GRASP: Guided Residual Adapters with Sample-wise Partitioning

Felix Nützel[1] **Mischa Dombrowski**[1] **Bernhard Kainz**[1,2]
[1] Friedrich-Alexander-Universität Erlangen-Nürnberg, GER
[2] Imperial College London, UK
felix.nuetzel@fau.de

Abstract

Text-to-image flow matching transformers degrade sharply in long-tail settings: tail-class outputs collapse in fidelity and diversity, limiting their value as synthetic augmentation for rare conditions. We trace this to low head-versus-tail gradient alignment during fine-tuning, an optimization-level pathology that conditioning- and sampling-side interventions do not address. We propose GRASP (Guided Residual Adapters with Sample-wise Partitioning): a deterministic partition of the conditioning space, paired with group-specific residual adapters in the transformer feedforward layers, that leaves the flow-matching objective and the sampler untouched. In conditional flow matching, condition values index distinct sets of probability paths, which makes the conditioning a structurally natural axis along which to factorize residual updates. Because the partition is static, every tail sample is guaranteed to update its assigned expert, which bypasses extreme long-tail failure modes. GRASP is non-invasive and composable: on MIMIC-CXR-LT, combining GRASP with self-guided minority sampling at inference outperforms its learned counterpart on 10 of 12 metrics. GRASP itself reduces overall Fréchet Inception Distance (FID) by up to 80% and lifts tail-class coverage by up to 44% over full fine-tuning, learned-routing mixture-of-experts and long-tail baselines. Performance is stable across random reassignments of labels to experts and across expert counts, as long as no expert is starved of samples. Used as training data for a downstream DenseNet classifier on NIH-CXR-LT, GRASP synthetics significantly outperform every non-GRASP alternative on macro F1 and yield nonzero F1 on 9 of 13 classes versus 3 of 13 from real data, while maintaining comparable macro F1. Results on ImageNet-LT confirm the mechanism is not tied to medical inductive bias. Code is available at https://github.com/Felix-012/GRASP-LT

1 Introduction

Text-to-image flow matching models have transformed visual generation by enabling controllable and photorealistic synthesis at scale. Yet, their performance degrades markedly in long-tail distributions, where rare classes dominate real-world scenarios, like medical imaging. In such settings, models prioritize frequent head classes, leading to mode collapse in tail classes: generated images for rare pathologies exhibit poor fidelity and severely limited diversity. The result is a systematic collapse of diversity in rare conditions, which limits the potential of synthetic data augmentation and undermines downstream robustness. This hampers critical applications, such as augmenting datasets for underrepresented diseases, where diverse, high-quality synthetics are essential for robust training.

Prior efforts to mitigate this fall into two camps. Post-hoc sampling strategies guide inference toward underrepresented modes [40, 43, 42, 32]; while effective for exploration, they leave the underlying

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

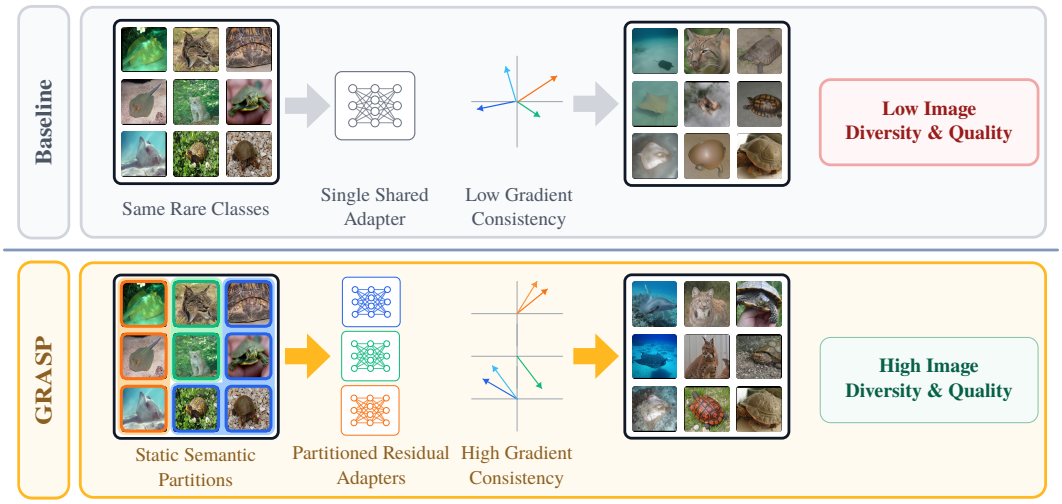


Figure 1: Our proposed method uses parallel GRASP adapters to gain more aligned gradient updates when training a diffusion transformer. This enables the model to generate images of higher quality and diversity, in particular for rare classes.

distribution unchanged and do not prevent conditional mode collapse during training [1]. Training-time methods alter the loss or the conditioning pathway, e.g., aligning conditional and unconditional objectives so that head-class signal can transfer to the tail [6], lesion-mask-controllable synthesis [49], or causal disentanglement of pathology and anatomy [34]. These approaches modify *what* the model learns; they do not directly target the optimization-level pathology that gradient analyses on classification and language tasks identify as the underlying cause of long-tail collapse, namely that head- and tail-class gradients are mismatched in norm and direction and average destructively at the parameter level [12, 10]. Dong et al. [10] address this in long-tail language modeling by learning a cluster-conditional gating network that routes samples to specialized experts, but to our knowledge no analogous mechanism exists for long-tail text-to-image diffusion training, despite the fact that conditional flow matching is the setting where conditioning-aligned routing is geometrically natural.

We bridge this gap with GRASP: Guided Residual Adapters with Sample-wise Partitioning. As displayed in Figure 1, GRASP uses a static external prior to partition training samples into groups in conditioning space (e.g., whole-label bins with a reserved healthy group, or clusters of text embeddings), so that each group's residual updates are isolated in its own expert. During fine-tuning of a frozen pre-trained backbone, cluster-specific residual adapters (lightweight non-linear skips inserted into feedforward layers) enable targeted refinement without learned routing. This preserves the original generation objective while enhancing stability and efficiency.

A natural objection is that this gives up the flexibility of learned gating. We argue the opposite holds in extreme long-tail. A learned gate is itself dependent on a gradient signal dominated by head classes: with hundreds of head examples per tail example, the gate has little incentive to route rare samples to dedicated experts. A static, conditioning-driven partition removes this failure mode by construction: every tail sample is guaranteed to update its assigned expert, at the cost of fixing the partition from an external prior.

Our approach delivers clear and consistent improvements across challenging long-tail benchmarks. On MIMIC-CXR-LT, GRASP reduces overall Fréchet Inception Distance (FID) by up to 80% and increases tail-class diversity metrics such as coverage by up to 44% compared to state-of-the-art baselines (Table 2). When GRASP-generated images are used for downstream classification on NIH-CXR-LT, the macro-average F1 is comparable to real-data baseline while extending nonzero-F1 coverage from 3 of 13 classes (real data), and 1 of 13 (full fine-tuning) to 9 of 13 (Table 5). Ablations indicate that both the partitioning and the residual adapter design contribute. The result does not depend on one grouping: random label-to-expert reassignments and other expert counts perform comparably, and only partitions that starve an expert degrade. With identical experts, fixed routing beats a learned gate on label-macro FID and downstream utility, while a gate initialized at the GRASP

assignment finds no better alternative (Table 4). On the optimization trajectory, conditioning-based routing yields higher gradient consistency than a learned gate or a shared adapter (Figure 4).

Our **contributions** are: **(1)** We transfer the gradient-consistency objective of Dong et al. [10], originally proposed for long-tail language modeling with learned cluster-conditional gating, to long-tail text-to-image flow matching, and replace the learned gate with a deterministic, conditioning-driven partition. In conditional flow matching, condition values index distinct sets of probability paths, so partitioning along the conditioning is a structurally natural factorization rather than a generic embedding-space heuristic. We argue that, whenever even coarse priors over the conditioning are available or easily obtainable (which is the common case in medical imaging), a static partition that gives every expert sufficient support should be the default. **(2)** GRASP is intentionally non-invasive: it leaves the flow-matching objective and the sampler unchanged, only restricting which parameters absorb which gradients. As a direct consequence it composes with orthogonal long-tail interventions: we demonstrate this by combining GRASP with self-guided minority sampling at inference time and observe that the composition favors static routing and outperforms minority guidance alone. **(3)** We provide comprehensive validation across medical (MIMIC-CXR-LT and NIH-CXR-LT) and general (ImageNet-LT) benchmarks, including a gradient consistency analysis, a partitioning sensitivity study, and a capacity-matched comparison between fixed and learned routing, showing consistent gains in image quality and diversity over full fine-tuning, generic adapters, and learned-routing mixture-of-experts (MoE) variants under deliberately minimal priors.

2 Related Work

Long-tail diffusion training. Most prior work on long-tail text-to-image generation acts at inference time, steering sampling with classifier-free guidance (CFG) variants [8, 39, 17] or minority-aware reweighting [40, 43, 42, 32, 31, 51]. These methods can improve exploration but do not change the learned distribution and therefore cannot prevent training-time conditional collapse [1]. Training-time alternatives modify the target objective or conditioning pathway, including conditional-unconditional alignment (CCUA) [6], failure-aware on-policy augmentation [36], lesion-mask-controllable synthesis [49], and causal pathology/anatomy disentanglement [34]. GRASP is orthogonal: it keeps the flow-matching objective fixed and changes only which parameters absorb which gradients.

Mixture-of-Experts and gradient consistency. Expert layers increase capacity through sparse routing [41], often with adapter-style experts such as LoRA [21, 47, 28] or cluster-conditional gating as in MoCLE [14]. Diffusion MoE variants similarly improve scaling or specialization through routed paths, timestep experts, dense-to-sparse conversion, or semantic superclass routing [48, 53, 13, 45]. Our starting point is Cluster-Guided Sparse Experts [10], which optimizes intra-expert gradient consistency for long-tail language modeling with a learned gate. GRASP transfers this objective to flow-matching text-to-image diffusion, where conditions index probability paths, and replaces learned gating with a deterministic partition π. This guarantees that tail samples update their assigned expert even when head-class gradients dominate the training signal, a guarantee learned routers do not provide.

