

# Boosting Latent Diffusion with Flow Matching

Johannes S. Fischer\*    Ming Gui\*    Pingchuan Ma\*  
 Nick Stracke    Stefan A. Baumann    Björn Ommer

CompVis @ LMU Munich, MCML

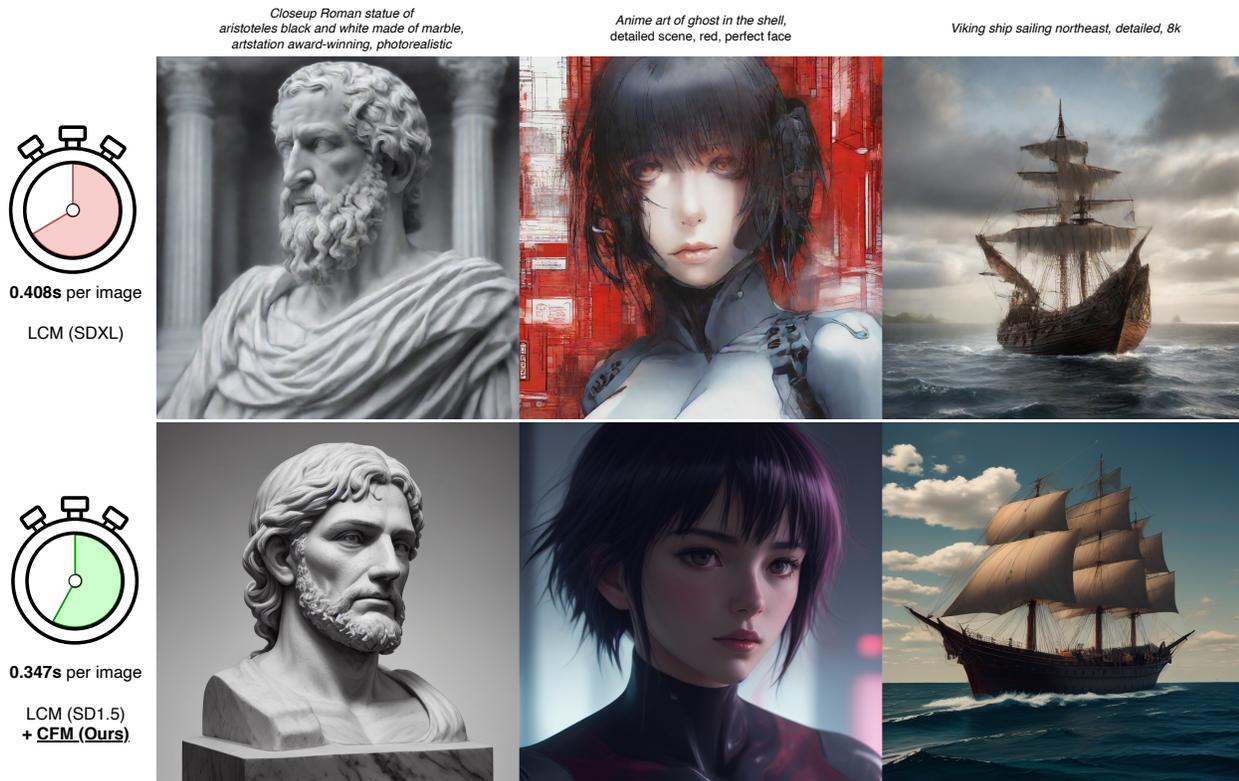


Figure 1. **Samples synthesized in  $1024^2$  px.** We elevate Diffusion Models (DMs) and similar architectures to a higher-resolution domain, achieving exceptionally rapid processing speeds. We use Latent Consistency Models (LCM) [43], distilled from SD1.5 [53] and SDXL [49], respectively. To achieve the same resolution as LCM-SDXL, we boost LCM-SD1.5 with our Coupling Flow Matching (CFM) model. The LCM-SDXL model fails to produce competitive results within this shortened timeframe, highlighting the effectiveness of our approach in achieving both speed and quality in image synthesis.

## Abstract

Visual synthesis has recently seen significant leaps in performance, largely due to breakthroughs in generative models. Diffusion models have been a key enabler, as they excel in image diversity. However, this comes at the cost of slow training and synthesis, which is only partially alleviated by latent diffusion. To this end, flow matching is an ap-

pealing approach due to its complementary characteristics of faster training and inference but less diverse synthesis. We demonstrate that introducing flow matching between a frozen diffusion model and a convolutional decoder enables high-resolution image synthesis at reduced computational cost and model size. A small diffusion model can then effectively provide the necessary visual diversity, while flow matching efficiently enhances resolution and detail by mapping the small to a high-dimensional latent space. These latents are then projected to high-resolution images by the

\*Equal Contribution

*subsequent convolutional decoder of the latent diffusion approach. Combining the diversity of diffusion models, the efficiency of flow matching, and the effectiveness of convolutional decoders, state-of-the-art high-resolution image synthesis is achieved at  $1024^2$  pixels with minimal computational cost. Further scaling up our method we can reach resolutions up to  $2048^2$  pixels. Importantly, our approach is orthogonal to recent approximation and speed-up strategies for the underlying model, making it easily integrable into the various diffusion model frameworks. Project page and code are available at <https://compvis.github.io/fm-boosting>.*

## 1. Introduction

Visual synthesis has recently witnessed unprecedented progress and popularity in computer vision and beyond. Various generative models have been proposed to address the diverse challenges in this field [71], including sample diversity, quality, resolution, training, and test speed. Among these approaches, diffusion models (DMs) [52, 53, 55] currently rank among the most popular and highest quality, defining the state of the art in numerous synthesis applications. While DMs excel in sample quality and diversity, they face challenges in high-resolution synthesis, slow sampling speed, and a substantial memory footprint.

Lately, numerous efficiency improvements to DMs have been proposed [12, 51, 62], but the most popular remedy has been the introduction of Latent Diffusion Models (LDMs) [53]. Operating only in a compact latent space, LDMs combine the strengths of DMs with the efficiency of a convolutional encoder-decoder that translates the latents back into pixel space. However, Rombach et al. [53] also showed that an excessively strong first-stage compression leads to information loss, limiting generation quality. Efforts have been made to expand the latent space [49] or stack a series of different DMs, each specializing in different resolutions [20, 55]. However, these approaches are still computationally costly, especially when synthesizing high-resolution images.

