ANATOMICALLY PLAUSIBLE HUMAN IMAGE GENERATION VIA SYNTHETIC LOCALIZED PREFERENCES

Anonymous authors
Paper under double-blind review

ABSTRACT

Large-scale text-to-image foundation models have substantially improved the visual realism of generated human images. However, their outputs may still exhibit anatomical artifacts in structurally complex regions. Preference-based alignment offers an architecture-preserving approach to improving anatomical plausibility. In practice, conventional preference pairs typically compare independently generated images whose differences extend beyond anatomy, making it unclear whether anatomical quality drives the preference. To provide direct anatomical supervision, we present Alignment via Synthetic Anatomical Preferences (ASAP), a data-centric framework built on synthetic localized preferences. ASAP constructs synthetic localized preferences by using verified anatomically plausible images as preferred anchors and creating dispreferred images via localized degradation, while largely preserving prompt semantics and non-target visual content. This pipeline yields the Human Anatomical Preference (HAP) dataset, comprising over 10K curated pairs with region masks and anatomical error metadata. To exploit their spatial attribution while preserving overall image quality, we adapt Direct Preference Optimization (DPO) with a region-aware preference gap that emphasizes modified regions and a margin-bounded regression objective that drives the gap toward a finite target. We further introduce HAF-Bench, a benchmark for systematic evaluation of anatomical fidelity across diverse scenarios. Extensive experiments across diffusion and flow-matching backbones demonstrate that ASAP consistently reduces anatomical errors while maintaining overall image quality.

1 INTRODUCTION

Human image generation has made remarkable progress in visual quality, prompt alignment, and stylistic diversity with the development of text-to-image foundation models (Esser et al., 2024; Labs, 2024; Podell et al., 2023; Wu et al., 2025). However, anatomical artifacts can still arise in challenging cases, ranging from malformed hands and incorrect finger counts to distorted facial structures (Ju et al., 2023b; Li et al., 2024; Wang et al., 2025c). Such errors are particularly detrimental because human observers are highly sensitive to subtle deviations in faces, hands, and body structures. Existing approaches address this problem through human-centric training data, structural conditions, part-specialized modules, or inference-time refinement (Fang et al., 2024; Li et al., 2024; Wang et al., 2025a; Zhu et al., 2024). While effective, these solutions often require task-specific components, additional control signals, or inference-time correction. Preference-based alignment offers a complementary, architecture-preserving alternative for teaching pretrained generators which anatomical structures should be preferred and which failures should be avoided.

The effectiveness of preference-based alignment depends critically on the supervision encoded in its preference pairs. Conventional text-to-image preference pairs typically compare independently generated images from the same prompt and subsequently ranked by users, annotators, reward models, or vision-language models (Kirstain et al., 2023; Ma et al., 2025; Wu et al., 2023a; Xu et al., 2023). Such pairs are well suited for learning holistic preferences such as aesthetics and prompt alignment, but are not constructed to target specific failure modes. Human-focused approaches such as HG-DPO (Na et al., 2025) narrow the data domain to human images, but their paired images may still differ across multiple semantic and visual attributes. Consequently, the preference label is not determined by anatomical quality, limiting its effectiveness as anatomy-specific supervision.

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



Figure 1: **Effectiveness of ASAP.** Foundation models such as FLUX and SDXL still struggle with complex human anatomy, ASAP significantly improves anatomical plausibility.

To address this problem, we present *Alignment via Synthetic Anatomical Preferences (ASAP)*, an anatomy-focused preference-alignment framework built on synthetic localized preferences. Rather than pairing independently generated images, we curate human-verified synthetic images with plausible anatomy as preferred anchors and construct corresponding negatives by introducing localized anatomical errors in image or skeleton space. Since each negative is constructed from its preferred anchor, the pair largely preserves prompt semantics and non-target visual content, making the localized anatomical quality the principal source of preference. Using this pipeline, we construct the Human Anatomical Preference (HAP) dataset, comprising over 10K curated preference pairs with region masks and anatomical error metadata. Although synthetic localized preferences provide clear anatomy-specific supervision, standard DPO does not exploit their spatial attribution and continues to reward larger preference gaps. Our experiments suggest that excessive preference gaps may compromise overall image quality despite improvements in anatomical plausibility. We therefore adapt DPO with a mask-aware, margin-bounded objective that emphasizes the modified regions and drives the preference gap towards a finite target margin. For systematic evaluation, we further introduce HAF-Bench, a benchmark comprising 500 prompts balanced across five anatomical categories, together with a standardized evaluation protocol and complementary absolute and relative metrics.

Our contributions are summarized as follows: (1) We introduce synthetic localized preferences for anatomical supervision and construct HAP, a dataset of over 10K pairs derived from human-verified synthetic anchors through localized degradations, with region masks and error metadata. (2) We propose a region-aware, margin-bounded preference objective that emphasizes modified regions while constraining excessive preference separation, improving the balance between anatomical plausibility and overall image quality. (3) We establish HAF-Bench, a 500-prompt benchmark for systematically evaluating anatomical plausibility. Experiments across diffusion- and flow-based generators demonstrate that ASAP improves anatomical plausibility over the pretrained generators and achieves a better balance between anatomical plausibility and global image quality.

2 Related Work

2.1 Human Image Generation

Human image generation is a central task in modern image synthesis, with broad applications in digital content creation and virtual humans (Jia et al., 2024). Despite substantial advances in visual realism, modern text-to-image models (Labs, 2024; Podell et al., 2023; Wu et al., 2025) still produce artifacts in structurally complex regions, particularly hands, faces, and articulated body parts. Existing

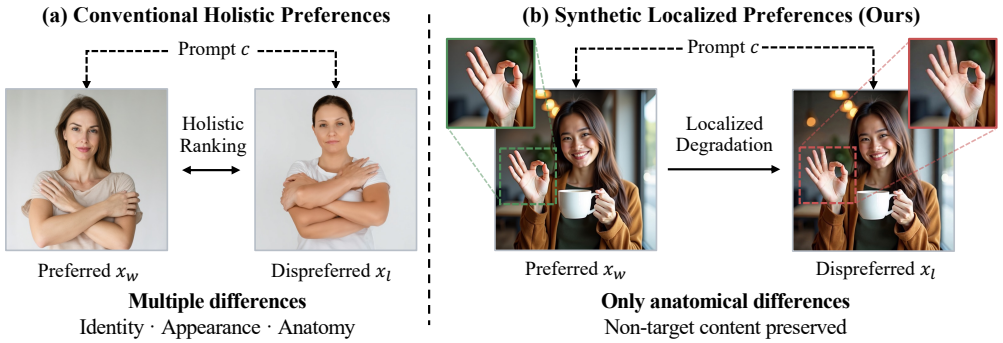

Figure 2: **Conventional holistic preferences vs. synthetic localized preferences.** Conventional pairs may differ across attributes, obscuring the source of preference. ASAP constructs synthetic localized preferences through localized degradations, providing anatomy-specific preference signals.

approaches address these failures by incorporating human-specific priors at different stages of the generation. At training time, HyperHuman and CosmicMan leverage large-scale human-centric data with detailed structural or semantic annotations, while HumanSD uses pose conditions to constrain body layout and articulation (Ju et al., 2023b; Li et al., 2024; Liu et al., 2023). MoLE and FairHuman instead focus on anatomically sensitive regions through specialized experts and spatial supervision (Wang et al., 2025c; Zhu et al., 2024). Other methods operate after generation: HumanRefiner combines pose-guided regeneration with localized inpainting, whereas RealisHuman regenerates malformed human parts and integrates them into the original image (Fang et al., 2024; Wang et al., 2025a). Our work explores a complementary, data-centric alignment approach that improves anatomical plausibility without modifying the generator architecture or inference pipeline.

