DECOMPOSING PRIVATE IMAGE GENERATION VIA COARSE-TO-FINE WAVELET MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative models trained on sensitive image datasets have been shown to risk memorizing and reproducing individual training examples, making strong privacy guarantees essential. While differential privacy (DP) provides a principled basis for such guarantees, applying standard DP finetuning (e.g., with DP-SGD) to generative models often results in severe degradation of image quality. To address this, we introduce a spectral DP framework that uses wavelet representations to control which image scales are included in private finetuning, motivated by the hypothesis that dataset-specific information needed for adaptation, such as face geometry and object shape, can be captured by low-frequency wavelet components while fine-scale details can be supplied by a public prior. Specifically, we privately finetune a publicly pretrained autoregressive spectral image model to generate intermediaries at a chosen level of detail, then use the frozen public model to complete any remaining wavelet scales without additional privacy loss. Experiments with coarse-prefix and full-sequence configurations on MS-COCO and MM-CelebA-HQ demonstrate strong distributional image quality relative to diffusion and raster-autoregressive baselines. Full-sequence adaptation generally yields better final-resolution quality while coarse-prefix adaptation achieves stronger results at the coarse evaluation scale with lower training cost.

1 INTRODUCTION

Recent advances in deep generative modeling have enabled high-quality image synthesis from textual descriptions, supporting applications such as creative image generation (Gafni et al., 2022; Rombach et al., 2022) and interactive editing (Nichol et al., 2021; Avrahami et al., 2022). However, training or finetuning these models on sensitive image datasets, including medical images and private photo collections, raises data-privacy concerns as generative models have been shown to memorize and reproduce information about individual training examples (Carlini et al., 2023; Buchanan et al., 2025).

Differential privacy (DP) provides a rigorous framework to bound this risk by ensuring that model outputs are statistically insensitive to the inclusion or exclusion of any single training sample. In practice, DP guarantees are commonly enforced during training or finetuning using variants of differentially private stochastic gradient descent (DP-SGD) (Abadi et al., 2016). While DP-SGD provides strong formal privacy guarantees, it introduces a privacy-utility trade-off because the noise required for privacy can overwhelm the clipped gradient signal during finetuning. This often leads to degradation in visual quality, particularly for complex image distributions with fine-grained structure, such as human faces (Liu et al., 2024; Li et al., 2024). As a result, existing DP image generation methods struggle to produce high-fidelity samples on challenging benchmarks, even when leveraging large pretrained models or parameter-efficient finetuning strategies.

To mitigate the privacy-utility trade-off, we ask which levels of image detail must be learned through private finetuning and which can be supplied by a public model. Our approach is based on two hypotheses regarding the spectral structure of images and the role of public generative priors. First, we hypothesize that low-frequency wavelet components retain substantial dataset-specific information about global image structure and appearance, including properties such as face geometry, object shape, and composition. This is motivated by the energy compaction property of wavelet representations, where low-frequency components capture coarse image structure while high-frequency coefficients are often sparse and concentrated around details like edges and textures (Mallat, 1989; Antonini

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.

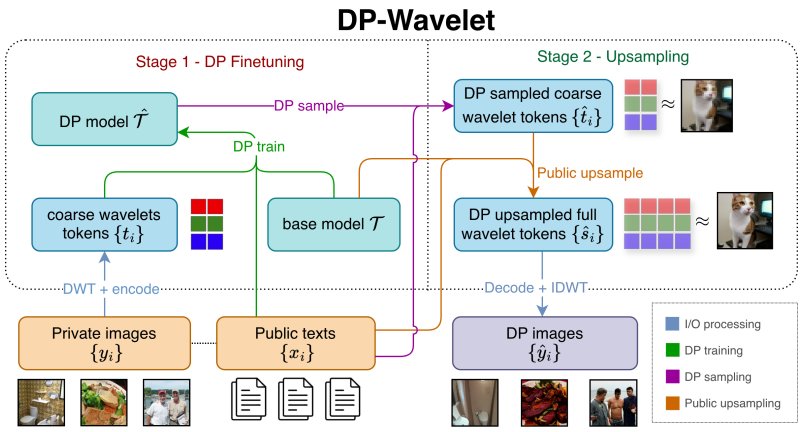


Figure 1: Illustration of DP-Wavelet in the coarse-prefix setting, where the privately finetuned model generates a coarse spectral prefix and the frozen public model completes the remaining scales.

et al., 1992). Second, we hypothesize that any remaining fine-scale details can be synthesized by a publicly pretrained model without direct access to the private images. Together, these hypotheses motivate choosing the number of wavelet scales included in private finetuning according to the target distribution, desired image fidelity, and available training resources.

To operationalize these insights, we propose *DP-Wavelet*, a coarse-to-fine method for DP text-to-image generation using wavelet representations at a chosen level of detail. Specifically, given a dataset of public texts $x = \{x_i\}$ (e.g., captions and prompts) paired with private images $y = \{y_i\}$, the goal is to generate high-fidelity synthetic images $\hat{y} = \{\hat{y}_i\}$ conditioned on the corresponding texts while ensuring that the overall generation mechanism satisfies differential privacy with respect to the private image dataset y. We focus on settings where textual descriptions are non-sensitive but the paired images contain private information (e.g., medical images paired with generic clinical descriptions). We discuss additional motivating settings, including style-preserving synthesis and extensions to private text prompts, in Appendix A.

DP-Wavelet follows a coarse-to-fine image synthesis procedure where a publicly pretrained autoregressive spectral image model is privately finetuned to generate text-conditioned wavelet representations at a chosen level of detail, with any remaining scales completed by the frozen public model. Our experiments consider both the coarse-prefix configuration illustrated in Figure 1 and a full-sequence configuration where the finetuned model generates all the tokens directly. Complementary membership-inference experiments on MS-COCO indicate that DP finetuning makes it harder for a hypothetical attacker to determine whether an image was used to train the coarse-prefix model compared with non-private finetuning.

Contributions. In summary, we make three methodological and empirical contributions towards private image generation.

1. We introduce *DP-Wavelet*, a spectral framework for DP image synthesis that controls which wavelet scales are included in private finetuning and preserves the image-level privacy guarantee established during finetuning throughout the generation procedure.

2. We empirically show that DP-Wavelet generates images with strong distributional quality (FID) under private finetuning on MS-COCO and MM-CelebA-HQ relative to diffusion and raster-autoregressive baselines across multiple privacy levels. Our qualitative results further illustrate how coarse intermediaries and final images capture salient structure, composition, and appearance.

3. We identify resolution-dependent tradeoffs in private spectral image generation, characterizing the relationship between the level of detail included in private finetuning, distributional image quality, and training cost. Comparisons using the same pretrained backbone show that coarse-prefix finetuning achieves lower FID at the coarse evaluation scale with modestly lower training cost while full-sequence finetuning generally yields lower final resolution FID.

2 Related Works

Image synthesis. VAEs (Kingma & Welling, 2013) and GANs (Goodfellow et al., 2014) established effective techniques for deep image generation, with learned autoencoder representations subsequently becoming a central component of many modern high-resolution generative pipelines. These pipelines often decouple image representation from generation by first compressing images into continuous latent or discrete token spaces and then modeling the resulting distribution. Diffusion-based approaches generate images through iterative denoising in pixel or learned latent spaces (Ho et al., 2020; Nichol & Dhariwal, 2021; Rombach et al., 2022) while many autoregressive (AR) methods formulate image synthesis as a sequence prediction problem over discrete image tokens (Esser et al., 2021; Yu et al., 2022). As image tokenizers and transformer architectures have improved, recent AR approaches have achieved image quality comparable to diffusion models on standard benchmarks, with performance continuing to improve as model capacity increases (Tian et al., 2024; Sun et al., 2024). These advances establish AR generation as a strong alternative to diffusion and provide a natural sequence-modeling foundation for our private spectral image generation framework.

Private image synthesis. Early research on DP image generation focused on GAN-based generators with evaluations concentrated on low-resolution image benchmarks (Xie et al., 2018; Torkzadehmahani et al., 2019; Bie et al., 2023; Chen et al., 2020). More recent work has adapted diffusion models under DP, initially through diffusion-specific training and sampling strategies for DP-SGD (Dockhorn et al., 2022) and subsequently by privately finetuning publicly pretrained diffusion models (Ghalebikesabi et al., 2023). Public pretraining allows private finetuning to adapt an existing generative model rather than learn the target distribution from scratch. Follow-up works improve private generation by selecting semantically matched public data for pretraining (Li et al., 2024), exploiting noise already present in the forward diffusion process (Wang et al., 2024), or restricting private adaptation to low-rank parameter updates (Tsai et al., 2025). FETA-Pro (Gong et al., 2026) instead incorporates frequency features as intermediate training shortcuts in a spatial frequency curriculum combined with DP-SGD. These works primarily evaluate unconditional or label conditioned generation often at lower resolutions whereas DP-LDM (Liu et al., 2024) demonstrates high-resolution text-to-image synthesis through private finetuning of a pretrained latent diffusion model. Private autoregressive generation is less explored, with recent work showing substantial utility degradation under private finetuning of Visual AutoRegressive models (Shaikh et al., 2025). Complementary training-free approaches instead keep foundation models fixed and apply DP to the selection or construction of synthetic intermediaries (Lin et al., 2023; Wang et al., 2025). Our work considers the distinct problem of directly adapting the generator to a private target distribution, using finetuning to learn visual characteristics that may be difficult to elicit from the fixed public model through prompting or sample selection alone.