The inherent stochasticity of DMs is key to their proficiency in generating diverse images. In the later stages of DM inference, as the global structure of the image has already been generated, the advantages of stochasticity diminish. Instead, the computational overhead due to the less efficient stochastic diffusion trajectories becomes a burden rather than helping in up-sampling to and improving higher resolution images [8]. At this stage, converse characteristics become beneficial: reduced diversity and a short and straight trajectory toward the high-resolution latent space of the decoder. These goals align precisely with the strengths of Flow Matching (FM) [5, 36, 39], another emerging family of generative models currently gaining significant atten-

tion. In contrast to DMs, Flow Matching enables the modeling of an optimal transport conditional probability path between two distributions that is significantly straighter than those achieved by DMs, making it more robust, and efficient to train. The deterministic nature of Flow Matching models also allows the utilization of off-the-shelf Ordinary Differential Equation (ODE) solvers, which are more efficient to sample from and can further accelerate inference.

We leverage the complementary strengths of DMs, FMs, and VAEs: the diversity of stochastic DMs, the speed of Flow Matching in training and inference stages, and the efficiency of a convolutional decoder when mapping latents into pixel space. This synergy results in a small diffusion model that excels in generating diverse samples at a low resolution. Flow Matching then takes a direct path from this lower-resolution representation to a higher-resolution latent, which is subsequently translated into a high-resolution image by a convolutional decoder. Moreover, the Flow Matching model can establish data-dependent couplings with the synthesized information from the DM, which automatically and inherently forms optimal transport paths from the noise to the data samples in the Flow Matching model [6, 65].

Note that our work is complementary to recent work on sampling acceleration of diffusion models like DDIM [62], DPM-Solver [41], and LCM-LoRA [42, 43]. Our approach can be directly integrated into any existing DMs architecture to increase the final output resolution efficiently.

## 2. Related Work

**Diffusion Models** Diffusion models [19, 61, 63] have shown broad applications in computer vision, spanning image [53], audio [37], and video [9, 21]. Albeit with high fidelity in generation, they do so at the cost of sampling speed compared to alternatives like Generative Adversarial Networks [16, 24, 26]. Hence, several works propose more efficient sampling techniques for diffusion models, including distillation [44, 57, 64], noise schedule design [28, 47, 50], and training-free sampling [27, 38, 40, 62]. Nonetheless, it is important to highlight that existing methods have not fully addressed the challenge imposed by the strong curvature in the sampling trajectory [31], which limits sampling step sizes and necessitates the utilization of intricately tuned solvers, making sampling costly.

**Flow Matching-based Generative Models** A recent competitor, known as Flow Matching [4, 36, 39, 45], has gained prominence for its ability to maintain straight trajectories during generation by modeling the synthesis process using an optimal transport conditional probability path with Ordinary Differential Equations (ODE), positioning it as an apt alternative for addressing trajectory straightness-related issues encountered in diffusion models. The versatility of Flow Matching has been showcased across various do-

mains, including image [13, 22, 36], video [7], audio [29]. This underscores its capacity to address the inherent trajectory challenges associated with diffusion models, mitigating the limitations of slow sampling in the current generation based on diffusion models. Considerable effort has been directed towards optimizing transport within Flow Matching models [39, 65], which contributes to enhanced training stability and accelerated inference speed by making the trajectories even straighter and thus enabling larger sampling step sizes. However, the generation capabilities of Flow Matching presently do not parallel those of diffusion models [13, 36]. We remedy this limitation by using a small diffusion model for synthesis quality.

**Image Super-Resolution** Image super-resolution (SR) is a fundamental problem in computer vision. Prominent methodologies include GANs [24, 30, 67, 74], diffusion models [32, 56, 73] and Flow Matching methods [6, 36].

Our methodology adopts the Flow Matching approaches, leveraging its objective to achieve faster training and inference compared to diffusion models. We take inspiration from latent diffusion models [53] and transition the training to the latent space, which further enhances computational efficiency. This enables the synthesis of images with significantly higher resolution, thereby advancing the capacity for image generation in terms of both speed and output resolution.

### 3. Method

We speed up and increase the resolution of existing LDMs by integrating Flow Matching in the latent space. The proposed architecture should not be limited to unconditional image synthesis but also be applicable to text-to-image synthesis [46, 52, 53, 55] and Diffusion models with other conditioning including depth maps, canny edges, etc. [15, 33, 75]. The main challenge

is not a deficiency in diversity within the Diffusion model; rather it is the slow convergence of the training procedure, the huge memory demand, and the slow inference [49, 55, 72].

While there are substantial efforts to accelerate inference speed of DMs either by distillation techniques [44], or by an ODE approximation at inference [40, 41, 62], we argue that we can achieve faster training and inference speed by training with an ODE assumption [36].

Flows characterized by straight paths without Wiener process inherently incur minimal time-discretization errors during numerical simulation [39] and can be simulated with only few ODE solver steps.

We employ a compact Diffusion model and a Flow Matching model aimed at high-resolution image generation (Sec. 3.1, Sec. 3.2). The combination of both models (Sec. 3.3) ensures efficient and detailed image generation.

### 3.1. From LDM to FM-LDM

Diffusion Models (DMs) [19] are generative models that learn a data distribution  $p(x)$  by learning to denoise noisy samples. During inference, they generate samples in a multi-step denoising process starting from Gaussian noise. Their inherent stochasticity allows them to effectively approximate the data manifold with high diversity, even in high-dimensional complex data domains such as images [47, 55] or videos [9, 21, 59], but makes generation inefficient, requiring many denoising steps at the data resolution. This problem has previously partially been addressed by *Latent Diffusion Models* (LDMs), which move the diffusion process to an autoencoder latent space, but efficiency is still a problem. While diffusion models’ stochasticity helps them generate high-quality samples, we propose that this stochasticity is not needed for later stages of generation and that the diffusion generation process can be separated into two parts without substantial loss in quality: one diffusion-based low-resolution stage for generating image semantics with high variation and a light-weight high-resolution stage with reduced stochasticity.

Recently, the formulation of generative processes as optimal transport conditional probability paths has gained much attraction [4, 36, 65], perfectly suiting this task of modeling straight trajectories between two distributions.

