HEAR: HIGH-FREQUENCY ENHANCED AUTOREGRESSIVE MODELING FOR IDENTITY-PRESERVING IMAGE GENERATION

Anonymous authors
Paper under double-blind review

Figure 1: **Showcases of the exceptional ability of HEAR** to preserve individual identity while maintaining high visual fidelity. Our method consistently retains identity-specific features across diverse conditions, including varying artistic styles, age groups, and skin tones.

ABSTRACT

Recent autoregressive models such as LlamaGen, VAR, and Infinity have demonstrated remarkable advancements in image generation, even surpassing popular diffusion models in several aspects. However, diffusion models still dominate in controllable image generation, particularly in identity-preserving (IP) text-to-image generation, where autoregressive approaches remain underexplored. To bridge this gap, we propose ***HEAR***, a high-frequency enhanced autoregressive identity-preserving text-to-image framework based on a coarse-to-fine next-scale prediction paradigm, which leverages the key property of VAR we discovered for separating high- and low-frequency features in image generation. Innovations of our method include: (1) A comprehensive identity data curation pipeline that integrates powerful open-source vision-language models (VLMs) for image filtering and recaptioning, along with diffusion models for generating high-quality synthetic training data; (2) A high-frequency identity feature tokenizer, fine-tuned with compound losses and face-specific masking, to enhance high-frequency features essential for identity preservation; (3) A dual-control strategy in the autoregressive backbone, incorporating global information into the cross-attention blocks and introducing a decoupled adapter operating in parallel to maintain high-frequency details. Extensive experiments demonstrate that HEAR surpasses mostly existing diffusion-based methods in identity-preserving image generation. This work presents a general and scalable autoregressive framework for controllable image generation.

1 INTRODUCTION

The rapid advancement of autoregressive (AR) models (Lee et al., 2022; Zheng et al., 2022; Huang et al., 2023; Yu et al., 2024c; Tian et al., 2024) has recently driven significant progress in text-to-image generation (Sun et al., 2024; Han et al., 2024). However, research on controllable generation, particularly identity-preserving (IP) image generation in autoregressive frameworks remains significantly underexplored compared to the remarkable success of diffusion models (Ye et al., 2023; Wang et al., 2024b; Li et al., 2024d). While diffusion-based paradigms (Sohl-Dickstein et al., 2015; Song et al., 2020; Dhariwal & Nichol, 2021; Betker et al., 2023; Esser et al., 2024) have dominated the field of controllable image generation (Zhang et al., 2023; Mou et al., 2024), their sequential denoising process and architectural heterogeneity fundamentally conflict with the requirements for unified multimodal modeling (Xie et al., 2024b; Yang et al., 2025). These limitations motivate our investigation into controllable image generation within autoregressive frameworks (Li et al., 2024c;e; Xiao et al., 2024), with a particular emphasis on identity-preserving image synthesis, which remains an open challenge in autoregressive-based methods.

Traditional autoregressive frameworks for image generation preliminary rely on next-token prediction for sequential modeling (Sun et al., 2024; Tang et al., 2024; Fan et al., 2024). However, this token-level sequential modeling paradigm poses significant challenges for image generation, as it lacks the flexibility to revise previously generated tokens based on subsequent tokens' information, creating critical limitations in achieving global coherence (Pang et al., 2024; Tian et al., 2024). Recent advances have introduced novel approaches to address these limitations in visual autoregressive modeling. VAR (Tian et al., 2024) proposes a novel next-scale prediction mechanism, redefining image generation as a hierarchical coarse-to-fine process. Its scale-wise image generation process shares structural similarities with the denoising process in diffusion models, as both employ progressive refinement from global structures to fine details, but it requires fewer steps. Building on this, Infinity (Han et al., 2024) extends VAR for scalable text-to-image generation. Next-scale prediction explicitly separates low-frequency macroscopic structures from high-frequency microscopic details during image generation. This leads us to posit that the paradigm is especially well-suited for identity-preserving image generation tasks, which demand precise control over high-frequency details.

We conducted experiments to validate our arguments by sampling 2,000 diverse text prompts and measuring both pixel-level and token-level reconstruction losses using Infinity (Han et al., 2024). At each scale, we sum the tokens predicted from all previous scales and reconstruct the intermediate image via detokenization. As shown in Fig. 2 (a), fundamental low-frequency features such as layout and color are largely established in the early scales, while later scales focus on refining details and reconstructing high-frequency components. Fig. 2 (b) further illustrates this high-low frequency separation through metric trends: in the later stages of image generation, pixel-level MSE exhibits smaller changes, indicating that these later scales primarily concentrate on fine-grained detail reconstruction. This empirical observation supports the effectiveness of the hierarchical VAR modeling paradigm, which performs coarse-to-fine feature decomposition with explicit separation of high- and low-frequency components, making it particularly well-suited for identity-preserving tasks that demand precise control over high-frequency information.

The separation of fine and coarse features in the hierarchical prediction framework renders it highly effective in identity-preserving text-to-image synthesis. To this end, we introduce HEAR, a high-frequency enhanced autoregressive identity-preserving text-to-image framework based on coarse-to-fine next-scale prediction paradigm for high-quality identity-preserving image generation. We first leverage the powerful open-sourced Vision-Language Models (VLMs) to perform precise face data filtering and recaption, while employing advanced generative models, including FLUX-dev(Labs, 2024) and SD3.5-large(Rombach et al., 2022) for high-quality synthetic data generation. Then we proposed a high-frequency identity encoder to specifically extract high-frequency face features. The training process employed multiple heterogeneous loss functions (including structural similarity loss and detail reconstruction loss) combined with a novel face-specific loss mask, which strategically weights facial regions through adaptive attention mechanisms during backpropagation. Finally, we implement a dual-controllable strategy for the backbone architecture by first injecting global information into the original cross-attention block and then incorporating a decoupled adapter that operates in parallel to preserve high-frequency features. Extensive qualitative and quantitative experiments demonstrate the effectiveness of our method and its significant improvement in identity-preserving image generation.

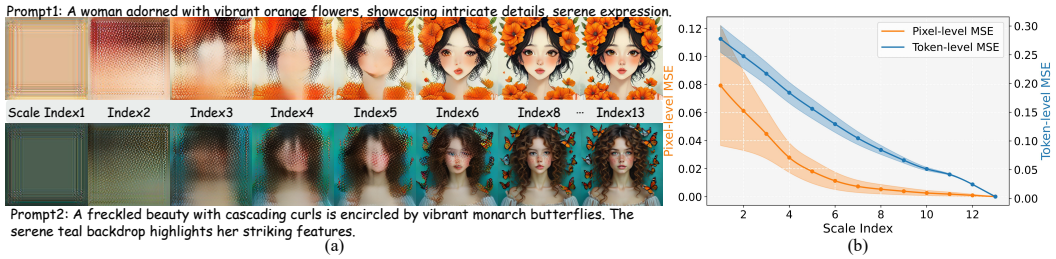

Figure 2: **Motivation of HEAR.** (a) In the early stages, smaller scales primarily determine low-frequency features such as shape, position, and color, while larger determine high-frequency features, including facial details. (b) This hypothesis is further validated through empirical experiments.

Our contributions are summarized as follows:

- An insightful discovery regarding the coarse-to-fine feature decomposition with explicit separation of high- and low-frequency in the next scale prediction paradigm.
- We propose HEAR, a new high-frequency enhanced visual autoregressive framework for identity-preserving image generation, and provide a new perspective for controllable autoregressive image generation.
- We introduce a novel identity data curation pipeline and train a high-frequency face encoder for the better construction of face details.
- Extensive qualitative and quantitative comparisons with previous powerful methods demonstrate the effectiveness and superiority of our method.

2 Related Work

Autoregressive Image Generation Autoregressive image generation models leveraged the GPT-style (Radford et al., 2018) paradigm to model the distribution of pixels or latent codes in a sequential manner (Esser et al., 2021; Razavi et al., 2019; Yang et al., 2025). Earlier autoregressive models, such as VQ-VAE (Van Den Oord et al., 2017), used discrete visual tokenizers to predict the next visual token. Parti (Yu et al., 2022) formulates high-resolution text-to-image generation as a sequence-to-sequence task, where the output is a sequence of image tokens. Open-MAGVIT2 (Luo et al., 2024) introduces asymmetric token decomposition and a next sub-token prediction mechanism to enhance generation quality. Autoregressive text-to-image generation has achieved remarkable advancement recently. Numerous works such as LlamaGen (Sun et al., 2024), which is LLM-based (Vaswani et al., 2017; Zhang et al., 2022; Devlin et al., 2019) architectures leveraging powerful scaling capabilities, enabling autoregressive models to rival or even surpass diffusion models in image generation quality. Beyond next-token prediction, some autoregressive models also shifted toward more diverse token representations. MAR (Li et al., 2024b) introduces a diffusion-based method (Ho et al., 2020) to model the probability distribution of each token in continuous space, replacing the conventional cross-entropy loss with a diffusion loss. VAR (Tian et al., 2024) redefines the conventional coarse-to-fine paradigm by shifting from next-token prediction to a novel next-scale prediction framework, demonstrating strong potential in image synthesis. xAR (Ren et al., 2025) proposes a generalized and more flexible next-X prediction framework, where X can represent tokens, scales, or spatial cells. Other innovations (Yu et al., 2024b; Li et al., 2024a;b) have also emerged in models such as DART (Gu et al., 2024) and Fluid (Fan et al., 2024). Built upon the next-scale prediction paradigm, Infinity (Han et al., 2024) employs an infinite-vocabulary tokenizer and classifier, along with a bit-level self-correction mechanism to achieve powerful text-to-image generation quality. We adopt Infinity as the backbone of our HEAR due to its flexibility in handling high-frequency enhancement.