Spectral and cascaded image generation. Coarse-to-fine synthesis strategies decompose generation by progressively resolving structure before details. Cascaded diffusion models such as Imagen (Saharia et al., 2022) implement this spatially through a pipeline of different models where a base generator synthesizes a low-resolution pixel grid that is subsequently refined by separate super-resolution upsamplers. In contrast, the spectral image tokenizer (SIT) (Esteves et al., 2025) applies a discrete wavelet transform and tokenizes the resulting subbands into a sequence ordered from coarse to fine, supporting partial decoding from coarse token prefixes. The resulting autoregressive model (AR-SIT) generates global low-frequency components before finer-scale details, unlike the raster-order generation used by Parti (Yu et al., 2022) and LlamaGen (Sun et al., 2024) which predicts tokens by row across the spatial token grid. This coarse to fine ordering is central to our framework as it allows the private training objective to cover only the coarse representation or extend to progressively finer wavelet scales.

3 Preliminaries

In this section, we review differential privacy and DP-SGD, multiresolution analysis and 2D wavelets, and the autoregressive spectral image model underlying DP-Wavelet.

3.1 Differential privacy and DP-SGD

Differential privacy (DP) is a rigorous framework for limiting the dependence of a randomized algorithm on any individual training example. Formally, it requires the algorithm's output dis-

tributions on neighboring datasets that differ by the addition or removal of one record to remain close. Consequently, observing the output of the algorithm should provide only limited evidence about whether a particular sample was used for training, which is precisely the question targeted by membership-inference attacks. Common formulations of DP include pure ε-DP and approximate (ε, δ)-DP (Dwork et al., 2006) and our work uses the latter.

Definition 1 ((ε, δ)-DP (Dwork et al., 2006)). *Given $\varepsilon \geq 0$ and $\delta \in [0, 1)$, a randomized mechanism $\mathcal{A} : \mathcal{D} \to \mathcal{R}$ is (ε, δ)-DP if, for all neighboring datasets D, D' and all measurable $S \subseteq \mathcal{R}$, we have*

$$\Pr[\mathcal{A}(D) \in S] \leq e^{\varepsilon} \Pr[\mathcal{A}(D') \in S] + \delta.$$

The privacy budget ε controls how strongly the output distributions on neighboring datasets may differ, with smaller values corresponding to stronger privacy. The parameter δ allows a small additive slack in this guarantee and the value is typically chosen below the inverse of the training dataset size.

DP-SGD. To achieve this guarantee in machine learning, it is common to use some variant of DP-SGD (Abadi et al., 2016), which we summarize in Algorithm 1.

Algorithm 1 DP-SGD (Abadi et al., 2016)

Require: dataset $D = \{x_1, \ldots, x_N\}$, loss $\mathcal{L}$, learning rate η, noise multiplier σ, clipping norm C, expected batch size B, iteration count T
Ensure: DP parameters θ_T

Initialize θ_0 independently of the private dataset, for example from a publicly pretrained model
for $t \in 0 \ldots T-1$ **do**
  Sample a Poisson batch $\mathcal{B}_t$ with rate $q = B/N$
  for $i \in \mathcal{B}_t$ **do**
    Compute $g_t(x_i) \leftarrow \nabla_{\theta_t} \mathcal{L}(x_i, \theta_t)$
    Clip $\bar{g}_t(x_i) \leftarrow g_t(x_i) / \max(1, \|g_t(x_i)\|_2 / C)$
  Sample $Z_t \sim \mathcal{N}(0, \sigma^2 C^2 I)$ and *set* $\tilde{g}_t \leftarrow \frac{1}{B} \left[\sum_{i \in \mathcal{B}_t} \bar{g}_t(x_i) + Z_t\right]$
  Update $\theta_{t+1} \leftarrow \theta_t - \eta \tilde{g}_t$
return θ_T

DP-SGD limits the influence of individual examples through per-example clipping and Gaussian noise. The noise multiplier σ controls the noise scale relative to the clipping threshold C and the resulting privatized update can be used with SGD or other first-order optimizers such as Adam (Ponomareva et al., 2023). Under Poisson sampling with rate $q = B/N$, the cumulative privacy guarantee depends on q, σ, and the number of optimization steps T. In our experiments, we fix a target (ε, δ) and use Privacy Loss Distribution (PLD) accounting (Doroshenko et al., 2022) to choose the corresponding σ.

Post-processing. A fundamental property of DP is invariance to post-processing, which guarantees that any computation applied to the output of a DP mechanism preserves its privacy guarantee provided it accesses the private data only through that output. Formally, this is stated as follows:

Proposition 1 (Post-Processing (Dwork et al., 2014)). *Let $\mathcal{M} : \mathcal{D} \to \mathcal{R}$ be an (ε, δ)-DP algorithm, and let $f : \mathcal{R} \to \mathcal{R}'$ be an arbitrary (possibly randomized) mapping. Then the composition $f \circ \mathcal{M} : \mathcal{D} \to \mathcal{R}'$ is (ε, δ)-DP.*

For DP-Wavelet, sampling under public text conditioning and subsequent image reconstruction constitute post-processing of the privately finetuned model, so the privacy guarantee is preserved whether the model generates a prefix that is completed publicly or the full token sequence.

3.2 MRA AND 2D WAVELETS

We ground our coarse-to-fine framework in Multiresolution Analysis (MRA) (Mallat, 1989), which decomposes an image into a low-frequency approximation together with progressively finer detail components. This provides a natural representation for separating global image structure from fine-scale image details.

We model a 2D image as $f \in L^2(\mathbb{R}^2)$, where $L^2(\mathbb{R}^2)$ denotes the space of square-integrable functions on $\mathbb{R}^2$. For a chosen approximation scale j_0, an orthonormal basis of scaling functions $\{\Phi_{j_0,\mathbf{k}}\}$ and wavelet functions $\{\Psi^{\lambda}_{j,\mathbf{k}}\}_{j \geq j_0}$ gives the following expansion with convergence in L^2:

$$f(x,y) = \underbrace{\sum_{\mathbf{k}\in\mathbb{Z}^2} c_{j_0,\mathbf{k}}\Phi_{j_0,\mathbf{k}}(x,y)}_{\text{Approximation at scale } j_0} + \underbrace{\sum_{j=j_0}^{\infty}\sum_{\mathbf{k}\in\mathbb{Z}^2}\sum_{\lambda\in\{H,V,D\}} d^{\lambda}_{j,\mathbf{k}}\Psi^{\lambda}_{j,\mathbf{k}}(x,y)}_{\text{Details at scale } j}, \tag{1}$$

where the approximation coefficients $c_{j_0,\mathbf{k}}$ and detail coefficients $d^{\lambda}_{j,\mathbf{k}}$ are obtained by projecting the image onto the corresponding scaling and wavelet functions:

$$c_{j_0,\mathbf{k}} = \langle f, \Phi_{j_0,\mathbf{k}}\rangle, \qquad d^{\lambda}_{j,\mathbf{k}} = \langle f, \Psi^{\lambda}_{j,\mathbf{k}}\rangle. \tag{2}$$

These projections separate the low-frequency image approximation from progressively finer details across scales and orientations. Natural images typically concentrate substantial signal energy in the low-frequency approximation and higher-frequency detail coefficients are often sparse or compressible (Mallat, 2008; Antonini et al., 1992). This energy compaction property motivates controlling the scope of private supervision through the retained wavelet scales, with longer spectral prefixes incorporating progressively finer image information.

For discrete images, the multiresolution decomposition is computed using the Discrete Wavelet Transform (DWT) through repeated low-pass and high-pass filtering followed by dyadic downsampling (Mallat, 2008). Each decomposition level produces four subbands: a low-frequency approximation LL and three orientation-selective detail subbands LH, HL, and HH. The transform is recursively applied to the LL component to obtain progressively coarser approximations. In this work, we denote the coarsest approximation by LL_0 and index the detail scales from coarse to fine. The approximation and detail coefficients together form the spectral representation encoded into discrete tokens by the tokenizer described in the next section. Appendix B gives the underlying MRA construction and Appendix C provides visualizations of this decomposition.