2.2 Alignment with Human Preference

As foundation models are typically trained with huge and diverse unlabeled datasets (Penedo et al., 2023; Raffel et al., 2020), the distributions they capture generally deviate from what humans prefer. In image generation, such misalignment is often observed as inferior aesthetics, text-image misalignment, image artifacts (Kirstain et al., 2023; Liang et al., 2024; Schuhmann et al., 2022). This misalignment is much more severe when it comes to human image generation since humans can easily identify visual artifacts of human bodies in generated images (Mori et al., 2012). Preference alignment is commonly performed through reinforcement learning with human feedback (RLHF), where a learned reward guides optimization, or direct preference optimization (DPO), which directly learns from preference pairs (Ouyang et al., 2022; Rafailov et al., 2023). Originally proposed for large language models, DPO has been adapted to diffusion and flow-matching models (Liu et al., 2026; Wallace et al., 2024). However, DPO adjusts the likelihood of all pixels in an image uniformly, which can be inefficient since human preferences often depend on only a small subset of regions. Recent patch- and region-aware alignment methods use local feedback or spatial constraints (Huang et al., 2025; Wang et al., 2025b; Xing et al., 2025). For anatomical alignment, our objective emphasizes modified regions using degradation masks and drives the preference gap toward a finite target, improving anatomical plausibility while mitigating global image-quality degradation.

2.3 Preference Data Construction for Visual Alignment

Visual preference datasets commonly compare images independently generated for the same prompt and collect image-level judgments from users or annotators (Kirstain et al., 2023; Ma et al., 2025; Wu et al., 2023b; Xu et al., 2023). Such pairs are effective for learning holistic preferences, but may differ across multiple semantic and visual attributes, obscuring the source of each preference. HG-DPO narrows the data domain and improves preferred-sample quality using real human images, yet retains a holistic relationship between the paired samples (Na et al., 2025). Recent studies move toward reducing non-target differences by constructing closely matched preference pairs. D-Fusion improves visual consistency for prompt alignment, while Di3PO, RealAlign, and LocalDPO develop task-specific preference construction strategies for text rendering, general visual alignment, and video generation, respectively (Chen et al., 2026; Hu et al., 2025; Huang et al., 2026; Reddy et al., 2026). Within this line of work, ASAP address how to construct preference supervision for anatomical failures, which are local in appearance yet structured in pose and connectivity.

3 PRELIMINARIES

Diffusion and Flow-Matching Models. The prevalent image generation models are mostly built with diffusion models (*e.g.*, Stable Diffusion (Rombach et al., 2022)) and flow-matching models (*e.g.*, FLUX (Labs, 2024)). Both can be described through time-dependent regression objectives. Let $\mathbf{x}_1$ denote a clean image, $\mathbf{x}_0 \sim \mathcal{N}(0, I)$ an initial noise sample, and $\mathbf{x}_t$ their intermediate state at time t. Using v_θ to denote the model prediction, the model is trained with the following loss:

$$\mathcal{L}(\theta) = \mathbb{E}_{t,\mathbf{x}_0,\mathbf{x}_1}\Big[w(t)\|v_\theta(\mathbf{x}_t, t) - u(\mathbf{x}_1, \mathbf{x}_0, t)\|^2\Big], \tag{1}$$

where $u(\mathbf{x}_1, \mathbf{x}_0, t)$ is the target velocity field induced by the reference path and $w(t)$ is a weighting factor depending on the model design (often set to 1). For instance, in diffusion models, $u(\mathbf{x}_t, \mathbf{x}_0, t) = \dot{\alpha}_t \mathbf{x}_0 + \dot{s}_t(\mathbf{x}_t - \alpha_t \mathbf{x}_0)/s_t$, where $\alpha(t)$ and $s(t)$ are the diffusion schedules; in a popular variant called rectified flow (Liu et al., 2022), $u(\mathbf{x}_1, \mathbf{x}_0, t) = \mathbf{x}_1 - \mathbf{x}_0$.

Direct Preference Optimization. Direct Preference Optimization (DPO) (Rafailov et al., 2023) aligns generative models with human preferences directly using preference datasets without any reinforcement learning process. Diffusion-DPO (Wallace et al., 2024) adapts this principle to visual generative models by utilizing the model training loss as a proxy for log-likelihoods and building an approximate DPO loss. Specifically, given a condition $\mathbf{c}$ and a preference pair $(\mathbf{x}_w, \mathbf{x}_l)$, the approximate preference gap is defined as (Liu et al., 2026; Ziv et al., 2025):

$$\delta_\theta(\mathbf{x}_w, \mathbf{x}_l, \mathbf{c}) \approx -\Big(\mathcal{L}^\theta(\mathbf{x}_w) - \mathcal{L}^\theta(\mathbf{x}_l) - \big(\mathcal{L}^{\text{ref}}(\mathbf{x}_w) - \mathcal{L}^{\text{ref}}(\mathbf{x}_l)\big)\Big), \tag{2}$$

where $\mathcal{L}$ implicitly conditions on $\mathbf{c}$ for brevity. Under the assumption of Bradley-Terry preference model (Bradley & Terry, 1952; Rafailov et al., 2023), the so-called Diffusion-DPO loss is

$$\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{(\mathbf{x}_w,\mathbf{x}_l,\mathbf{c})\sim\mathcal{D}}\big[\log\sigma(\beta\delta_\theta(\mathbf{x}_w, \mathbf{x}_l, \mathbf{c}))\big], \tag{3}$$

where σ is the sigmoid function and β controls the deviation from the reference model. Flow-matching models can be similarly fine-tuned with it (Liu et al., 2026) to yield satisfactory results.

4 METHODS

As illustrated in Fig. 3, ASAP comprises two stages. We first construct the HAP dataset through synthetic localized preferences and then adapt DPO with mask-aware weighting and a finite target margin to exploit the spatial attribution provided by these pairs.

4.1 CONSTRUCTION OF SYNTHETIC LOCALIZED PREFERENCES

Preferred Sample Construction. We first construct a pool of human-verified synthetic images with plausible anatomy as preferred samples. Instead of using existing datasets in the domain of human image generation, we adopt a pose-conditioned generation pipeline to create synthetic images for this dataset, so as to avoid the domain gap between photorealistic images and synthetic images during preference learning. Given a source human image, we extract its pose skeleton $\mathbf{p}$ using DWPose (Yang et al., 2023) and use a vision-language model (Comanici et al., 2025) to refine the corresponding prompt into a detailed caption $\mathbf{c}_{\text{aug}}$. These conditions guide a pretrained FLUX-ControlNet (Shakker-Labs, 2025) to generate candidate preferred samples $\mathbf{x}_w \sim G(\mathbf{c}_{\text{aug}}, \mathbf{p})$, from which annotators remove images with implausible anatomy. Each retained sample subsequently serves as the anchor for constructing its dispreferred counterpart.

Dispreferred Sample Construction. For each preferred sample $\mathbf{x}_w$, we construct a corresponding dispreferred sample $\mathbf{x}_l$ by introducing a localized anatomical error in either image or skeleton space. For image-space gradation, GroundingSAM2 (Ren et al., 2024) extracts a binary mask $\mathbf{M}$ for the target region, which is regenerated using an inpainting model conditioned on an anatomical-error instruction. For skeleton-space degradation, we perturb $\mathbf{p}$ by modifying anatomical keypoints or connections to obtain $\mathbf{p}_{\text{error}}$, and regenerate $\mathbf{x}_l$ using the same caption and FLUX-ControlNet. The affected region and error type are retained as the region mask and anatomical error metadata.