Generative models for long-tail medical imaging. Medical diffusion work addresses rare pathology synthesis through text-guided augmentation [37], sparse feature-space mixing [11], lesion-focused objectives [49], and causal disentanglement [34]; adjacent systems scale chest-X-ray, segmentation, MRI, or 3D synthesis with larger domain-specific pipelines [4, 50, 46, 15]. These approaches rely on stronger medical priors or large curated corpora. GRASP instead operates at fine-tuning time on a generic flow-matching backbone and is evaluated on both MIMIC-CXR-LT and ImageNet-LT, separating the routing mechanism from any single medical inductive bias.

3 Method

GRASP is a non-invasive architectural intervention. It leaves the flow-matching loss and the sampler unchanged and only modifies which parameters absorb which per-sample gradients. The contribution is therefore not a different generative objective but a routing strategy adapted to long-tail training.

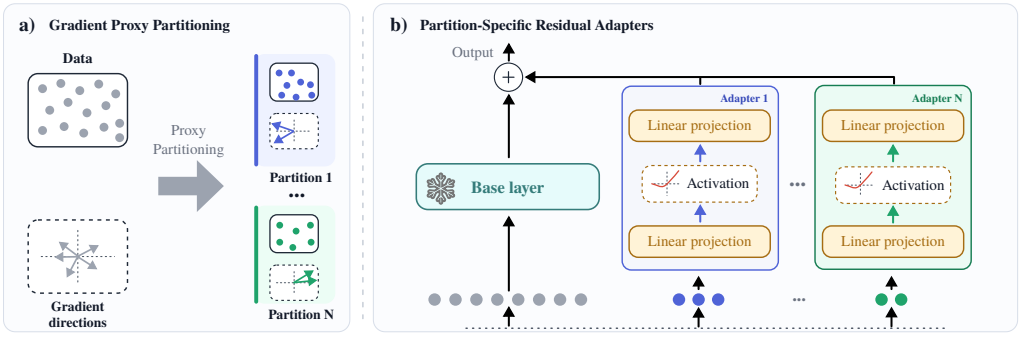


Figure 2: Overview of the GRASP architecture: **a)** We want to maximize gradient consistency during training by partitioning the samples so that each expert's dominant update direction aligns with that of the samples routed to it. We achieve this by partitioning based on flow-matching conditioning. **b)** Based on this partitioning, we deterministically route samples to their designated expert, while keeping the base model frozen.

Sample-wise Partitioning. We measure how well an update serves a group of samples with the gradient consistency of Dong et al. [10](Figure 2):

$$\mathrm{GC}_\theta(X') = \frac{g_\theta(X) \cdot g_\theta(X')}{\|g_\theta(X)\| \, \|g_\theta(X')\|} \tag{1}$$

where $g_\theta(X')$ is the dominant update direction of a sample set X', the leading singular vector of its per-sample gradients and X is the set of samples that update the same parameters. Directly optimizing this objective would require online computation and clustering of per-sample gradients, which is infeasible for large flow-matching transformers.

Instead of approximating gradient similarity directly, we partition the model's conditioning space. The key claim of this section is that, in conditional flow matching with class-style conditions, this is not a generic embedding-space heuristic: the conditioning is a structurally natural axis along which to factorize residual updates. This claim only concerns the domain of the factorization, so it does not assert that any particular partition is optimal, or that conditions grouped into one expert have similar gradients.

Conditional flow matching does not learn a single velocity field but a family of condition-indexed fields $\{v_\theta(\cdot, t, c)\}_{c \in \mathcal{C}}$, where each c specifies a conditional target $p_{\mathrm{data}}(\cdot \mid c)$ and a probability path to the base distribution [29, 2, 26, 5, 3]. With class-style conditions, each value of c thus indexes a *distinct set of probability paths*. A static partition $\pi : \mathcal{C} \to \{1, \ldots, K\}$ is the simplest mechanism that respects this structure: it groups samples by path family and gives each group a dedicated expert. The proxy claim is therefore not that semantically similar prompts have similar gradients in a generic embedding sense; it is that conditioning indexes path geometry in conditional flow matching, and we hypothesize that gradients inherit that geometry [7]. The partition need not be sharp: conditions in the same expert only need to be more path-homogeneous within than across, the regime in which the expert's residual adapter can absorb the shared local dynamics. Because consistency (Eq. 1) compares each subset with the samples routed to the same expert, not with each other, an expert can also mix frequent and rare labels, as our label experts do to balance support (Table 1), and its rare labels can still score high. Figure 4 supports this on the optimization trajectory: GRASP reaches higher gradient consistency than MoCLE and a shared adapter, and expert-wise gradient consistency increases with expert-wise conditioning homogeneity.

Formally, let $c_i \in \mathcal{C}$ denote the condition associated with sample x_i. Depending on the setting, c_i may be obtained from a class-label embedding, $c_i = \phi(\ell_i)$, or from a text/conditioning encoder, $c_i = \psi(t_i)$. We define a static routing function

$$\pi : \mathcal{C} \to \{1, \ldots, K\}, \tag{2}$$

which assigns each condition to one of K experts, and thus routes each sample x_i to expert $\pi(c_i)$.

A straightforward instantiation is label-based partitioning which assigns every label wholesale to one expert, with several labels sharing one expert to ensure sufficient support. Medical imaging adds a

Table 1: Exact label-to-expert lookup π used for MIMIC-CXR-LT (all 87,493 training samples). Ranks are pathology training-frequency ranks (1 = most frequent pathology, *No Finding* excluded). Every label is assigned wholesale to one expert. E0–E3 correspond to C0–C3 in Figure 5 and in Table 7.

Expert	Samples	Ranks	Labels
E0	11,506	5–11	Support Devices, Edema, Pneumonia, Pneumothorax, Lung Lesion, Fracture, Enlarged Cardiomediastinum
E1	9,243	1, 12–18	Lung Opacity, Consolidation, Pleural Other, Calcification of the Aorta, Tortuous Aorta, Pneumoperitoneum, Subcutaneous Emphysema, Pneumomediastinum
E2	13,484	2–4	Cardiomegaly, Atelectasis, Pleural Effusion
E3	53,260	–	No Finding (reserved healthy expert)

domain-specific inductive bias: the healthy/pathological dichotomy. On MIMIC-CXR-LT [18], we therefore augment the partition with an additional cluster π_{healthy} containing healthy cases. Then we split the remaining 18 pathology labels into three groups, a number that follows the head/medium/tail convention [52, 19], giving $K = 4$ experts in total. To get non-starved experts, we use a balanced graph partition for construction, with the single labels treated as concept sets. This results in an approximately frequency-ordered whole label binning which serves as our π (Table 1): ranked by training frequency, the first expert holds classes 2 to 4, the second 5 to 11 and the third 8 to 12, plus the largest pathology class (*Lung Opacity*). The latter exception to the ordering reduces the imbalance factor from $16:1$ to $1.46:1$, giving 13,484/11,506/9,243 samples. This partition is merely used as a simple example, it was neither tuned nor curated. Section 4 shows that random reassignments perform comparably.

While label-based routing is simple and interpretable, it ignores within-group conditioning variation. We therefore also consider text-based partitioning. Specifically, we map each sample to a semantic feature representation $f(x_i) = \psi(t_i)$, where t_i is the paired textual prompt and $\psi(\cdot)$ is the frozen text encoder of the diffusion backbone. We then compute π_{text} with bisecting K-means on $\{f(x_i)\}$, using $K = 4$ and cosine similarity. In contrast to label-based routing, this groups samples by conditioning-space proximity rather than label identity or frequency alone. However, it is still a minimalist approach by design. The resulting cluster compositions are compared in Appendix A.

For ImageNet-LT, where label-level partitioning is impractical due to label cardinality and no analogue of the healthy/pathological dichotomy exists, we rely exclusively on text-embedding clustering with no domain-specific cluster. This is a deliberate choice: it shows that the routing framework generalizes beyond the medical setting and does not depend on handcrafted domain partitions, while still producing competitive long-tail behavior (Section 4).

We emphasize that both choices of π are deliberately minimal. The label-based partition reuses the noisy labels released with MIMIC-CXR-LT, with a single domain-prior split (healthy vs. pathological) and a balance constraint; we do not denoise or curate them. The text-based partition is vanilla bisecting K-means on the frozen text-encoder embeddings, with no hand-tuned distance metric, weighting scheme, or post-hoc correction. The point of GRASP is not to engineer a high-quality prior. It is to show that even very coarse, off-the-shelf priors are sufficient to recover specialized adapters, as long as every expert receives sufficient support (Section 4). This robustness positions the static partition as the default under extreme imbalance whenever such priors are at hand, which covers most medical imaging applications and is cheap to construct elsewhere via vanilla clustering; learned cluster-conditional gating remains the right choice when no usable prior exists. The deployment criterion is therefore practical: take whatever partition you can construct, give it to GRASP, and you receive the update isolation by construction, without modifying the loss or the sampler.

Guided Residual Adapters. Given the static routing function $\pi : \mathcal{C} \to \{1, \ldots, K\}$, we condition the architecture via guided residual adapters. Unlike conventional MoE approaches with learned gating [41, 10], each sample x_i is deterministically assigned to expert $\pi(c_i)$. This removes routing noise and makes specialization directly interpretable in terms of conditional subproblems.

The hidden representation at transformer block l is denoted by $h_l \in \mathbb{R}^d$, and the corresponding base transformation by $F_l(h_l)$. We augment this block with a residual adapter $A_{k,l}$ specific to expert

$k = \pi(c_i)$. The layer output for sample x_i becomes

$$h_{l+1} = F_l(h_l) + A_{\pi(c_i),l}(h_l). \tag{3}$$

This yields a decomposition of the learned vector field into a shared backbone and an expert-specific residual correction,

$$\hat{v}_\theta(z, t, c) = s_\phi(z, t, c) + a_{\pi(c),\psi_{\pi(c)}}(z, t, c), \tag{4}$$

where s_ϕ captures broadly shared dynamics and $a_{\pi(c),\psi_{\pi(c)}}$ captures the correction associated with the selected condition group. The key assumption is not that all conditions within one expert are identical, but that after factoring out globally shared structure, the remaining residual dynamics are similar enough to justify parameter sharing within that expert.