3.3 Spectral image tokenization and AR-SIT

Our generative framework builds on vector-quantized image tokenization, where an encoder $\mathcal{E}$ maps images to sequences of discrete codebook tokens and a decoder $\mathcal{D}$ reconstructs images from these tokens (Esser et al., 2021). An autoregressive transformer $\mathcal{T}$ then models the distribution of the resulting token sequences. We specifically use the Spectral Image Tokenizer (SIT) and its associated autoregressive model AR-SIT (Esteves et al., 2025). SIT first applies the DWT described above, patchifies each wavelet subband separately, and encodes the resulting patches into discrete tokens. In our implementation, this decomposition uses the Haar wavelet basis.

In conventional autoregressive image models, discrete image tokens are commonly arranged in spatial raster order. In contrast, SIT organizes its spectral tokens from coarse to fine, beginning with the low-frequency approximation and followed by progressively finer detail components. Let $s = (s_1, \ldots, s_K)$ denote a complete spectral token sequence. For text-conditioned generation with input x, AR-SIT models the distribution of this sequence autoregressively as

$$p_\theta(s \mid x) = \prod_{k=1}^{K} p_\theta(s_k \mid s_{<k}, x), \tag{3}$$

where θ denotes the transformer parameters and $s_{<k} = (s_1, \ldots, s_{k-1})$ contains the preceding tokens. Let K_0 denote the number of tokens representing LL_0 and write $t = s_{1:K_\mathrm{p}}$ for a prefix ending at a complete wavelet scale. The choices $K_\mathrm{p} = K_0$ and $K_\mathrm{p} = K$ retain the coarse approximation and the full sequence respectively.

A key property of the SIT configuration is partial decoding, whereby a prefix containing the approximation and any retained detail scales can be decoded without the remaining finer-scale tokens. AR-SIT can condition on this prefix and autoregressively complete the sequence (Esteves et al., 2025). These capabilities allow DP-Wavelet to vary the level of detail included in private finetuning and use public completion only when the privately generated prefix does not cover the full sequence. Appendix C provides further details on the SIT and AR-SIT architectures.

4 DP-Wavelet

In this section, we present DP-Wavelet, our coarse-to-fine framework for DP image generation.

Training. The private objective is defined over a prefix of the spectral token sequence which allows supervision to stop at a chosen wavelet scale. Given pairs $\{(x_i, y_i)\}_{i=1}^{N}$ of public texts and private images, we use the public SIT encoder $\mathcal{E}$ to obtain the spectral token sequence $s_i = (s_{i,1}, \ldots, s_{i,K})$ for each image y_i. We select a fixed prefix length K_{p} ending at a complete wavelet scale and retain $t_i = s_{i,1:K_{\mathrm{p}}}$ as the private prediction target. We then finetune the autoregressive transformer $\mathcal{T}$ to predict these prefixes conditioned on the corresponding texts, computing the private prediction loss only over the selected tokens. More specifically, the per-image prediction loss is

$$\ell_{K_{\mathrm{p}}}(\theta, x_i, t_i) = -\frac{1}{K_{\mathrm{p}}} \sum_{k=1}^{K_{\mathrm{p}}} \log p_\theta(t_{i,k} \mid t_{i,<k}, x_i). \tag{4}$$

Applying a DP optimizer to this objective yields the privately finetuned transformer $\hat{\mathcal{T}}$ while the public encoder $\mathcal{E}$, decoder $\mathcal{D}$, and original transformer $\mathcal{T}$ remain fixed. In our experiments, we evaluate the coarse-prefix configuration $K_{\mathrm{p}} = K_0$, which supervises only the LL_0 approximation and the full-sequence configuration $K_{\mathrm{p}} = K$, which includes all spectral tokens in private finetuning.

Inference. At inference time, the privately finetuned model $\hat{\mathcal{T}}$ generates a prefix of length K_{p} conditioned on the public text prompt. When $K_{\mathrm{p}} < K$, the frozen public transformer $\mathcal{T}$ autoregressively completes the remaining tokens from this prefix and the same text conditioning. When $K_{\mathrm{p}} = K$, the transformer $\hat{\mathcal{T}}$ generates the complete sequence directly and no public completion is required. The public decoder $\mathcal{D}$ then maps the complete spectral sequence to an image. Algorithm 2 summarizes the full procedure.

Algorithm 2 DP-Wavelet

Require: Fixed prefix length K_{p}, public text-private image pairs $\{(x_i, y_i)\}_{i=1}^{N}$, generation prompt $x^\star$, public pretrained SIT encoder $\mathcal{E}$, decoder $\mathcal{D}$, AR-SIT transformer $\mathcal{T}$, privacy parameters (ε, δ)

Ensure: Synthetic image $\hat{y}$ generated under (ε, δ)-DP

Stage 1: Private spectral finetuning

Encode each private image y_i with $\mathcal{E}$ and retain $t_i = s_{i,1:K_{\mathrm{p}}}$.

DP finetune a copy of $\mathcal{T}$ using equation 4 to obtain $\hat{\mathcal{T}}$ under (ε, δ)-DP.

Stage 2: Generation and reconstruction

Generate a prefix $\hat{t}$ of length K_{p} using $\hat{\mathcal{T}}$ conditioned on $x^\star$.

if $K_{\mathrm{p}} < K$ **then**

    Complete $\hat{t}$ into a full spectral sequence $\hat{s}$ using the frozen transformer $\mathcal{T}$ conditioned on $x^\star$.

else

    Set $\hat{s} \leftarrow \hat{t}$.

Decode $\hat{s}$ using $\mathcal{D}$ to obtain $\hat{y}$.

return $\hat{y}$

Privacy guarantees. Private finetuning is performed with a DP optimizer calibrated to satisfy an (ε, δ)-DP guarantee. Once $\hat{\mathcal{T}}$ has been trained under (ε, δ)-DP, any remaining steps such as sampling using public text, completing remaining tokens with the frozen public transformer, and decoding the resulting sequence access the private dataset only through $\hat{\mathcal{T}}$. By the post-processing property of DP in Proposition 1, the complete generation procedure therefore preserves the same (ε, δ)-DP guarantee as the one established during finetuning. Appendix D provides the formal argument for this.

5 Experiments

We evaluate DP-Wavelet on two text-conditioned image-generation benchmarks: MS-COCO (Lin et al., 2014), which contains diverse natural images paired with captions, and MM-CelebA-HQ (Xia et al., 2021), which contains high-resolution celebrity face images paired with textual descriptions.

DP-Wavelet configurations. Our experiments consider two levels of private spectral supervision using the same pretrained AR-SIT backbone. *DP-Wavelet-Coarse* finetunes on the coarse LL_0 prefix

Table 1: FID and LPIPS at the 32×32 and 256×256 resolution where each entry is reported as Metric@32 / Metric@256. Bold indicates the best result and lower values are better.

FID ↓					
Dataset	Privacy	DP-LDM	DP-LlamaGen	DP-Wavelet-Coarse	DP-Wavelet-Full
MS-COCO	Pretrained	12.0 / **15.0**	20.2 / 17.7	**10.3** / 17.2	**10.3** / 17.2
	$\varepsilon = \infty$	13.5 / 14.9	14.2 / 11.6	**4.1** / 10.1	6.5 / **8.8**
	$\varepsilon = 10$	11.8 / 14.8	16.8 / 13.8	**6.8** / 13.9	6.9 / **8.9**
	$\varepsilon = 1$	12.0 / 15.0	19.2 / 16.1	**7.5** / 14.2	7.9 / **10.3**
MM-CelebA-HQ	Pretrained	36.5 / 64.3	63.2 / 86.5	**30.7** / **55.3**	**30.7** / **55.3**
	$\varepsilon = \infty$	19.5 / 29.6	27.8 / 32.5	**9.9** / 15.3	10.4 / **9.4**
	$\varepsilon = 10$	20.9 / 24.9	49.1 / 63.7	**13.9** / 19.3	16.4 / **17.8**
	$\varepsilon = 1$	26.7 / **35.1**	54.9 / 74.2	**24.1** / 36.7	27.2 / 36.6

LPIPS ↓					
Dataset	Privacy	DP-LDM	DP-LlamaGen	DP-Wavelet-Coarse	DP-Wavelet-Full
MS-COCO	Pretrained	0.600 / 0.745	**0.463** / **0.728**	0.585 / 0.748	0.585 / 0.748
	$\varepsilon = \infty$	0.593 / 0.753	0.448 / 0.715	0.424 / **0.665**	**0.421** / 0.717
	$\varepsilon = 10$	0.599 / 0.763	**0.457** / **0.723**	0.584 / 0.749	0.583 / 0.731
	$\varepsilon = 1$	0.600 / 0.745	**0.464** / **0.728**	0.587 / 0.748	0.585 / 0.732
MM-CelebA-HQ	Pretrained	0.497 / 0.735	**0.373** / **0.672**	0.486 / 0.712	0.486 / 0.712
	$\varepsilon = \infty$	0.418 / 0.693	**0.276** / 0.586	0.289 / **0.581**	0.332 / 0.602
	$\varepsilon = 10$	0.461 / 0.725	**0.329** / **0.649**	0.393 / 0.658	0.394 / 0.655
	$\varepsilon = 1$	0.476 / 0.732	**0.349** / **0.659**	0.421 / 0.677	0.437 / 0.681

and uses the frozen public model to complete the remaining scales and *DP-Wavelet-Full* finetunes on the complete spectral token sequence and generates all tokens directly.