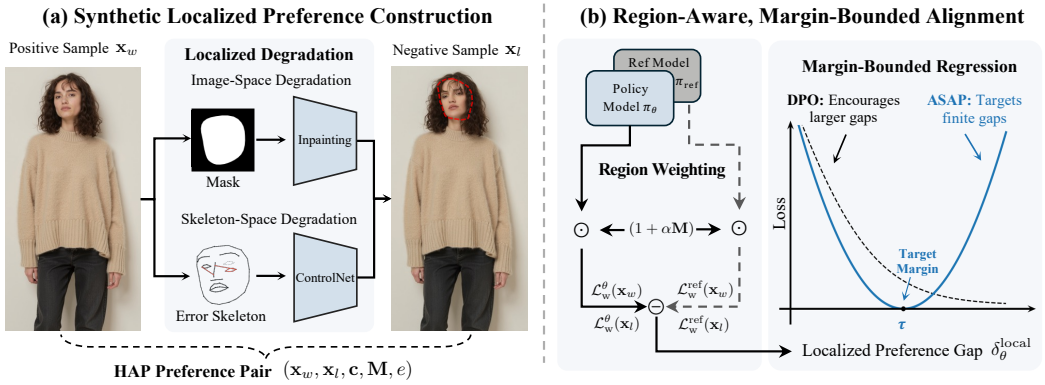

Figure 3: **Overview of the ASAP. Phase I (left): Synthetic localized preference construction.** Starting from a verified preferred anchor $\mathbf{x}_w$, we introduce localized anatomical errors through image-space inpainting or skeleton-space perturbation to construct the corresponding negative $\mathbf{x}_l$. **Phase II (Right): Region-aware and margin-bounded preference optimization** . The region mask defines the spatial weighting $1 + \alpha\mathbf{M}$ used to compute the localized preference gap. Rather than rewarding larger gaps as in standard DPO, our margin-bounded objective regresses the gap toward a finite target τ, emphasizing the modified regions while mitigating over-optimization.

HAP Dataset Curation. Applying the two construction strategies produces approximately 25K candidate preference pairs. We employ a panel of 11 human annotators to manually screen these pairs according to three criteria: (1) the preferred sample exhibits plausible anatomy and satisfactory overall quality, (2) the dispreferred sample contains a clearly identifiable anatomical error, and (3) prompt semantics and non-target visual content remain sufficiently consistent within the pair. A pair is retained only when all three criteria are satisfied. After screening, the Human Anatomical Preference (HAP) dataset contains over 10K curated tuples $(\mathbf{x}_w, \mathbf{x}_l, \mathbf{c}, \mathbf{M}, \mathbf{e})$, where $\mathbf{e}$ denotes the anatomical error metadata. Further curation details are provided in *Sup.Mat.*'s Sec. B. Tab. 1 confirms this locality, with substantially higher SSIM and lower LPIPS outside the degradation masks.

Table 1: **Spatial locality of HAP pairs.**

Region	SSIM	LPIPS
Masked	0.5315	0.3446
Non-masked	0.9827	0.0109

4.2 Region-Aware and Margin-Bounded Preference Optimization

Each synthetic localized preference pair includes a region mask $\mathbf{M}$ that identifies where the anatomical error is introduced. Standard image-level DPO neither explicitly exploits this spatial attribution nor specifies a finite preference gap at which the desired correction is considered sufficient. We therefore adapt DPO with a region-aware preference gap and a margin-bounded objective.

Localized Preference Gap. Given an image $\mathbf{x}$, its text condition $\mathbf{c}$, and the corresponding region mask $\mathbf{M}$, we define the mask-weighted training loss as

$$\mathcal{L}_{\text{w}}(\theta; \mathbf{x}, \mathbf{c}, \mathbf{M}) = \mathbb{E}_{t,\mathbf{x}_0}\left[\left\|\sqrt{1+\alpha\mathbf{M}} \odot [v_\theta(\mathbf{x}_t, t, \mathbf{c}) - (\mathbf{x}_1 - \mathbf{x}_0)]\right\|_2^2\right], \tag{4}$$

where $\odot$ denotes element-wise multiplication, and α controls the weighting strength. The factor $1+\alpha\mathbf{M}$ retains unit weight outside the mask and assigns additional weight to the target region.

The localized preference gap is thus ($\mathcal{L}_w^\theta(\mathbf{x})$ stands for $\mathcal{L}_w(\theta; \mathbf{x}, \mathbf{c})$):

$$\delta_\theta^{\text{local}}(\mathbf{x}_w, \mathbf{x}_l, \mathbf{c}, \mathbf{M}) \approx -\Big(\mathcal{L}_{\text{w}}^\theta(\mathbf{x}_w) - \mathcal{L}_{\text{w}}^\theta(\mathbf{x}_l) - \big(\mathcal{L}_{\text{w}}^{\text{ref}}(\mathbf{x}_w) - \mathcal{L}_{\text{w}}^{\text{ref}}(\mathbf{x}_l)\big)\Big). \tag{5}$$

Margin-Bounded Regression. To constrain the localized preference gap to a finite target, we adopt the margin-constrained formulation to the DPO variant for flow matching models (Azar et al., 2024; Liu et al., 2026) and define the following regression objective:

$$\mathcal{L}_{\text{Bounded}}(\theta) = \mathbb{E}_{(\mathbf{x}_w, \mathbf{x}_l, \mathbf{c}, \mathbf{M})\sim\mathcal{D}}\left[\left(\delta_\theta^{\text{local}}(\mathbf{x}_w, \mathbf{x}_l, \mathbf{c}, \mathbf{M}) - \tau\right)^2\right], \tag{6}$$

where τ denotes the finite target margin. By penalizing deviations from τ, the objective prevents excessive preference gaps.

5 HAF-BENCH

We introduce HAF-Bench (Human Anatomical Fidelity Benchmark), a 500-prompt benchmark for evaluating anatomical plausibility in generated human images. Complementing prior human-image benchmarks designed for artifact localization or broad quality assessment, HAF-Bench targets structural integrity, articulation, and physically plausible interactions across diverse scenarios.

Prompts and Taxonomy. The prompts are generated with an LLM and subsequently filtered to remove ambiguous descriptions. HAF-Bench contains 100 prompts in each of five complementary categories: (1) Basic Anatomy, covering the integrity of individual body parts; (2) Semantic Gestures, covering meaning-specific hand gestures; (3) Self-Interaction, covering contact between body parts; (4) Object Interaction, covering human–object contact; and (5) Full Body, covering global pose and anatomical connectivity.

Evaluation Protocol. We use Gemini-3-Pro as an automated judge under a fixed anatomy-focused instruction. For absolute evaluation, the judge assigns each generated image a Pass label if no anatomical artifact is detected and a Fail label otherwise. For relative evaluation, it compares two images generated from the same prompt and returns Win, Tie, or Lose according to their anatomical plausibility. All methods use identical instructions; judge prompts and protocol details are provided in *Sup.Mat.*'s Sec. C. On the full evaluation, Gemini-3-Pro agrees with human judgments in 82% of cases, supporting its use as a scalable anatomy evaluator.

Metrics. For absolute evaluation, we report the Anatomical Error Rate (AER), defined as the proportion of generated images judged to contain anatomical artifact. Let $\mathcal{D}_{\mathrm{HAF}}$ denote the set of evaluated images generated from HAF-Bench prompts, and let $\mathcal{S}_{\mathrm{VLM}}(\mathbf{x}) \in \{\text{Pass}, \text{Fail}\}$ denote the judgment for image $\mathbf{x}$. AER is computed as

$$\mathrm{AER} = \frac{1}{|\mathcal{D}_{\mathrm{HAF}}|} \sum_{\mathbf{x} \in \mathcal{D}_{\mathrm{HAF}}} \mathbb{I}(\mathcal{S}_{\mathrm{VLM}}(\mathbf{x}) = \text{Fail}), \tag{7}$$

where each image contributes a single binary outcome regardless of the number or severity of its artifacts. A lower AER indicates better anatomical plausibility.

For relative evaluation, we define the Relative Superiority Index (RSI) to summarize pairwise outcomes while retaining ties:

$$\mathrm{RSI} = \frac{N_{\mathrm{win}} + N_{\mathrm{tie}}}{N_{\mathrm{lose}} + N_{\mathrm{tie}}}, \tag{8}$$

where N_{win}, N_{tie}, and N_{lose} denote the corresponding outcomes from the evaluated model's perspective. RSI is greater than 1 exactly when wins outnumber losses. Because ties enter the numerator and denominator symmetrically, tie-heavy comparisons produce values closer to 1.

