., Banerjee, K., Avancha, S., Vooturi, D. T., Jammalamadaka, N., Huang, J., Yuen, H., Yang, J., Park, J., Heinecke, A., Georganas, E., Srinivasan, S., Kundu, A., Smelyanskiy, M., Kaul, B., and Dubey, P. A study of bfloat16 for deep learning training, 2019.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation, 2018.

Karthik, S., Roth, K., Mancini, M., and Akata, Z. If at first you don't succeed, try, try again: Faithful diffusion-based text-to-image generation by selection, 2023.

Ke, Z., Shao, Y., Lin, H., Konishi, T., Kim, G., and Liu, B. Continual pre-training of language models. In *The Eleventh International Conference on Learning Representations*, 2023.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollar, P., and Girshick, R. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, October 2023.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., and Hadsell, R. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114. URL https://www.pnas.org/doi/abs/10.1073/pnas.1611835114.

Kopiczko, D. J., Blankevoort, T., and Asano, Y. M. VeRA: Vector-based Random Matrix Adaptation, October 2023. URL http://arxiv.org/abs/2310.11454. arXiv:2310.11454 [cs].

Kornblith, S., Shlens, J., and Le, Q. V. Do Better ImageNet Models Transfer Better?, June 2019. URL http://arxiv.org/abs/1805.08974. arXiv:1805.08974 [cs, stat].

Köpf, A., Kilcher, Y., von Rütte, D., Anagnostidis, S., Tam, Z.-R., Stevens, K., Barhoum, A., Duc, N. M., Stanley, O., Nagyfi, R., ES, S., Suri, S., Glushkov, D., Dantuluri, A., Maguire, A., Schuhmann, C., Nguyen, H., and Mattick, A. Openassistant conversations – democratizing large language model alignment, 2023.

Lee, J., Cho, K., and Kiela, D. Countering language drift via visual grounding. In Inui, K., Jiang, J., Ng, V., and Wan, X. (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4385–4395, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1447. URL https://aclanthology.org/D19-1447.

Lester, B., Al-Rfou, R., and Constant, N. The Power of Scale for Parameter-Efficient Prompt Tuning. pp. 3045–3059, January 2021. doi: 10.18653/v1/2021.emnlp-main.243.

Li, X., Grandvalet, Y., and Davoine, F. Explicit Inductive Bias for Transfer Learning with Convolutional Networks, June 2018. URL http://arxiv.org/abs/1802.01483. arXiv:1802.01483 [cs].

Li, X. L. and Liang, P. Prefix-tuning: Optimizing continuous prompts for generation, 2021.

Lin, S., Hilton, J., and Evans, O. Truthfulqa: Measuring how models mimic human falsehoods, 2022.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollár, P. Microsoft coco: Common objects in context, 2015.

Liu, W., Qiu, Z., Feng, Y., Xiu, Y., Xue, Y., Yu, L., Feng, H., Liu, Z., Heo, J., Peng, S., Wen, Y., Black, M. J., Weller, A., and Schölkopf, B. Parameter-efficient orthogonal finetuning via butterfly factorization, 2023a.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach, July 2019. URL https://arxiv.org/abs/1907.11692v1.

Liu, Z., Feng, R., Zhu, K., Zhang, Y., Zheng, K., Liu, Y., Zhao, D., Zhou, J., and Cao, Y. (Cones) Cones: Concept Neurons in Diffusion Models for Customized Generation, March 2023b. URL http://arxiv.org/abs/2303.05125. arXiv:2303.05125 [cs].

Longpre, S., Hou, L., Vu, T., Webson, A., Chung, H. W., Tay, Y., Zhou, D., Le, Q. V., Zoph, B., Wei, J., and Roberts, A. The Flan Collection: Designing Data and Methods for Effective Instruction Tuning.

Lu, Y., Singhal, S., Strub, F., Courville, A., and Pietquin, O. Countering language drift with seeded iterated learning. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6437–6447. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/lu20c.html.

Mangrulkar, S., Gugger, S., Debut, L., Belkada, Y., Paul, S., and Bossan, B. Peft: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft, 2022.

Mehta, S. V., Patil, D., Chandar, S., and Strubell, E. An empirical investigation of the role of pre-training in lifelong learning, 2022. URL https://openreview.net/forum?id=D9E8MKsfhw.

Mou, C., Wang, X., Xie, L., Zhang, J., Qi, Z., Shan, Y., and Qie, X. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023.