Identity Preserving Text-to-Image Generation Identity-preserving text-to-image (T2I) generation extends conventional text-to-image generation by enforcing strict identity (ID) consistency between the generated image and a reference subject. Numerous methods (Yu et al., 2024a; Wang et al., 2024b; Liang et al., 2024; Zhang et al., 2024c;a;b) leverage diffusion models to achieve remarkable success in this area. LoRA (Hu et al., 2022) and ControlNet (Zhang et al., 2023) augment

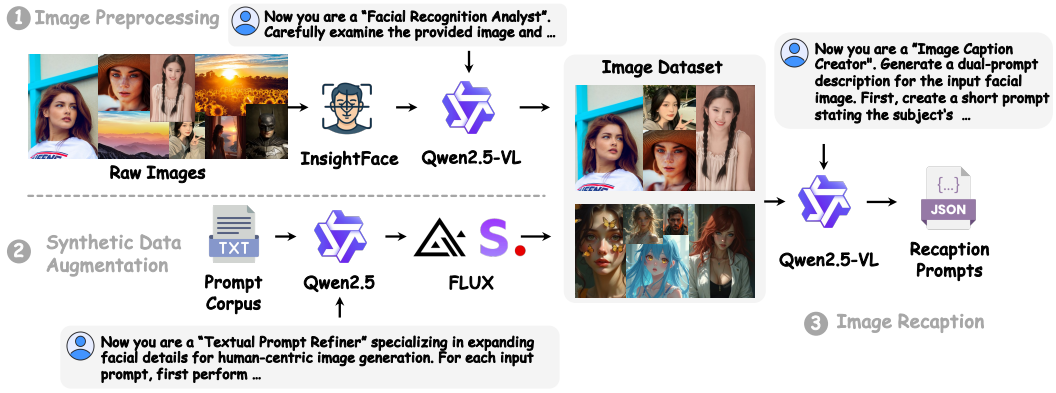

Figure 3: An overview of Identity-Preserving Dataset Curation Pipeline.

base diffusion models with trainable layers to enable controllable generation conditioned on inputs like pose, masks, edges, and depth. PhotoMaker (Li et al., 2024d) preserves identity preservation by directly merging text embeddings with image embeddings. Other methods, including IP-Adapter (Ye et al., 2023), InstantID (Wang et al., 2024b), and ConsistentID (Huang et al., 2024) freeze backbone parameters and inject identity features through a decoupled cross-attention mechanism, achieving strong ID preservation with minimal additional training. Furthermore, methods like UniPortrait (He et al., 2024) and ID-Adapter (Chen et al., 2024) continue to push the boundaries of generalization and effectiveness in identity-preserving generation. However, identity-preserving text-to-image generation has been rarely addressed using autoregressive models. Some methods such as ControlVAR (Li et al., 2024c), ControlAR (Li et al., 2024e), and OmniGen (Xiao et al., 2024) mainly explore controllable generation under different input conditions. Compared to diffusion models, autoregressive methods offer superior inference efficiency and stronger multimodal fusion capabilities. Our method trains a high-frequency face encoder for extracting high-frequency image features, and adopts the next-scale prediction autoregressive framework and employs a decoupled cross-attention mechanism to inject both global visual features and high-frequency identity features into the transformer layers.

3 Method

In this section, we present HEAR, a high-frequency enhanced autoregressive text-to-image framework designed for high-quality identity-preserving image generation with a coarse-to-fine next-scale prediction paradigm, as illustrated in Fig. 4.

3.1 Curation of High-Quality Identity-Preserving Data

A high-quality training dataset is essential for achieving identity-preserving generation. However, datasets curated specifically for identity-preserving tasks often contain significant noise, including profile views and heavily occluded faces. To overcome these limitations, we propose an automated data curation pipeline that harnesses the capabilities of advanced Vision-Language Models (VLMs). This system efficiently filters out low-quality samples, using synthetic data to augment the original dataset and provide fine-grained captions. The overall pipeline is illustrated in Fig. 3.

Image Preprocessing We begin by applying InsightFace (Deng et al., 2019) to the raw images, retaining only those in which a face can be reliably detected. Subsequently, we perform a second-round filtering using Qwen2.5-VL-32B (Wang et al., 2024a) to ensure high data purity. Through this preprocessing pipeline, we obtain a collection of high-quality face images, free from profile views and heavy occlusions.

Synthetic Data Augmentation The dataset primarily consists of authentic photographic content, which often lacks aesthetic quality. To enhance the visual appeal of the generated outputs of our model, we augment the training corpus with a substantial synthetic dataset produced by state-of-the-art image generation models. Specifically, we randomly select 50K identity-related text prompts and use Qwen2.5-32B (Yang et al., 2024) to enrich them with fine-grained facial descriptions, including attributes such as ethnicity, gender, facial features, expressions, and accessories. These finalized prompts are then fed into high-fidelity image generation models, including Stable Diffusion 3.5

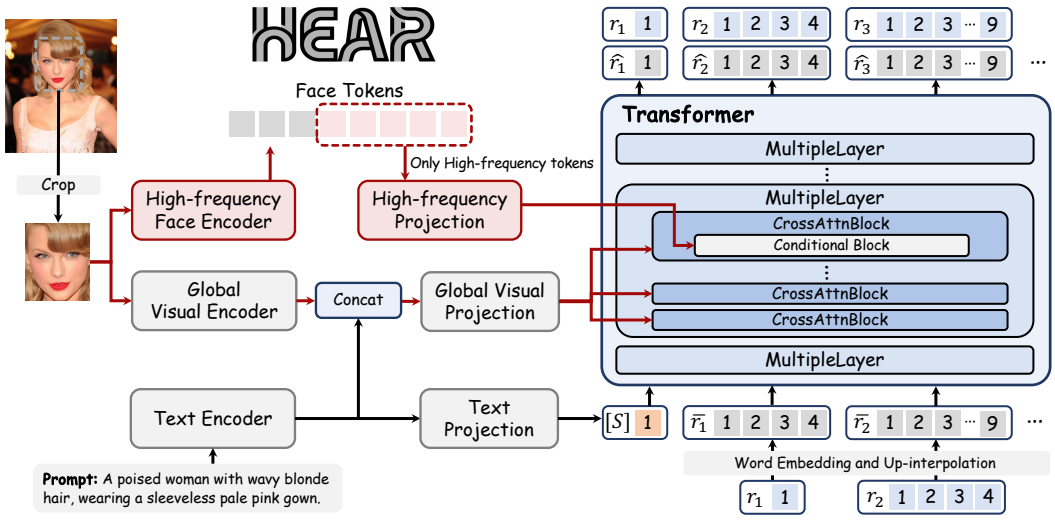

Figure 4: An overview of HEAR pipeline.

Large (Rombach et al., 2022) and FLUX-dev (Labs, 2024), to produce photorealistic synthetic face images.

Image Recaption Image Recaption Identity preserving T2I models depend on accurate captions, as generic descriptions miss crucial identity cues. For each image we first use Qwen2.5-VL-32B to produce a brief prompt that specifies gender, ethnicity, action, and environment; the model then expands this summary with subtle facial expressions and gaze, detailed clothing characteristics, and quantitative spatial relations between the subject and the scene. During training, each image is paired with either the short or the expanded caption with equal probability, allowing the network to learn from both coarse and fine descriptions and thereby improving robustness.

3.2 High-frequency Face Encoder

The fundamental goal of identity preservation in facial reconstruction is to retain identity-specific details, which are primarily encoded in high-frequency components. However, existing visual tokenizers struggle to capture these high-frequency features accurately, as they are typically pre-trained on large-scale, general-purpose image datasets. These datasets prioritize broad semantic coverage rather than fine-grained facial attributes, and thus lack domain-specific optimization for identity-related features. To address this issue, we present a high-frequency face encoder, specifically designed and trained for facial identity reconstruction tasks, to enhance sensitivity to high-frequency identity features. This extractor is trained with composite losses, including reconstruction loss $\mathcal{L}_{\text{recon}_{L_1}}$ and $\mathcal{L}_{\text{recon}_{L_2}}$, vector quantization loss $\mathcal{L}_{\text{VQ}}$, Perceptual loss $\mathcal{L}_{\text{lpips}}$, CLIP loss $\mathcal{L}_{\text{clip}}$ and adaface loss $\mathcal{L}_{\text{adaface}}$:

$$\mathcal{L} = \lambda_{\text{recon}_{L_1}}\mathcal{L}_{\text{recon}_{L_1}} + \lambda_{\text{recon}_{L_2}}\mathcal{L}_{\text{recon}_{L_2}} + \lambda_{\text{VQ}}\mathcal{L}_{\text{VQ}} + \lambda_{\text{lpips}}\mathcal{L}_{\text{lpips}} + \lambda_{\text{clip}}\mathcal{L}_{\text{clip}} + \lambda_{\text{adaface}}\mathcal{L}_{\text{adaface}} \quad (1)$$

Specifically, the reconstruction loss measures the L_1 and L_2 distances between the reconstructed image and the ground truth, measuring pixel-level fidelity. The vector quantization loss encourages alignment between the encoded features and their corresponding codebook vectors. To capture perceptual similarity, the perceptual loss compares high-level feature representations extracted by the pre-trained LPIPS (Zhang et al., 2018). The CLIP loss enforces semantic consistency by regularizing the semantic tokens using features from the pre-trained DINOv2 (Oquab et al., 2023). Finally, an AdaFace recognition module is integrated with the AdaFace loss (Kim et al., 2022) to ensure facial similarity between the reconstructed and ground truth images.