Baselines. For comparison, we use representative diffusion and autoregressive image generation methods under the same adaptation regimes. *DP-LDM* (Liu et al., 2024) finetunes the attention modules and conditioning embedder of a pretrained latent diffusion model initialized from Stable Diffusion v1.4. *DP-LlamaGen* finetunes the LlamaGen-XL autoregressive model (Sun et al., 2024) using its standard next-token prediction objective over the complete raster-ordered token sequence.

Evaluation protocol. We evaluate performance across three adaptation regimes: (i) the publicly pretrained model without target-domain finetuning (*Pretrained*), (ii) non-private finetuning with $\varepsilon = \infty$, (iii) and private finetuning with $\varepsilon \in \{1, 10\}$. All private runs use DP-Adam with PLD privacy accounting to determine the corresponding noise multiplier σ. To quantify distributional image quality, we use the Fréchet Inception Distance (FID) (Heusel et al., 2017), which compares feature distributions between 30,000 generated and reference images. To measure paired perceptual similarity, we use LPIPS (Zhang et al., 2018), which averages the perceptual distance between each generated image and the reference image associated with the same conditioning text. We use *Metric*@k to denote evaluation at the spatial resolution $k \times k$. For DP-Wavelet-Coarse, evaluation at 32×32 uses the decoded coarse intermediary while outputs from DP-Wavelet-Full and the external baselines are resized to the same resolution. At 256×256, all methods are evaluated on their final generated images.

Implementation details. DP-Wavelet uses the publicly pretrained "SIT-SC-5" AR-SIT checkpoint (Esteves et al., 2025) where LL_0 corresponds to a 32×32 coarse representation. We use DP-Adam for private finetuning with PLD accounting through JAX Privacy (Balle et al., 2022) and Appendix E gives further information about the training procedure, hyperparameters, and runtimes.

5.1 Results and image samples

Quantitative results. We first compare how the amount of spectral supervision affects distributional quality and paired perceptual similarity across resolutions. Table 1 reports FID and LPIPS values at the 32×32 coarse representation scale and the final 256×256 resolution with additional results at 16×16



Figure 2: Example generations on MS-COCO (left) and MM-CelebA-HQ (right) across adaptation regimes for all compared methods. All generations use the same conditioning prompt with the corresponding reference image shown in the leftmost column.

and 64×64 resolutions given in Appendix F.1. DP-Wavelet-Coarse achieves substantially lower FID at the coarse evaluation scale than the diffusion and raster-autoregressive baselines after target-domain finetuning with modestly lower training cost than DP-Wavelet-Full as described in Appendix E. At the final resolution, DP-Wavelet-Full improves FID@256 over DP-Wavelet-Coarse in every finetuned comparison and achieves the lowest FID@256 in most settings. This suggests that including the complete spectral sequence in the private objective can be beneficial when final-resolution distribution matching is the main target, although this trend is not uniform across all baselines and metrics. In particular, DP-LDM obtains the marginally lowest FID@256 on MM-CelebA-HQ at $\varepsilon = 1$ while DP-LlamaGen often achieves the lowest LPIPS under privacy, indicating stronger paired perceptual similarity to the corresponding reference images. Overall, these results demonstrate that DP-Wavelet is especially effective at modeling the target distribution under private finetuning, with the preferred supervision level depending on the evaluation objective and target resolution.

Qualitative results. Next we examine the samples qualitatively to provide a complementary view of how spectral supervision translates into intermediate and final generations. To this end, Figure 2 gives sample generations from all methods on MS-COCO and MM-CelebA-HQ across the pretrained, non-private, and private regimes with additional samples in Appendix F.2. These samples show that all methods generally produce prompt-relevant images and that finetuning changes the visual attributes and target-domain appearance captured by each model. We further inspect the coarse-to-fine mechanism in Figure 3, which shows decoded DP-Wavelet-Coarse intermediaries compared with their completed images. The coarse intermediaries capture broad target-domain structure, including object layout on MS-COCO and facial pose and geometry on MM-CelebA-HQ, while leaving fine-scale texture unresolved. Public completion adds local details and produces coherent final images, although some artifacts and detail shifts may remain. Appendix F.3 examines this dependence by replacing AR-SIT spectral completion with Gemini-based image refinement, which produces smoother outputs but higher FID and LPIPS at every evaluated resolution. This behavior aligns with the quantitative results where coarse-prefix finetuning performs strongest by low-resolution quality and final-resolution quality depends on the remaining spectral scales.

5.2 Privacy Auditing

To assess detectable membership leakage in the finetuned models, we run a JAX Privacy audit of DP-Wavelet-Coarse on MS-COCO (McKenna et al., 2026). The audit compares seen canaries, which are samples included during finetuning, with unseen canaries drawn from the same distribution but held out from training. For each canary, we compute a reference-calibrated score that measures how much finetuning increases its likelihood relative to the pretrained model using the coarse token sequence. Larger scores provide stronger evidence that a sample was used for training and thresholding these scores gives membership predictions. In Table 2, we report the resulting empirical estimate $\hat{\varepsilon}$ together



Figure 3: DP-Wavelet-Coarse examples on MS-COCO (left) and MM-CelebA-HQ (right). From left to right, each row shows the reference image, completed final generation, and the decoded low-resolution intermediary generated before public completion.

with a confidence-adjusted Clopper-Pearson lower bound $\varepsilon_{\mathrm{CP}}^{\mathrm{LB}}$ (Nasr et al., 2023) and a one-run f-DP lower bound $\varepsilon_{f\mathrm{DP}}^{\mathrm{LB}}$ (Mahloujifar et al., 2025).

Table 2: Membership-inference audit of DP-Wavelet-Coarse on MS-COCO.

Training regime	$\hat{\varepsilon}$	$\varepsilon_{\mathrm{CP}}^{\mathrm{LB}}$	$\varepsilon_{f\mathrm{DP}}^{\mathrm{LB}}$
$\varepsilon = \infty$	5.20	7.04	12.93
$\varepsilon = 10$	0.25	0.00	0.00
$\varepsilon = 1$	0.15	0.00	0.00

The non-private model exhibits clear detectable membership leakage as both estimators give positive lower bounds on ε. In contrast, neither estimator returns a positive lower bound for the privately finetuned models and the raw estimates are much smaller than the non-private estimate and accounted privacy budgets $\varepsilon \in \{1, 10\}$. Thus, the audit indicates that DP finetuning substantially reduces the membership signal detected by this attack. A zero empirical lower bound does not imply zero privacy loss but it provides additional evidence that is consistent with the formal DP guarantee.

6 Conclusion

In this work, we introduced DP-Wavelet, a spectral framework for differentially private text-to-image generation that uses the coarse-to-fine ordering of wavelet tokens to control the level of detail included in private finetuning. The framework supports both coarse-prefix finetuning with public completion and full-sequence spectral finetuning and sampling, optional completion, and decoding preserve the privacy guarantee established during training. Empirically, DP-Wavelet achieves strong distributional image quality under varying privacy budgets on the MS-COCO and MM-CelebA-HQ datasets relative to the baselines. Comparing the two configurations further reveals a resolution-dependent tradeoff, with coarse-prefix finetuning giving stronger coarse-scale FID for lower training cost and full-sequence finetuning generally improving the final-resolution FID.

Overall, these results show that spectral representations provide not only a coarse-to-fine generative ordering, but also a natural way to make the scope of private supervision explicit. DP-Wavelet therefore reframes the private training objective as a choice over how much image detail should be learned from sensitive data and how much can be supplied by a public generative prior. More broadly, designing representations that expose this choice offers a promising route toward higher-fidelity generation under strong privacy constraints.

AI USE STATEMENT

We used generative AI tools to assist with the following tasks: proposing and refining hypotheses, assist in the writing of proofs, interpreting results, and implementing methods. We did not use generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, design the research methodology or experiments, assist with translation, clean or reformat datasets, or support qualitative or thematic data analysis. Additionally, we used generative AI tools to create and edit code, draft and edits parts of the paper, edit the paper draft to improve readability, and identify and understand relevant literature. We have carefully reviewed all AI-assisted work. For instance, all hypotheses and result interpretations have been manually checked. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies differentially private image generation with the goal of reducing privacy leakage during finetuning. Our experiments use existing image-generation benchmarks and do not involve the collection of new personal data, human-subject experiments, or user studies. We evaluate on the publicly available MS-COCO and MM-CelebA-HQ datasets and conduct membership-inference attacks only with respect to membership in these benchmark datasets. The work is intended to reduce privacy risks arising from adaptation on sensitive image data by providing formal guarantees on the influence of individual training records. Broader risks associated with generative image models, including misuse, bias, or misleading generated content, remain application dependent and should be assessed separately in the relevant context.