Mukhopadhyay, S., Gwilliam, M., Agarwal, V., Padmanabhan, N., Swaminathan, A., Hegde, S., Zhou, T., and Shrivastava, A. Diffusion Models Beat GANs on Image Classification, July 2023. URL http://arxiv.org/abs/2307.08702. arXiv:2307.08702 [cs].

OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023.

Pfeiffer, J., Kamath, A., Rücklé, A., Cho, K., and Gurevych, I. *AdapterFusion: Non-Destructive Task Composition for Transfer Learning*. May 2020.

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Muller, J., Penna, J., and Rombach, R. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *ArXiv*, abs/2307.01952, 2023a.

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis, July 2023b. URL http://arxiv.org/abs/2307.01952. arXiv:2307.01952 [cs].

Qiu, Z., Liu, W., Feng, H., Xue, Y., Feng, Y., Liu, Z., Zhang, D., Weller, A., and Schölkopf, B. Controlling text-to-image diffusion by orthogonal finetuning. *arXiv preprint arXiv:2306.07280*, 2023.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning Transferable Visual Models From Natural Language Supervision, February 2021. URL http://arxiv.org/abs/2103.00020. arXiv:2103.00020 [cs].

Richardson, E., Goldberg, K., Alaluf, Y., and Cohen-Or, D. ConceptLab: Creative Generation using Diffusion Prior Constraints, August 2023. URL http://arxiv.org/abs/2308.02669. arXiv:2308.02669 [cs].

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

Roth, K., Thede, L., Koepke, A. S., Vinyals, O., Henaff, O. J., and Akata, Z. Fantastic gains and where to find them: On the existence and prospect of general knowledge transfer between any pretrained model. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=m50eKHCttz.

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and Aberman, K. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *CVPR*, 2022.

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and Aberman, K. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation, March 2023. URL http://arxiv.org/abs/2208.12242. arXiv:2208.12242 [cs].

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., Salimans, T., Ho, J., Fleet, D. J., and Norouzi, M. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding, May 2022. URL http://arxiv.org/abs/2205.11487. arXiv:2205.11487 [cs].

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 06 2014.

Stojanovski, Z., Roth, K., and Akata, Z. Momentum-based weight interpolation of strong zero-shot models for continual learning, 2022.

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., and Hashimoto, T. B. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971, 2023a.

Touvron, H., Martin, L., Stone, K. R., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D. M., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A. S., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I. M., Korenev, A. V., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288, 2023b.

Valipour, M., Rezagholizadeh, M., Kobyzev, I., and Ghodsi, A. DyLoRA: Parameter Efficient Tuning of Pre-trained Models using Dynamic Search-Free Low-Rank Adaptation, April 2023. URL http://arxiv.org/abs/2210.07558. arXiv:2210.07558 [cs].

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Linzen, T., Chrupała, G., and Alishahi, A. (eds.), *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL https://aclanthology.org/W18-5446.

Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., and Hajishirzi, H. Self-Instruct: Aligning Language Models with Self-Generated Instructions, May 2023. URL http://arxiv.org/abs/2212.10560. arXiv:2212.10560 [cs].

Xiao, T., Liu, Y., Zhou, B., Jiang, Y., and Sun, J. Unified perceptual parsing for scene understanding, 2018.

Xu, Y., Xie, L., Gu, X., Chen, X., Chang, H., Zhang, H., Chen, Z., Zhang, X., and Tian, Q. QA-LoRA: Quantization-Aware Low-Rank Adaptation of Large Language Models, September 2023. URL http://arxiv.org/abs/2309.14717. arXiv:2309.14717 [cs].

Zhang, G., Wang, L., Kang, G., Chen, L., and Wei, Y. Slca: Slow learner with classifier alignment for continual learning on a pre-trained model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19148–19158, October 2023a.

Zhang, L. and Agrawala, M. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023.

Zhang, L., Rao, A., and Agrawala, M. Adding Conditional Control to Text-to-Image Diffusion Models, September 2023b. URL http://arxiv.org/abs/2302.05543. arXiv:2302.05543 [cs].

Zhang, Q., Chen, M., Bukharin, A., He, P., Cheng, Y., Chen, W., and Zhao, T. Adaptive Budget Allocation for Parameter-Efficient Fine-Tuning, March 2023c. URL http://arxiv.org/abs/2303.10512. arXiv:2303.10512 [cs].

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric, April 2018. URL http://arxiv.org/abs/1801.03924. arXiv:1801.03924 [cs].