To focus on identity-specific details within facial regions during training, we introduce a face-specific spatial weighting mask that amplifies the loss contribution of facial regions. Specifically, for each training image, we use InsightFace (Deng et al., 2019) to detect facial bounding boxes with coordinates (x_1, x_2, y_1, y_2). A position-dependent weighting factor $\alpha(\alpha > 1)$ is applied to all pixels within the facial region $\mathcal{R} = \{(i, j) \mid x_1 \leq i \leq x_2,\ y_1 \leq j \leq y_2\}$, while pixels outside this region are assigned a default weight of 1. This spatial weighting mechanism compels the encoder to focus

more on discriminative, high-frequency facial patterns that are critical for identity preservation:

$$w_{i,j} = \begin{cases} \alpha, & (i,j) \in \mathcal{R}, \\ 1, & \text{otherwise.} \end{cases} \tag{2}$$

Consequently, facial regions within the bounding box receive significantly higher attention, effectively guiding the model to concentrate on identity-relevant features during training.

3.3 HEAR: High-frequency Enhanced Autoregressive Model

The overview of HEAR is illustrated in Fig. 4. Given a reference image, the global visual encoder and local high-frequency face encoder respectively inject global and high-frequency facial feature into the model via cross-attention blocks and the decoupled adapter.

3.3.1 Global View: Cross-Attention Blocks

In our proposed global view framework, we begin by extracting text embeddings from a text encoder and global facial image embeddings from a global visual encoder. We leverage a pre-trained and frozen word embedding layer to directly align the global visual embeddings $\mathbf{v}_\text{g}$ with the text embeddings $\mathbf{t}$. We then concatenate the text and global facial image embeddings and passed through a global projection layer. The fused features $\mathbf{f}$ can be formulated as:

$$\mathbf{f} = \text{Concat}(\mathbf{t}, \mathbf{v}_\text{g}) \cdot \mathbf{W}_\text{g} \tag{3}$$

where $\text{Concat}(\cdot)$ means concatenating the text embeddings $\mathbf{t}$ and global visual embeddings $\mathbf{v}_\text{g}$ along the sequence (length) dimension. $\mathbf{W}_\text{g}$ is the global projection matrix. Given the query features $\mathbf{Z}$ and the fusion features $\mathbf{f}$, the output of global view cross-attention $\mathbf{Z}_g$ is computed as follows:

$$\begin{gathered} \mathbf{Z}_\text{g} = \text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right) \cdot \mathbf{V} \\ \mathbf{Q} = \mathbf{W}_Q \cdot \mathbf{Z},\ \mathbf{K} = \mathbf{W}_K \cdot \mathbf{f},\ \mathbf{V} = \mathbf{W}_V \cdot \mathbf{f} \end{gathered} \tag{4}$$

where $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$ are the query, key, and values matrices of the attention operation respectively, and $\mathbf{W}_Q$, $\mathbf{W}_K$, $\mathbf{W}_V$ are the weight matrices of the trainable linear projection layers.

3.3.2 Local View: A Decoupled Adapter In Parallel

For each distinct identity, high-frequency features capture unique and discriminative characteristics more effectively than low-frequency components, making them critical for distinguishing one identity from another. To leverage this, after injecting global facial information through cross-attention blocks, we reintroduce high-frequency facial features extracted by a dedicated high-frequency identity feature extractor. Inspired by prior works (Ye et al., 2023; Wang et al., 2024b; Huang et al., 2024), we incorporate these features using a lightweight decoupled adapter to avoid redundant control modules and excessive trainable parameters.

Leveraging the unique quantization mechanism of the next-scale prediction paradigm, the encoder naturally organizes latent image embeddings into frequency-ordered embeddings (from low to high frequency) after quantization. Once a cropped facial image is encoded into its corresponding face embedding, we apply parameter FT to truncate the visual embedding $\mathbf{v}$ and extract the high-frequency visual embedding $\mathbf{v}_\text{hf}$. This process is formally defined as follows:

$$\mathbf{v}_\text{hf} = \mathbf{v}[\text{FT}:] \cdot \mathbf{W}_\text{hf} \tag{5}$$

where $[\text{FT}:]$ denotes slicing for tokens from the latter high-frequency scales. $\mathbf{W}_\text{hf}$ is the high-frequency projection matrix. Given the query features $\mathbf{Z}$ and the high-frequency face embedding $\mathbf{v}_\text{hf}$, the output of local view high-frequency cross-attention $\mathbf{Z}_\text{hf}$ can be defined by the following equation:

$$\begin{gathered} \mathbf{Z}_\text{hf} = \text{Attention}(\mathbf{Q}, \mathbf{K}', \mathbf{V}') = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}'^\top}{\sqrt{d}}\right) \cdot \mathbf{V}' \\ \mathbf{Q} = \mathbf{W}_Q \cdot \mathbf{Z},\ \mathbf{K}' = \mathbf{W}'_K \cdot \mathbf{v}_\text{hf},\ \mathbf{V}' = \mathbf{W}'_V \cdot \mathbf{v}_\text{hf} \end{gathered} \tag{6}$$

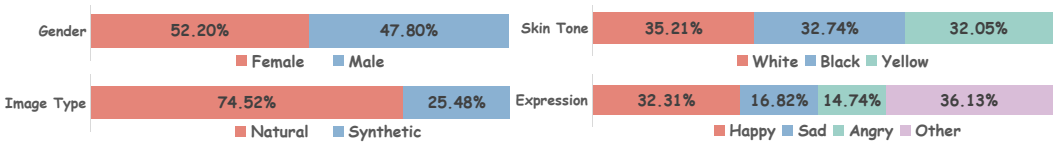

Figure 5: The statistical characteristics distribution in the training dataset

where $\mathbf{Q}$, $\mathbf{K}'$, and $\mathbf{V}'$ denote the query, key, and value matrices used in the attention operation, respectively. The query matrix $\mathbf{W}_Q$ is shared between the global cross-attention and the high-frequency cross-attention modules. The matrices $\mathbf{W'}_K$ and $\mathbf{W'}_V$ are the corresponding weight matrices for the key and value projections in the high-frequency pathway. As a result, only two additional parameters, $\mathbf{W'}_K$ and $\mathbf{W'}_V$, are introduced per cross-attention layer. We initialize $\mathbf{W'}_K$ and $\mathbf{W'}_V$ with the weights of $\mathbf{W}_K$ and $\mathbf{W}_V$ respectively.

3.3.3 A DUAL-CONTROL STRATEGY

The global and high-frequency face features are respectively integrated into the backbone via the inherent and decoupled cross-attention. In the original Infinity model, the text features from the CLIP text encoder are plugged into transformer by feeding into the cross-attention blocks. Given the global view cross-attention $\mathbf{Z}_{\text{g}}$ and the local view high-frequency cross-attention $\mathbf{Z}_{\text{hf}}$, the final output of cross-attention $\mathbf{Z}'$ is defined as follows:

$$\mathbf{Z}' = \mathbf{Z}_{\text{g}} + \mathbf{Z}_{\text{hf}} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right) \cdot \mathbf{V} + \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K'}^\top}{\sqrt{d}}\right) \cdot \mathbf{V}' \tag{7}$$

Ultimately, global facial features $\mathbf{Z}_{\text{g}}$ are still injected through the original cross-attention blocks of the transformer, while high-frequency details $\mathbf{Z}_{\text{hf}}$ are enhanced via the insertion of lightweight decoupled image cross-attention layers. This design ensures minimal parameter overhead and avoids introducing additional heavy modules.

4 EXPERIMENT

4.1 EXPERIMENTAL SETUP

Implementation Details Our experiments are based on Infinity (Han et al., 2024), where we fine-tune a custom high-frequency face encoder on top of the original Infinity image encoder. During the training of this face encoder, we set the position-dependent weighting factor α to 2. We set $\lambda_{\text{recon}_{L_1}} = 0.2$,$\lambda_{\text{recon}_{L_2}} = \lambda_{\text{VQ}} = 1.0$, $\lambda_{\text{lpips}} = 0.5$ and $\lambda_{\text{clip}} = \lambda_{\text{adaface}} = 0.1$.

During transformer training, we define multiple aspect ratio templates, and all images are resized to match one of the predefined ratios. This allows the model to generate outputs of varying aspect ratios during inference. To improve robustness, each image is paired with both a long and a short caption, with a 50% probability of either being selected during training. In our design, we use 13 scales, and FT refers to the total token length of the first 6 scales. We adopt the AdamW optimizer with a fixed learning rate of 0.0001 and a weight decay of 0.01. Our model is trained for 500K steps on a single machine equipped with 8*NVIDIA A100-80G GPUs, using a batch size of 8 per GPU.

Data Composition and Distribution We sampled 51 k, 98 k, and 482 k images from CelebA, LAION-Face, and X2I, then applied the Fig. 3 filter, yielding 150 k photos with clear, unobstructed faces and distinct identities. We augmented this set with synthetic images generated by Stable Diffusion 3.5-large and FLUX-dev, maintaining a 3:1 natural-to-synthetic ratio. Gender, skin tone, and other facial attributes were distribution-balanced (Fig. 5), giving the final corpus both demographic parity and strong photorealism, which in turn boosts model generalization.

Experimental Metrics To comprehensively evaluate the effectiveness and efficiency of HAER, we adopt five widely recognized metrics (Ruiz et al., 2023): CLIP-T (Gal et al., 2022), CLIP-I (Radford et al., 2021), DINO (Cong et al., 2020), FaceSim (Schroff et al., 2015), and computational efficiency (inference speed). All experiments were conducted under standardized hardware conditions to ensure fairness and reproducibility.