6 EXPERIMENTS

6.1 EXPERIMENTAL SETTINGS

Base Models. We evaluate ASAP on two text-to-image backbones: FLUX.1-dev (Labs, 2024), based on flow matching, and SDXL (Podell et al., 2023), based on diffusion. Together, they evaluate the applicability of ASAP across two widely used generative formulations and model scales.

Baselines. We compare supervised fine-tuning and preference optimization methods. The supervised baselines include: (1) *SFT*, which fine-tunes each backbone using only preferred HAP samples; and (2) *FairHuman*† (Wang et al., 2025c), which jointly optimizes global, hand, and face objectives. The preference optimization baselines include: (3) *HG-DPO*† (Na et al., 2025), which constructs preference pairs by selecting higher- and lower-reward images from generated candidates for each prompt; and (4) *ASAP-DPO*, which applies vanilla DPO to the same HAP pairs without the localized preference gap or margin-bounded regression. *ASAP* denotes our complete framework. Further baseline and implementation details are provided in *Sup.Mat.*'s Sec. D.2.

Table 2: **Quantitative comparison on HAF-Bench.** AER and RSI assess anatomy quality; the remaining metrics assess global image quality. Base provides the reference; adapted methods are grouped into supervised fine-tuning and preference optimization for each backbone.

Method	*Anatomy Quality*		*Global Image Quality*				
	AER ↓	RSI ↑	HPS v3 ↑	HPS v2.1 ↑	PickScore ↑	ImageReward ↑	AesScore ↑
FLUX.1-dev:							
Base	0.417	1.00 (Ref.)	10.307	0.303	22.797	1.037	5.793
Supervised fine-tuning							
SFT	0.391	1.03	10.320	0.303	22.808	1.097	**5.822**
FairHuman†	0.346	1.09	10.324	**0.304**	22.792	1.093	5.806
Preference optimization							
HG-DPO†	0.371	1.05	10.323	0.302	**22.967**	1.098	5.772
ASAP-DPO (Ours)	0.293	1.13	9.680	0.299	22.476	1.062	5.752
ASAP (Ours)	**0.289**	**1.17**	**10.360**	**0.304**	22.814	**1.102**	5.779
SDXL:							
Base	0.873	1.00 (Ref.)	6.371	0.277	22.205	0.737	5.709
Supervised fine-tuning							
SFT	0.820	1.23	7.233	0.289	22.369	0.770	**5.766**
FairHuman†	0.791	1.42	7.319	0.286	22.341	0.759	5.763
Preference optimization							
HG-DPO†	0.804	1.31	7.368	0.287	**22.512**	**0.781**	5.750
ASAP-DPO (Ours)	0.775	1.49	6.526	0.279	22.254	0.751	5.723
ASAP (Ours)	**0.764**	**1.58**	**7.450**	**0.291**	22.364	0.778	5.754

Evaluation Benchmarks and Metrics. We evaluate anatomical plausibility on HAF-Bench using AER for absolute evaluation and RSI for pairwise comparison. RSI compares each adapted model against its corresponding base model, whose value is 1.00 by definition. To assess whether anatomical improvements preserve general visual quality and prompt alignment, we additionally report HPSv3 (Ma et al., 2025), HPSv2.1 (Wu et al., 2023b), PickScore (Kirstain et al., 2023), ImageReward (Xu et al., 2023), and Aesthetic Score (Schuhmann et al., 2022). All methods use identical prompts and sampling settings, with each metric averaged over five seeds.

Implementation Details. All adapted models are fine-tuned using rank-32 LoRA (Hu et al., 2022) for 3,000 steps with an effective batch size of 16. For FLUX.1-dev, we use a learning rate of 5×10^{-6}. For ASAP, we set the spatial weighting coefficient to $\alpha = 10$ and the target margin to $\tau = 0.01$. Detailed hyperparameter settings for the SDXL are provided in *Sup.Mat.*'s Sec. D.1.

6.2 General Results

Quantitative Analysis. Tab. 2 compares ASAP with supervised fine-tuning and preference-optimization baselines on HAF-Bench. Across both backbones, ASAP achieves the lowest AER and highest RSI while maintaining competitive global image-quality scores. Applying vanilla DPO to the same synthetic localized preferences (ASAP-DPO) already substantially reduces anatomical errors, demonstrating the effectiveness of HAP as preference supervision. However, this variant consistently degrades global quality. ASAP retains these anatomical improvements while largely recovering the global metrics, indicating that the region-aware, margin-bounded objective better balances localized anatomical correction and global quality preservation.

Qualitative Analysis. As shown in Fig. 1, ASAP consistently reduces anatomical artifacts in images generated by both FLUX and SDXL, particularly malformed hands and faces, while preserving the overall image quality. To compare different fine-tuning strategies, Fig. 4 further presents samples generated by the base FLUX model, SFT, ASAP-DPO, and ASAP. The base model exhibits visible anatomical errors, which SFT often leaves unresolved. ASAP-DPO improves anatomical plausibility but at the cost of lower overall visual quality. In contrast, ASAP retains these anatomical improvements while better preserving global visual quality.

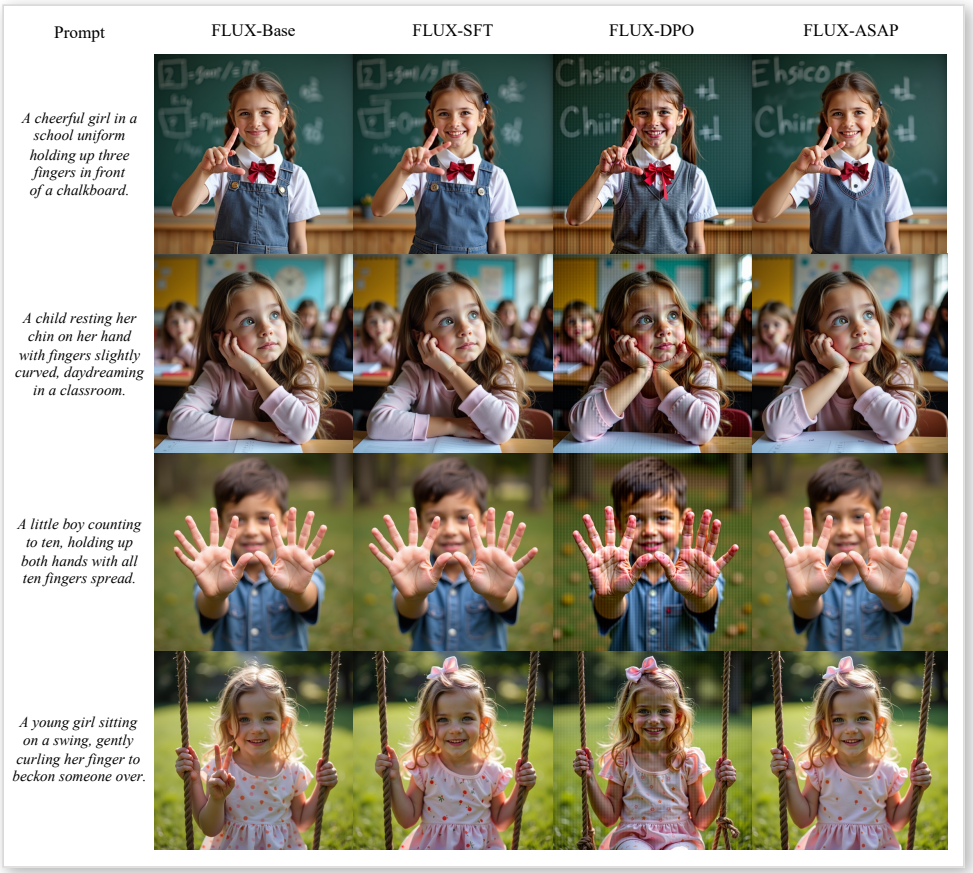

Figure 4: **Qualitative comparison of fine-tuning strategies on FLUX.** SFT uses the preferred images, ASAP-DPO applies vanilla DPO to the synthetic preference pairs, and ASAP further introduces region-aware, margin-bounded objective. ASAP corrects localized anatomical errors while better preserving non-target content and overall image quality.