Zhang, S., Dong, L., Li, X., Zhang, S., Sun, X., Wang, S., Li, J., Hu, R., Zhang, T., Wu, F., and Wang, G. Instruction Tuning for Large Language Models: A Survey, October 2023d. URL http://arxiv.org/abs/2308.10792. arXiv:2308.10792 [cs].

Zhang, Y., Huang, N., Tang, F., Huang, H., Ma, C., Dong, W., and Xu, C. Inversion-Based Style Transfer with Diffusion Models, March 2023e. URL http://arxiv.org/abs/2211.13203. arXiv:2211.13203 [cs].

Zhao, J., Zhang, Z., Chen, B., Wang, Z., Anandkumar, A., and Tian, Y. Galore: Memory-efficient llm training by gradient low-rank projection, 2024.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.-Y., and Wen, J.-R. A Survey of Large Language Models, September 2023. URL http://

arxiv.org/abs/2303.18223. arXiv:2303.18223 [cs].

Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., and Torralba, A. Semantic Understanding of Scenes through the ADE20K Dataset, October 2018. URL http://arxiv.org/abs/1608.05442. arXiv:1608.05442 [cs].

Appendix

In this appendix, we augment the main paper with additional, qualitative evidence for the learning rate robustness of *ETHER* transformations in Appendix A. In addition, we also provide benchmark-specific qualitative examples for subject-driven and controllable image generation in Appendix B. For all experiments - both those in the main paper and supplementary results, we then list all relevant details in Appendix C for our studies on finetuning in subject-driven image generation (§C.1), controllable image synthesis (§C.2), natural language understanding tasks (§C.3) and instruction tuning (§C.4). We then provide two additional *ETHER* ablations in Appendix D - for the number of block-diagonals and the specific double-sided application in *ETHER+*. Finally, we discuss LoRA shortcomings for adaptation in the controllable image synthesis task (§F).

A. Qualitative Evidence of Learning Rate Robustness

As introduced in Sec. 3, when finetuning with *ETHER* transformation, by construction, the learning rate only controls the speed with which reflection angels change. As a consequence, *ETHER* methods are much more robust to learning rate choices, and less likely to diverge and cause model deterioration. This allows for user control over the convergence speed while minimizing the risk of model collapse during training. To demonstrate this, Sec. 4 introduced both a qualitative example comparing the impact of minimal and maximal perturbation strength on the model output in Fig. 3, and quantitative evaluations on the Semantic Map to Image task against learning rate choices in Figs. 5 and 6.

In this section, we augment Sec. 4 and provide additional qualitative results and impressions to highlight the non-deteriorating nature of *ETHER* transformation. For this, we showcase subject-driven generation results using different finetuning methods in Fig. 8, with default generations using the best learning rate. We then systematically increase the finetuning learning rate by 10 and by 100 times, and visualize the correspondingly generated output. As can be seen, for $10\times$ higher learning rates OFT and Naive fail to follow the text prompt, while LoRA finetuning quickly collapses. With $10\times$ lower learning rates instead, OFT, Naive and *ETHER* are not able to generate the subject correctly in the predefined number of iterations.