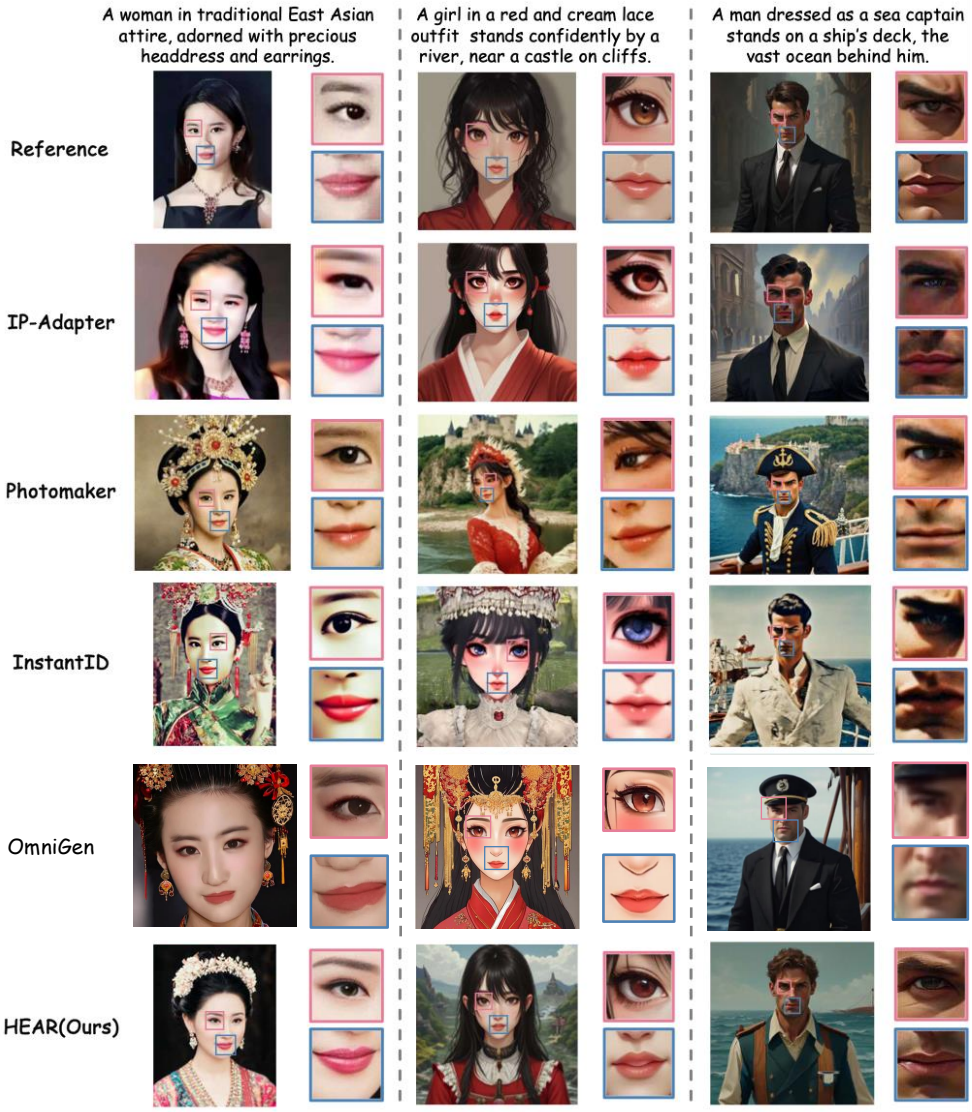

Figure 6: **Comparison of facial feature details between HEAR and existing methods.** The characters generated by our method demonstrate enhanced identity consistency, particularly in facial features such as the eyes, nose, and mouth.

4.2 Main Results

4.2.1 Quantitative Comparison

We assembled a test set of 40 distinct identities spanning a wide variety of appearances. Consistent with PhotoMaker (Li et al., 2024d), the set also includes the images associated with MyStyle (Nitzan et al., 2022) identities. HEAR consistently outperforms other methods across most evaluation metrics, particularly in CLIP-T, FaceSim, and inference speed, as shown in Table 1. The strong performance on CLIP-T and FaceSim can be attributed to HEAR's high-frequency enhancement strategy, which enables more precise control over fine-grained details. Its superior speed results from both the inherent efficiency of the autoregressive framework and HEAR's streamlined design, which avoids heavy modules and excessive parameter growth. These combined strengths allow HEAR to preserve fine-grained identity features while remaining an efficient, lightweight multimodal face prompt generator.

4.2.2 Qualitative Comparison

To intuitively demonstrate the advantages of HEAR, we conducted a qualitative evaluation using a diverse set of images with varying types and styles, comparing our method against IP-Adapter

Table 1: Comparative Evaluation of Identity Preservation Methods

Method	CLIP-T(↑)	CLIP-I(↑)	DINO(↑)	FaceSim(↑)	Speed(↓)
Photomaker Li et al. (2024d)	30.1	67.4	73.8	50.8	20.18
IP-Adapter Ye et al. (2023)	29.2	68.2	74.5	52.1	12.48
InstantID Wang et al. (2024b)	30.4	70.2	78.1	55.1	17.51
OmniGen Xiao et al. (2024)	32.5	69.3	78.1	53.4	40.32
HEAR (Ours)	**32.6**	**72.1**	**78.9**	**56.3**	**6.62**

Table 2: Comparative Analysis of Image Tokenization Methods: Performance on MSE, SSIM and PSNR.

Method	MSE↓	SSIM↑	PSNR↑
Open-MAGVIT2 Luo et al. (2024)	57.43	0.72	22.94
LlamaGen Sun et al. (2024)	52.33	0.76	23.15
Show-o Xie et al. (2024a)	42.35	0.76	23.63
Infinity Han et al. (2024)	15.15	0.94	34.31
HEAR (Ours)	**13.87**	**0.95**	**35.32**

(Ye et al., 2023), PhotoMaker (Li et al., 2024d), and InstantID (Wang et al., 2024b). We selected reference images from five different identities to showcase the text-driven generation results for each method, as illustrated in Fig. 6.

Both IP-Adapter and InstantID exhibit a certain degree of failure in guiding image generation effectively with textual prompts. In contrast, HEAR leverages a Dual-Control Strategy that allows textual input to participate more actively in the generation process, resulting in better controllability. Furthermore, IP-Adapter and ControlNet fall short in preserving facial details compared to our approach, primarily because HEAR incorporates a high-frequency enhancement mechanism to retain fine-grained identity features. In terms of visual appeal, InstantID also underperforms relative to HEAR, as our method benefits from a curated high-quality dataset that significantly improves aesthetic quality. In summary, our model surpasses existing methods in textual guidance, high-frequency-controlled generation, as well as in overall image quality and aesthetic fidelity.

4.3 Comparison of High-frequency Face Encoder

Quantitative Comparison To rigorously evaluate the reconstruction capabilities of our high-frequency face encoder, we conducted comprehensive comparisons against several tokenizers using the MyStyle (Nitzan et al., 2022) facial dataset. The evaluation focused on image–condition alignment, employing two principal metrics SSIM and PSNR.

As shown in Table 2, our method outperforms competing approaches across both metrics. This technical advantage stems from high-frequency face encoder's specialized high-frequency feature extraction mechanism, combined with a dataset-specific fine-tuning strategy optimized for facial characteristics. This synergistic design enables more accurate preservation of critical biometric details while maintaining strong alignment with visual-textual conditions.

5 Conclusion

In this paper, we present HEAR, a high-frequency enhanced autoregressive identity-preserving (IP) text-to-image framework based on a coarse-to-fine next-scale prediction paradigm. We introduce targeted improvements across the entire training pipeline, including dataset curation, high-frequency face encoder training, and transformer-based architectural design. Experimental results demonstrate HEAR's superior performance in identity-preserving image generation, outperforming several powerful diffusion-based models. However, HEAR also has limitations due to its current model size (2B parameters), and we will use a larger model size for better performance.

REFERENCES

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

Weifeng Chen, Jiacheng Zhang, Jie Wu, Hefeng Wu, Xuefeng Xiao, and Liang Lin. Id-aligner: Enhancing identity-preserving text-to-image generation with reward feedback learning. *arXiv preprint arXiv:2404.15449*, 2024.

Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. Dovenet: Deep image harmonization via domain verification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8394–8403, 2020.

Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

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.

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.

Lijie Fan, Tianhong Li, Siyang Qin, Yuanzhen Li, Chen Sun, Michael Rubinstein, Deqing Sun, Kaiming He, and Yonglong Tian. Fluid: Scaling autoregressive text-to-image generative models with continuous tokens. *arXiv preprint arXiv:2410.13863*, 2024.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

Jiatao Gu, Yuyang Wang, Yizhe Zhang, Qihang Zhang, Dinghuai Zhang, Navdeep Jaitly, Josh Susskind, and Shuangfei Zhai. Dart: Denoising autoregressive transformer for scalable text-to-image generation. *arXiv preprint arXiv:2410.08159*, 2024.

Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. *arXiv preprint arXiv:2412.04431*, 2024.

Junjie He, Yifeng Geng, and Liefeng Bo. Uniportrait: A unified framework for identity-preserving single-and multi-human image personalization. *arXiv preprint arXiv:2408.05939*, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

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.

Jiehui Huang, Xiao Dong, Wenhui Song, Zheng Chong, Zhenchao Tang, Jun Zhou, Yuhao Cheng, Long Chen, Hanhui Li, Yiqiang Yan, et al. Consistentid: Portrait generation with multimodal fine-grained identity preserving. *arXiv preprint arXiv:2404.16771*, 2024.

Mengqi Huang, Zhendong Mao, Zhuowei Chen, and Yongdong Zhang. Towards accurate image coding: Improved autoregressive image generation with dynamic vector quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22596–22605, 2023.

Minchul Kim, Anil K Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18750–18759, 2022.

Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11523–11532, 2022.

Haopeng Li, Jinyue Yang, Kexin Wang, Xuerui Qiu, Yuhong Chou, Xin Li, and Guoqi Li. Scalable autoregressive image generation with mamba. *arXiv preprint arXiv:2408.12245*, 2024a.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems*, 37: 56424–56445, 2024b.

Xiang Li, Kai Qiu, Hao Chen, Jason Kuen, Zhe Lin, Rita Singh, and Bhiksha Raj. Controlvar: Exploring controllable visual autoregressive modeling. *arXiv preprint arXiv:2406.09750*, 2024c.

Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8640–8650, 2024d.