Each adapter is a two-layer transformation $A_{k,l}(h) = W^{(2)}_{k,l}\,\sigma(W^{(1)}_{k,l}h)$ with expert-specific projections $(W^{(1)}_{k,l}, W^{(2)}_{k,l})$. Residual adapters fit our setting because they combine strong parameter sharing with small condition-specific blocks [38, 20, 25]: allocating a full expert per rare condition is statistically brittle, whereas grouped residual adapters share strength across related conditions while isolating condition-group-specific updates. In contrast to LoRA [21], which adds a linear low-rank update to existing weights, our adapter is a non-linear residual module whose width is not tied to a low rank constraint: $W^{(1)}_{k,l} \in \mathbb{R}^{d' \times d}$, $W^{(2)}_{k,l} \in \mathbb{R}^{d \times d'}$, with d' ranging from $d' \ll d$ to $d' > d$. We empirically trade off adapter width d' against the number of attached layers: either a few large adapters ($d' > d$) on the last dual-stream blocks (GRASP-HighDim) or lightweight adapters ($d' < d$) on all blocks (GRASP-LowDim).

By construction, expert-specific parameters ψ_k are updated only by samples with $\pi(c) = k$, so $\nabla_{\psi_k}\mathcal{L} = \mathbb{E}[\mathbf{1}\{\pi(C) = k\}\nabla_{\psi_k}\ell]$. This isolates group-specific gradients while the shared backbone ϕ provides transferable structure across conditions, the classical MoE intuition [23].

Our backbone is Flux [29, 32]: a flow-matching transformer with 19 dual-stream blocks (text+image) followed by 38 single-stream blocks (image only), $L = 57$ in total. Because dual blocks mediate multi-modal fusion, we instantiate $A_{k,l}$ in their FeedForward submodules; configurations that adapt all blocks (e.g., GRASP-LowDim) additionally attach $A_{k,l}$ to the MLP output projection of every single-stream block.

4 Experiments

Datasets. Motivated by our central goal of enabling more diverse and faithful image generation in medical imaging, we adopt MIMIC-CXR-LT [18] as our primary benchmark. This dataset extends the large-scale MIMIC-CXR collection [24] by introducing additional rare disease categories and constructing single-label splits to emphasize long-tail behavior. The resulting distribution poses a realistic challenge for generative modeling, capturing the severe frequency imbalance observed in real clinical data. We follow the official MIMIC-CXR-LT protocol, using 87,493 training samples to train our text-to-image diffusion model and 23,550 test-set prompts to generate synthetic images for evaluation. Both splits exhibit similar long-tailed characteristics, enabling consistent assessment of performance across head and tail classes.

Likewise, we base the evaluation of the classifier in our downstream task on NIH-CXR-LT [18]. This chest X-ray dataset is a split of ChestXray14 [44] with additional long-tail classes added. For evaluation, we take the proposed test split (20,279 images) and filter any labels that are not part of MIMIC-CXR-LT, resulting in 16,506 samples. The downstream classifier is trained only on images generated from the MIMIC-CXR-LT test prompts and evaluated on this real NIH-CXR-LT set. We chose this cross-dataset design to limit memorization effects from report-to-image generation.

To show the generalizability of our method, we conduct further experiments on the widely used Imagenet-LT [30] dataset. This allows us to verify our approach on a domain-unspecific long-tail setting. The train split consists of 115,846 image-label pairs, while the test split consists of 50,000 pairs.

Metrics. For the image generation, we want to reliably assess both the quality and diversity of the generated samples. To quantify the quality of the synthetic images, we utilize the common metrics Fréchet Inception Distance (FID) [16]. To assess the diversity of our generated samples, we employ Coverage [33], which is generally more robust to outliers than Improved Recall [27].

Given real features Φ_r and generated features Φ_g, the k-nearest-neighbor radius around $x \in \Phi$ is $\rho_\Phi(x) := \|x - \mathrm{NN}_k(x; \Phi \setminus \{x\})\|$, and the corresponding ball-membership indicator is $f(x, \Phi) := \mathbf{1}[\exists y \in \Phi : \|x - y\| \leq \rho_\Phi(y)]$. Coverage then estimates the fraction of real samples whose neighborhood contains a generated sample, $\mathrm{Cov} = N^{-1} \sum_{i=1}^{N} \max_j \mathbf{1}[\|\varphi_r^{(i)} - \varphi_g^{(j)}\| \leq \rho_r(\varphi_r^{(i)})]$. Since diversity is our primary concern, we report an additional robust diversity metric, called the adjusted Image Retrieval Score (IRS) [9]. For each image, this intuitive metric tries to retrieve the closest image in the training and test datasets, estimating which percentage of the datasets the synthetic images cover. The adjusted score is the test score normalized by the train score, which results in a metric that does not reward memorization, while keeping in mind the maximum possible coverage.

All image generation metrics are computed for a domain-specific DenseNet121 [22] as a feature extractor, which aligns with our downstream task and provides insights into radiographic properties of the generated images. For broader comparability, we also report features computed by the well-studied feature extractor DINOv2 [35], giving us insights into domain-unspecific features of our synthetic dataset. To gain robust estimates of our metrics, we generate 50,000 samples with resolution 1024×1024 per model and training seed. For the evaluation of the classifier in the downstream task, we report the per-label results for the F1-score.

Implementation. We fine-tune Flux (FLUX.1-dev) with the flow-matching objective and classifier-free guidance (CFG) (scale 5) for 10,000 steps at batch size 8 on 8 H100 GPUs; full fine-tuning takes $\sim$30 h, adapters 7 to 12 h. The MoCLE baseline reuses our partitions and number of experts at LoRA rank 64. Per-seed downstream synthesis is 50,000 images at 1024^2, the DenseNet-121 classifier is trained for 100 epochs over three classifier seeds and three dataset seeds (nine runs per configuration). Full setup and compute breakdown is in Appendix E.

Image Metrics Results. We evaluate generation quality and diversity in domain-specific (DenseNet, Table 2) and domain-unspecific (DINOv2, Table 3) feature spaces. Adapter-based methods generally outperform full fine-tuning as well as CCUA, and within adapter-based methods static GRASP outperforms MoCLE and the shared adapter on aggregate metrics. The low-dimensional, all-layers GRASP variant gives the most uniform gains and the best global FID under CXR features; the high-dimensional, last-layer variant is slightly weaker globally but wins on some labels (e.g., *Atelectasis*, *Edema*). Under DINOv2 features the high-dimensional variant clearly leads.

The MoE-GRASP variants share GRASP's experts and capacity, but use a learned gate. They are behind unguided GRASP on label-macro FID in both feature spaces, and on most other metrics, while recovering less metrics downstream (Tables 2, 3, 5). The learned variants only lead on two (LowDim) and four (HighDim) of the twelve feature-space metrics, mostly all-label ones. Additionally, their gates never make use of all experts and always route the healthy majority together with pathology groups. We read the loss in label-wise fidelity and tail utility as the signature of a gate that re-allocates routing capacity toward head-dominated gradients (Section 1). CCUA [6], which changes the training objective instead, performs worse than GRASP-LowDim on 11 of 12 feature space metrics and both downstream metrics. On ImageNet-LT, the high-dimensional GRASP variant again leads on aggregate; MoCLE recovers competitive performance on this dataset, indicating learned routing remains viable when no strong domain-prior partition is available.

GRASP improves most metrics, especially on rare labels such as *Pneumoperitoneum*; qualitatively, it preserves the baseline prompt structure while recovering class-specific features that baselines smooth out (Figure 3).

Composability with inference-time interventions. Because GRASP only changes which parameters take which gradients, it composes with self-guided minority sampling [43] (MG), a post-hoc sampler modification. With MG applied to both, fixed routing beats the learned router (MoE-GRASP-LowDim+MG) on 10 of 12 metrics, including every label-macro metric and CXR IRS (.68 vs. .63; Tables 2 and 3). GRASP-LowDim+MG attains the best all-labels IRS (.68) and macro IRS (.61) under CXR features (Table 2). MG on full fine-tune recovers only a small fraction of this gain, indicating that the lever is GRASP's training-time distribution shift, with MG re-allocating sampling effort within it. This is consistent with the intended role of static partitioning: it makes the architecture long-tail-friendly without precluding standard inference- or training-side recipes.

Partition Sensitivity. Table 4 swaps the submitted partition for alternatives of GRASP-LowDim width. Random reassignment randomizes only which labels share an expert. Its results are comparable

Table 2: CXR feature-space evaluation on MIMIC-CXR-LT. We report coverage (Cov., ↑), IRS (↑), and FID (↓) for all labels and macro averages. All-label ±: SD over three generator seeds; macro ±: SD across labels. Best values per column are shown in bold. *Adapter* refers to fine-tuning using one shared sequential residual adapter without specific routing and Moe-GRASP refers to the learned matched comparison to our designs specified in Section 3.