REPRODUCIBILITY STATEMENT

We describe the method and privacy-preserving generation procedure in Section 4 with the formal privacy argument given in Appendix D. Section 5 provides information about all datasets, compared methods, evaluation protocol, and metrics used in this work. Appendix E further reports the optimization hyperparameters, implementation and hardware details, and data-processing procedure for all evaluated methods. These sections are intended to provide the information needed to reproduce our training and evaluation procedures.

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 308–318, 2016.

Marc Antonini, Michel Barlaud, Pierre Mathieu, Ingrid Daubechies, et al. Image coding using wavelet transform. *IEEE Trans. Image Processing*, 1, 1992.

Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18208–18218, 2022.

Borja Balle, Leonard Berrada, Zachary Charles, Christopher A Choquette-Choo, Soham De, Vadym Doroshenko, Dj Dvijotham, Andrew Galen, Arun Ganesh, Sahra Ghalebikesabi, Jamie Hayes, Peter Kairouz, Ryan McKenna, Brendan McMahan, Aneesh Pappu, Natalia Ponomareva, Mikhail Pravilov, Keith Rush, Samuel L Smith, and Robert Stanforth. JAX-Privacy: Algorithms for privacy-preserving machine learning in jax, 2022. URL `http://github.com/google-deepmind/jax_privacy`.

Alex Bie, Gautam Kamath, and Guojun Zhang. Private gans, revisited. *arXiv preprint arXiv:2302.02936*, 2023.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL `http://github.com/google/jax`.

Sam Buchanan, Druv Pai, Yi Ma, and Valentin De Bortoli. On the edge of memorization in diffusion models. *arXiv preprint arXiv:2508.17689*, 2025.

Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium*, pp. 5253–5270, 2023.

Dingfan Chen, Tribhuvanesh Orekondy, and Mario Fritz. Gs-wgan: A gradient-sanitized approach for learning differentially private generators. *Advances in Neural Information Processing Systems*, 33:12673–12684, 2020.

Tim Dockhorn, Tianshi Cao, Arash Vahdat, and Karsten Kreis. Differentially private diffusion models. *arXiv preprint arXiv:2210.09929*, 2022.

Vadym Doroshenko, Badih Ghazi, Pritish Kamath, Ravi Kumar, and Pasin Manurangsi. Connect the dots: Tighter discrete approximations of privacy loss distributions. *arXiv preprint arXiv:2207.04380*, 2022.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 486–503. Springer, 2006.

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12873–12883, 2021.

Carlos Esteves, Mohammed Suhail, and Ameesh Makadia. Spectral image tokenizer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17181–17190, 2025.

Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *European Conference on Computer Vision*, pp. 89–106. Springer, 2022.

Sahra Ghalebikesabi, Leonard Berrada, Sven Gowal, Ira Ktena, Robert Stanforth, Jamie Hayes, Soham De, Samuel L Smith, Olivia Wiles, and Borja Balle. Differentially private diffusion models generate useful synthetic images. *arXiv preprint arXiv:2302.13861*, 2023.

Chen Gong, Kecen Li, Zinan Lin, and Tianhao Wang. From easy to hard++: Promoting differentially private image synthesis through Spatial-Frequency curriculum. In *35th USENIX Security Symposium*, 2026.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2014.

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 *Advances in Neural Information Processing Systems*, volume 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

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

Kecen Li, Chen Gong, Zhixiang Li, Yuzhong Zhao, Xinwen Hou, and Tianhao Wang. {PrivImage}: Differentially private synthetic image generation using diffusion models with {Semantic-Aware} pretraining. In *33rd USENIX Security Symposium*, pp. 4837–4854, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755. Springer, 2014.

Zinan Lin, Sivakanth Gopi, Janardhan Kulkarni, Harsha Nori, and Sergey Yekhanin. Differentially private synthetic data via foundation model apis 1: Images. *arXiv preprint arXiv:2305.15560*, 2023.

Michael F Liu, Saiyue Lyu, Margarita Vinaroz, and Mijung Park. Dp-ldms: Differentially private latent diffusion models. *Transactions on Machine Learning Research*, 2024.

Saeed Mahloujifar, Luca Melis, and Kamalika Chaudhuri. Auditing f-differential privacy in one run. In *International Conference on Machine Learning*, volume 267, pp. 42615–42641, 2025.

Stéphane Mallat. *A wavelet tour of signal processing: the sparse way*. Academic Press, 2008.

Stephane G Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, 1989.

Ryan McKenna, Galen Andrew, Borja Balle, Vadym Doroshenko, Arun Ganesh, Weiwei Kong, Alex Kurakin, Brendan McMahan, and Mikhail Pravilov. Jax-privacy: A library for differentially private machine learning, feb 2026.

Milad Nasr, Jamie Hayes, Thomas Steinke, Borja Balle, Florian Tramèr, Matthew Jagielski, Nicholas Carlini, and Andreas Terzis. Tight auditing of differentially private machine learning, feb 2023.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pp. 8162–8171, 2021.

Natalia Ponomareva, Hussein Hazimeh, Alex Kurakin, Zheng Xu, Carson Denison, H Brendan McMahan, Sergei Vassilvitskii, Steve Chien, and Abhradeep Guha Thakurta. How to dp-fy ml: A practical guide to machine learning with differential privacy. *Journal of Artificial Intelligence Research*, 77:1113–1201, 2023.

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

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

Kaif Shaikh, Franziska Boenisch, and Adam Dziedzic. Implementing adaptations for vision autoregressive model. *arXiv preprint arXiv:2507.11441*, 2025.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bernhard Schölkopf, and Ze Liu. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

Kauldron Team. Kauldron: A stable and flexible codebase for research. `https://github.com/google-research/kauldron`, 2024.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in Neural Information Processing Systems*, 37:84839–84865, 2024.

Reihaneh Torkzadehmahani, Peter Kairouz, and Benedict Paten. Dp-cgan: Differentially private synthetic data and label generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 0–0, 2019.

Yu-Lin Tsai, Yizhe Li, Chia-Mu Yu, Xuebin Ren, Po-Yu Chen, Zekai Chen, and Francois Buet-Golfouse. Differentially private fine-tuning of diffusion models. In *2025 IEEE/CVF International Conference on Computer Vision*, pp. 4561–4571. IEEE, 2025.

Haichen Wang, Shuchao Pang, Zhigang Lu, Yihang Rao, Yongbin Zhou, and Minhui Xue. dp-promise: Differentially private diffusion probabilistic models for image synthesis. In *33rd USENIX Security Symposium*, pp. 1063–1080, 2024.

Haoxiang Wang, Zinan Lin, Da Yu, and Huishuai Zhang. Synthesize privacy-preserving high-resolution images via private textual intermediaries. *arXiv preprint arXiv:2506.07555*, 2025.

Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2256–2265, 2021.

Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. Differentially private generative adversarial network. *arXiv preprint arXiv:1802.06739*, 2018.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research*, 2022. Featured Certification.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

A MOTIVATING SETTINGS AND EXTENSIONS

The public-text/private-image setting that we consider in this work arises in scenarios where textual descriptions are non-sensitive but the associated images contain confidential information. One example is with medical imaging where generic or clinical descriptions may be publicly shareable while the underlying scans are protected. A second use case is style-preserving synthesis, e.g., modeling an artist's style from public descriptions while treating individual images as the private records. In such cases, DP-Wavelet can adapt the transformer to aggregate visual characteristics while limiting the influence of any individual image.

The framework can also be extended to settings where the conditioning texts are private. One approach is to train a separate DP text model C on the private texts and use samples from C as conditioning prompts at generation time while DP-Wavelet is finetuned on the private image-text pairs. The quality of the resulting images would then additionally depend on how well the DP text model reproduces the relevant conditioning distribution. Our experiments focus on the public-text/private-image setting in order to isolate the effect of private image finetuning from errors introduced by privately generating the conditioning text.

B MULTIRESOLUTION ANALYSIS AND 2D WAVELET BASES

Section 3.2 introduced the multiresolution decomposition used by DP-Wavelet. In this section, we briefly review the orthonormal MRA construction underlying the approximation and detail spaces in Equations equation 1 and equation 2.

Following Mallat (1989; 2008), an orthonormal multiresolution analysis of $L^2(\mathbb{R})$ consists of nested closed approximation spaces $\{V_j\}_{j\in\mathbb{Z}}$ satisfying

$$V_j \subset V_{j+1}, \qquad \overline{\bigcup_{j\in\mathbb{Z}} V_j} = L^2(\mathbb{R}), \qquad \bigcap_{j\in\mathbb{Z}} V_j = \{0\},$$

together with the dyadic scaling and translation structure. A scaling function ϕ generates an orthonormal basis of each V_j and the associated wavelet function ψ generates the detail space W_j, defined as the orthogonal complement of V_j in V_{j+1}. Thus,

$$V_{j+1} = V_j \oplus W_j.$$

For 2D images, we construct a separable basis using tensor products of the 1D spaces. The approximation space at scale j is

$$V_j^{2D} = V_j \otimes V_j,$$

and the next finer approximation space decomposes as

$$V_{j+1}^{2D} = (V_j \otimes V_j) \oplus (V_j \otimes W_j) \oplus (W_j \otimes V_j) \oplus (W_j \otimes W_j).$$

The first term contains the approximation coefficients and the remaining three terms correspond to the horizontal, vertical, and diagonal detail components. The associated separable scaling function is $\Phi(x, y) = \phi(x)\phi(y)$ with three 2D wavelets $\{\Psi^H, \Psi^V, \Psi^D\}$ formed from the products $\phi\psi$, $\psi\phi$, and $\psi\psi$.