Zongming Li, Tianheng Cheng, Shoufa Chen, Peize Sun, Haocheng Shen, Longjin Ran, Xiaoxin Chen, Wenyu Liu, and Xinggang Wang. Controlar: Controllable image generation with autoregressive models. *arXiv preprint arXiv:2410.02705*, 2024e.

Chao Liang, Fan Ma, Linchao Zhu, Yingying Deng, and Yi Yang. Caphuman: Capture your moments in parallel universes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6400–6409, 2024.

Zhuoyan Luo, Fengyuan Shi, Yixiao Ge, Yujiu Yang, Limin Wang, and Ying Shan. Open-magvit2: An open-source project toward democratizing auto-regressive visual generation. *arXiv preprint arXiv:2409.04410*, 2024.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, pp. 4296–4304, 2024.

Yotam Nitzan, Kfir Aberman, Qiurui He, Orly Liba, Michal Yarom, Yossi Gandelsman, Inbar Mosseri, Yael Pritch, and Daniel Cohen-Or. Mystyle: A personalized generative prior. *ACM Transactions on Graphics (TOG)*, 41(6):1–10, 2022.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Ziqi Pang, Tianyuan Zhang, Fujun Luan, Yunze Man, Hao Tan, Kai Zhang, William T Freeman, and Yu-Xiong Wang. Randar: Decoder-only autoregressive visual generation in random orders. *arXiv preprint arXiv:2412.01827*, 2024.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.

Sucheng Ren, Qihang Yu, Ju He, Xiaohui Shen, Alan Yuille, and Liang-Chieh Chen. Beyond next-token: Next-x prediction for autoregressive visual generation. *arXiv preprint arXiv:2502.20388*, 2025.

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.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22500–22510, 2023.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

Haotian Tang, Yecheng Wu, Shang Yang, Enze Xie, Junsong Chen, Junyu Chen, Zhuoyang Zhang, Han Cai, Yao Lu, and Song Han. Hart: Efficient visual generation with hybrid autoregressive transformer. *arXiv preprint arXiv:2410.10812*, 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.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024b.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. *arXiv preprint arXiv:2409.11340*, 2024.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024a.

Rongchang Xie, Chen Du, Ping Song, and Chang Liu. Muse-vl: Modeling unified vlm through semantic discrete encoding. *arXiv preprint arXiv:2411.17762*, 2024b.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Ling Yang, Xinchen Zhang, Ye Tian, Chenming Shang, Minghao Xu, Wentao Zhang, and Bin Cui. Hermesflow: Seamlessly closing the gap in multimodal understanding and generation. *arXiv preprint arXiv:2502.12148*, 2025.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Cheng Yu, Haoyu Xie, Lei Shang, Yang Liu, Jun Dan, Liefeng Bo, and Baigui Sun. Facechain-fact: Face adapter with decoupled training for identity-preserved personalization. *arXiv preprint arXiv:2410.12312*, 2024a.

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. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.

Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Randomized autoregressive visual generation. *arXiv preprint arXiv:2411.00776*, 2024b.

Qihang Yu, Mark Weber, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. An image is worth 32 tokens for reconstruction and generation. *Advances in Neural Information Processing Systems*, 37:128940–128966, 2024c.

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

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

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Xinchen Zhang, Ling Yang, Yaqi Cai, Zhaochen Yu, Kai-Ni Wang, Ye Tian, Minkai Xu, Yong Tang, Yujiu Yang, Bin Cui, et al. Realcompo: Balancing realism and compositionality improves text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 37:96963–96992, 2024a.

Xinchen Zhang, Ling Yang, Guohao Li, Yaqi Cai, Jiake Xie, Yong Tang, Yujiu Yang, Mengdi Wang, and Bin Cui. Itercomp: Iterative composition-aware feedback learning from model gallery for text-to-image generation. *arXiv preprint arXiv:2410.07171*, 2024b.

Youliang Zhang, Ronghui Li, Yachao Zhang, Liang Pan, Jingbo Wang, Yebin Liu, and Xiu Li. A plug-and-play physical motion restoration approach for in-the-wild high-difficulty motions. *arXiv preprint arXiv:2412.17377*, 2024c.

Chuanxia Zheng, Tung-Long Vuong, Jianfei Cai, and Dinh Phung. Movq: Modulating quantized vectors for high-fidelity image generation. *Advances in Neural Information Processing Systems*, 35:23412–23425, 2022.

REFERENCES

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

Weifeng Chen, Jiacheng Zhang, Jie Wu, Hefeng Wu, Xuefeng Xiao, and Liang Lin. Id-aligner: Enhancing identity-preserving text-to-image generation with reward feedback learning. *arXiv preprint arXiv:2404.15449*, 2024.

Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. Dovenet: Deep image harmonization via domain verification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8394–8403, 2020.

Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

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.

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.

Lijie Fan, Tianhong Li, Siyang Qin, Yuanzhen Li, Chen Sun, Michael Rubinstein, Deqing Sun, Kaiming He, and Yonglong Tian. Fluid: Scaling autoregressive text-to-image generative models with continuous tokens. *arXiv preprint arXiv:2410.13863*, 2024.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

Jiatao Gu, Yuyang Wang, Yizhe Zhang, Qihang Zhang, Dinghuai Zhang, Navdeep Jaitly, Josh Susskind, and Shuangfei Zhai. Dart: Denoising autoregressive transformer for scalable text-to-image generation. *arXiv preprint arXiv:2410.08159*, 2024.

Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. *arXiv preprint arXiv:2412.04431*, 2024.

Junjie He, Yifeng Geng, and Liefeng Bo. Uniportrait: A unified framework for identity-preserving single-and multi-human image personalization. *arXiv preprint arXiv:2408.05939*, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

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.

Jiehui Huang, Xiao Dong, Wenhui Song, Zheng Chong, Zhenchao Tang, Jun Zhou, Yuhao Cheng, Long Chen, Hanhui Li, Yiqiang Yan, et al. Consistentid: Portrait generation with multimodal fine-grained identity preserving. *arXiv preprint arXiv:2404.16771*, 2024.

Mengqi Huang, Zhendong Mao, Zhuowei Chen, and Yongdong Zhang. Towards accurate image coding: Improved autoregressive image generation with dynamic vector quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22596–22605, 2023.

Minchul Kim, Anil K Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18750–18759, 2022.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11523–11532, 2022.

Haopeng Li, Jinyue Yang, Kexin Wang, Xuerui Qiu, Yuhong Chou, Xin Li, and Guoqi Li. Scalable autoregressive image generation with mamba. *arXiv preprint arXiv:2408.12245*, 2024a.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems*, 37:56424–56445, 2024b.

Xiang Li, Kai Qiu, Hao Chen, Jason Kuen, Zhe Lin, Rita Singh, and Bhiksha Raj. Controlvar: Exploring controllable visual autoregressive modeling. *arXiv preprint arXiv:2406.09750*, 2024c.

Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8640–8650, 2024d.

Zongming Li, Tianheng Cheng, Shoufa Chen, Peize Sun, Haocheng Shen, Longjin Ran, Xiaoxin Chen, Wenyu Liu, and Xinggang Wang. Controlar: Controllable image generation with autoregressive models. *arXiv preprint arXiv:2410.02705*, 2024e.

Chao Liang, Fan Ma, Linchao Zhu, Yingying Deng, and Yi Yang. Caphuman: Capture your moments in parallel universes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6400–6409, 2024.

Zhuoyan Luo, Fengyuan Shi, Yixiao Ge, Yujiu Yang, Limin Wang, and Ying Shan. Open-magvit2: An open-source project toward democratizing auto-regressive visual generation. *arXiv preprint arXiv:2409.04410*, 2024.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, pp. 4296–4304, 2024.

Yotam Nitzan, Kfir Aberman, Qiurui He, Orly Liba, Michal Yarom, Yossi Gandelsman, Inbar Mosseri, Yael Pritch, and Daniel Cohen-Or. Mystyle: A personalized generative prior. *ACM Transactions on Graphics (TOG)*, 41(6):1–10, 2022.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Ziqi Pang, Tianyuan Zhang, Fujun Luan, Yunze Man, Hao Tan, Kai Zhang, William T Freeman, and Yu-Xiong Wang. Randar: Decoder-only autoregressive visual generation in random orders. *arXiv preprint arXiv:2412.01827*, 2024.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.

Sucheng Ren, Qihang Yu, Ju He, Xiaohui Shen, Alan Yuille, and Liang-Chieh Chen. Beyond next-token: Next-x prediction for autoregressive visual generation. *arXiv preprint arXiv:2502.20388*, 2025.

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.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22500–22510, 2023.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

Haotian Tang, Yecheng Wu, Shang Yang, Enze Xie, Junsong Chen, Junyu Chen, Zhuoyang Zhang, Han Cai, Yao Lu, and Song Han. Hart: Efficient visual generation with hybrid autoregressive transformer. *arXiv preprint arXiv:2410.10812*, 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.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024b.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. *arXiv preprint arXiv:2409.11340*, 2024.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024a.

Rongchang Xie, Chen Du, Ping Song, and Chang Liu. Muse-vl: Modeling unified vlm through semantic discrete encoding. *arXiv preprint arXiv:2411.17762*, 2024b.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Ling Yang, Xinchen Zhang, Ye Tian, Chenming Shang, Minghao Xu, Wentao Zhang, and Bin Cui. Hermesflow: Seamlessly closing the gap in multimodal understanding and generation. *arXiv preprint arXiv:2502.12148*, 2025.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Cheng Yu, Haoyu Xie, Lei Shang, Yang Liu, Jun Dan, Liefeng Bo, and Baigui Sun. Facechain-fact: Face adapter with decoupled training for identity-preserved personalization. *arXiv preprint arXiv:2410.12312*, 2024a.

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. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.

Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Randomized autoregressive visual generation. *arXiv preprint arXiv:2411.00776*, 2024b.