	Method	All labels			Macro Avg		
		Cov. ↑	IRS ↑	FID ↓	Cov. ↑	IRS ↑	FID ↓
Baseline	**Full Fine-Tune (class)**	$.55_{\pm.01}$	$.11_{\pm0}$	$.06_{\pm.01}$	$.62_{\pm.16}$	$.17_{\pm.16}$	$.14_{\pm.11}$
	Full Fine-Tune (text+class)	$.65_{\pm.01}$	$.16_{\pm.01}$	$.05_{\pm.01}$	$.70_{\pm.13}$	$.20_{\pm.15}$	$.12_{\pm.10}$
	MoCLE [14]	$.84_{\pm.01}$	$.32_{\pm.01}$	$.04_{\pm0}$	$.85_{\pm.12}$	$.35_{\pm.17}$	$.09_{\pm.07}$
	CCUA [6]	$.64_{\pm.03}$	$.13_{\pm.01}$	$.07_{\pm0}$	$.67_{\pm.15}$	$.20_{\pm.18}$	$.15_{\pm.11}$
	Adapter (768d)	$.86_{\pm.01}$	$.33_{\pm.02}$	$.03_{\pm0}$	$.85_{\pm.12}$	$.37_{\pm.17}$	$.07_{\pm.07}$
	MoE-GRASP-LowDim	$.927_{\pm.003}$	$.454_{\pm.007}$	$.012_{\pm.001}$	$.903_{\pm.007}$	$.498_{\pm.013}$	$.051_{\pm.002}$
	MoE-GRASP-HighDim	$.925_{\pm.004}$	$.462_{\pm.010}$	$.017_{\pm.002}$	$.912_{\pm.003}$	$.514_{\pm.015}$	$.051_{\pm.001}$
Ours	**GRASP (768d)**	$.90_{\pm.01}$	$.35_{\pm.01}$	$.02_{\pm0}$	$\mathbf{.92_{\pm.10}}$	$.53_{\pm.17}$	$.05_{\pm.06}$
	GRASP-LowDim	$.94_{\pm0}$	$.50_{\pm0}$	$\mathbf{.01_{\pm0}}$	$\mathbf{.92_{\pm.09}}$	$.56_{\pm.12}$	$\mathbf{.04_{\pm.05}}$
	GRASP-HighDim	$.92_{\pm.01}$	$.41_{\pm.01}$	$.02_{\pm0}$	$.91_{\pm.10}$	$.52_{\pm.10}$	$\mathbf{.04_{\pm.05}}$
+MG	**Vanilla + MG** [43]	$.62_{\pm.02}$	$.15_{\pm.01}$	$.05_{\pm.01}$	$.67_{\pm.02}$	$.19_{\pm.02}$	$.13_{\pm.01}$
	MoE-GRASP-LowDim + MG [43]	$.94_{\pm0}$	$.63_{\pm.02}$	$\mathbf{.01_{\pm0}}$	$\mathbf{.92_{\pm.01}}$	$.60_{\pm.01}$	$.05_{\pm0}$
	GRASP-LowDim + MG [43]	$\mathbf{.95_{\pm0}}$	$\mathbf{.68_{\pm0}}$	$\mathbf{.01_{\pm0}}$	$\mathbf{.92_{\pm.01}}$	$\mathbf{.61_{\pm0}}$	$\mathbf{.04_{\pm0}}$

Table 3: DINOv2 feature-space evaluation on MIMIC-CXR-LT (all labels and macro over 19 labels) and ImageNet-LT (all labels); ± as in Table 2. Text-conditioned full fine-tuning does not apply to ImageNet-LT; CCUA, MoE-GRASP-HighDim and guided runs were evaluated on MIMIC-CXR-LT only. Best per column in bold.

	Method	MIMIC-CXR-LT						ImageNet-LT		
		All labels			Macro Avg			All labels		
		Cov. ↑	IRS ↑	FID ↓	Cov. ↑	IRS ↑	FID ↓	Cov. ↑	IRS ↑	FID ↓
Baseline	**Full Fine-Tune (class)**	$.01_{\pm0}$	$.08_{\pm.01}$	$106.6_{\pm14.2}$	$.07_{\pm.05}$	$.22_{\pm.11}$	$145.5_{\pm28.5}$	$.71_{\pm.01}$	$.76_{\pm.01}$	$155.1_{\pm5.5}$
	Full Fine-Tune (text+class)	$.02_{\pm.01}$	$.12_{\pm.02}$	$88.3_{\pm12.6}$	$.11_{\pm.07}$	$.27_{\pm.14}$	$128.3_{\pm29.1}$	–	–	–
	MoCLE [14]	$.02_{\pm0}$	$.13_{\pm0}$	$87.5_{\pm.43}$	$.15_{\pm.11}$	$.29_{\pm.14}$	$123.0_{\pm29.2}$	$\mathbf{.75_{\pm0}}$	$.76_{\pm0}$	$\mathbf{109.4_{\pm1.0}}$
	CCUA [6]	$.01_{\pm0}$	$.10_{\pm.01}$	$117.7_{\pm39.5}$	$.11_{\pm.08}$	$.25_{\pm.10}$	$157.5_{\pm26.4}$	–	–	–
	Adapter (768d)	$.02_{\pm0}$	$.15_{\pm.01}$	$80.9_{\pm4.2}$	$.14_{\pm.1}$	$.34_{\pm.17}$	$119.9_{\pm33.2}$	$.57_{\pm0}$	$.39_{\pm0}$	$249.9_{\pm0.4}$
	MoE-GRASP-LowDim	$.007_{\pm.000}$	$.176_{\pm.002}$	$98.6_{\pm3.8}$	$.095_{\pm.010}$	$.362_{\pm.017}$	$155.2_{\pm7.3}$	$.690_{\pm.004}$	$\mathbf{.817_{\pm.002}}$	$136.0_{\pm4.6}$
	MoE-GRASP-HighDim	$.024_{\pm.003}$	$.186_{\pm.002}$	$73.4_{\pm2.1}$	$.176_{\pm.017}$	$.375_{\pm.005}$	$114.5_{\pm1.7}$	–	–	–
Ours	**GRASP (768d)**	$.02_{\pm0}$	$.17_{\pm0}$	$76.6_{\pm.54}$	$.17_{\pm.12}$	$.37_{\pm.17}$	$115.4_{\pm39.2}$	$.67_{\pm.07}$	$.70_{\pm.22}$	$164.4_{\pm60.2}$
	GRASP-LowDim	$.01_{\pm0}$	$.17_{\pm0}$	$91.4_{\pm1.6}$	$.12_{\pm.1}$	$.38_{\pm.17}$	$150.6_{\pm76.2}$	$.70_{\pm0}$	$.81_{\pm0}$	$132.3_{\pm2.2}$
	GRASP-HighDim	$\mathbf{.03_{\pm0}}$	$.19_{\pm0}$	$\mathbf{67.3_{\pm1.6}}$	$\mathbf{.23_{\pm.16}}$	$.39_{\pm.18}$	$\mathbf{107.7_{\pm33.3}}$	$\mathbf{.75_{\pm0}}$	$.80_{\pm.01}$	$\mathbf{109.4_{\pm1.0}}$
+MG	**Vanilla + MG** [43]	$.01_{\pm0}$	$.11_{\pm.02}$	$100.5_{\pm12.3}$	$.08_{\pm.04}$	$.25_{\pm.03}$	$140.7_{\pm12.3}$	–	–	–
	MoE-GRASP-LowDim + MG [43]	$.01_{\pm0}$	$\mathbf{.22_{\pm0}}$	$90.6_{\pm4.2}$	$.11_{\pm.01}$	$.41_{\pm.01}$	$151.4_{\pm9.1}$	–	–	–
	GRASP-LowDim + MG [43]	$.01_{\pm0}$	$.20_{\pm0}$	$84.0_{\pm1.6}$	$.14_{\pm0}$	$\mathbf{.43_{\pm0}}$	$142.0_{\pm4.6}$	–	–	–

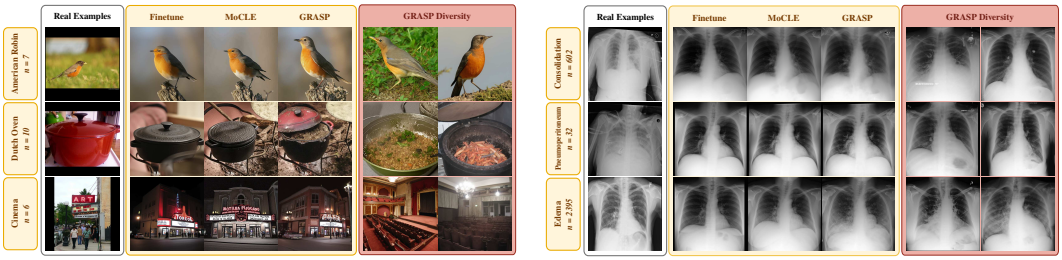


Figure 3: Qualitative comparison on tail classes of MIMIC-CXR-LT (random samples from heuristic pathologies) and ImageNet-LT (random pick from the bottom 2%). Real counterparts are from NIH-CXR. Comparisons between vanilla, MoCLE [14], GRASP are made using the same seeds. The architectural intervention is intentionally non-invasive: GRASP outputs remain visually close to baselines on the same prompt, but show more distinct class-specific appearance that the baselines collapse on.

to the submitted version, so the results do not depend on our specific grouping. Using two or four pathology groups also stays close to ours, only versions that starve experts degrade: official head/medium/tail routing (a 71-sample tail expert) and one expert per label (12 samples for the rarest).

Table 4: Partition sensitivity **(a)** and routing controls **(b)** on MIMIC-CXR-LT with fixed backbone, optimizer, schedule and evaluation budget. *min. expert*: training samples of the smallest expert. The partition variants in (a) and the hybrid use one generator seed $\times$ three classifier seeds (GRASP and MoE-GRASP rows: three each). (b) compares learned with fixed routing at each adapter width; the fixed LowDim reference is the submitted partition in (a). Best value per column in bold. All 12 feature metrics: Appendix D.