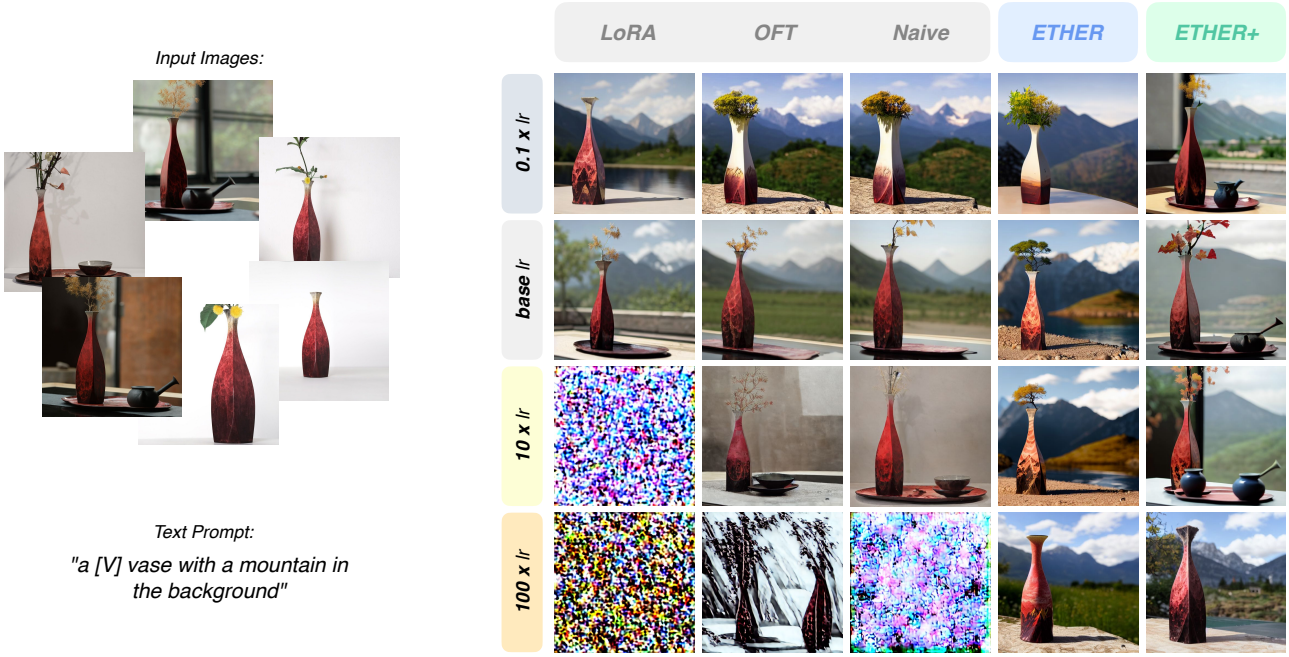

Figure 8. Qualitative visualization of learning rate robustness of *ETHER* and *ETHER+* in subject-driven generation finetuning. We see how *ETHER* methods are able to consistently produce good results avoiding model deterioration. Specifically, *ETHER+* shows impressive capabilities, being able to follow the subject-prompt instructions in the widest learning rate range.

B. Qualitative Examples for *ETHER* Finetuning

We show some qualitative results by using the finetuning methods proposed in this paper.

B.1. Subject-driven Generation.

In Figure 9 we report subject-driven generation examples. In particular, for a fair comparison, we report images which come from the same noise vector in the Stable Diffusion latent space. For the *sunglasses* images, we see how non-*ETHER* methods manage to reproduce the subject, but fail to follow the text prompt in most cases. Interestingly in the first row, we notice how *ETHER+* is able to properly control the generation, by transforming the yellow area (associated to a beer in other models) in an enlightened Eiffel Tower. For the *teapot* images instead, we see how *ETHER+* is able to better keep the appearances of the subject.