### 3.2. Flow Matching

Flow Matching models are generative models that regress vector fields based on fixed conditional probability paths. Let  $\mathbb{R}^d$  be the data space with data points  $x$ . Let  $u_t(x) : [0, 1] \times \mathbb{R}^d \rightarrow \mathbb{R}^d$  be the time-dependent vector field, which defines the ODE in the form of  $dx = u_t(x)dt$ , and let  $\phi_t(x)$  denote the solution to this ODE with the initial condition  $\phi_0(x) = x$ .

The probability density path  $p_t : [0, 1] \times \mathbb{R}^d \rightarrow \mathbb{R}_{>0}$  depicts the probability distribution of  $x$  at timestep  $t$  with  $\int p_t(x)dx = 1$ . The pushforward function  $p_t = [\phi_t]_{\#}(p_0)$  then transports the probability density path  $p$  along  $u$  from timestep 0 to  $t$ . Assuming that  $p_t(x)$  and  $u_t(x)$  are known, and the vector field  $u_t(x)$  generates  $p_t(x)$ , we can regress a vector field  $v_\theta(t, x)$  parameterized by a neural network with learnable parameters  $\theta$  using the Flow Matching objective

$$\mathcal{L}_{FM}(\theta) = \mathbb{E}_{t, p_t(x)} \|v_\theta(t, x) - u_t(x)\|. \quad (1)$$

While we generally do not have access to a closed form of  $u_t$  because this objective is intractable, Lipman et al. [36] showed that we can acquire the same gradients and therefore efficiently regress the neural network using the coupling Flow Matching (CFM) objective, where we can compute  $u_t(x|z)$  by efficiently sampling  $p_t(x|z)$ ,

$$\mathcal{L}_{CFM}(\theta) = \mathbb{E}_{t, q(z), p_t(x|z)} \|v_\theta(t, x) - u_t(x|z)\|, \quad (2)$$

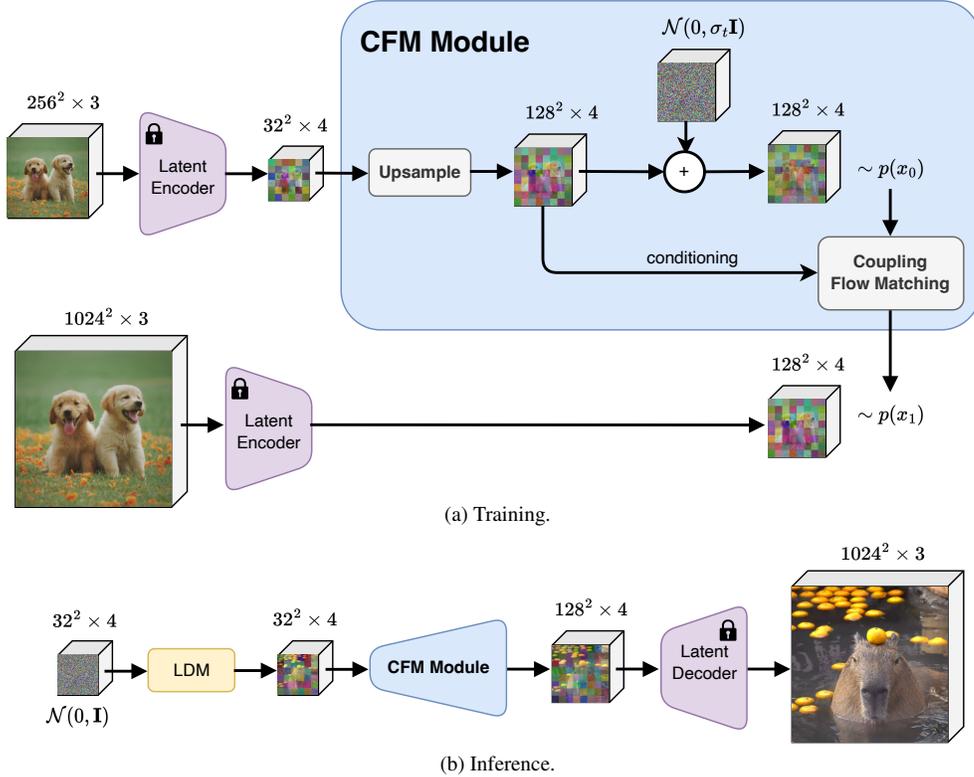


Figure 2. Approach overview. **a)** During training we feed both a low- and a high-res image through the pre-trained encoder to obtain a low- and a high-res latent code, respectively. Based on the concatenated low-res latent code and a noisy version of it, the model regresses a vector field within  $t \in [0, 1]$ . **b)** During inference we can take any Latent Diffusion Model, generate the low-res latent, and then use our coupling flow matching model to synthesize the higher dimensional latent code. Finally, the pre-trained decoder projects the latent code back to pixel space.

with  $z$  as a conditioning variable and  $q(z)$  the distribution of that variable. We parameterize  $v_\theta$  as a U-Net [54], which takes the data sample  $x$  as input and  $z$  as conditioning information.

### 3.2.1 Naïve Flow Matching

We first assume that the probability density path starts from  $p_0$  with standard Gaussian distribution and ends up in a Gaussian distribution  $\mathcal{N}(x_1, \sigma_{\min}^2 \mathbf{I})$  that is smoothed around a data sample  $x_1$  with minimal variance. In this case, the conditioning signal  $z$  would be  $x_1$ , and the optimal transportation path would be formulated as follows [36],

$$p_t(x|z) = \mathcal{N}(x|tx_1, (t\sigma_{\min} - t + 1)^2 \mathbf{I}), \quad (3)$$

$$u_t(x|z) = \frac{x_1 - (1 - \sigma_{\min})x}{1 - (1 - \sigma_{\min})t}; \quad (4)$$

$$\phi_t(x|z) = (1 - (1 - \sigma_{\min})t)x + tx_1. \quad (5)$$

The resulting FM loss takes the form of

$$\begin{aligned} \mathcal{L}_{FM}(\theta) &= \mathbb{E}_{t,z,p_t(x|z)} \|v_\theta(t, \phi_t(x_0)) - \frac{d}{dt} \phi_t(x_0)\| \\ &= \mathbb{E}_{t,z,p(x_0)} \|v_\theta(t, \phi_t(x_0)) - (x_1 - (1 - \sigma_{\min})x_0)\|. \end{aligned} \quad (6)$$

### 3.2.2 Data-Dependent Couplings

In our case, we also have access to the representation of a low-resolution image generated by a DM at inference time. It seems intuitive to incorporate the inherent relationship between the conditioning signal and our target within the Flow Matching objective, as is also stated in [6]. Let  $x_1$  denote a high-resolution image. The conditioning signal  $z := x_1$  remains unchanged from the previous formulation. Instead of randomly sampling from a Gaussian distribution in the naïve Flow Matching method, the starting point  $x_0 = \mathcal{E}(x_1)$  corresponds to an encoded representation of the image, with  $\mathcal{E}$  being a fixed encoder.