				DINOv2 FID ↓		CXR FID ↓			
	Configuration	K	**min. expert**	**All**	**Macro**	**All**	**Macro**	**Macro F1 ↑**	**#nz ↑**
(a) Partition	Submitted partition (GRASP-LowDim)	4	9,243	91.4	150.6	.014	**.037**	.077	**9**
	Random reassignment (mean of 3)	4	7,927	84.2	139.6	.012	**.037**	.080	**9**
	Fewer pathology groups	3	17,080	89.5	149.7	.014	.041	.075	8
	More pathology groups	5	8,431	94.2	150.7	.014	.039	**.085**	8
	Literal official head/medium/tail [18]	3	71	80.6	317.6	.018	.075	.071	2
	One expert per label	19	12	106.6	447.2	**.011**	.050	.073	5
(b) Routing	MoE-GRASP-LowDim (learned gate)	4	–	98.6	155.2	.012	.051	.076	7
	Static-initialized learned gate (LowDim)	4	–	94.8	150.0	.013	.038	.081	6
	MoE-GRASP-HighDim (learned gate)	4	–	73.4	114.5	.017	.051	.070	7
	GRASP-HighDim (fixed)	4	9,243	**67.3**	**107.7**	.019	.039	.080	8

Table 5: Downstream DenseNet-121 F1 on real NIH-CXR-LT (13 shared labels, 16,506 images), trained only on images generated from MIMIC-CXR-LT test prompts (Real: real MIMIC-CXR-LT images). **#nz**: classes with nonzero mean F1. p: paired Wilcoxon over nine seed cells (exploratory). Learned-gate MoE-GRASP: Table 4(b). Per-label results: Appendix C.

Training data	**#nz**	**Macro F1 (13) ↑**	**No Finding**	p **vs GRASP-HD**
Real MIMIC-CXR-LT (reference)	3	$.079_{\pm.007}$	$.881_{\pm.002}$	–
FT (class)	1	$.068_{\pm.000}$	$\mathbf{.882_{\pm.000}}$	.004
FT (text+class)	1	$.068_{\pm.000}$	$\mathbf{.882_{\pm.000}}$	.004
MoCLE [14]	5	$.069_{\pm.002}$	$.881_{\pm.001}$	.004
CCUA [6]	1	$.068_{\pm.000}$	$\mathbf{.882_{\pm.000}}$	.004
Adapter (768d)	1	$.068_{\pm.000}$	$\mathbf{.882_{\pm.000}}$	.031
GRASP (768d)	9	$.076_{\pm.004}$	$.872_{\pm.006}$	.203
GRASP-LowDim	9	$.077_{\pm.005}$	$.868_{\pm.010}$	.359
GRASP-HighDim	8	$\mathbf{.080_{\pm.006}}$	$.848_{\pm.020}$	–

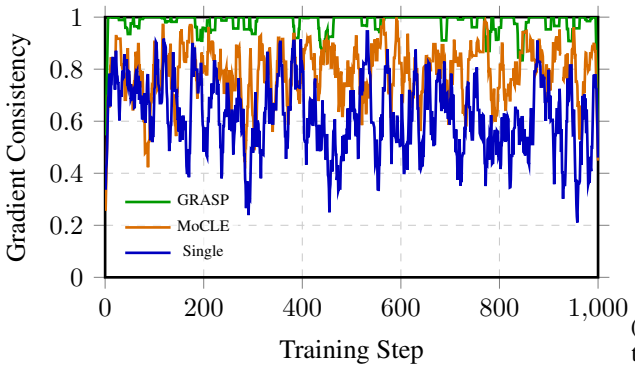
Every configuration whose smallest expert holds sufficient support behaves like ours, so minimum expert support matters more than which labels share an expert.

Fixed Versus Learned Routing. Table 4 isolates routing: MoE-GRASP is identical to its corresponding GRASP variant, with the only difference being an added 512-parameter top-1 gate. In both widths, fixed routing improves label-macro FID in both feature spaces, label-macro DINOv2 coverage and IRS, all-label DINOv2 FID and the number of recovered classes, and matches or improves macro F1, while the learned gate is ahead mainly on all-label metrics; with minority guidance applied to both, fixed routing also leads on CXR IRS. In a hybrid, we initialize the gate at the GRASP assignment and train the adapters from scratch. The gate does not drift across all checkpoints, and macro F1 is similar (.081), but recovery falls from 9/13 to 6/13. The learned gate thus neither finds a better assignment nor improves on a viable static one.

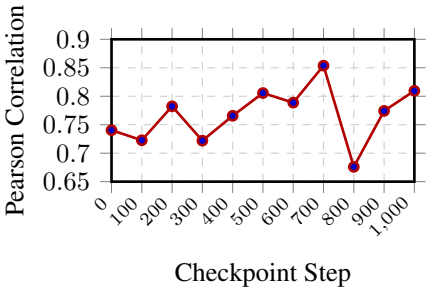
Downstream Classification Task. Table 5 summarizes DenseNet-121 F1 on real NIH-CXR-LT images for classifiers trained only on synthetic images (nine runs per cell; the full per-label results in Appendix C). The task is intrinsically hard: even real MIMIC-CXR-LT training data yields nonzero F1 only on *No Finding*, *Pleural Effusion*, and *Edema*. The informative result is therefore breadth rather than only macro F1. Full fine-tuning, CCUA and the shared adapter learn only the head class; MoCLE's learned cluster-conditional gate recovers limited tail signal. The static-partition GRASP variants improve further, producing nonzero mean F1 on eight or nine of thirteen classes while being comparable to the real-data macro F1. GRASP-LowDim recovers six classes on which real-data training has zero F1 and loses none in the reverse direction (one-sided: $p = .016$, two-sided $p = .031$). The remaining missed classes are extreme-tail labels with only a handful of training examples, where the real-data baseline also fails. Per-label gains over MoCLE are modest, often a few percentage points, and come with a small head-class reduction; the key result is that useful signal is spread across more labels. A paired two-sided Wilcoxon test on macro F1, paired by synthetic-dataset and classifier seed, separates GRASP-HighDim from every non-GRASP baseline ($p \leq 0.031$), but not from the other static-partition GRASP variants ($p \geq 0.20$).

Table 6: Ablation on resampling and clustering. Higher is better for Cov/IRS; lower is better for FID.

Label	Resample	Embed	Cov ↑	IRS ↑	FID ↓
✓	–	–	$.91_{\pm 1}$	$.52_{\pm 1}$	$\mathbf{.04_{\pm 05}}$
✓	✓	–	$\mathbf{.92_{\pm 09}}$	$\mathbf{.52_{\pm 12}}$	$\mathbf{.04_{\pm 05}}$
–	✓	✓	$.91_{\pm 1}$	$.52_{\pm 16}$	$.05_{\pm 06}$



(a) Gradient consistency (Eq. 1) over the first 1,000 training steps. GRASP stays highest, followed by MoCLE and the single-adapter baseline.



(b) Checkpoint-wise Pearson correlation between mean conditioning and mean gradient values across experts (GRASP, fixed checkpoints $0, 100, \ldots, 1000$).

Figure 4: **Gradient analysis.** (a) Conditioning-aware partitioning yields stronger optimization coherence, with GRASP maintaining the highest gradient consistency (Eq. 1) during training. (b) At fixed checkpoints, Pearson correlation remains consistently positive, supporting the link between conditioning homogeneity and gradient alignment.

Partitioning Source and Resampling. Table 6 ablates partition choice and round-robin resampling, which guarantees one sample per expert per step. Label partitions slightly outperform vanilla K-means, and resampling gives a small diversity boost at no quality cost; we keep it off in headline runs to keep the method minimal but recommend enabling it.

Gradient Analysis. Figure 4 supports the mechanism on the optimization trajectory: GRASP has the highest within-expert gradient consistency over the first 1,000 steps, with MoCLE intermediate and the single-adapter baseline lowest. Checkpoint-wise conditioning homogeneity also correlates positively with expert-wise gradient alignment.

Limitations. Downstream gains remain modest because the classifier task itself is extremely long-tailed, as shown by the real-data baseline collapsing on most labels. Also, the partition π is fixed before training and cannot be corrected from gradients; we accept this for the tail-routing guarantee, and empirically GRASP tolerates coarse partitions, including random label reassignments and vanilla K-means embeddings (Tables 4 and 6). Finally, the FLUX-scale backbone makes the fine-tuning compute-intensive (Appendix E, whereas GRASP itself adds little overhead.

5 Conclusion

GRASP is a non-invasive intervention for long-tail text-to-image flow matching. A static conditioning-space partition routes samples to group-specific residual adapters, improving gradient consistency without changing the objective or sampler. Unlike a learned gate, this routing guarantee does not depend on an optimization signal dominated by head classes, which is the failure mode we target in extreme long-tail regimes. Across MIMIC-CXR-LT, NIH-CXR-LT, and ImageNet-LT, GRASP improves rare-class quality and diversity, supports stronger downstream classifiers than non-GRASP synthetic baselines, and composes with inference-time minority guidance. Learned routing remains appropriate when no usable prior exists, but when such priors are available, static routing is the more stable default.

Funding

The authors gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) under the NHR project b180dc. NHR funding is provided by federal and Bavarian state authorities. NHR@FAU hardware is partially funded by the German Research Foundation (DFG) – 440719683.

Competing Interests

The authors declare no competing interests.

Ethics Statement

This work uses only de-identified, publicly available research datasets. MIMIC-CXR and its long-tailed split MIMIC-CXR-LT [24, 18] were accessed through PhysioNet under its credentialed-access data use agreement; NIH ChestX-ray14 / NIH-CXR-LT [44, 18] and ImageNet-LT [30] are publicly released. No new data were collected from human subjects. Improving rare-condition synthesis could make model development and stress testing less dependent on scarce medical data and may help expose failures on underrepresented pathologies. However, synthetic images are not a substitute for clinically grounded cohorts: they can preserve or amplify dataset biases, encode spurious correlations, and create misleading evidence if used without provenance and downstream validation. Our evaluation relies on feature-space metrics and a downstream classifier, not on expert radiological reading, so it does not establish clinical fidelity of individual images. We therefore frame GRASP as a controlled augmentation and analysis tool, not as an automatic replacement for real data or expert review, and synthetic images produced with it should not be used for clinical decision-making.

References

[1] Sumukh K Aithal, Pratyush Maini, Zachary Chase Lipton, and J Zico Kolter. Understanding hallucinations in diffusion models through mode interpolation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=aNTnHBkw4T.

[2] Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. In *arXiv preprint arXiv:2303.08797*, 2023.