Qihang Yu, Mark Weber, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. An image is worth 32 tokens for reconstruction and generation. *Advances in Neural Information Processing Systems*, 37: 128940–128966, 2024c.

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

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

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Xinchen Zhang, Ling Yang, Yaqi Cai, Zhaochen Yu, Kai-Ni Wang, Ye Tian, Minkai Xu, Yong Tang, Yujiu Yang, Bin Cui, et al. Realcompo: Balancing realism and compositionality improves text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 37:96963–96992, 2024a.

Xinchen Zhang, Ling Yang, Guohao Li, Yaqi Cai, Jiake Xie, Yong Tang, Yujiu Yang, Mengdi Wang, and Bin Cui. Itercomp: Iterative composition-aware feedback learning from model gallery for text-to-image generation. *arXiv preprint arXiv:2410.07171*, 2024b.

Youliang Zhang, Ronghui Li, Yachao Zhang, Liang Pan, Jingbo Wang, Yebin Liu, and Xiu Li. A plug-and-play physical motion restoration approach for in-the-wild high-difficulty motions. *arXiv preprint arXiv:2412.17377*, 2024c.

Chuanxia Zheng, Tung-Long Vuong, Jianfei Cai, and Dinh Phung. Movq: Modulating quantized vectors for high-fidelity image generation. *Advances in Neural Information Processing Systems*, 35:23412–23425, 2022.

This supplementary material is organized into several sections, each offering additional details and analysis related to HEAR. The topics covered include:

- In Appendix A, we provide a preliminary about next-scale prediction and Infinity.
- In Appendix B, we provide more results of HEAR including ablation study, visualized comparisons of high-frequency encoder and high-quality visualization.
- In Appendix C, we provide the prompts used in the ID dataset curation pipeline in Fig. 3.

A PRELIMINARY

A.1 NEXT-SCALE PREDICTION

VAR (Tian et al., 2024) reconceptualizes autoregressive modeling for images by shifting from a next-token prediction strategy to a next-scale prediction strategy. In this formulation, the basic autoregressive unit is an entire token map rather than a single token. We begin by quantizing a feature map $\mathbf{f} \in \mathbb{R}^{h \times w \times C}$ into K multi-scale token maps $(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_K)$, each corresponding to progressively higher resolutions $h_k \times w_k$, where $\mathbf{r}_K$ matches the original feature map's resolution $h \times w$. The autoregressive likelihood is then defined as:

$$p(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_K) = \prod_{k=1}^{K} p(\mathbf{r}_k \mid \mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_{k-1}), \tag{8}$$

where each autoregressive unit $\mathbf{r}_k \in [V]^{h_k \times w_k}$ represents the token map at scale k, consisting of $h_k \times w_k$ tokens. The sequence $(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_{k-1})$ serves as the prefix for predicting $\mathbf{r}k$. At the k-th autoregressive step, the distributions over all $h_k \times w_k$ tokens are generated in parallel, conditioned on $\mathbf{r}_k$'s prefix and the corresponding k-th position embedding map. During training, a block-wise causal attention mask is applied to ensure that each $\mathbf{r}_k$ only attends to its prefix $\mathbf{r}_{\leq k}$. During inference, key-value caching can be used and no masking is required.

A.2 AUTOREGRESSIVE TEXT-TO-IMAGE GENERATION: INFINITY

Infinity (Han et al., 2024) introduces a bitwise visual autoregressive framework that significantly enhances high-resolution image synthesis through two core innovations:

Infinite-Vocabulary Tokenizer & Classifier Employing LFQ or BSQ, infinity quantizes residual features R_k into binary bit sequences through dimension-independent encoding, theoretically scaling vocabulary size to 2^{32} or 2^{64}. Given K scales in the multi-scale quantizer, at the k-th scale, the input continuous residual vector $z_k \in \mathbb{R}^d$ is quantized into a binary output q_k as illustrated below:

$$q_k = \mathcal{Q}(z_k) = \begin{cases} \text{sign}(z_k) & \text{if LFQ} \\ \frac{1}{\sqrt{d}}\text{sign}(\frac{z_k}{|z_k|}) & \text{if BSQ} \end{cases} \tag{9}$$

The Infinite-Vocabulary Classifier (IVC) decomposes traditional V_d class prediction into d parallel binary classifiers, reducing classifier parameters by 99.95% while maintaining exponential vocabulary capacity. The index label $\mathbf{y}_k(m, n)$ is obtained by multiplying the positive elements with their corresponding bases and summing the results:

$$\mathbf{y}_k(m, n) = \sum_{p=0}^{d-1} \mathbb{I}_{\mathbf{R}_k(m,n,p)>0} \cdot 2^p \tag{10}$$

where $m \in [0, h_k)$ and $n \in [0, w_k)$. The next-scale residual $\mathbf{R}_k(m, n, p)$ is predicted to be positive or negative by d binary classifiers operating in parallel.

Bitwise Self-Correction During training, random bit-flipping ($p \in [0, 30\%]$) on R_k generates perturbed features R_k^{flip} , followed by re-quantization of subsequent residuals. This forces the model to learn to correct its own errors, effectively reducing error accumulation in teacher-forcing training:

$$R_k^{\text{flip}} = \text{Random\_Flip}(R_k, p) \tag{11}$$

Table 3: Quantitative comparative analysis of Global Cross-Attention and Local High-Frequency Cross-Attention across multiple evaluation metrics.

Model	CLIP-T(↑)	CLIP-I(↑)	DINO(↑)	FaceSim(↑)
w/o Local High-frequency Cross-attention	30.5	70.4	77.2	52.8
w/o Global Cross-attention	28.6	69.3	74.1	50.2
HEAR	**32.6**	**72.1**	**78.9**	**56.3**

$$F_k^{\text{flip}} = \sum_{i=1}^{k} \text{up}(R_i^{\text{flip}}, (h, w)) \tag{12}$$

where Random_Flip is uniformly sampled from the interval $[0, p]$ to simulate varying levels of prediction errors at the k-th scale and obtain R_k^{flip} by randomly flipping the bits in R_k.

B More Results

B.1 Ablation Study

B.1.1 Comparison of Global Cross-Attention and Local High-Frequency Cross-Attention

As shown in the table 3, the absence of either module leads to a noticeable decline across multiple evaluation metrics. Combined with Figure 8 in Supplementary Material, we can confidently conclude that our modules effectively contribute to both the preservation of global features and the enhancement of high-frequency details.

In addition, we provide several other ablation studies to further ensure the rigor and validity of our experimental results.

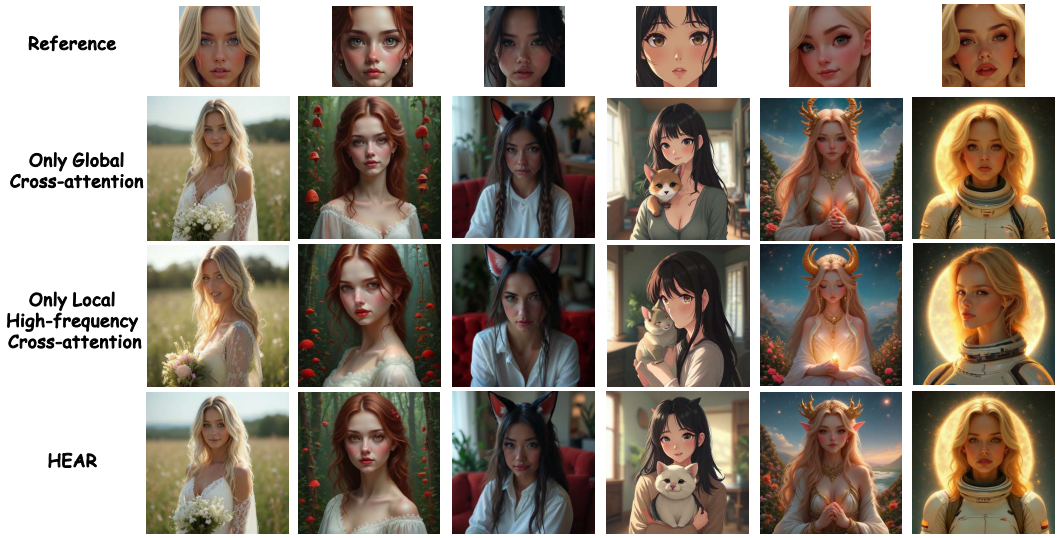

Figure 7: **Ablation study on the two key components of HEAR.** We compare global cross-attention and local high-frequency cross-attention. The results clearly demonstrate that HEAR significantly outperforms High-frequency Enhancement only in terms of global control, while also achieving superior high-frequency detail preservation compared to Global Control only.

To validate the effectiveness of the dual-control strategy, we also compare it against two ablated variants: one with only global cross-attention and another with only local high-frequency cross-attention. In the global-only setting, image features are concatenated with text features and injected with global information via cross-attention blocks. In the local high-frequency-only setting, only the high-frequency components of image features are retained and injected via a decoupled adapter. For a fair comparison, both adapters are trained under the same configuration for 500K steps. Figure 7

Table 4: Effect of Reconstruction Losses (RLs), Vector Quantization Loss (VQL), Perceptual Loss (PL), CLIP Loss (CL), and AdaFace Loss (AL) as measured by automatic metrics.

Method	MSE(↓)	SSIM(↑)	PSNR(↑)
w/o VQL	21.87	0.93	34.73
w/o PL	14.57	0.94	34.82
w/o CL	14.05	0.95	35.26
w/o AL	13.98	0.95	35.28
HEAR	**13.87**	**0.95**	**35.32**

presents qualitative examples comparing HEAR with the two ablated baselines. As shown, HEAR's dual-control strategy enables it to not only preserve global facial identity information effectively, but also excel in capturing fine-grained local high-frequency details.