Similar to the previously described case, we smooth around the data samples within a minimal variance to acquire the corresponding data distribution  $\mathcal{N}(x_0, \sigma_{\min}^2 \mathbf{I})$  and

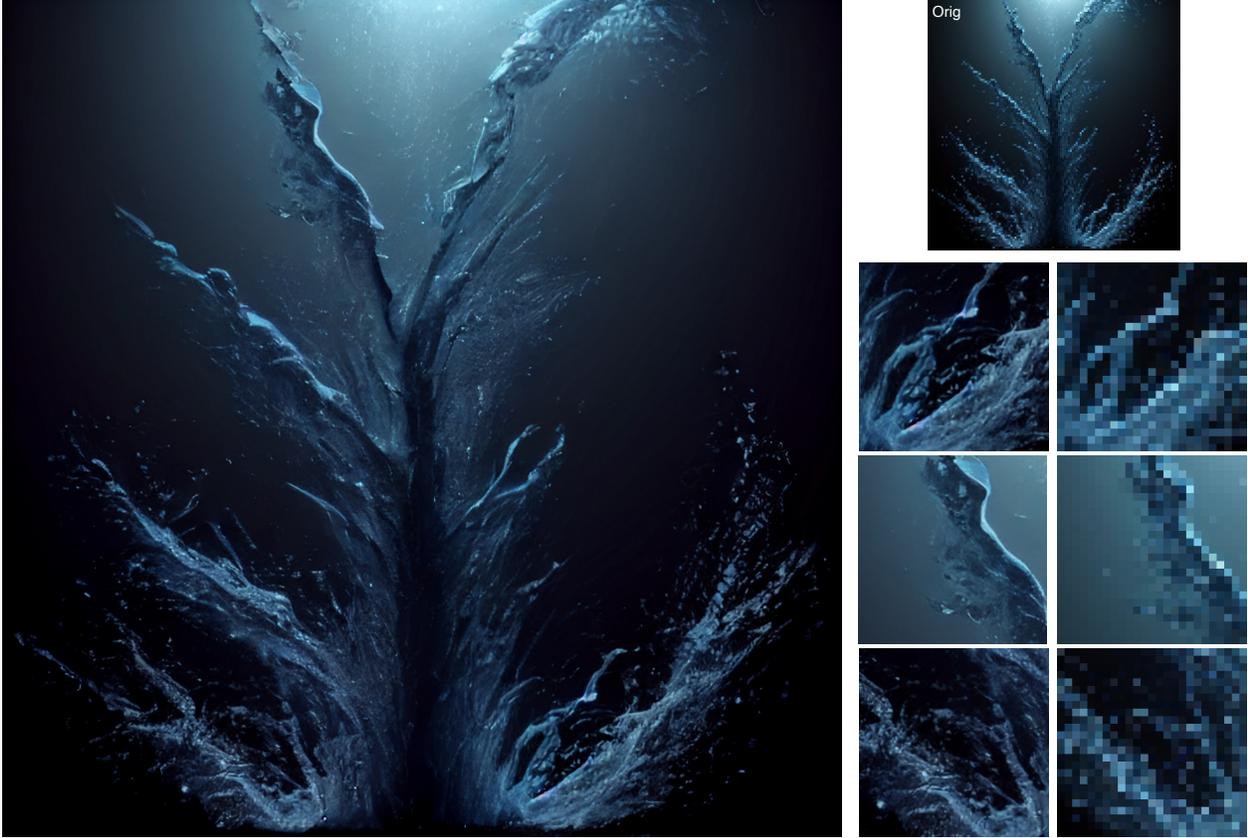


Figure 3. Chaining our models enables elevating the image resolution from  $128^2$  to  $2048^2$  px. The contrast before and after upsampling is presented in the right column, with the original low-resolution image positioned in the top-right corner for reference.

$\mathcal{N}(x_1, \sigma_{\min}^2)$ . The Gaussian flows can be defined by the equations

$$p_t(x|z) = \mathcal{N}(x|tx_1 + (1-t)x_0, \sigma_{\min}^2 \mathbf{I}), \quad (7)$$

$$u_t(x|z) = x_1 - x_0; \quad \phi_t(x|z) = tx_1 + (1-t)x_0. \quad (8)$$

Notably, the optimal transport condition between the probability distributions  $p_0(x|z)$  and  $p_1(x|z)$  is inherently satisfied due to the data coupling. This automatically solves the dynamic optimal transport problem in the transition from low to high resolution within the Flow Matching paradigm and enables more stable and faster training [65]. We name these Flow Matching models with data-dependent couplings *Coupling Flow Matching* (CFM) models, and the CFM loss then takes the form of

$$\mathcal{L}_{CFM}(\theta) = \mathbb{E}_{t,z,p(x_0)} \|v_\theta(t, \phi_t(x_0)) - (x_1 - x_0)\|. \quad (9)$$

### 3.2.3 Noise Augmentation

Noise Augmentation is a technique for boosting generative models' performance introduced for cascaded Diffusion models [20]. The authors found that applying random

Gaussian noise or Gaussian blur to the conditioning signal in super-resolution Diffusion models results in higher-quality results during inference. Drawing inspiration from this, we also implement Gaussian noise augmentation on  $x_0$ . Following variance-preserving DMs, we noise  $x_0$  according to the cosine schedule first proposed in [47]. In line with [20], we empirically discover that incorporating a specific amount of Gaussian noise enhances performance. We hypothesize that including a small amount of Gaussian noise smoothes the base probability density  $p_0$  so that it remains well-defined over the higher-dimensional space. Note that this noise augmentation is only applied to  $x_0$  but not to the conditioning information  $z$ , since the model relies on the precise conditioning information to construct the straight path.

### 3.2.4 Latent Flow Matching

In order to reduce the computational demands associated with training FM models for high-resolution image synthesis, we take inspiration from [13, 53] and utilize an autoencoder model that provides a compressed latent space that aligns perceptually with the image pixel space similar

to LDMs. By training in the latent space, we get a two-fold advantage: i) The computational cost associated with the training of flow-matching models is reduced substantially, thereby enhancing the overall training efficiency. ii) Leveraging the latent space unlocks the potential to synthesize images with significantly increased resolution efficiently and with a faster inference speed.