6.3 Analysis and Ablation Studies

Comparison of Preference Construction Strategies. To assess preference-pair construction, we compare four strategies under vanilla DPO objective: (1) *Standard Pairwise Ranking*, which ranks two independently generated images for each prompt; (2) *HG-DPO-Easy* (Na et al., 2025), which selects the highest- and lowest-scoring samples from a larger candidate pool; (3) *Real-Anchor Localized Preferences*, which construct localized negatives from real human images; and (4) our *Synthetic Localized Preferences*, constructed from human-verified synthetic images. All variants use the same prompt set, number of training pairs, and optimization settings, construction details are provided in *Sup.Mat.*'s Sec. D.3. As shown in Tab. 3, synthetic localized preferences achieve the lowest AER and highest RSI across both backbones. Their advantage over ranking-based strategies demonstrates the benefit of reducing non-target differences, while the comparison with real anchors suggests that matching the preference data to the model's generative domain improves anatomical alignment.

Ablation of the Alignment Objective. Tab. 4 evaluates the two objective components while keeping the HAP dataset and training settings fixed. Using the localized preference gap alone achieves the lowest AER, showing that spatially focused supervision strengthens anatomical correction. Bounded regression instead substantially improves HPSv3 but produces more conservative anatomical correction. Combining both components improves AER over vanilla DPO while achieving the highest HPSv3, balancing between anatomical plausibility and overall generation quality.

Sensitivity to the Target Margin. We examine how the target margin τ affects anatomical plausibility and overall generation quality, measured by AER and HPSv3, respectively. As shown in Fig. 5, small margins ($\tau \leq 0.001$) constrain the preference gap too tightly, resulting in insufficient anatomical correction and higher AER. In contrast, large margins ($\tau \geq 0.1$) provide little further

Table 3: **Comparison of preference construction strategies.** All variants use vanilla DPO with matched data and training settings.

	FLUX.1-dev		SDXL	
Preference Strategy	AER ↓	RSI ↑	AER ↓	RSI ↑
Base	0.417	1.00	0.873	1.00
Standard Pairwise Ranking	0.415	0.98	0.823	1.13
HG-DPO-Easy	0.371	1.05	0.804	1.31
Real-Anchor Localized	0.356	1.08	0.783	1.43
Synthetic Localized (Ours)	**0.293**	**1.13**	**0.775**	**1.49**

Table 4: **Ablation of the alignment objective.** All variants use the same HAP pairs with matched training settings on FLUX.1-dev.

Localized	Bounded	AER ↓	HPSv3 ↑
×	×	0.293	9.680
✓	×	*0.283*	*9.717*
×	✓	*0.337*	*10.355*
✓	✓	0.289	10.360

Note: The first row corresponds to vanilla DPO, and the last row to the complete ASAP objective.

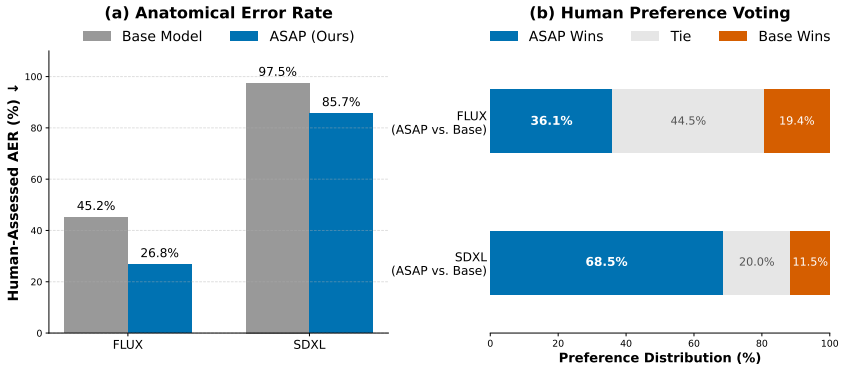

Figure 6: **Human Evaluation.** (a) Human-assessed AER for the base and ASAP-tuned models. ASAP reduces anatomical errors on both FLUX and SDXL. (b) Pairwise preference voting between ASAP and the base model. ASAP wins more frequently than the base model for both backbones.

improvement in AER while substantially reducing HPSv3. Among the evaluated settings, $\tau = 0.01$ achieves the highest HPSv3 score of 10.36 with a near-minimal AER of 0.289, providing the best empirical balance between anatomical plausibility and overall generation quality.

6.4 Human Evaluation

To assess whether the automated results align with human perception, we conduct human evaluation on the full HAF-Bench using the same absolute and pairwise protocols. As shown in Fig. 6(a), ASAP reduces human-assessed AER for both FLUX and SDXL, confirming that its anatomical improvements are perceptible to human observers. Pairwise voting in Fig. 6(b) further favors ASAP over the corresponding base models: ASAP wins more frequently than the base model for both backbones. Together, these results corroborate the automated evaluation and provide consistent human evidence for the effectiveness of ASAP.

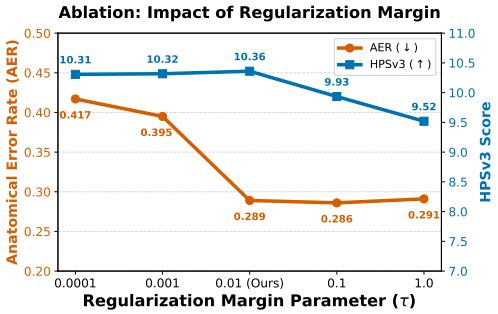

Figure 5: **Sensitivity to the Target Margin.**AER (↓) and HPSv3 (↑) across values of τ. Small margins result in insufficient anatomical correction, whereas large margins reduce HPSv3 with little further improvement in AER. $\tau = 0.01$ achieves the highest HPSv3 while maintaining near-minimal AER, providing the best trade-off.

7 Conclusion

In this paper, we present ASAP, a data-centric preference-alignment framework for improving anatomical plausibility in human image generation. ASAP constructs synthetic localized preferences by introducing localized anatomical degradations into human-verified synthetic anchors, and couples them with a region-aware, margin-bounded objective that balances anatomical correction with overall visual quality. We further introduce HAF-Bench for systematic evaluation across diverse anatomy-sensitive scenarios. Experiments on flow-matching and diffusion backbones show that ASAP reduces anatomical errors while preserving overall image quality, with human evaluations corroborating these improvements. These results demonstrate the value of anatomy-specific synthetic localized preferences for fine-grained alignment of text-to-image models.