[3] Lazar Atanackovic, Xi Zhang, Brandon Amos, Mathieu Blanchette, Leo J Lee, Yoshua Bengio, Alexander Tong, and Kirill Neklyudov. Meta flow matching: Integrating vector fields on the wasserstein manifold. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=9SYczU3Qgm.

[4] Pierre Chambon, Christian Bluethgen, Jean-Benoit Delbrouck, Rogier Van der Sluijs, Małgorzata Połacin, Juan Manuel Zambrano Chaves, Tanishq Mathew Abraham, Shivanshu Purohit, Curtis P Langlotz, and Akshay Chaudhari. A vision–language foundation model for the generation of realistic chest x-ray images. *Nature Biomedical Engineering*, 2025.

[5] Jannis Chemseddine, Paul Hagemann, Gabriele Steidl, and Christian Wald. Conditional wasserstein distances with applications in bayesian ot flow matching. *Journal of Machine Learning Research*, 26(141):1–47, 2025. URL http://jmlr.org/papers/v26/24-0586.html.

[6] Fang Chen, Alex Villa, Gongbo Liang, Xiaoyi Lu, and Meng Tang. Contrastive conditional-unconditional alignment for long-tailed diffusion model, 2025. URL https://arxiv.org/abs/2507.09052.

[7] Ho Kei Cheng and Alexander Schwing. The curse of conditions: Analyzing and improving optimal transport for conditional flow-based generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15875–15884, October 2025.

[8] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf.

[9] Mischa Dombrowski, Weitong Zhang, Sarah Cechnicka, Hadrien Reynaud, and Bernhard Kainz. Image generation diversity issues and how to tame them. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 3029–3039, June 2025.

[10] Fang Dong, Mengyi Chen, Jixian Zhou, Yubin Shi, Yixuan Chen, Mingzhi Dong, Yujiang Wang, Dongsheng Li, Xiaochen Yang, Rui Zhu, Robert Dick, Qin Lv, Fan Yang, Tun Lu, Ning Gu, and Li Shang. Once read is enough: Domain-specific pretraining-free language models with cluster-guided sparse experts for long-tail domain knowledge. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 88956–88980. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/a1f12d8d3cc1b4789ff4ebec46e27609-Paper-Conference.pdf.

[11] Rafael Elberg, Denis Parra, and Mircea Petrache. Long Tail Image Generation Through Feature Space Augmentation and Iterated Learning. *arXiv preprint arXiv:2405.01705*, 2024.

[12] Emanuele Francazi, Marco Baity-Jesi, and Aurelien Lucchi. A theoretical analysis of the learning dynamics under class imbalance. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10285–10322. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/francazi23a.html.

[13] Alireza Ganjdanesh, Yan Kang, Yuchen Liu, Richard Zhang, Zhe Lin, and Heng Huang. Mixture of Efficient Diffusion Experts Through Automatic Interval and Sub-Network Selection. In *European Conference on Computer Vision (ECCV)*, 2024.

[14] Yunhao Gou, Zhili Liu, Kai Chen, Lanqing Hong, Hang Xu, Aoxue Li, Dit-Yan Yeung, James T Kwok, and Yu Zhang. Mixture of cluster-conditional lora experts for vision-language instruction tuning. *arXiv preprint arXiv:2312.12379*, 2023.

[15] Pengfei Guo, Can Zhao, Dong Yang, Ziyue Xu, Vishwesh Nath, Yucheng Tang, Benjamin Simon, Mason Belue, Stephanie Harmon, Baris Turkbey, et al. Maisi: Medical ai for synthetic imaging. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4430–4441. IEEE, 2025.

[16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Neural Information Processing Systems*, 2017. URL https://api.semanticscholar.org/CorpusID:326772.

[17] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. URL https://openreview.net/forum?id=qw8AKxfYbI.

[18] Gregory Holste, Song Wang, Ziyu Jiang, Thomas C. Shen, George Shih, Ronald M. Summers, Yifan Peng, and Zhangyang Wang. Long-tailed classification of thorax diseases on chest x-ray: A new benchmark study. In Hien V. Nguyen, Sharon X. Huang, and Yuan Xue, editors, *Data Augmentation, Labelling, and Imperfections*, pages 22–32, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-17027-0.

[19] Yaxin Hou and Yuheng Jia. A square peg in a square hole: Meta-expert for long-tailed semi-supervised learning. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=h0ZeiDRN8A.

[20] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/houlsby19a.html.

[21] 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.

[22] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

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

[24] Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathan R. Greenbaum, Matthew P. Lungren, Catherine Y. Deng, Roger G. Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, 2019. doi: 10.1038/s41597-019-0322-0.

[25] Rabeeh Karimi Mahabadi, Sebastian Ruder, Mostafa Dehghani, and James Henderson. Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 565–576, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.47. URL https://aclanthology.org/2021.acl-long.47/.

[26] Gavin Kerrigan, Giosue Migliorini, and Padhraic Smyth. Dynamic conditional optimal transport through simulation-free flows. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 93602–93642. Curran Associates, Inc., 2024. doi: 10.52202/079017-2968. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/aa1d8cf866c4a684ef2e066dd200f8e4-Paper-Conference.pdf.

[27] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *CoRR*, abs/1904.06991, 2019.

[28] Yinghao Li, Vianne R. Gao, Chao Zhang, and MohamadAli Torkamani. Ensembles of low-rank expert adapters. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=l0gZS0sAlf.

[29] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[30] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. Large-scale long-tailed recognition in an open world. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[31] Zichen Miao, Jiang Wang, Ze Wang, Zhengyuan Yang, Lijuan Wang, Qiang Qiu, and Zicheng Liu. Training diffusion models towards diverse image generation with reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10844–10853, June 2024.

[32] Mashrur M. Morshed and Vishnu Boddeti. Diverseflow: Sample-efficient diverse mode coverage in flows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23303–23312, June 2025.

[33] Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7176–7185. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/naeem20a.html.

[34] Weizhi Nie, Zichun Zhang, Weijie Wang, Bruno Lepri, Anan Liu, and Nicu Sebe. Causal disentanglement for robust long-tail medical image generation, 2025. URL https://arxiv.org/abs/2504.14450.

[35] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

[36] Daehee Park, Monu Surana, Pranav Desai, Ashish Mehta, Reuben John, and Kuk-Jin Yoon. Generative active learning for long-tail trajectory prediction via controllable diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 27839–27850, 2025.

[37] Sivaramakrishnan Rajaraman, Zhaohui Liang, Zhiyun Xue, and Sameer K. Antani. Addressing class imbalance with latent diffusion-based data augmentation for improving disease classification in pediatric chest x-rays. In *BIBM*, pages 5059–5066, 2024. URL https://doi.org/10.1109/BIBM62325.2024.10822172.

[38] S-A Rebuffi, H. Bilen, and A. Vedaldi. Learning multiple visual domains with residual adapters. In *Advances in Neural Information Processing Systems*, 2017.

[39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

[40] Vikash Sehwag, Caner Hazirbas, Albert Gordo, Firat Ozgenel, and Cristian Canton Ferrer. Generating High Fidelity Data from Low-density Regions using Diffusion Models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11482–11491, New Orleans, LA, USA, June 2022. IEEE. ISBN 978-1-6654-6946-3. doi: 10.1109/CVPR52688.2022.01120. URL https://ieeexplore.ieee.org/document/9878507/.

[41] 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. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=B1ckMDqlg.

[42] Soobin Um and Jong Chul Ye. Self-Guided Generation of Minority Samples Using Diffusion Models. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 414–430, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73113-6. doi: 10.1007/978-3-031-73113-6_24.

[43] Soobin Um, Suhyeon Lee, and Jong Chul Ye. Don't play favorites: Minority guidance for diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3Nm09lY4Jn.

[44] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jul 2017. doi: 10.1109/CVPR.2017.369.

[45] Yujie Wei, Shiwei Zhang, Hangjie Yuan, Yujin Han, Zhekai Chen, Jiayu Wang, Difan Zou, Xihui Liu, Yingya Zhang, Yu Liu, et al. Routing matters in moe: Scaling diffusion transformers with explicit routing guidance. In *International Conference on Learning Representations*, 2026.

[46] Haoning Wu, Ziheng Zhao, Ya Zhang, Yanfeng Wang, and Weidi Xie. Mrgen: Segmentation data engine for underrepresented mri modalities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19903–19913, 2025.

[47] Xun Wu, Shaohan Huang, and Furu Wei. Mixture of loRA experts. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=uWvKBCYh4S.

[48] Zeyue Xue, Guanglu Song, Qiushan Guo, Boxiao Liu, Zhuofan Zong, Yu Liu, and Ping Luo. Raphael: Text-to-image generation via large mixture of diffusion paths. *Advances in Neural Information Processing Systems*, 36:41693–41706, 2023.

[49] Hantao Zhang, Yuhe Liu, Jiancheng Yang, Shouhong Wan, Xinyuan Wang, Wei Peng, and Pascal Fua. LeFusion: Controllable Pathology Synthesis via Lesion-Focused Diffusion Models. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=9H1N6nC8aF.

[50] Li Zhang, Basu Jindal, Ahmed Alaa, Robert Weinreb, David Wilson, Eran Segal, James Zou, and Pengtao Xie. Generative ai enables medical image segmentation in ultra low-data regimes. *Nature Communications*, 16(1):6486, 2025.

[51] Tianjiao Zhang, Huangjie Zheng, Jiangchao Yao, Xiangfeng Wang, Mingyuan Zhou, Ya Zhang, and Yanfeng Wang. Long-tailed diffusion models with oriented calibration. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=NW2s5XXwXU.

[52] Zhe Zhao, Haibin Wen, Zikang Wang, Pengkun Wang, Fanfu Wang, Song Lai, Qingfu Zhang, and Yang Wang. Breaking long-tailed learning bottlenecks: A controllable paradigm with hypernetwork-generated diverse experts. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 7493–7520. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/0e21697f3b9385f6c4100aad94c26bed-Paper-Conference.pdf.