### 3.3. High-Resolution Image Synthesis

Overall, our approach integrates all the components discussed above into a cohesive synthesis pipeline, as depicted in detail in Fig. 2. We start from a DM for content synthesis and move the generation to a latent space with a pretrained VAE encoder, which optimizes memory usage and enhances inference speed. To further alleviate the computational load of the DM and achieve additional acceleration, we adopt a relatively compact DM that produces compressed information. Subsequently, the FM model projects the compressed information to a high-resolution latent image with a straight conditional probability path. Finally, we decompress the latent space using a pre-trained VAE decoder. Note that the VAE decoder performs well across various resolutions, we show further proof in the appendix.

The integration of FM with DMs in the latent space presents a promising approach to address the trade-off between flexibility and efficiency in modeling the dynamic image synthesis process. The inherent stochasticity within a DM’s sampling process allows for a more nuanced representation of complex phenomena, while the FM model exhibits greater computational efficiency, which is useful when handling high-resolution images, but lower flexibility and image fidelity as of yet when it comes to image synthesis [36]. By combining them in the pipeline, we benefit from the flexibility of the DM while capitalizing on the efficiency of FM as well as a VAE.

## 4. Experiments

### 4.1. Metrics and datasets

For quantitative evaluation, we use the standard Fréchet Inception Distance (FID)[17], SSIM[70], and PSNR to measure the realism of the output distribution and the quality of the image. The general dataset we use for initial experiments and ablations is FacesHQ, a compilation of CelebA-HQ [25] and FFHQ [26], as used in previous work [14, 56] for high-resolution synthesis tasks. However, as highlighted in [11], FID struggles to capture detail and measure fidelity at higher resolutions. To remedy this, we also report p-FID [17] for a more comprehensive evaluation, especially when images contain objects at different scales, such as LHQ [60], which contains 90k high-resolution landscape images and offers a more diverse scale of scenes/objects presented in the image compared to FacesHQ. These two

datasets serve as the basis for the evaluation.

For the general T2I image synthesis task, we train on the Unsplash dataset [2], which provides diverse and high-quality images for training our model. Later, we evaluate on a high-resolution subset of LAION-5B [58] to check how well the model generalizes to unseen data.

### 4.2. Boosting LDM with CFM

Combining LDM with CFM achieves an optimal trade-off between computational efficiency and visual fidelity. We visualize the time taken by LDM and FM, respectively, to synthesize 1k resolution images in Fig. 5, where LDM’s inference time scales quadratically with increasing resolution, and inference is nearly impractical for real-time inference for a latent space of  $128^2$ . To ensure a fair comparison within the limited time frame, we compare our combination to the LCM-LoRA SDXL model [43, 64], which is known for its significantly faster inference than the original SDXL model. Tab. 1 shows that our approach with a standard SD baseline model yields superior performance in terms of FID and inference speed. Note that we apply attention scaling [23] on SD to synthesize images for varying resolutions and finetune the models, with more details in the Appendix. We present a selection of image samples from the baseline SD1.5 model and CFM  $64^2 \rightarrow 128^2$  in Fig. 4. We can equally upscale the LCM-LoRA SD1.5 model from 512 to 1k resolution images with our CFM model. We present our synthesized results in Fig. 1. The inference time for a batch of four samples is 1.388 seconds on an NVIDIA A100 GPU. The LCM-LoRA SDXL model fails to produce images with similar fidelity at the same resolution within the same time.

We further demonstrate the effectiveness of our approach by comparing it to state-of-the-art models [48, 49, 76] in image synthesis on COCO  $1024 \times 1024$ , including CogView3 [76]. We reduce the computational cost of the diffusion component by using a lower resolution and fewer steps, while offloading the remaining steps to our CFM module. This approach significantly reduces the inference time and maintains a good trade-off between speed and accuracy, as shown by the FID in Tab. 2. In summary, we achieve a competitive FID at a faster inference speed than the counterpart diffusion models.

### 4.3. Baseline Comparison

We compare our CFM model to three baseline methods on the FacesHQ and LHQ datasets. For a fair comparison, we fix the UNet architecture and hyperparameters so that the models only differ in their respective training objectives.

**Regression Baseline.** Similar to [56], we compare simple one-step regression models with  $L1$  and  $L2$  loss, respectively. The input is the low-resolution latent code and the target is the corresponding high-resolution latent code of the pre-trained KL autoencoder. In contrast, our method is