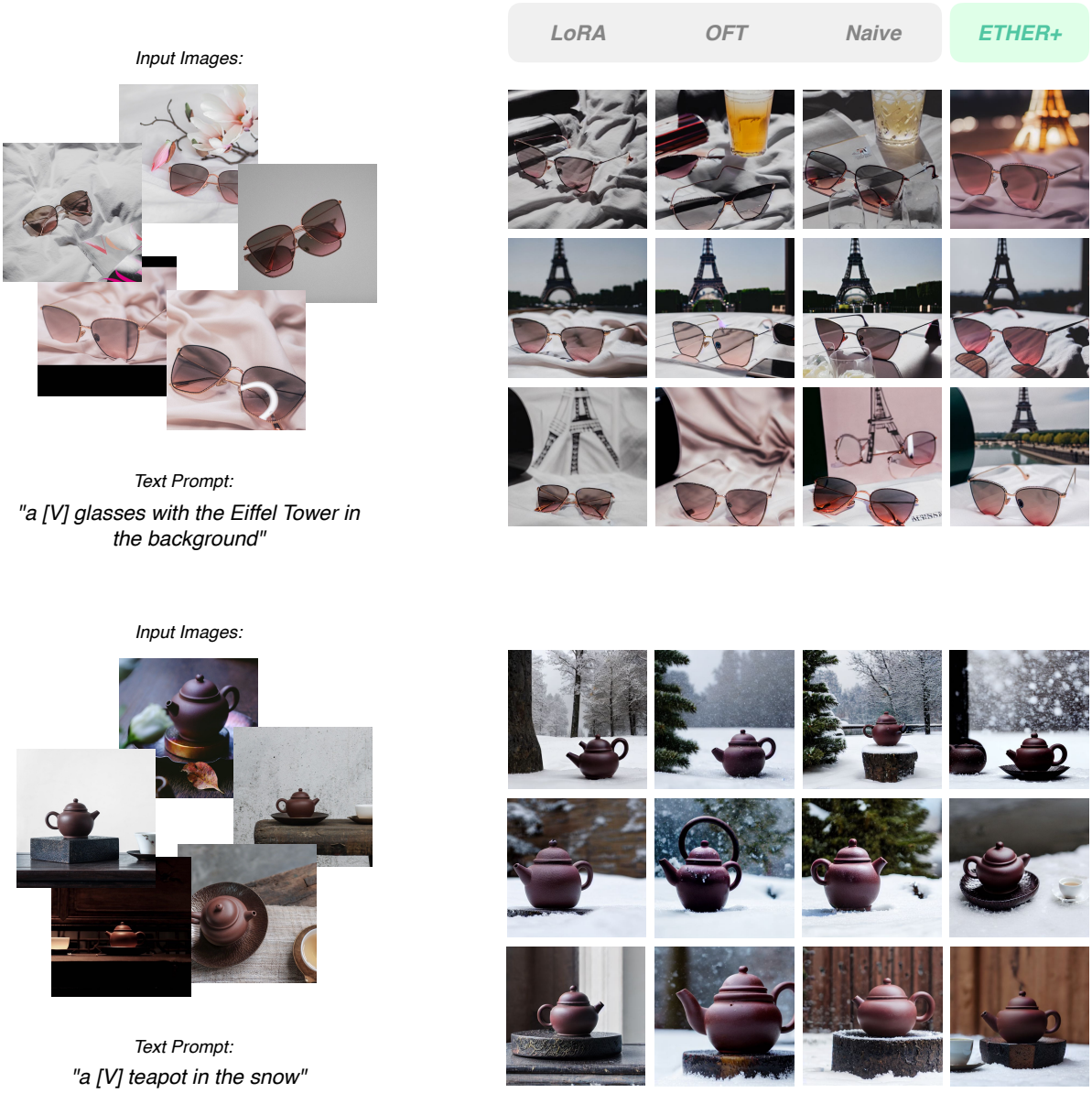

Figure 9. Subject-driven Generation results. Each row shares initial latent noise (notice row-wise similarities). We can see that *ETHER+* method is better at adapting the model to the subjects. Notice how for the pink sunglasses, OFT and Naive fail in following the prompt.

B.2. Controllable Generation.

In Figure 10 we show some examples from the Semantic Map to Image task. In particular, we notice how in the first row all models but *ETHER+* fail to control the image correctly, not being able to separate the land from the water. Additionally, in the second row OFT fails to generate the sky, while Naive presents a halo effect. These examples showcase the abilities of *ETHER+* finetuning over the other methods.