REFERENCES

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pp. 4447–4455. PMLR, 2024. 5

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. 4

Weiyan Chen, Weijian Deng, Yao Xiao, Weijie Tu, ZiYi Dong, Ibrahim Radwan, Liang Lin, and Pengxu Wei. When preference labels fall short: Aligning diffusion models from real data. *arXiv preprint arXiv:2605.19839*, 2026. 3

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 4

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 1

Guian Fang, Wenbiao Yan, Yuanfan Guo, Jianhua Han, Zutao Jiang, Hang Xu, Shengcai Liao, and Xiaodan Liang. Humanrefiner: Benchmarking abnormal human generation and refining with coarse-to-fine pose-reversible guidance. In *European Conference on Computer Vision*, pp. 201–217. Springer, 2024. 1, 3

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 7, 15

Zijing Hu, Fengda Zhang, and Kun Kuang. D-fusion: Direct preference optimization for aligning diffusion models with visually consistent samples. *arXiv preprint arXiv:2505.22002*, 2025. 3

Qihan Huang, Long Chan, Jinlong Liu, Wanggui He, Hao Jiang, Mingli Song, and Jie Song. Patchdpo: Patch-level dpo for finetuning-free personalized image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18369–18378, 2025. 3

Zitong Huang, Kaidong Zhang, Yukang Ding, Chao Gao, Rui Ding, Ying Chen, and Wangmeng Zuo. Mind the generative details: Direct localized detail preference optimization for video diffusion models. *arXiv preprint arXiv:2601.04068*, 2026. 3

Zhen Jia, Zhang Zhang, Liang Wang, and Tieniu Tan. Human image generation: A comprehensive survey. *ACM Computing Surveys*, 56(11):1–39, 2024. 2

Xuan Ju, Ailing Zeng, Jianan Wang, Qiang Xu, and Lei Zhang. Human-art: A versatile human-centric dataset bridging natural and artificial scenes. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 618–629. IEEE, 2023a. 16

Xuan Ju, Ailing Zeng, Chenchen Zhao, Jianan Wang, Lei Zhang, and Qiang Xu. Humansd: A native skeleton-guided diffusion model for human image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15988–15998, 2023b. 1, 3

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023. 1, 3, 7

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 1, 2, 4, 6, 15

Shikai Li, Jianglin Fu, Kaiyuan Liu, Wentao Wang, Kwan-Yee Lin, and Wayne Wu. Cosmicman: A text-to-image foundation model for humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6955–6965, 2024. 1, 3

Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, et al. Rich human feedback for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19401–19411, 2024. 3

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, et al. Improving video generation with human feedback. *Advances in Neural Information Processing Systems*, 38:82155–82192, 2026. 3, 4, 5

Xian Liu, Jian Ren, Aliaksandr Siarohin, Ivan Skorokhodov, Yanyu Li, Dahua Lin, Xihui Liu, Ziwei Liu, and Sergey Tulyakov. Hyperhuman: Hyper-realistic human generation with latent structural diffusion. *arXiv preprint arXiv:2310.08579*, 2023. 3

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 4

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 15

Yuhang Ma, Yunhao Shui, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. *arXiv preprint arXiv:2508.03789*, 2025. 1, 3, 7

Masahiro Mori, Karl F MacDorman, and Norri Kageki. The uncanny valley [from the field]. *IEEE Robotics & automation magazine*, 19(2):98–100, 2012. 3

Sanghyeon Na, Yonggyu Kim, and Hyunjoon Lee. Boost your human image generation model via direct preference optimization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 23551–23562, 2025. 1, 3, 6, 8, 15

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022. 3

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*, 2023. 3

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1, 2, 6

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023. 3, 4

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 3

Sanjana Reddy, Ishaan Malhi, Sally Ma, and Praneet Dutta. Di3po-diptych diffusion dpo for targeted improvements in image generation. *arXiv preprint arXiv:2602.06355*, 2026. 3

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 4

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022. 4, 15

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022. 3, 7

Shakker-Labs. Controlnet-union. https://huggingface.co/Shakker-Labs/FLUX.1-dev-ControlNet-Union-Pro-2.0, 2025. 4

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8228–8238, 2024. 3, 4

Benzhi Wang, Jingkai Zhou, Jingqi Bai, Yang Yang, Weihua Chen, Fan Wang, and Zhen Lei. Realishuman: A two-stage approach for refining malformed human parts in generated images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7509–7517, 2025a. 1, 3

Shuting Wang, Haihong Tang, Zhicheng Dou, and Chenyan Xiong. Harness local rewards for global benefits: Effective text-to-video generation alignment with patch-level reward models. *arXiv preprint arXiv:2502.06812*, 2025b. 3

Yuxuan Wang, Tianwei Cao, Huayu Zhang, Zhongjiang He, Kongming Liang, and Zhanyu Ma. Fairhuman: Boosting hand and face quality in human image generation with minimum potential delay fairness in diffusion models. *arXiv preprint arXiv:2507.02714*, 2025c. 1, 3, 6, 15

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025. 1, 2

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023a. 1

Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2096–2105, 2023b. 3, 7

Xiaoying Xing, Avinab Saha, Junfeng He, Susan Hao, Paul Vicol, Moonkyung Ryu, Gang Li, Sahil Singla, Sarah Young, Yinxiao Li, et al. Focus-n-fix: Region-aware fine-tuning for text-to-image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18486–18496, 2025. 3

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023. 1, 3, 7

Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4210–4220, 2023. 4

Jie Zhu, Yixiong Chen, Mingyu Ding, Ping Luo, Leye Wang, and Jingdong Wang. Mole: Enhancing human-centric text-to-image diffusion via mixture of low-rank experts. *Advances in Neural Information Processing Systems*, 37:29354–29386, 2024. 1, 3

Alon Ziv, Sanyuan Chen, Andros Tjandra, Yossi Adi, Wei-Ning Hsu, and Bowen Shi. Mr-flowdpo: Multi-reward direct preference optimization for flow-matching text-to-music generation. *arXiv preprint arXiv:2512.10264*, 2025. 4

Appendix

Table of Contents

A Use of Large Language Models **14**

B HAP Dataset Construction and Validation **14**
B.1 Data Curation Protocol 14
B.2 Representative Failure Modes 14

C HAF-Bench Evaluation Protocols **14**
C.1 Automated Evaluation 14
C.2 Human Evaluation and Agreement 14

D Implementation Details **15**
D.1 Training Settings 15
D.2 Baseline Adaptations 15
D.3 Preference Construction Strategies 16

E Additional Benchmark Results **16**

F DPO Hyperparameter and Early-Stopping Analysis **17**

G Limitations **17**

A USE OF LARGE LANGUAGE MODELS

We used a large language model to assist with copy editing—grammar checking, wording suggestions, and minor style and clarity improvements—after the scientific content, methodology, analyses, and conclusions had been written by the authors.

B HAP DATASET CONSTRUCTION AND VALIDATION

B.1 DATA CURATION PROTOCOL

The localized degradation pipelines initially produce approximately 25K candidate preference pairs. A panel of 11 trained annotators independently screens the candidate pool, retaining a pair only when all three of the following criteria are satisfied:

- **Plausible Preferred Sample.** The preferred image $\mathbf{x}_w$ must exhibit plausible anatomy and satisfactory overall visual quality. Pairs are discarded if the original synthetic anchor contains perceptible anatomical artifacts.
- **Identifiable Anatomical Degradation.** The dispreferred image $\mathbf{x}_l$ must contain a clearly identifiable anatomical error. Pairs with imperceptible modifications or without meaningful structural degradation are discarded.
- **Preserved Non-Target Content.** Prompt semantics and visual content outside the degraded region must remain sufficiently consistent within the pair. Pairs exhibiting noticeable unintended changes in identity, composition, background, or other non-target regions are discarded.

After filtering, 10,652 pairs, corresponding to approximately 42.6% of the candidate pool, form the final HAP dataset. Each retained tuple $(\mathbf{x}_w, \mathbf{x}_l, \mathbf{c}, \mathbf{M}, e)$ contains a preferred image, a dispreferred image, the corresponding caption, a degradation mask, and anatomical error metadata.

B.2 REPRESENTATIVE FAILURE MODES

We further examine whether the constructed negatives resemble anatomical failures that arise naturally in text-to-image generation. As shown in Fig. 7, untuned model outputs and synthetic negatives exhibit analogous defects, including extra or missing fingers and malformed facial structures. Beyond this visual correspondence, ASAP-DPO, trained on HAP pairs using the vanilla DPO objective, substantially reduces AER on unseen HAF-Bench prompts, as reported in Tab. 2. This downstream generalization indicates that HAP captures transferable anatomical failure patterns rather than arbitrary image corruptions. Together, the qualitative examples and quantitative improvements support the validity of HAP as anatomy-specific preference supervision, although the constructed negatives are not intended to cover the complete distribution of anatomical failures.