For a spatial index $\mathbf{k} = (k_1, k_2) \in \mathbb{Z}^2$, the normalized 2D scaling basis functions are

$$\Phi_{j,\mathbf{k}}(x, y) = 2^j \Phi(2^j x - k_1, 2^j y - k_2), \tag{5}$$

with the corresponding wavelet families $\{\Psi^\lambda_{j,\mathbf{k}}\}$ defined analogously for $\lambda \in \{H, V, D\}$. These bases give the 2D approximation and detail expansion in Equations equation 1 and equation 2.

C ADDITIONAL DETAILS ON SPECTRAL TOKENIZATION AND AR-SIT

Building on the multiresolution construction in Appendix B, this section provides visual and architectural details of the spectral representation used by DP-Wavelet. Figure 4 illustrates how the 2D DWT separates an image into a coarse approximation and progressively finer detail components

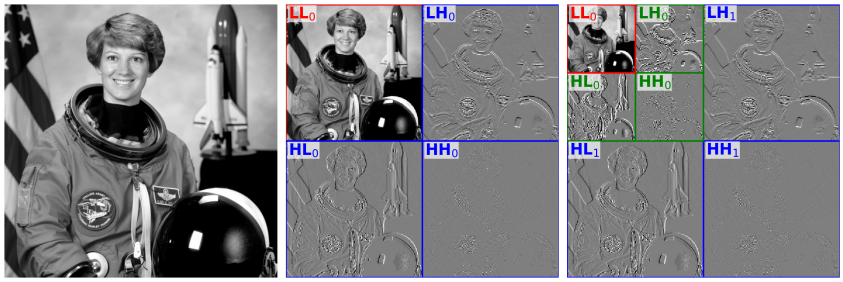


Figure 4: Visualization of the 2D Discrete Wavelet Transform. The original image is shown on the left, the center shows a one-level decomposition into the approximation LL_0 and detail subbands $\{LH_0, HL_0, HH_0\}$, and the right shows a two-level decomposition obtained by recursively decomposing the first-level approximation (Esteves et al., 2025)

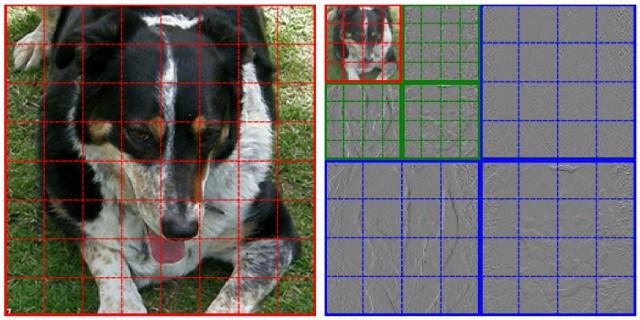


Figure 5: The left side shows standard ViT patchification into equal sized spatial patches and the right side gives the SIT patchification of DWT coefficients, with colors indicating different wavelet scales and the detail scales comprising horizontal, vertical, and diagonal subbands (Esteves et al., 2025).

and Figures 5 and 6 show how SIT organizes these components into spectral tokens and supports coarse-to-fine decoding (Esteves et al., 2025).

SIT applies the DWT and organizes the resulting coefficients by scale, beginning with the low-frequency approximation and followed by progressively finer detail components. The approximation and each detail scale are patchified separately before being encoded into discrete tokens. Figure 5 contrasts this spectral patchification with standard spatial patchification.

Since the approximation and detail coefficients have different statistics, SIT uses an Approximation-Details Transformer that applies separate attention projections, layer-normalization parameters, and MLP parameters to approximation and detail tokens while processing them within a shared sequence. The decoder additionally uses scale-causal attention, allowing a prefix ending at a complete scale to be decoded without requiring any finer-scale tokens. This partial-decoding property enables AR-SIT to generate images progressively from coarse to fine. Early prefixes reconstruct coarse image representations and longer prefixes incorporate increasingly fine spectral detail, as illustrated in Figure 6. DP-Wavelet uses this ordering and partial-decoding capability to control how much of the spectral sequence is included in private finetuning and, when the selected prefix does not cover the full sequence, to generate the remaining scales with the frozen public model.



Figure 6: An example of coarse-to-fine generation with AR-SIT. From left to right, the first 25%, 50%, 75%, and 100% of the spectral token sequence are generated and decoded, producing progressively finer image representations from 32×32 to 256×256 (Esteves et al., 2025).

D Privacy guarantee of DP-Wavelet

This section gives the formal argument that the privacy guarantee established during private finetuning is preserved through generation, optional public completion, and decoding.

Theorem 1. *Let $\mathcal{A}_{\mathrm{DP}}$ denote the complete randomized finetuning procedure used to obtain $\hat{\mathcal{T}}$. Suppose $\mathcal{A}_{\mathrm{DP}}$ is calibrated to satisfy (ε, δ)-DP under addition or removal of one training record. Then any outputs generated by DP-Wavelet using public conditioning and without further access to the private training data satisfy (ε, δ)-DP with respect to the private image dataset.*

Proof. Let Y and Y' be neighboring private image datasets that differ by the addition or removal of one image and let the conditioning information be public. For a fixed prefix length K_p, define the record-wise preprocessing map

$$P_{K_p}(y) = \mathcal{E}(y)_{1:K_p},$$

where the encoder $\mathcal{E}$ is fixed and public. Applying P_{K_p} independently to every image produces transformed datasets

$$P_{K_p}(Y) = \{P_{K_p}(y) : y \in Y\} \quad \text{and} \quad P_{K_p}(Y') = \{P_{K_p}(y) : y \in Y'\}.$$

Because Y and Y' differ in at most one record and P_{K_p} is applied independently to each record, the transformed datasets also differ in at most one record.

The privately finetuned transformer is

$$\hat{\mathcal{T}} = \mathcal{A}_{\mathrm{DP}}(P_{K_p}(Y)).$$

Since $\mathcal{A}_{\mathrm{DP}}$ is (ε, δ)-DP on neighboring transformed datasets, for every measurable set S,

$$\Pr[\mathcal{A}_{\mathrm{DP}}(P_{K_p}(Y)) \in S] \leq e^{\varepsilon} \Pr[\mathcal{A}_{\mathrm{DP}}(P_{K_p}(Y')) \in S] + \delta.$$

Hence $\hat{\mathcal{T}}$ is (ε, δ)-DP with respect to the original private image dataset.

At inference time, $\hat{\mathcal{T}}$ generates a spectral prefix conditioned only on public information. When $K_p < K$, the frozen public transformer $\mathcal{T}$ completes the remaining spectral tokens and for $K_p = K$ no additional completion is required. The public decoder $\mathcal{D}$ then maps the complete spectral sequence to an image. These operations access the private training data only through the DP model $\hat{\mathcal{T}}$ and are therefore post-processing. By Proposition 1, generation, optional public completion, and decoding preserve the same (ε, δ)-DP guarantee. □

E Training summary and hyperparameters

This section gives the training and implementation details used for the experiments in Section 5. Tables 3 and 4 report the dataset-specific hyperparameter values for all evaluated methods.

Table 3: Training hyperparameters for MS-COCO.

	DP-LDM	DP-LlamaGen	DP-Wavelet Coarse	DP-Wavelet Full
Max LR, $\varepsilon = \infty$	1×10^{-5}	2×10^{-5}	2×10^{-5}	1×10^{-5}
Max LR, $\varepsilon < \infty$	1×10^{-4}	2×10^{-5}	2×10^{-5}	1×10^{-5}
Updates, $\varepsilon = \infty$	3,230	4,500	100,000	100,000
Updates, $\varepsilon < \infty$	3,230	1,125	100,000	100,000
LR warmup steps	100	500	1,000	1,000
Batch size, $\varepsilon = \infty$	256	256	1,024	128
Batch size, $\varepsilon < \infty$	256	1,024	128	128
ℓ_2 clipping norm, $\varepsilon < \infty$	0.1	0.1	0.1	0.01

Table 4: Training hyperparameters for MM-CelebA-HQ.