B.2 Validity of Motivation and High-Frequency Enhancement

Figure 8: **Visualization of HEAR's generation process across different scales.** This illustrates how HEAR controls the generation of image identities, particularly in terms of high-frequency variations.

As shown in Figure 8, we visualize the generation process of HEAR across different scales. It can be observed that when the process reaches the stage of high-frequency detail generation (specifically after Index 6 in the figure), each scale's residual summation results in progressively greater similarity to the reference image. This indicates that every high-frequency control adapter contributes meaningfully to the generation. Moreover, through the accumulation of incremental changes, a qualitative transformation is eventually achieved.

B.3 Ablation study and visualized comparisons of Reconstruction Fidelity

As illustrated in the table 4, the exclusion of any individual loss component results in a noticeable increase in MSE and a decrease in SSIM and PSNR scores. The relatively smaller impact of the Adaface loss may be attributed to its specific role in enhancing the VAE's facial reconstruction capability by computing facial similarity loss. Both the perceptual loss and CLIP loss compute image-level similarity to enhance the encoder's reconstruction capability, and their impact is slightly greater than that of the AdaFace loss.

Thank you for your feedback. We will incorporate ablation studies into the revised version of the paper.

To visually demonstrate the reconstruction fidelity of the high-frequency face encoder, we conducted visualization experiments using several tokenizers, including Open-MAGVIT2 Luo et al. (2024), LlamaGen Sun et al. (2024), Show-o Xie et al. (2024a), and Infinity Han et al. (2024). Fig. 9 presents side-by-side comparisons of reconstruction results from all evaluated methods, using reference images from two distinct identities. The visual evidence highlights high-frequency

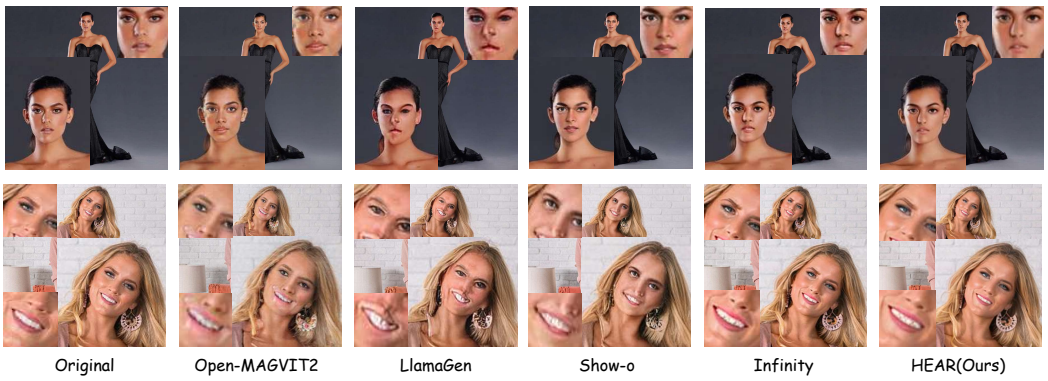

Figure 9: **Reconstruction Comparison Across Different Visual Tokenizers.** The proposed high-frequency visual tokenizer consistently outperforms existing approaches in both pixel-level accuracy and perceptual fidelity.

face encoder's ability to produce highly accurate facial reconstructions, preserving both color consistency and fine-grained facial morphology. To further validate these observations, we magnified and compared specific facial features across both identities. By leveraging fine-grained multimodal cues and identity-specific details within key regions, particularly the eyes, nose, and lips, our model exhibits exceptional identity preservation. These results emphasize high-frequency face encoder's superiority in maintaining anatomical precision and textural authenticity, clearly surpassing existing methods in visual fidelity.

B.3.1 Exploring HEAR's Potential Across Diverse ID Types and Prompt Variations

We selected a wide range of identity (ID) types without restricting to any single category as controlled reference images, and used a diverse set of prompts to provide textual guidance. We aimed to evaluate HEAR's ability to preserve facial identity, follow prompt instructions, and maintain aesthetic quality across various scenarios. As illustrated in Fig. 10 and 11, the results were truly remarkable: HEAR consistently retained the facial ID while satisfying most prompt requirements, and demonstrated exceptionally high aesthetic quality.

C Prompts Used in the Identity-Preserving Dataset Curation Pipeline

To provide a clearer understanding of our Identity-Preserving Dataset Curation Pipeline, we present the complete prompts that were partially omitted in Figure 3 as follow.

Prompt of Image Preprocessing

Your Role: Facial Recognition Analyst
Objective: Determine if the subject's face in the provided image meets frontal orientation and clarity requirements by verifying some specific criteria.
Process Steps:

1. Measure the face's side/profile angles to ensure they do not exceed 30 degrees.
2. Verify eyes, nose, and mouth are mostly visible (e.g., no hair, accessories, or hands blocking features). Ensure no reflections (e.g., glasses) obscure key areas.
3. Evaluate Image Clarity: Look for blurring, pixelation, or compression artifacts affecting facial details. Check resolution quality (e.g., edges of facial features must be sharp).
4. Ensure even illumination across the face (no shadows over eyes/nose/mouth). Confirm no overexposure or underexposure distorting features.

Examples:
- Example 1 (Non-Compliant):

  [leftmargin=2em]

  User Prompt: Now you are a "Facial Recognition Analyst". Carefully examine the provided image and ...

  Reasoning: Face has a 45-degree side angle. Left eye obscured by hair. Blurring around the mouth. Harsh shadows under the nose.

  Objects: [("face", ["45° side angle", "obscured eye", "blurring", "harsh shadows"])]

  Negation: True

- Example 2 (Compliant):

  [leftmargin=2em]

  User Prompt: Now you are a "Facial Recognition Analyst". Carefully examine the provided image and ...

  Reasoning: Full frontal view (5° tilt). All features visible. Sharp resolution. Even lighting with no shadows.

  Objects: [("face", [None, None, None, None])]

  Negation: False

Your Current Task:
Follow the Process Steps to analyze the provided image. Present results in the format:

[leftmargin=2em]

Reasoning: Detailed analysis of each criterion.

Objects: List attributes (or None if compliant).

Negation: Boolean (True/False) indicating compliance.

User prompt: {Now you are a "Facial Recognition Analyst". Carefully examine the provided image and determine whether the subject's face meets frontal orientation and clarity requirements by verifying the following criteria: Full frontal view with no side/profile angles exceeding 30 degrees, all facial features (eyes, nose, mouth) fully visible and unobstructed, no significant blurring or low-resolution artifacts, and proper illumination without harsh shadows obscuring features. }

Prompt of Synthetic Data Augmentation

Your Role: Textual Prompt Refiner
Objective: Enhance human-centric image generation prompts by diversity injection through 3-4 specific physical attributes.
Process Steps:

1. Inject descriptors like "high nasal root typical of East Asian ancestry" or "broad alar base common in West African phenotypes".
2. Add gender-fluid traits (e.g., "softened mandibular angle with stubble shadow").
3. Specify hormonal influences (e.g., "post-adolescent acne scarring along the jawline").
4. Map muscle movements to emotional states: "subtle crow's feet from frequent laughter" or "vertical glabellar lines indicating chronic focus".
5. Use dynamic modifiers (e.g., "partially contracted corrugator supercilii muscles").
6. Embed region-specific adornments (e.g., "Maasai beadwork collar", "Balinese temple ear cuffs").
7. Reference symbolic body modifications (e.g., "Yakuza-inspired fingertip tattoos").

Examples
- Example 1 (Original $\rightarrow$ Enhanced):

  [leftmargin=2em]

  Input Prompt: "A young woman smiling."

  Enhanced Prompt: "A Southeast Asian woman in her 20s with epicanthic folds and a low nasal bridge, displaying asymmetrical nasolabial folds from a half-suppressed grin. Traditional sihn skirt drapes over her knees, complemented by a sak yant tattoo peeking above her collarbone."

- Example 2 (Original $\rightarrow$ Enhanced):

  [leftmargin=2em]

  Input Prompt: "An old man with a beard."

  Enhanced Prompt: "A Kurdish man in his late 60s with salt-and-pepper şal û şapik mustache, deep nasojugial grooves from decades of squinting in sunlight, and deq facial tattoos fading into sagging jowls." Objects: [("face", ["ethnic wrinkles", "cultural facial hair", "tribal ink"])]

Your Current Task:
Building upon the original input prompt, we enhance human-centric image generation by injecting descriptive attributes, adding gender-fluid traits, specifying hormonal influences, mapping muscle movements to emotional states, using dynamic modifiers, embedding region-specific adornments, and referencing symbolic body modifications.
User prompt: {Now you are a "Textual Prompt Refiner" specializing in expanding facial details for human-centric image generation. For each input prompt, first perform diversity injection by adding 3-4 specific physical attributes including racial features (like epicanthic folds or nose bridge height), gender characteristics (such as androgynous jawline or beard density), micro-expressions (for example nasolabial folds when smiling), and cultural accessories (like tribal scarification patterns). }

Prompt of Image Recaption

Your Role: Image Caption Creator
Objective: Enhance facial image prompts by generating a dual-description structure that transitions from concise demographic-action-environment tagging to richly detailed, photo-realistic portrayals emphasizing expression, clothing, and spatial context.
Process Steps:

1. **Short-Form Prompt Construction**: "[Gender] [Ethnicity] [Action] in [Environment]" (e.g., "Middle-aged Hispanic man adjusting tie in a city street at dusk")
2. **Long-Form Prompt Expansion**:
    - Facial Expression Expansion: Describe micro-expressions and gaze direction with anatomical precision. (e.g.,"slight levator labii contraction from a half-smirk," "glance directed 30° rightward," "orbicularis oculi tension from narrowed gaze.")
    - Clothing Detailing: Highlight fabric type, texture, and interaction with light. (e.g.,"wool-blend blazer catching low-angle sunlight," "crimson silk scarf loosely looped at clavicle.")
    - Spatial Elements: Specify physical orientation and spatial relationships with the surrounding environment. (e.g.,"leaning 15° toward a reflective window 1 meter to his left," "background softened by a 3-meter depth of field with blurred pedestrians.")
    - Photorealism Enforcement: Include optical effects that simulate real-world imaging conditions. (e.g.,"light scattering," "depth blur," "lens distortion.")