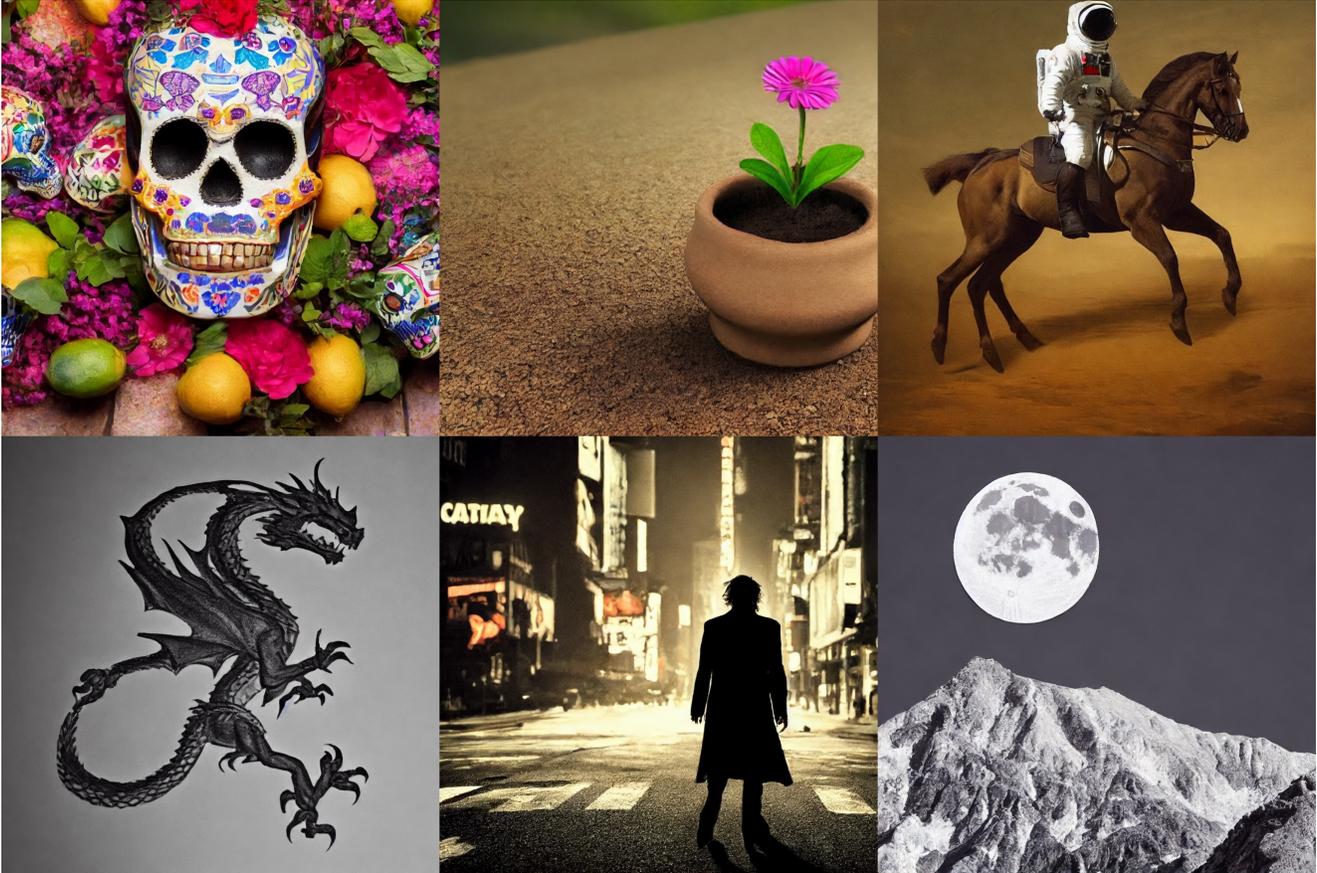


Figure 4. Uncurated samples from the Coupling Flow Matching model on top of SD 1.5 [53] using a classifier-free guidance scale of 7.5. Samples are generated in *latent space*  $64^2$  and up-sampled with CFM from  $64^2$  to  $128^2$ . The resulting images have a resolution of  $1024 \times 1024$  pixels. Best viewed via zoomed in.

Zero-shot LAION-5k $1024 \times 1024$				
Model	CLIP $\uparrow$	FID $\downarrow$	p-FID $\downarrow$	time (s/im) $\downarrow$
SD1.5 + CFM $256^2 \rightarrow 1024^2$	23.75	<u>25.47</u>	<u>23.31</u>	<b>0.62</b>
SD1.5 + CFM $512^2 \rightarrow 1024^2$	<b>26.14</b>	<b>21.67</b>	<b>15.96</b>	3.16
LCM-LoRA SDXL [43]	<u>24.51</u>	28.98	24.00	<u>1.83</u>

Table 1. Quantitative comparison for  $1024^2$  image synthesis using SD v1.5 [53] plus our Coupling Flow Matching (CFM) method against a state-of-the-art diffusion speed-up method. The numbers after CFM symbolize the starting and ending resolutions in pixel space. FID and Patch-FID are computed for 5k samples. We use the fixed step-size Euler ODE solver with 40 number of function evaluations for CFM. For LCM-LoRA SDXL [43] we use 4 sampling steps.

trained with  $L2$  loss on intermediate vector fields. Tab. 3 shows that CFM yields superior metric results. This is also reflected qualitatively, as visualized in Fig. 8, where the images from the regression baseline are visually blurry due to the mode-averaging behavior of the MSE regression. CFM excels at adding fine-grained, high-resolution detail to the image. We conclude that simple regression models trained with  $L1$  or  $L2$  loss are not sufficient to increase resolution in latent space.

**Diffusion Models.** Based on optimal transport theory, the training of a constant velocity field presents a more straightforward training objective when contrasted with the intricate high-curvature probability paths found in DMs [13, 36]. This distinction often translates to slower training convergence and potentially sub-optimal trajectories for DMs, which could detrimentally impact both training duration and overall model performance. Fig. 9 shows that within 100k iterations and for different numbers of function eval-

Zero-shot COCO-5k 1024×1024			
Model	Steps	Time Cost	FID↓
SDXL [49]	50	19.67s	<b>26.29</b>
StableCascade [48]	20+10	10.83s	36.59
CogView3 [76]	50+10	10.33s	31.63
<b>256<sup>2</sup> → 1024<sup>2</sup></b>			
SD1.5 + CFM	25+20	4.07s	33.48
SD1.5 + CFM	40+40	5.88s	30.64
<b>512<sup>2</sup> → 1024<sup>2</sup></b>			
SD1.5 + CFM	25+20	8.79s	<b>28.81</b>

Table 2. Metric results on 5k samples from the COCO dataset [35]. All samples are generated on 1024 × 1024. The time cost is measured with a batch size of 4. Table adapted from [76].

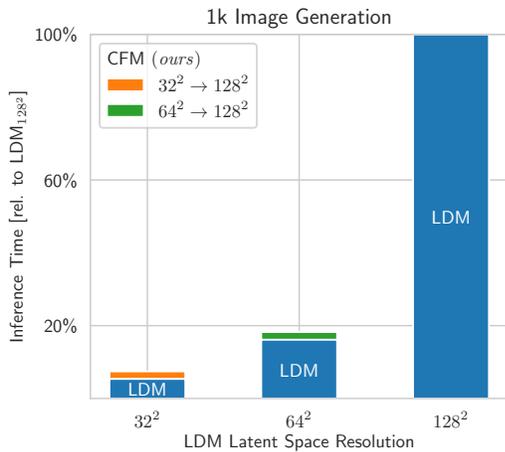


Figure 5. Comparison of 1k image synthesis performance using different architectures. We utilize SD v1.5 as our base model for LDM and adapt its resolution based on [23]. LDM’s inference time grows quadratically with higher resolutions, making real-time inference nearly impractical at a 128<sup>2</sup> resolution latent space. In contrast, the integration of Coupling Flow Matching (CFM) with 50 function evaluations exhibits consistently faster inference, highlighting its efficiency in high-resolution image synthesis.

uations (NFE) after convergence, we consistently achieve a lower FID compared to the DM. In particular, the CFM model shows a faster reduction of the FID and provides better results.