[53] Youwei Zheng, Yuxi Ren, Xin Xia, Xuefeng Xiao, and Xiaohua Xie. Dense2MoE: Restructuring Diffusion Transformer to MoE for Efficient Text-to-Image Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 18661–18670, October 2025.

A Cluster compositions

We visualize the cluster compositions induced by the two partition strategies in Figure 5: the label-based partition (frequency-binned MIMIC-CXR labels with a healthy split) and the text-embedding partition (vanilla bisecting K-means on frozen text-encoder features). The label-based partition assigns each sample to a single cluster by frequency band, while the text-based partition produces softer, conditioning-space groupings whose label distributions overlap.

B Expert specialization and resampling

Figure 6(a) reports per-expert label-wise diversity for the label-based partition. Experts inherit the conditioning composition of their cluster: the expert assigned to the (single-label) healthy cluster performs best on *No Finding* and worse on most pathologies, while mixed experts perform best on the labels that dominate their cluster (e.g., the Expert 0 cluster has the largest share of *Enlarged Mediastinum* and *Fracture*, and Expert 0 is best on those labels). The relationship is consistent but noisy, with occasional outliers. Figure 6(b) shows the effect of resampling on expert-wise diversity: resampling redistributes utilization (Table 7) and yields a small but consistent overall improvement, at the cost of a slight per-expert quality reduction in the previously dominant healthy expert.

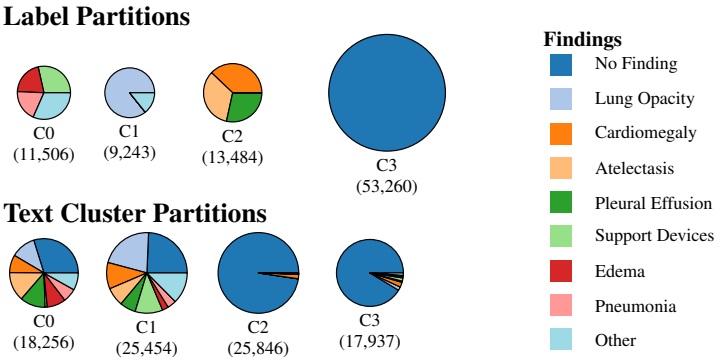
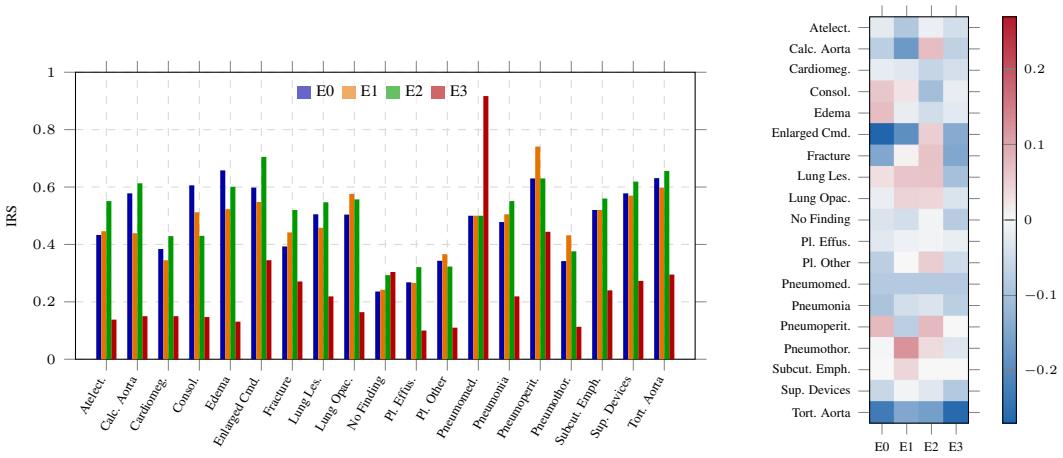


Figure 5: Composition of the partitioning based on labels (top) compared to the partitioning based on text clusters (bottom).



(a) Expert specialization shown by label-wise IRS scores for each expert (no-resample setting).

(b) ΔIRS (no-resample $-$ resample).

Figure 6: Comparison of expert specialization and resampling impact.

C Per-label downstream classification results

Table 8 reports the full per-label F1 scores summarized in Table 5 of the main text. We include this table for completeness; the main-text discussion (Section 4) is based on these values, in particular the nonzero-coverage count (**#nz**) and the macro F1 over all 13 classes. The four classes *Calc. of the Aorta*, *Pneumomediastinum*, *Pneumoperitoneum*, and *Tortuous Aorta* collapse to true zero F1 across every method (including the real-data baseline), reflecting the extreme tail of the MIMIC-CXR-LT distribution.

Table 7: Expert usage percentages by partitioning and sampling method for MIMIC-CXR-LT.

Label	Embed	Resample	Expert 0 (%)	Expert 1 (%)	Expert 2 (%)	Expert 3 (%)
✓			13.01	10.34	15.44	61.20
✓		✓	24.03	19.78	17.08	39.11
	✓		21.82	28.79	29.29	20.10
	✓	✓	25.64	26.46	25.81	22.09

Table 8: Downstream DenseNet-121 classification, full per-label F1 on the real NIH-CXR-LT test set restricted to the 13 labels shared with MIMIC-CXR-LT (16,506 images). Classifiers are trained only on images generated from the 23,550 MIMIC-CXR-LT test prompts (Real: real MIMIC-CXR-LT images). Mean±SD across the nine runs per cell (three synthetic-dataset seeds × three classifier seeds; three classifier seeds for Real). Best values per row are shown in bold. Entries reported as $.000_{\pm.000}$ are true zeros (no correct predictions across any of the nine runs), not rounded near-zeros. The body summary view is in Table 5.

	Real	Baselines					Learned gate		GRASP		
Label	**Real**	**FT (class)**	**FT (text+class)**	**MoCLE**	**CCUA**	**Adapter (768d)**	**MoE-GRASP-LowDim**	**MoE-GRASP-HighDim**	**GRASP (768d)**	**GRASP-LowDim**	**GRASP-HighDim**
No Finding	$.881_{\pm.002}$	$\mathbf{.882_{\pm.000}}$	$\mathbf{.882_{\pm.000}}$	$.881_{\pm.001}$	$\mathbf{.882_{\pm.000}}$	$\mathbf{.882_{\pm.000}}$	$.870_{\pm.005}$	$.879_{\pm.004}$	$.872_{\pm.006}$	$.868_{\pm.010}$	$.848_{\pm.020}$
Pleural Effusion	$\mathbf{.130_{\pm.091}}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.004_{\pm.004}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.081_{\pm.046}$	$.013_{\pm.019}$	$.023_{\pm.029}$	$.052_{\pm.046}$	$.082_{\pm.061}$
Edema	$.011_{\pm.009}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.012_{\pm.023}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.008_{\pm.014}$	$.005_{\pm.014}$	$\mathbf{.061_{\pm.027}}$	$.032_{\pm.019}$	$.044_{\pm.036}$
Atelectasis	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.001}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.007_{\pm.011}$	$.004_{\pm.009}$	$.004_{\pm.006}$	$.016_{\pm.026}$	$\mathbf{.022_{\pm.025}}$
Cardiomegaly	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.007_{\pm.010}$	$.001_{\pm.002}$	$.012_{\pm.012}$	$.003_{\pm.005}$	$\mathbf{.020_{\pm.014}}$
Consolidation	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.003_{\pm.007}$	$.000_{\pm.000}$	$.002_{\pm.002}$	$\mathbf{.014_{\pm.018}}$	$.012_{\pm.025}$
Pneumonia	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.006_{\pm.009}$	$.004_{\pm.012}$	$.004_{\pm.012}$	$\mathbf{.007_{\pm.008}}$	$.007_{\pm.012}$
Pneumothorax	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.001}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.002_{\pm.003}$	$.003_{\pm.006}$	$.004_{\pm.008}$	$\mathbf{.007_{\pm.008}}$
Subcut. Emphysema	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$\mathbf{.004_{\pm.012}}$	$.004_{\pm.012}$	$.000_{\pm.000}$
Calc. of the Aorta	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$
Pneumomediastinum	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$
Pneumoperitoneum	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$
Tortuous Aorta	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$	$.000_{\pm.000}$
Macro Avg	$.079_{\pm.007}$	$.068_{\pm.000}$	$.068_{\pm.000}$	$.069_{\pm.002}$	$.068_{\pm.000}$	$.068_{\pm.000}$	$.076_{\pm.003}$	$.070_{\pm.003}$	$.076_{\pm.004}$	$.077_{\pm.005}$	$\mathbf{.080_{\pm.006}}$

D Partition sensitivity and routing controls: full results

Table 9 reports all 12 feature-space metrics for the controls summarized in Table 4, and for learned versus fixed routing under minority guidance. Random reassignments are generated within the construction of Section 3: each map keeps the four experts, the reserved healthy expert and whole-label assignment, and fills the three pathology groups toward the same target loads as the submitted map; the map is fixed across training and evaluation. The $K = 3$ and $K = 5$ variants use two and four pathology groups plus the healthy expert. The literal head/medium/tail variant routes the official MIMIC-CXR-LT frequency strata [18] to three experts, and the per-label variant uses one expert for each of the 19 labels. The two starved designs degrade mainly in label-macro FID, whose large across-label SD reflects the collapse of the rarest labels. For downstream recovery, the random maps are pooled over three maps × three classifier seeds; individually they recover 9, 7 and 5 of 13 classes.

E Compute Resources

All experiments were conducted on a shared academic research cluster using NVIDIA H100 GPUs with 80GB memory per GPU. Training and sampling were implemented in PyTorch and used mixed precision. We report compute in H100 GPU-hours, computed as wall-clock runtime multiplied by the number of GPUs.