	DP-LDM	DP-LlamaGen	DP-Wavelet Coarse	DP-Wavelet Full
Max LR, $\varepsilon = \infty$	5×10^{-5}	2×10^{-5}	2×10^{-5}	1×10^{-5}
Max LR, $\varepsilon < \infty$	2×10^{-5}	2×10^{-5}	2×10^{-5}	1×10^{-5}
Updates, $\varepsilon = \infty$	4,680	1,125	100,000	100,000
Updates, $\varepsilon < \infty$	4,680	1,125	100,000	100,000
LR warmup steps	100	500	1,000	1,000
Batch size, $\varepsilon = \infty$	1,024	1,024	1,024	128
Batch size, $\varepsilon < \infty$	1,024	1,024	128	128
ℓ_2 clipping norm, $\varepsilon < \infty$	0.1	0.1	0.1	0.01

Optimization. All runs use a linear learning rate warmup followed by cosine decay to 5% of the max learning rate. Private runs use DP-Adam with the per-example gradients clipped before noise addition as described in Section 3.1. Since the compared architectures differ in optimization dynamics and per-update cost, we select the training horizon separately for each method based on preliminary convergence behavior rather than enforcing a common number of updates and fix these horizons for the reported experiments. DP-Wavelet and DP-LDM are implemented in JAX (Bradbury et al., 2018) using Kauldron (Team, 2024) and DP-LlamaGen is implemented in PyTorch. The two DP-Wavelet configurations use identical 32-chip TPU v4 hardware for the direct training-cost comparison in Table 5.

Data processing. We finetune all methods using random cropping and horizontal flipping. For datasets with multiple captions per image, we uniformly sample a caption during training. At evaluation time, all methods use the first caption associated with each test image to give a fixed conditioning text for the paired LPIPS comparison.

Table 5: Private training cost of the two DP-Wavelet configurations.

	DP-Wavelet-Coarse	DP-Wavelet-Full
Training step time	0.1135 s	0.1305 s
Training throughput	1127.5 examples/s	980.5 examples/s
Training time, 100K updates	3.15 h	3.63 h
TPU chip-hours, 100K updates	100.8	116.2

F Additional experimental results

This section provides additional quantitative and qualitative results to complement results in 5.1 as well as experiments with an alternative model for public refinement for DP-Wavelet-Coarse.

Table 6: Additional FID and LPIPS results at 16×16 and 64×64 where each entry reports Metric@16 / Metric@64.

FID ↓					
Dataset	Privacy	DP-LDM	DP-LlamaGen	DP-Wavelet Coarse	DP-Wavelet Full
MS-COCO	Pretrained	13.0 / 11.8	17.3 / 17.8	**7.5 / 11.3**	**7.5 / 11.3**
	$\varepsilon = \infty$	14.2 / 13.2	11.4 / 12.7	**3.4 / 4.5**	3.7 / 7.9
	$\varepsilon = 10$	12.9 / 11.6	13.9 / 14.7	4.3 / 8.1	**4.0 / 7.8**
	$\varepsilon = 1$	13.0 / 11.8	16.6 / 16.5	5.1 / **8.4**	**4.8** / 8.6
MM-CelebA-HQ	Pretrained	42.1 / 44.9	57.9 / 69.6	**28.6 / 38.9**	**28.6 / 38.9**
	$\varepsilon = \infty$	29.7 / 21.7	28.0 / 32.1	12.5 / 12.1	**11.5 / 11.0**
	$\varepsilon = 10$	30.1 / 22.0	47.6 / 55.0	**15.3 / 16.1**	18.2 / 16.7
	$\varepsilon = 1$	39.4 / 28.7	51.9 / 61.9	**24.9 / 28.6**	28.1 / 30.2

LPIPS ↓					
Dataset	Privacy	DP-LDM	DP-LlamaGen	DP-Wavelet Coarse	DP-Wavelet Full
MS-COCO	Pretrained	0.430 / 0.667	0.448 / **0.509**	**0.417** / 0.651	**0.417** / 0.651
	$\varepsilon = \infty$	0.421 / 0.661	0.431 / **0.490**	0.259 / 0.523	**0.258** / 0.521
	$\varepsilon = 10$	0.430 / 0.666	0.441 / **0.501**	0.416 / 0.653	**0.415** / 0.652
	$\varepsilon = 1$	0.430 / 0.667	0.449 / **0.508**	0.419 / 0.653	**0.418** / 0.652
MM-CelebA-HQ	Pretrained	0.396 / 0.578	0.410 / **0.377**	**0.380** / 0.562	**0.380** / 0.562
	$\varepsilon = \infty$	0.334 / 0.500	0.306 / **0.283**	**0.194** / 0.384	0.238 / 0.424
	$\varepsilon = 10$	0.373 / 0.544	0.366 / **0.332**	**0.310** / 0.484	0.312 / 0.486
	$\varepsilon = 1$	0.388 / 0.556	0.386 / **0.351**	**0.333** / 0.513	0.345 / 0.526

F.1 Additional quantitative results

We give Table 6 which provides supplementary information to the main results in Table 1, and reports the FID and LPIPS for all methods evaluated at the 16×16 and 64×64 resolutions.

F.2 Additional qualitative results

This section provides additional qualitative results complementing the examples in Section 5.1. Figure 7 gives additional generations from DP-LDM, DP-LlamaGen, and DP-Wavelet-Coarse on MS-COCO and MM-CelebA-HQ across the adaptation regimes considered in Section 5. Figure 8 provides additional decoded intermediaries generated by DP-Wavelet-Coarse before public completion, illustrating the image structure captured by the privately finetuned spectral prefix across datasets and privacy regimes. Finally, Figure 9 gives additional DP-Wavelet-Coarse examples in the non-private finetuning setting, pairing each reference image with its final generation after public completion and the corresponding decoded coarse intermediary.



Figure 7: Additional generations on MS-COCO (left) and MM-CelebA-HQ (right) from DP-LDM, DP-LlamaGen, and DP-Wavelet-Coarse across the evaluated adaptation regimes.

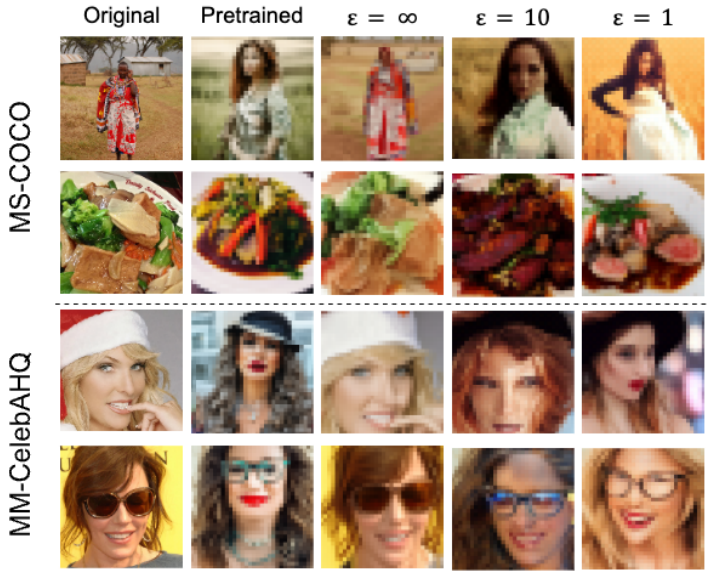


Figure 8: Additional decoded coarse intermediaries from DP-Wavelet-Coarse on MS-COCO and MM-CelebA-HQ. The first column shows reference images and the remaining columns show decoded intermediaries from the pretrained, non-private, and private adaptation regimes before public completion.