Tab. 4 shows that the combination of DM and CFM outperforms the cascaded DMs across the board.

Taken together, these results underscore the training efficiency of our CFM model over DMs and its superior performance on the up-sampling task after fewer training steps.

**Naïve Flow Matching.** Finally, we compare with Naïve Flow Matching (FM). Similar to DM, FM is conditioned on the low-resolution latent code and starts with Gaus-

sian noise, but uses an optimal transport-based objective to regress the vector fields. In contrast, our CFM method starts directly from the low-resolution latent code and regresses the vector field to the high-resolution counterpart. Due to the presence of data-dependent coupling, we have guaranteed optimal transport during training. For all methods, the low-resolution latent code is available as conditioning information throughout the entire generation trajectory. We evaluated the aforementioned two variants quantitatively (Tab. 3) and qualitatively (Fig. 8), where we observed that the CFM model with data-dependent coupling readily outperforms the ones without. We provide more information about the noise augmentation in Fig. 10. Notably, in the specific upsampling scenario from 256<sup>2</sup> to 1024<sup>2</sup>, we observe an optimal configuration with a noising timestep of 400. The introduction of Gaussian noise proves beneficial as it imparts a smoothing effect on the input probability path, resulting in improved performance. However, excessive Gaussian noise can lead to the loss of valuable information, subsequently deteriorating the data-dependent coupling and reverting the model’s behavior to FM’s Gaussian assumption of  $p(x_0)$ . This finding underscores the delicate balance required in incorporating noise for optimal model performance.

#### 4.4. CFM for Degraded Image Super-Resolution

Our model is originally intended to render image synthesis with existing diffusion models more effectively by enabling them to operate on a lower resolution while increasing pixel-level resolution. However, our method can also be generalized to work on super-resolution tasks which usually include image degradations [67] for low-resolution images. By fine-tuning our method, we can achieve state-of-the-art results on two common benchmark datasets on a 4× upsampling task from 128<sup>2</sup> to 512<sup>2</sup> pixels. We provide quantitative (Tab. 5) and qualitative (Fig. 13) results in the appendix.

#### 4.5. CFM Model Ablations

**Upsampling Methods** Since the dimensionality of the samples from both terminal distributions must be consistent for CFM, we need to upsample the low-resolution latent code  $x_0$  to match the resolution at  $x_1$ . In this context, we perform an ablation study comparing three different upsampling methods: nearest neighbor upsampling, bilinear upsampling, and pixel space upsampling (PSU). The first two methods operate in latent space, while PSU requires the use of the KL autoencoder to upsample in pixel space. Denoting the latent encoder as  $\mathcal{E}$ , the decoder as  $\mathcal{D}$ , and the bilinear upsampling operation as  $UP$ , the upsampling operation PSU can be represented as  $\mathcal{E}(UP(\mathcal{D}(\cdot)))$ . We empirically find that upsampling in latent space works well, but introduces artifacts that make distribution matching with CFM more difficult. In contrast, PSU yields faster model convergence



Figure 6. Samples synthesized in  $2048^2$  px. The base diffusion model is SD1.5 synthesizing images in  $512^2$  px. By plugging in our CFM module, we can quickly boost the resolution to 2k and generate high-fidelity images. Best viewed when zoomed in.



Figure 7. Sample quality for different number of function evaluations (NFE). From left to right, 1st column represents the ground truth, high-resolution image. From the 2nd column on, we show the results for  $NFE = 1, 2, 4, 10, 50$  with the Euler ODE solver.

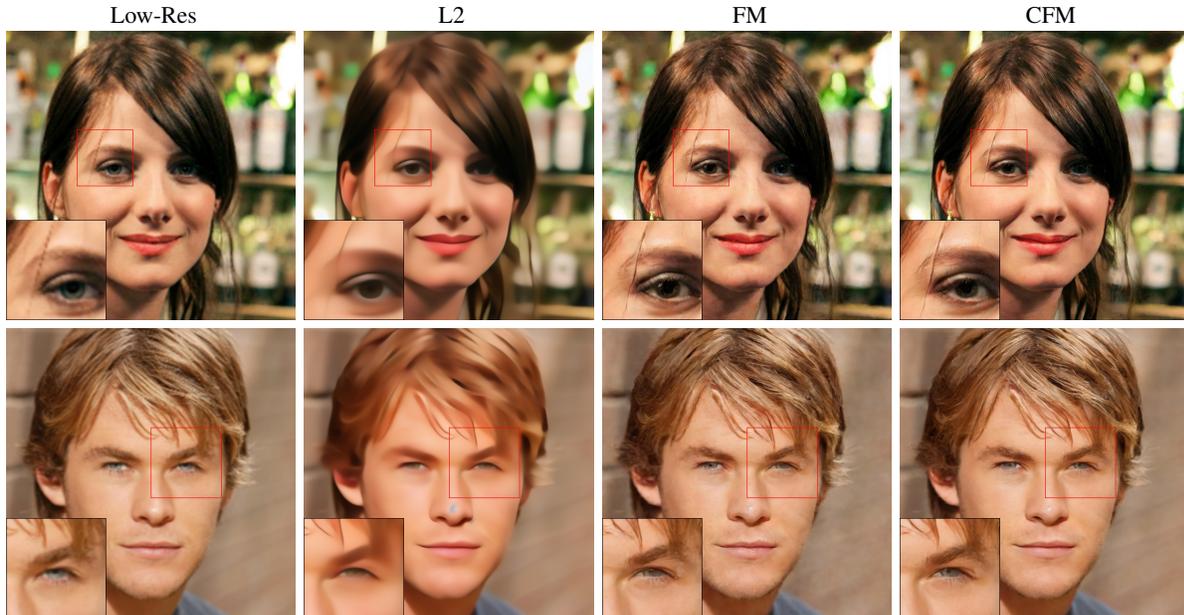


Figure 8. Results for different baseline methods, increasing resolution from  $256^2$  px to  $1024^2$  px. *Low-Res* corresponds to bi-linear upsampling of the low-resolution image, *L2* refers to the L2 regression baseline. *FM* and *CFM* correspond to Flow Matching and Coupling Flow Matching, respectively. Best viewed when zoomed in.