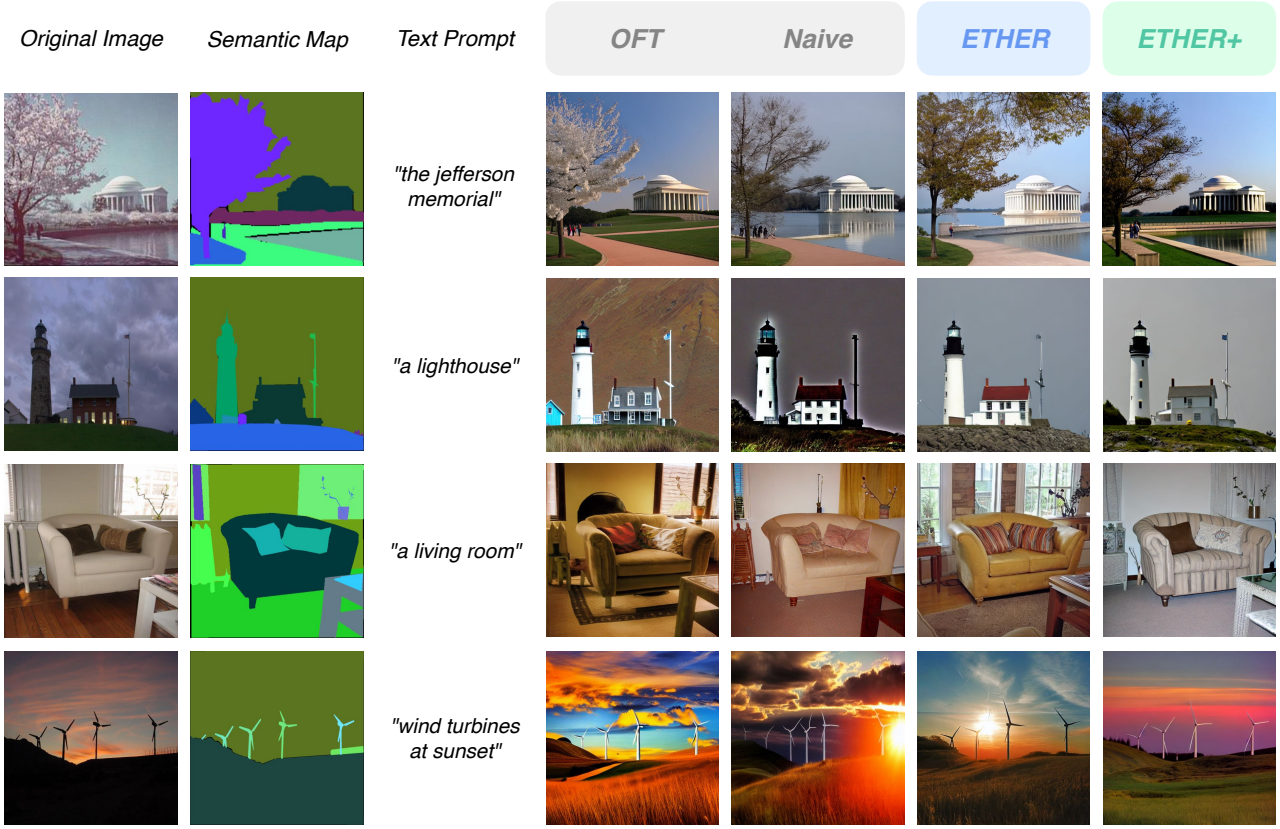

Figure 10. Semantic Map to Image Qualitative Results. We notice how in the first row all models but *ETHER+* fail to control the image correctly. Overall *ETHER+* controlled images show better control.

To show broader controllable capabilities, we also report few qualitative examples with *ETHER* methods trained with Landmarks and Canny Edge Maps control signals on CelebA-HQ (Karras et al., 2018) and COCO 2017 (Lin et al., 2015) datasets respectively.

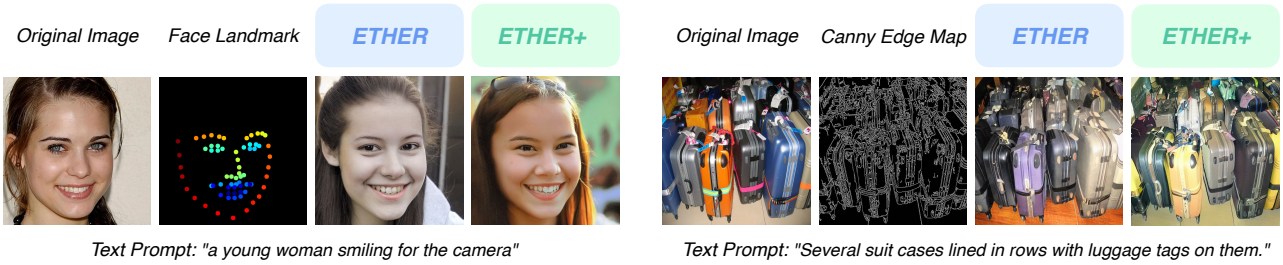

Figure 11. Examples of Landmark to Face (left) and Canny Edge Map to Image (right) controlled generation with *ETHER* methods.

C. Experimental Details

This section provides additional experimental details for replication not listed in the main benchmark experimental section 5. It is worth noting that while in most of our experiments we do not employ regular dropout (Srivastava et al., 2014), Liu et al. (2023a) proposes a multiplicative dropout form specifically designed for multiplicative finetuning methods, which we did not test in this study. We hypothesize that this specialized dropout technique could potentially work better than regular dropout for *ETHER* and *ETHER+* as well. We also note that Qiu et al. (2023) report OFT's number of parameters as half of the actual trainable parameters due to the redundancy in the skew symmetric matrices S^B in the Cayley parametrization of Q^B. Basically, we they report the storage parameters for Q^B rather than the training parameters. For consistency and fair comparisons, we follow the same convention for OFT throughout our paper.

C.1. Subject-driven Generation

For subject-driven generation, we follow the same setting listed in DreamBooth (Ruiz et al., 2023), using DreamBooth and OFT (Qiu et al., 2023) baselines as implemented in official OFT GitHub repository. The additional trainable layers follow (Qiu et al., 2023) and are added to the Q,K,V layers and the projection layer inside every attention module. The training is performed over 1400 iterations for each method, evaluating the generation results every 200 iterations at selecting the best one (typically around 1200 iterations). For DreamBooth and OFT, we follow the original implementations and use a learning rate of 5×10^{-6} and 6×10^{-5} respectively, with a batch size of 1. For Naive - the non-orthogonal OFT variant - we use the same setting of OFT for a fair comparison. For LoRA we select a learning rate of 6×10^{-4}. For *ETHER* and *ETHER+*, we use a learning rate of 6×10^{-3}. We perform the training on a Tesla V100-32GB GPU.