Each generative model was fine-tuned for 10,000 steps (CCUA: 20,000) with a batch size of 8 on 8 H100 GPUs. A full fine-tuning run required approximately 30 wall-clock hours, corresponding to about 240 H100 GPU-hours per seed. Adapter-based runs, including GRASP variants, required approximately 7-12 wall-clock hours, corresponding to about 56-96 H100 GPU-hours per seed. For each model and seed, we generated 50,000 images at 1024×1024 resolution. This required approximately 12 wall-clock hours on 8 H100 GPUs, plus additional feature extraction and metric computation.

The primary GRASP configuration, evaluated over three generative seeds, requires approximately 550-650 H100 GPU-hours including training, image generation, and metric computation. Reproducing the primary comparison against the strongest baseline requires approximately 1,100-1,300 H100 GPU-hours.

Reproducing the main MIMIC-CXR-LT generation results, including full fine-tuning baselines, MoCLE, sequential adapters, GRASP variants, learned-routing variants, and minority-guided sampling evaluations, requires approximately 5,500-6,500 H100 GPU-hours. Including the ImageNet-LT experiments increases the estimate to approximately 9,500-11,500 H100 GPU-hours.

The downstream DenseNet-121 evaluation uses three classifier seeds on each of three synthetic dataset seeds, yielding nine classifier runs per configuration. Including these downstream classifier runs, the ablation experiments, gradient-analysis runs, and qualitative sample generation, we estimate that reproducing all reported quantitative results requires approximately 11,000-14,000 H100 GPU-hours.

Table 9: Full feature-space metrics for the partition-sensitivity, routing and external-baseline controls on MIMIC-CXR-LT (GRASP-LowDim width unless stated). All-label ± is the SD over three generator seeds (submitted GRASP-LowDim and CCUA) or over the three random maps; other variants use generator seed 0 and are point estimates. Macro ± is the SD across the 19 label-wise values, as in Tables 2 and 3. CCUA is trained for 20k steps at per-GPU batch size 2 (half of GRASP's sample exposures; ∼28 h on 8 GPUs per seed, 2–4× GRASP's adapter training). The two +MG rows compare learned and fixed routing under minority guidance (three generator seeds each).

		All labels			**Macro Avg**		
Configuration	K	**Cov.** ↑	**IRS** ↑	**FID** ↓	**Cov.** ↑	**IRS** ↑	**FID** ↓
DINOv2 features							
Submitted partition (GRASP-LowDim)	4	$.009_{\pm.001}$	$.169_{\pm.003}$	$91.36_{\pm1.59}$	$.116_{\pm.098}$	$.376_{\pm.166}$	$150.59_{\pm76.15}$
Random reassignment (mean of 3)	4	$.012_{\pm.000}$	$.196_{\pm.001}$	$84.22_{\pm4.11}$	$.128_{\pm.097}$	$.394_{\pm.150}$	$139.63_{\pm66.06}$
Random map 0	4	.012	.194	87.26	$.126_{\pm.093}$	$.381_{\pm.138}$	$142.91_{\pm78.86}$
Random map 1	4	.012	.195	86.98	$.119_{\pm.082}$	$.397_{\pm.156}$	$142.46_{\pm75.37}$
Random map 2	4	.011	.198	78.41	$.139_{\pm.128}$	$.403_{\pm.165}$	$133.51_{\pm62.54}$
Fewer pathology groups	3	.012	.189	89.48	$.119_{\pm.078}$	$.404_{\pm.167}$	$149.65_{\pm72.57}$
More pathology groups	5	.010	.186	94.16	$.110_{\pm.101}$	$.421_{\pm.192}$	$150.67_{\pm75.88}$
Literal official head/medium/tail	3	.012	.182	80.61	$.071_{\pm.060}$	$.388_{\pm.218}$	$317.55_{\pm490.91}$
One expert per label	19	.009	.188	106.59	$.035_{\pm.023}$	$.384_{\pm.184}$	$447.19_{\pm468.42}$
Static-initialized learned gate	4	.010	.187	94.80	$.106_{\pm.100}$	$.408_{\pm.170}$	$149.96_{\pm74.13}$
MoE-GRASP-LowDim + MG (learned)	4	$.012_{\pm.001}$	$.218_{\pm.002}$	$90.55_{\pm4.17}$	$.112_{\pm.096}$	$.412_{\pm.173}$	$151.35_{\pm65.46}$
GRASP-LowDim + MG (fixed)	4	$.014_{\pm.001}$	$.200_{\pm.004}$	$83.99_{\pm1.58}$	$.140_{\pm.107}$	$.429_{\pm.191}$	$142.04_{\pm75.33}$
CCUA [6]	–	$.010_{\pm.003}$	$.100_{\pm.007}$	$117.72_{\pm39.49}$	$.107_{\pm.083}$	$.247_{\pm.099}$	$157.50_{\pm26.43}$
CXR features							
Submitted partition (GRASP-LowDim)	4	$.942_{\pm.001}$	$.503_{\pm.003}$	$.0139_{\pm.0009}$	$.924_{\pm.086}$	$.555_{\pm.125}$	$.0368_{\pm.0465}$
Random reassignment (mean of 3)	4	$.953_{\pm.003}$	$.687_{\pm.006}$	$.0119_{\pm.0002}$	$.926_{\pm.094}$	$.617_{\pm.099}$	$.0368_{\pm.0386}$
Random map 0	4	.955	.695	.0116	$.934_{\pm.088}$	$.601_{\pm.126}$	$.0354_{\pm.0432}$
Random map 1	4	.948	.685	.0122	$.921_{\pm.088}$	$.625_{\pm.122}$	$.0376_{\pm.0372}$
Random map 2	4	.954	.680	.0118	$.923_{\pm.114}$	$.626_{\pm.134}$	$.0375_{\pm.0373}$
Fewer pathology groups	3	.952	.660	.0139	$.922_{\pm.094}$	$.665_{\pm.157}$	$.0406_{\pm.0379}$
More pathology groups	5	.941	.656	.0138	$.920_{\pm.108}$	$.580_{\pm.137}$	$.0393_{\pm.0384}$
Literal official head/medium/tail	3	.944	.624	.0181	$.887_{\pm.121}$	$.503_{\pm.165}$	$.0754_{\pm.0788}$
One expert per label	19	.956	.678	.0105	$.903_{\pm.119}$	$.540_{\pm.170}$	$.0496_{\pm.0683}$
Static-initialized learned gate	4	.948	.675	.0133	$.938_{\pm.085}$	$.652_{\pm.160}$	$.0383_{\pm.0464}$
MoE-GRASP-LowDim + MG (learned)	4	$.941_{\pm.005}$	$.632_{\pm.019}$	$.0118_{\pm.0009}$	$.920_{\pm.096}$	$.604_{\pm.166}$	$.0512_{\pm.0622}$
GRASP-LowDim + MG (fixed)	4	$.949_{\pm.000}$	$.679_{\pm.002}$	$.0135_{\pm.0008}$	$.921_{\pm.093}$	$.614_{\pm.122}$	$.0392_{\pm.0478}$
CCUA [6]	–	$.636_{\pm.031}$	$.126_{\pm.006}$	$.0671_{\pm.0035}$	$.673_{\pm.153}$	$.202_{\pm.178}$	$.1457_{\pm.1144}$

The additional controls (from Slurm accounting) required approximately 4,100 GPU-hours: about 2,300 for the partition-sensitivity and static-initialized learned-gate runs (eight GRASP-LowDim trainings of about 150 GPU-hours each, one generator seed, plus generation, evaluation and classifiers), about 1,000 for CCUA (three seeds of 20k steps, about 28 wall-clock hours on 8 GPUs per seed, plus generation and classifiers), and about 840 for the fixed-router minority-guidance regenerations. Reproducing all reported quantitative results therefore requires approximately 15,000-18,000 GPU-hours.

Including exploratory development, debugging runs, failed configurations, preliminary hyperparameter choices, and intermediate ablations, we estimate the total compute used over the course of the project to be approximately 22,000-32,000 GPU-hours.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the proposed long-tail generative modeling approach and present the empirical scope of the claims. The claims are tied to the reported ImageNet-LT and MIMIC-CXR-LT experiments.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations of the empirical setting in a specified section, including the focus on the evaluated long-tailed image benchmarks, as well as the inherent trade-off between routing guarantees and adaptability is discussed. Additionally, the dependence on the existence of (imperfect) priors is acknowledged in the paper.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper is primarily empirical and does not introduce formal theoretical results requiring theorem statements or proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The experimental sections and appendix describe the datasets, baselines, model variants, training setup, evaluation metrics, and seeds needed to reproduce the main results. The implementation scripts in the provided source code further specify the exact training and evaluation commands.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide code and reproduction scripts for the proposed method and baselines. The raw datasets are established third-party resources and must be obtained through their original access procedures and terms rather than redistributed by us.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the relevant data splits, baselines, optimization details, hyperparameters, model choices, and evaluation protocols in the experimental section and appendix. Additional command-level details are provided in the accompanying code.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: Main quantitative results are reported over multiple random seeds with variability indicated in the tables. The paper states that the reported uncertainty corresponds to variation across repeated runs. Significance tests are specified and applied on the main results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: In the appendix, we specify estimates for GPU hours needed for each experiment, as well as the estimated total amount of GPU hours needed for this paper. Additionally, we specify details such as hardware requirements, numerical precision and so on.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research uses established datasets and pretrained models under their respective access procedures and terms. We do not collect new human-subject data, perform user studies, or deploy the resulting models; the work is limited to offline evaluation of long-tail generative modeling methods.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the potential benefits of improved data synthesis for rare classes and the risks of generative misuse or biased downstream deployment.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release a high-risk scraped dataset or unrestricted general-purpose image generator. Any released code relies on existing gated or third-party model and dataset access procedures.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper cites the existing datasets, pretrained models, and software components used in the experiments and follows their access terms. Raw third-party data are not redistributed.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The new asset introduced by the paper is the accompanying implementation and experiment scripts, which are documented with reproduction commands and configuration files. No new human-subject dataset is introduced.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing experiments, new data collection from participants, or human-subject user studies.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or human-subject research requiring IRB or equivalent approval.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not used as an important, original, or non-standard component of the core methodology or experimental evaluation.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.