C HAF-BENCH EVALUATION PROTOCOLS

C.1 AUTOMATED EVALUATION

Gemini-3-Pro serves as the automated judge under fixed anatomy-focused instructions shared across methods. In the absolute protocol, each generated image receives a Pass or Fail label as in Tab. 5. AER is the fraction labeled Fail, with each image contributing one binary outcome irrespective of the number or severity of its artifacts. In the pairwise protocol, two images generated from the same prompt receive a Win, Tie, or Lose judgment from the evaluated model's perspective. RSI is computed as $(N_{\text{win}} + N_{\text{tie}})/(N_{\text{lose}} + N_{\text{tie}})$. The base model is the reference for each adapted model.

C.2 HUMAN EVALUATION AND AGREEMENT

Eleven human raters assess anatomical plausibility under protocols paralleling the automated evaluation. For absolute evaluation, they provide Yes/No judgments on anatomical artifacts. For pairwise evaluation, they select the more anatomically plausible image or a tie, with image order randomized. The main paper reports human-assessed AER and preference distributions for ASAP versus the

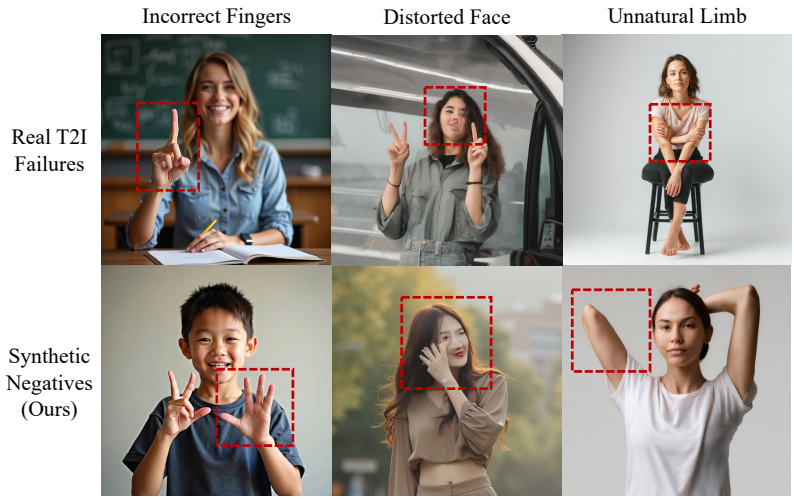

Figure 7: **Synthetic negatives capture representative anatomical failure patterns.** Each column pairs an anatomical failure observed in real T2I generation (top) with a synthetic negative from HAP (bottom) exhibiting the same error type. Highlighted regions indicate the anatomical artifacts.

corresponding base model on both backbones. To measure human–VLM agreement, we compare the human evaluation results from Sec. 6.4 with the corresponding Gemini-3-Pro automated judgments across the entire HAF-Bench. We pool absolute Yes/No and pairwise Win/Lose/Tie decisions, counting agreement only when the labels match exactly. This comparison yields an overall agreement rate of 82%. Although human and automated AER differ in value, both consistently favor ASAP. Human-assessed AER decreases from 45.2% to 26.8% for FLUX and from 97.5% to 85.7% for SDXL; automated AER decreases from 41.7% to 28.9% and from 87.3% to 76.4%, respectively. These results support a consistent improvement in anatomical plausibility.

D Implementation Details

D.1 Training Settings

As introduced in the main paper, we use LoRA (Hu et al., 2022) with a rank of 32. All experiments are conducted on 8 NVIDIA A100 (80GB) GPUs. All images are processed at a resolution of 1024×1024. For the optimization process, we employ the AdamW (Loshchilov & Hutter, 2017) optimizer with a weight decay of 0.01. Both models are trained for 3,000 steps using an effective batch size of 16, accompanied by a linear learning rate warm-up of 200 steps. Specifically for the FLUX (Labs, 2024) backbone, we apply a learning rate of $5e-6$, and set the spatial weighting scalar $\alpha = 10$ along with a regularization margin of $\tau = 0.01$ for our alignment objective. For SDXL (Rombach et al., 2022) backbone, we apply a learning rate of $5e-5$, and set the spatial weighting scalar $\alpha = 10$ along with a regularization margin of $\tau = 0.002$ for our alignment objective.

D.2 Baseline Adaptations

SFT fine-tunes each backbone using only the preferred images from HAP, without preference optimization. *ASAP-DPO* instead applies vanilla DPO to exactly the same HAP pairs used by ASAP, but removes the localized preference gap and margin-bounded regression. Their comparison therefore isolates the contribution of preference supervision, while the comparison between ASAP-DPO and ASAP isolates the proposed alignment objective. The objective ablation in Tab. 4 further evaluates localized weighting and margin-bounded regression separately using the same training settings. The superscript † denotes baselines reproduced in our training framework rather than results or checkpoints taken directly from the original papers. Specifically, FairHuman† (Wang et al., 2025c) and HG-DPO† (Na et al., 2025) are trained using our data and the same backbone-specific optimization budget as the other methods. For FairHuman, the training code was publicly released, but the training data were unavailable. We retain its multi-objective supervision over global images, hands, and faces and

Table 5: **Prompt for anatomical evaluation.** The instruction supplied to Gemini-3-Pro for absolute evaluation. A "Yes" response is mapped to *Fail*, whereas a "No" response is mapped to *Pass*.

ROLE
You are a Senior Visual QA Specialist specializing in Human Anatomy. Your specific objective is to identify **Severe and Obvious** anatomical artifacts in AI-generated human images.

TASK
Conduct a general anatomical plausibility check on the person(s) in the image. Identify **glaring** anatomical errors in any visible part of the human body.

CRITICAL TOLERANCE SETTINGS
1. **IGNORE** minor artifacts, artistic stylization, and slight blur.
2. **IGNORE** background characters or heavily occluded body parts.
3. **ONLY FLAG** distortions that are physically impossible and immediately visible. If unsure, return "No".

FAIL CRITERIA
Analyze the image for the following defects. If any are found, return "Yes" (*has distortion*).

1. Hands and Fingers
- **Obvious Polydactyly:** Explicitly visible extra fingers (*e.g.*, six or more fingers on one hand).
- **Unexplained Oligodactyly:** Fingers that are missing or cut off without a plausible cause. Do not flag fingers hidden by natural occlusion, such as when holding an object, forming a fist, appearing behind another body part, or being obscured by perspective. Flag only when the hand structure itself is clearly incomplete.
- **Severe Malformation:** Fingers that branch incorrectly or exhibit claw-like or melted distortions.
- **Fusion:** Fingers fused without visible separation or merged unnaturally into objects.

2. Face and Head
- **Severe Facial Collapse:** Facial features are merged, melted, or missing.
- **Gross Misplacement:** Eyes, ears, or other facial features appear in anatomically implausible positions. Do not flag asymmetric pupils or an unusual gaze unless the eye structure itself is visibly malformed.

3. Body and Limbs
- **Physiologically Impossible Joints:** Limbs bend in impossible directions or attach to the body at structurally implausible angles.
- **Surplus or Missing Limbs:** Extra limbs emerge from the body, required limbs are absent, or limbs appear detached.
- **Severe Blending:** Limbs merge indistinguishably into clothing, furniture, or other objects.

OUTPUT RULES
1. Output only a valid JSON string without introductory text or Markdown formatting.
2. The JSON must contain exactly two keys: reasoning, containing a brief observation of the face, hands, and limbs, and has_distortion," whose value must be either "Yes" or "No".

OUTPUT EXAMPLES Distortion found: reasoning: [Briefly describe the severe anatomical errors observed], has_distortion: "Yes"
No distortion found: reasoning: [Briefly describe why the anatomical structures are plausible], has_distortion: "No"

apply it to our training samples. For HG-DPO, the original training data and official implementation of its data construction pipeline were unavailable to us. We therefore implement its easy-stage preference construction strategy using our data pool.