C.2. Controllable Generation

For our experiments on controllable image generation we follow the setting of Qiu et al. (2023), using the signal encoder from ControlNet (Zhang & Agrawala, 2023) (comprising 8 trainable convolutional layers, accounting for 3.1M additional learnable parameters). Finetuning parameters are added to the Q,K,V layers as well as the projection layer of the attention modules and the subsequent feedforward layers. As baselines, we use the official implementation of OFT. Similarly to Qiu et al. (2023), for OFT and Naive we use a learning rate of 1×10^{-5}. For *ETHER* and *ETHER+* we use a larger learning rate of 1×10^{-3}. For all experiments, we upper bound the learning rate of the signal encoder to 1×10^{-4}. We perform all the training runs on a single Nvidia-A100-40GB with a batch size of 10. As listed in Sec. 5.1.2 and expanded in Sec. F, we tried to utilize LoRA for controllable generation as well but found no comparable results even after extensive trials with different hyperparameters.

C.3. Natural Language Understanding

For our GLUE benchmark experiments finetuning DeBERTaV3-base (He et al., 2023), we make use of the peft Hugging Face repository (Mangrulkar et al., 2022) as the basis for our implementations. To compare our results with those of Liu et al. (2023a), we follow their implementation and apply *ETHER* and *ETHER+* to all the linear layers in every transformer block. The relevant hyperparameters for each task are reported in Tab. 8. All training runs are conducted on a single Nvidia-A100-40GB GPU.

Table 7. GLUE benchmark hyperparameters.

Method	Hyperparameters	MNLI	SST-2	CoLA	QQP	QNLI	RTE	MRPC	STS-B
ETHER	Learning Rate	8e-4	1e-3	1e-3	3e-4	1e-3	1e-3	3e-4	2e-3
	Batch Size	32	32	32	8	8	32	32	8
	Num. Epochs	9	14	10	20	7	13	14	8
	Dropout	1e-3	1e-3	1e-1	1e-1	1e-3	1e-2	1e-1	1e-1
	Max Seq. Len.	256	128	64	320	512	320	320	128
ETHER+	Learning Rate	8e-4	1e-4	1e-3	3e-3	3e-3	3e-4	8e-4	8e-4
	Batch Size	8	8	8	32	32	8	32	8
	Num. Epochs	8	10	6	16	5	35	17	11
	Dropout	1e-3	1e-3	1e-1	1e-3	1e-3	1e-3	1e-2	1e-3
	Max Seq. Len.	256	128	64	320	512	320	320	128

C.4. Instruction Tuning

For our Instruction Tuning experiments, we use the LoRA (Hu et al., 2022) finetuning implementation in the lit-gpt repository (AI, 2023) as baseline. For evaluations, we make use of Gao et al. (2023)'s benchmark implementations. For the recently proposed VeRA (Kopiczko et al., 2023) baseline, we reproduce the model implementation following their best performing method as described in the paper: sampling random A and B matrices with uniform kaiming initialization scaled by the matrix dimension, and a learnable, non-zero diagonalized vector initialized as a vector of all zeros apart for one element equal to 0.1. Same for OFT, for which we follow the implementation in the official repository oft, selecting the number of block-diagonal matrices such that the overall number of parameters becomes comparable with *ETHER+* and LoRA rank 8. For all experiments, we use a cosine annealing learning rate scheduler, no dropout, and 1000 warmup steps. For LoRA, VeRA, and OFT we use AdamW optimizer with a weight decay of 0.01, while for *ETHER* methods, given the normalization happening on the parameters, weight decay would have limited impact and thus we set it to 0. For LoRA and VeRA, we keep α fixed with respect to the learning rate by setting it equal to the rank. For all experiments, we conduct an extensive grid search over learning rates and batch sizes. For each combination, we perform the LLama-2-7B (Touvron et al., 2023b) finetuning over Alpaca (Taori et al., 2023) for one epoch. All training runs are conducted on a single Nvidia-A100-40GB GPU, but could also be run on a consumer NVIDIA GeForce-RTX-3090-24G GPU.

Table 8. Instruction Tuning hyperparameters.