Examples

- Short Prompt: "Young East Asian woman sipping coffee in a sunlit cafe corner."
- Long Prompt: "A young East Asian woman sipping from a white ceramic mug, with a soft smile indicated by gentle zygomaticus major activation and a downward gaze angled 45°, showing mild orbicularis oculi engagement. She wears a cream cashmere sweater with visible ribbing, catching warm sunlight across the sleeves, and a loosely tied silk scarf with floral tones at her collarbone. Her posture leans 10° forward, right elbow resting on a polished wooden table; a frosted window 80 cm to her left reflects ambient light. The background fades with a 2-meter depth blur, and subtle lens bloom and light scattering reinforce the scene's photorealism."

Your Current Task: Generate both a short and a detailed caption for the following image using a full attribute structure. Present the results in the following format:

- Short Prompt: "[Gender] [Ethnicity] [Action] in [Environment]"
- Long Prompt: Detailed expansion based on the short prompt, covering facial expression, clothing, spatial elements and photorealism enforcemen.

User prompt: {Now you are a "Image Caption Creator". Generate a dual-prompt description for the input facial image. First, create a short prompt stating the subject's gender, ethnicity, action, and surrounding environment in a concise format (e.g., '[Gender] [Ethnicity] [Action] in [Environment]'). Then, expand this into a long prompt by adding: 1) Detailed facial expressions (e.g., micro-expressions, eye direction, muscle tension), 2) Clothing specifics (textures, colors, material interactions), and 3) Spatial relationships between the subject and objects or environment (quantify distances and angles where applicable). Ensure both prompts maintain photographic realism and avoid artistic stylization. }

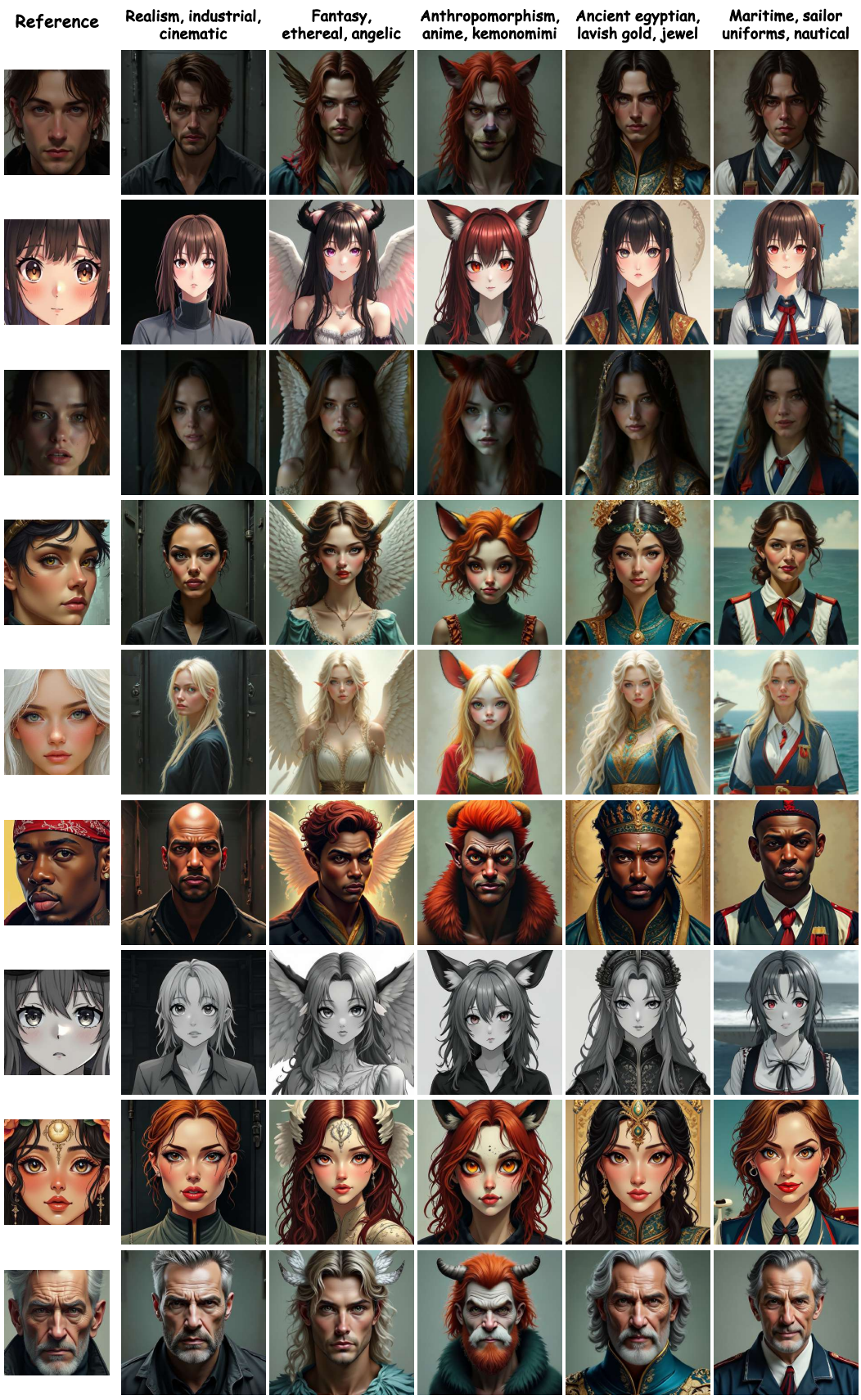

Figure 10: **Remarkable capability of HEAR to preserve individual identity while achieving high visual fidelity.** Our method consistently retains identity-specific features across a wide range of conditions, including diverse artistic styles, age groups, and skin tones.

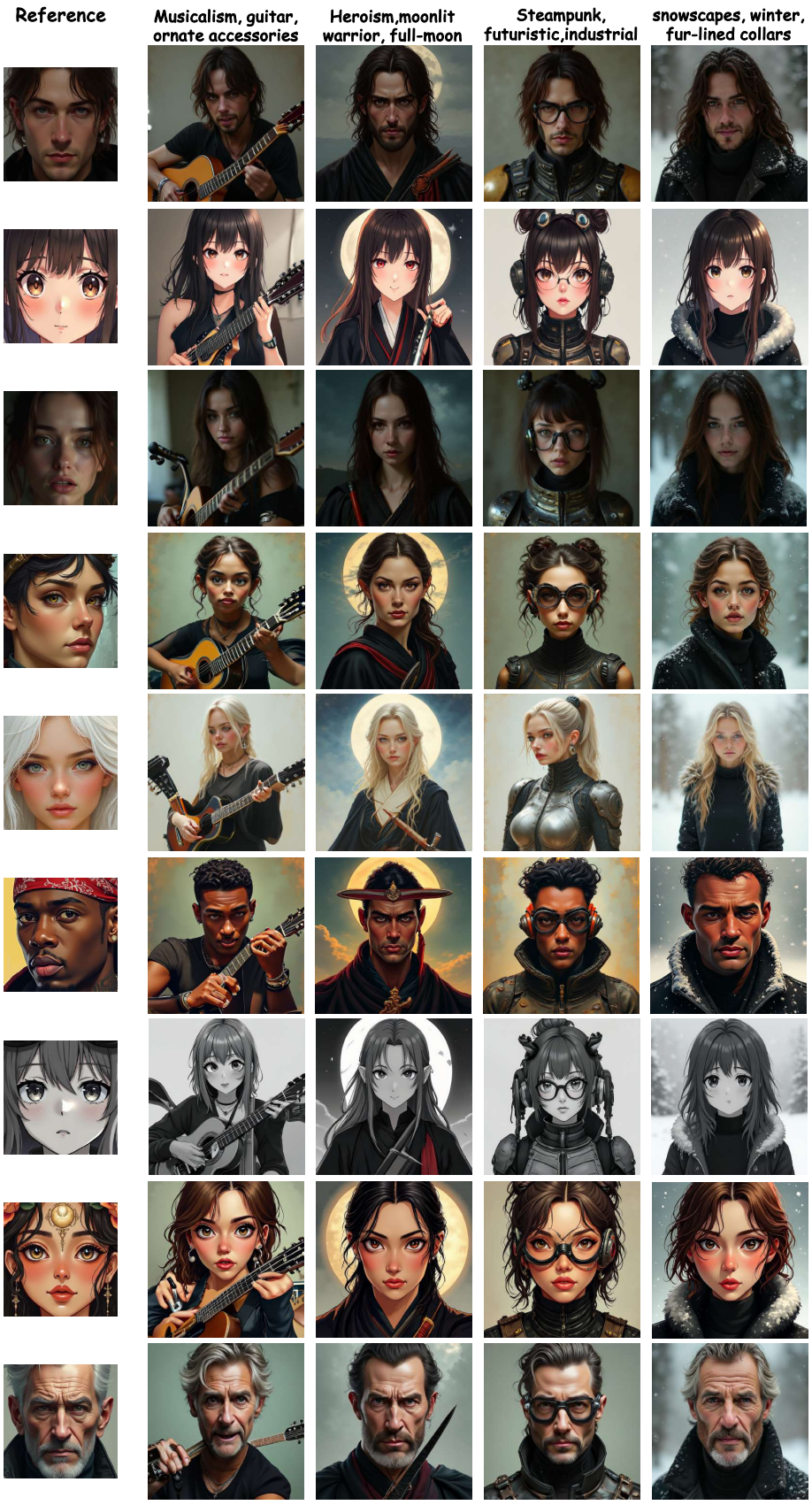

Figure 11: **More Results.**