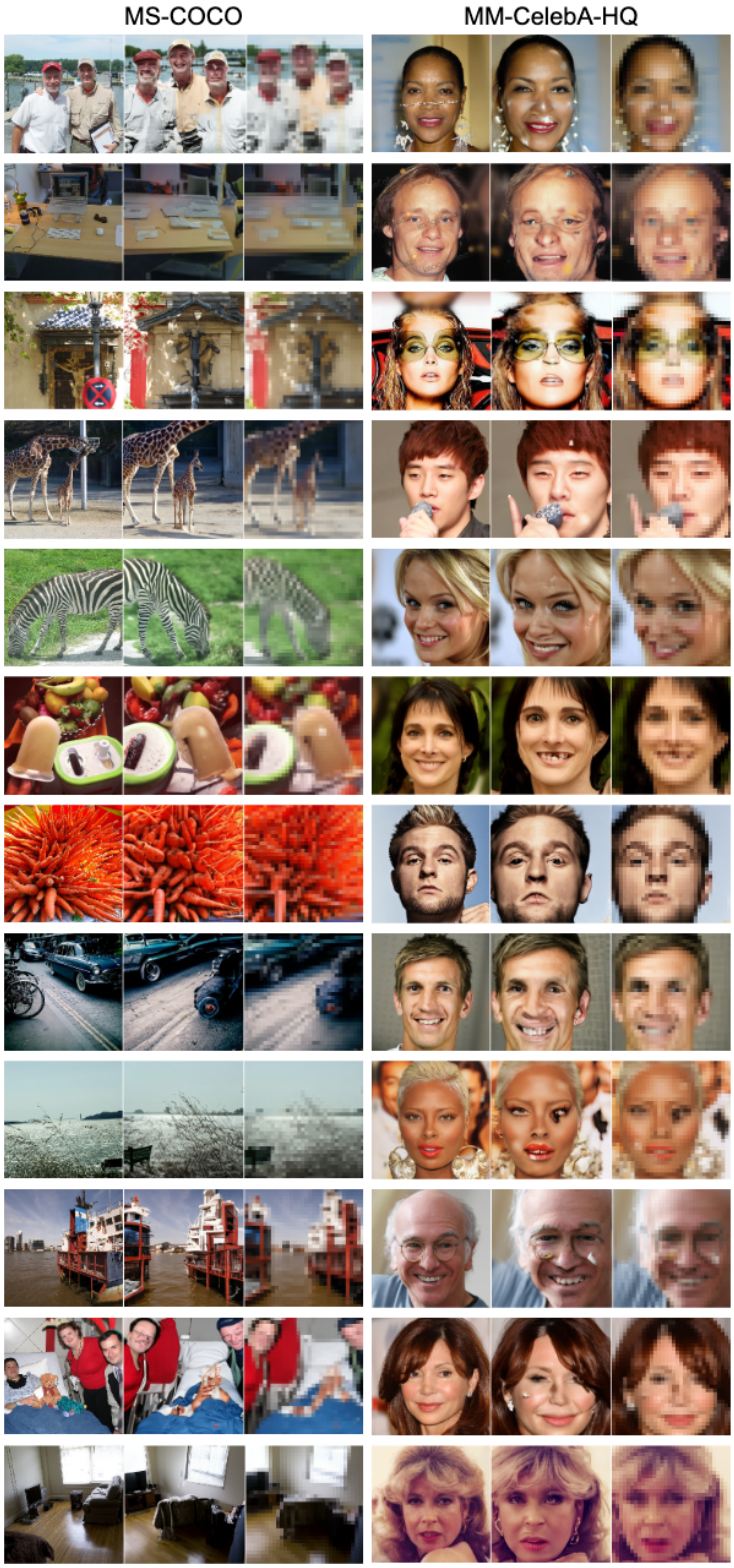


Figure 9: Additional DP-Wavelet-Coarse examples on MS-COCO and MM-CelebA-HQ with $\varepsilon = \infty$. Each triplet shows, from left to right, the reference image, final generation after public completion, and decoded coarse intermediary.

F.3 Alternative public refinement for DP-Wavelet-Coarse

For DP-Wavelet-Coarse, the frozen public AR-SIT transformer $\mathcal{T}$ completes the spectral sequence generated by the privately finetuned model $\hat{\mathcal{T}}$ before the public decoder $\mathcal{D}$ reconstructs the final image. Because this completion stage receives only the DP-generated intermediary and public text conditioning, replacing it with another public model remains post-processing and does not change the privacy guarantee. To examine how the choice of public refinement model affects the final output, we additionally decode the coarse DP-Wavelet representation and refine the resulting image using Gemini 3.1 Flash Image (Nano Banana 2) conditioned on the same public text.

We use the following prompt:

```
You are an image inpainting model.  Your task is to generate
a 512x512 version of the input image based on the caption:
<caption>.  Respond with ONLY the image data.  Do not
include any conversational text, explanations, or code
blocks.  Just the image.  Image:  <image>
```

Gemini produces a 512×512 refined image which we resize to each evaluation resolution before computing FID and LPIPS. Table 7 compares this variant with the standard AR-SIT completion stage for the non-private ($\varepsilon = \infty$) MS-COCO experiment.

Table 7: AR-SIT vs Gemini-based refinement for DP-Wavelet-Coarse on MS-COCO with $\varepsilon = \infty$.

	FID ↓		LPIPS ↓	
Resolution	AR-SIT	Gemini	AR-SIT	Gemini
16×16	3.4	4.0	0.259	0.289
32×32	4.1	5.6	0.424	0.457
64×64	4.5	6.8	0.523	0.546
256×256	10.1	19.8	0.665	0.672

Figures 10 and 11 show sample outputs from the Gemini refinement variant. Qualitatively, Gemini produces smoother high-resolution images and removes some artifacts visible after AR-SIT completion. However, this visual improvement does not translate to the evaluated metrics as AR-SIT achieves lower FID and LPIPS at every resolution in Table 7. These results show that the choice of public refinement model can materially affect the final appearance and that reduced visible artifacts do not necessarily correspond to better distributional or paired perceptual agreement with the reference data.

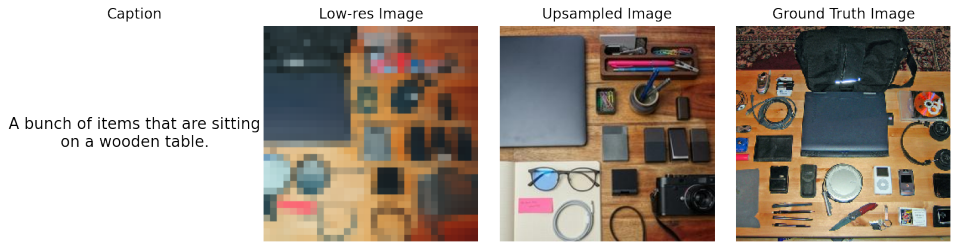


Figure 10: Representative DP-Wavelet-Coarse example on MS-COCO with $\varepsilon = \infty$ using Gemini 3.1 Flash Image for public refinement. From left to right, the figure shows the conditioning text, decoded coarse intermediary, Gemini-refined image, and reference image.

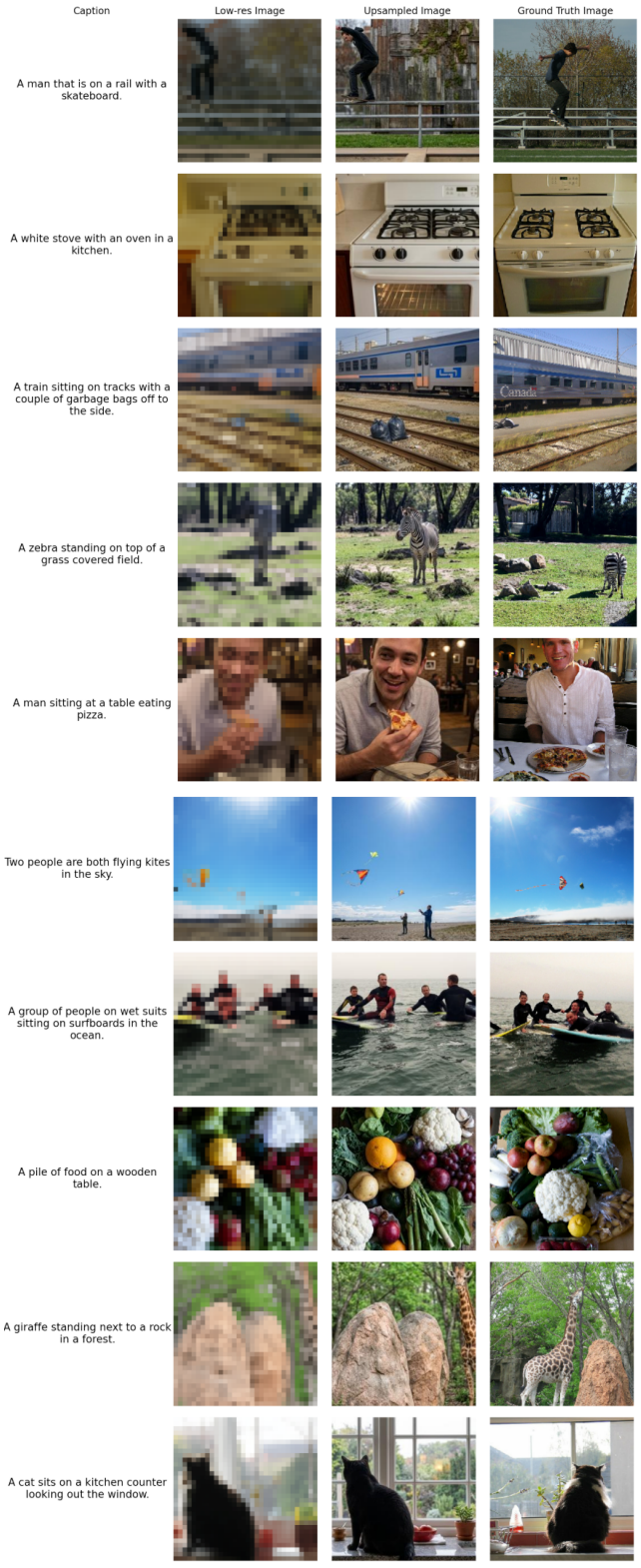


Figure 11: Additional DP-Wavelet-Coarse examples on MS-COCO with $\varepsilon = \infty$ using Gemini 3.1 Flash Image for public refinement. Each row shows the conditioning text, decoded coarse intermediary, Gemini-refined image, and reference image.