	VeRA$_{r=64}$	VeRA$_{r=256}$	LoRA$_{r=1}$	LoRA$_{r=8}$	OFT$_{n=256}$	*ETHER*$_{n=32}$	*ETHER+*$_{n=32}$
Learning Rate	5e-3	1e-3	3e-3	5e-4	5e-4	2e-3	5e-3
Batch Size	32	32	8	8	16	8	16

D. *ETHER* Ablations

This section details additional ablation experiments on the impact of the block-diagonality degree on the final performance, as well as experimental support to the theoretical motivation in Sec. 3.3 to apply the relaxed Householder transformation on both the left and right side of the weight matrix.

D.1. Block-diagonal *ETHER* Performances

In Table 9 and Table 10, we compare the usage of multiple diagonal blocks for *ETHER* finetuning to allow for fast performance, especially in large models domain. Both tables augment our method description in Sec. 3.4 and the shortened results in Tab. 1. In all cases, we notice that performance remains almost unaffected by the choice of block number, while on the other hand, the computational efficiency consistently increases (8.22 TFLOPs for $n = 32$ versus 25.26 TFLOPs for $n = 1$ for Llama-2-7B). It is worth noting that results for *ETHER+* with $n = 32$ show better performance with respect to less diagonalized counterparts.

Table 9. Semantic Map to Image (S2I) results for different number of diagonal blocks n on *ETHER* finetuning at epoch 10

ETHER	#params	mIoU ↑	Acc ↑	FID↓
$n = 1$	0.1M	23.1	61.23	31.7
$n = 4$	0.1M	22.9	60.92	30.5
$n = 16$	0.1M	22.3	60.35	30.7

Table 10. Instruction Tuning results for different number of diagonal blocks n on *ETHER* finetuning

ETHER+	#params	TFLOPs	MMLU ↑	ARC ↑	Tru-1↑	Tru-2 ↑
$n = 1$	1.04M	51.65	43.75	46.76	28.03	41.06
$n = 4$	1.04M	18.66	43.91	45.73	27.54	40.46
$n = 32$	1.04M	9.04	44.87	46.50	29.38	43.51

D.2. Double-sided Application of *ETHER+*

Finally, we provide a brief ablation study in Tab. 11, comparing the *ETHER+* performance when applying the relaxed Householder transformations H^+ on only one side versus both sides. Although the parameter count doubles, we observe a significant increase in performance (e.g. 0.666 vs 0.618 in DINO score) as higher transformation distances can be achieved.

Table 11. Subject-driven Generation image quality results comparison (at iteration 1200) among standard *ETHER+* and its version only applied on one side of the weight matrix.

	#params	DINO ↑	CLIP-I ↑
ETHER+ (one-sided)	0.2M	0.618	0.777
ETHER+	0.4M	**0.666**	**0.800**

E. VTAB preliminary results

We also perform a small evaluation over a subset of the popular Visual Task Adaptation Benchmark (VTAB), using an ImageNet-21k pretrained ViT-B. As can be seen, *ETHER* and *ETHER+* perform comparably to OFT with $n = 256$ and LoRA rank 8, while using a fraction of the trainable parameters.

Table 12. VTAB results

	#params	**Natural**				**Specialized**	**Structured**
		Caltech101	DTD	Flowers102	SVHN	EuroSAT	sNORB-Elev
Full Finetuning	85.8M	96.26	73.03	98.71	73.71	96.16	63.36
Linear Probing	0	95.96	72.34	**99.12**	52.55	95.03	34.09
LoRA$_{r=8}$	1.33M	97.69	**77.50**	99.10	**97.40**	98.92	74.89
OFT$_{n=256}$	0.29M	96.95	75.80	98.60	96.58	98.83	74.37
ETHER	0.08M	97.64	75.85	98.83	95.81	98.80	74.17
ETHER+	0.33M	**98.27**	76.92	98.88	96.84	**99.15**	**78.41**

F. LoRA failures for S2I task

As described at the beginning of our benchmark experimental section (Sec. 5.1.2, we excluded LoRA from our experimental results listed in Tab. 3, as it did not convincingly converge to good results. In this section, we provide additional experimental details on our application of LoRA to controllable image synthesis, given which still no suitable results were able to be achieved. In particular, LoRA finetuning for the Semantic Map to Image task (S2I) task was tested on a large, randomly subsampled grid of learning rate {1e-5, 5e-5, 1e-4, 1e-3, 1e-2}, batch size {4, 8, 16, 32, 64}, rank {4, 8, 64}, scaling factor {0.1, 1}, dropout {0, 0.1}, for which we also visualize reflective samples in Fig. 12.

Figure 12. LoRA randomly sampled generated samples from validation set control. Control signal reported on the left.