D.3 Preference Construction Strategies

To isolate the effect of preference-pair construction, all strategies start from the same prompt set and retain the same number of training pairs. The resulting datasets are used to train the corresponding backbone with vanilla DPO under identical optimization settings. The four strategies differ only in how their preferred and dispreferred samples are obtained: *Standard Pairwise Ranking* independently generates two images for each prompt and assigns their preference order according to reward scoring. *HG-DPO-Easy* selects highest and lowest reward samples from a larger candidate pool. *Real-Anchor Localized Preferences* use real human images as preferred anchors and construct the corresponding dispreferred samples through localized anatomical degradation. *Synthetic Localized Preferences* use human-verified synthetic images as preferred anchors and apply the same localized degradation principle to construct anatomically implausible negatives.

E Additional Benchmark Results

To evaluate whether ASAP generalizes beyond our proposed HAF-Bench, we conduct an additional experiment on the publicly released Human-Art (Ju et al., 2023a) validation split. We use all 1,250 text descriptions from its five natural-scene categories: Acrobatics, Cosplay, Dance, Drama, and Movie. These prompts are independently collected and were not involved in the construction of HAP or HAF-Bench. We compare FLUX with its ASAP-tuned model using identical prompts, seeds, and sampling settings. We report AER to evaluate anatomical plausibility, and HPSv3 to mesure overall image quality and prompt alignment.

Table 6: **Evaluation on the Human-Art natural-scene validation subset.** We use 1,250 text descriptions from five real-human categories. AER evaluate anatomical plausibility, while HPSv3 measures overall image quality and prompt alignment.

Method	AER ↓	HPSv3 ↑
FLUX	0.223	10.140
FLUX-ASAP (Ours)	0.176	10.215

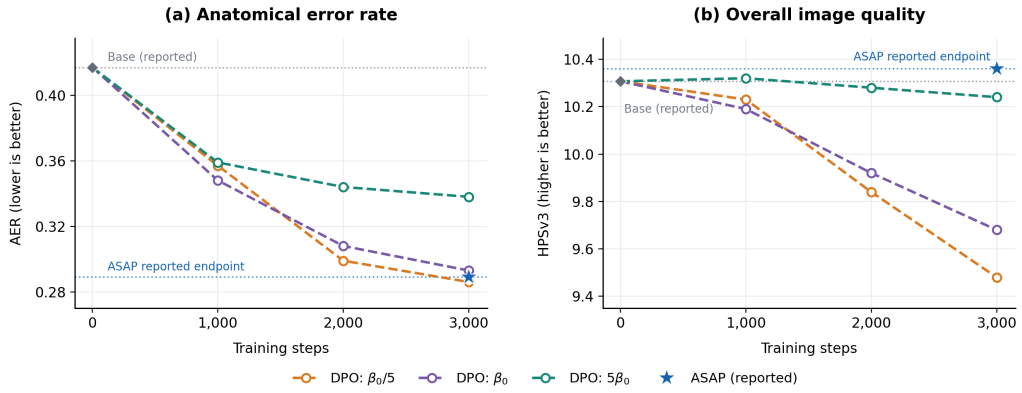

Figure 8: **DPO tuning and early stopping for image-quality preservation.** We vary β and training duration on FLUX.1-dev using the same HAP pairs. Adjusting β or stopping early mitigates image-quality degradation but reduces anatomical gains. Within the evaluated configurations, ASAP achieves a better balance between anatomical plausibility and overall image quality.

F DPO Hyperparameter and Early-Stopping Analysis

To examine whether DPO's quality degradation can be mitigated through hyperparameter tuning or early stopping, we train FLUX on the same HAP pairs with $\beta \in \{\beta_0/5, \beta_0, 5\beta_0\}$, where $\beta_0 = 100$ is our default DPO coefficient for FLUX. We evaluate checkpoints at 1,000, 2,000, and 3,000 steps under identical sampling conditions. As shown in Fig. 8, larger β or earlier stopping generally preserves higher HPSv3 scores but yields less anatomical improvement. Further training reduces AER with diminishing gains, while overall image quality can deteriorate. Within the tested configurations, ASAP achieves a more favorable balance between anatomical plausibility and overall image quality.

G Limitations

While our proposed ASAP framework significantly reduces anatomical errors, we observe two primary failure modes, both of which stem from the inherent limitations of the text-to-image foundation models, illustrated in Fig. 9.

Text-Image Misalignment. As illustrated in the left column, when prompts require precise numerical counting (*e.g.*, "counting to five") or specific constraint-based gestures (*e.g.*, "single-hand"), the ASAP-tuned models successfully generate highly realistic and anatomically plausible human images. However, they may occasionally fail to follow the precise semantic constraints of the text. This indicates that while ASAP effectively aligns the generative model toward anatomical correctness, it does not overcome the base model's inherent limitations in text-image compositional alignment.

Failure to Extreme Artifacts. We observe that, if the image generated by the base model with some seed contains extreme anatomical artifacts, the finetuned model may still generate artifacts in the corresponding human body parts. For instance, as demonstrated in the right column, when the base model produces severe anatomical artifacts across extensive regions, ASAP still enforces noticeable rectifications (*e.g.*, separating fused fingers), but frequently falls short of achieving perfect

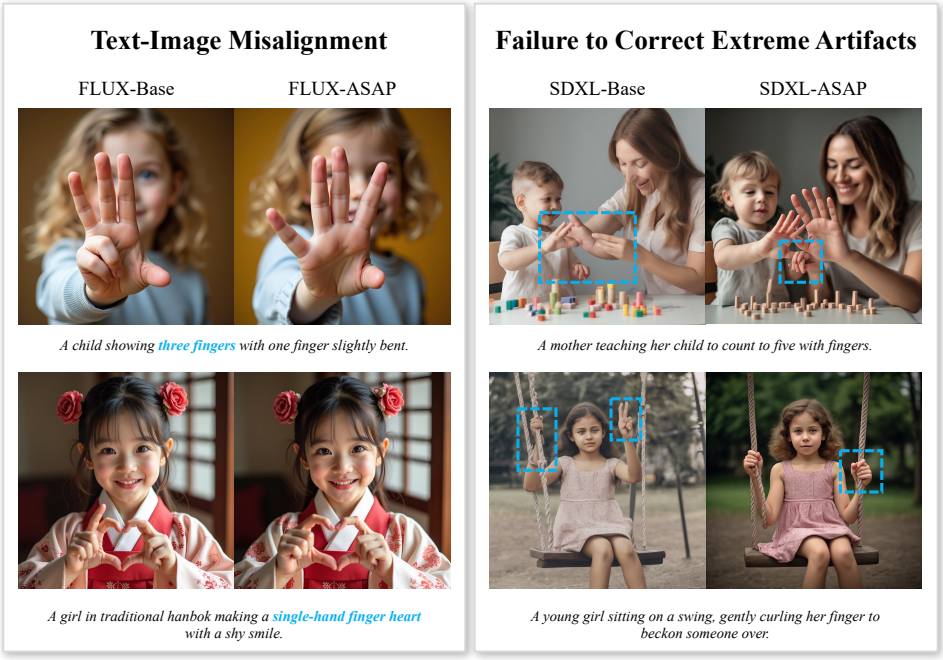

Figure 9: **Failure Case Analysis. (Left) Text-Image Misalignment:** While the ASAP-tuned models successfully generate anatomically plausible human images, they may occasionally fail to follow precise semantic constraints (*e.g.*, numerical counting or specific gestures). **(Right) Failure to Extreme Artifacts:** In instances where the base model produces severe structural degradation when failing to resolve specific complex details, ASAP still enforces noticeable rectifications (*e.g.*, separating severely fused fingers) but may fall short of achieving perfect anatomical plausibility.

anatomical plausibility. This indicates that, as a fine-grained alignment strategy, ASAP fundamentally requires a minimally plausible anatomical foundation from the base generative model.