Model	FacesHQ				LHQ			
	SSIM $\uparrow$	PSNR $\uparrow$	FID $\downarrow$	p-FID $\downarrow$	SSIM $\uparrow$	PSNR $\uparrow$	FID $\downarrow$	p-FID $\downarrow$
<i>L1</i>	<b>0.86</b>	<b>31.78</b>	4.52	6.51	<b>0.72</b>	<b>26.99</b>	4.88	6.54
<i>L2</i>	<u>0.85</u>	<u>31.48</u>	5.73	9.07	<b>0.72</b>	<u>26.87</u>	6.02	8.59
<i>DM</i>	0.73	23.68	2.72	4.71	0.61	19.94	4.29	4.55
<i>FM</i>	0.82	30.46	<u>1.37</u>	<u>2.10</u>	0.68	25.50	<u>2.31</u>	<u>2.61</u>
<i>CFM (ours)</i>	0.82	30.40	<b>1.36</b>	<b>1.62</b>	<u>0.69</u>	25.69	<b>2.27</b>	<b>2.38</b>

Table 3. Metric results for *L1* and *L2* regression, diffusion-based (*DM*) similar to [56], Flow Matching (*FM*) [36], and our Coupling Flow Matching (*CFM*) on 5k samples from FacesHQ and LHQ high-resolution datasets, respectively.

at minimal additional cost (see Fig. 11) and also makes our approach invariant to the autoencoder used. Therefore, we use PSU unless otherwise stated.

**Noise augmentation** We systematically investigate the impact of varying levels of noise augmentation. Fig. 10 shows the FID and Patch-FID for different noise augmentation, with higher values corresponding to more noise.

Our findings suggest that noise augmentation is crucial for model performance, albeit being quite robust to the amount of noise. Empirically, we discovered that  $t = 400$  yields the best results overall.

**Intermediate Results along the ODE Trajectory** In Fig. 14 we show intermediate results along the ODE trajectory. It can be seen that the CFM model gradually trans-

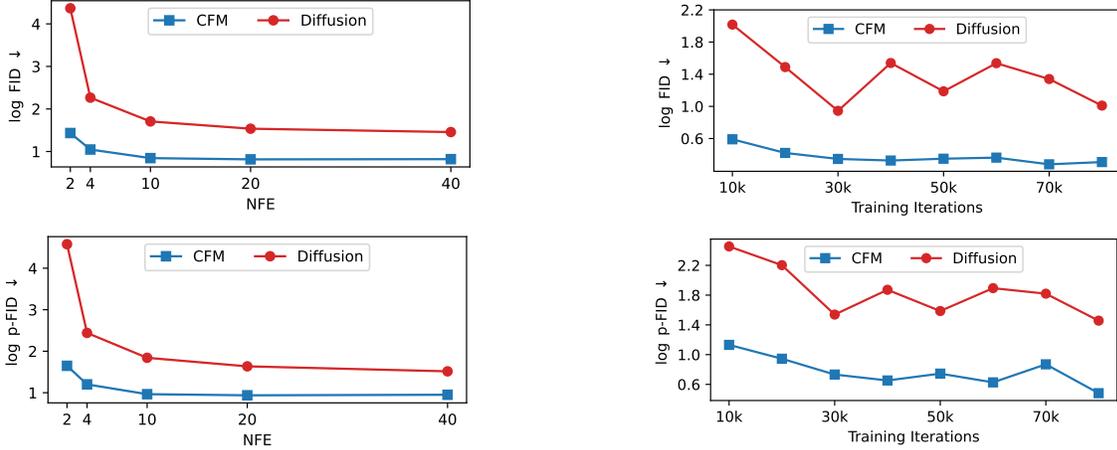


Figure 9. Comparison of a diffusion-based [56] and our Coupling Flow Matching (CFM) module over the training for  $4\times$  up-sampling of the latent codes from  $32^2 \rightarrow 128^2$ . The decoded output resolution is  $1024^2$ . We report FID and p-FID for **a)** different numbers of function evaluations (NFE) and **b)** throughout training. Architecture and hyperparameters are kept fixed. FID evaluated on 5k samples from the LHQ validation set. We use 50 steps for both DDIM [62] sampling and the Euler ODE solver.

Zero-shot LAION-5k $1024 \times 1024$									
Model	Steps	$256^2 \rightarrow 1024^2$				$512^2 \rightarrow 1024^2$			
		CLIP $\uparrow$	FID $\downarrow$	p-FID $\downarrow$	time (s/im) $\downarrow$	CLIP $\uparrow$	FID $\downarrow$	p-FID $\downarrow$	time (s/im) $\downarrow$
Diffusion	4	22.92	35.22	50.53	0.19	25.68	26.53	25.20	2.72
CFM ( <i>ours</i> )	4	23.77	27.54	24.02	0.19	26.16	21.61	15.83	2.72
Diffusion	40	23.55	26.67	24.16	0.62	26.05	22.29	16.36	3.16
CFM ( <i>ours</i> )	40	23.75	25.47	23.31	0.62	26.14	21.67	15.96	3.16

Table 4. Quantitative comparison for  $1024^2$  px image synthesis using SD1.5 [53] for sampling and either our Coupling Flow Matching (CFM) method or a diffusion-based latent space up-sampling model (DM) [56]. FID and Patch-FID are computed for 5k samples. We use the Euler ODE solver for CFM and DDIM sampling for the DM.

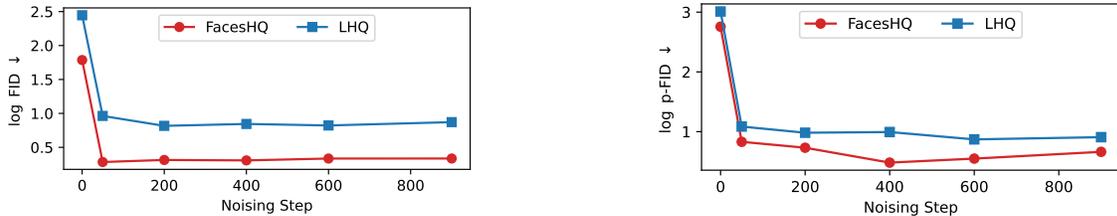


Figure 10. FID (*left*) and Patch-FID (*right*) for our model when applying different degrees of noise augmentation. Evaluated on 5k samples.

forms the noise-augmented image representation to its high-resolution image counterpart.

## 5. Conclusion

Our work introduces a novel and effective approach to high-resolution image synthesis, combining the generation diversity of Diffusion Models, the efficiency of Flow Matching, and the effectiveness of convolutional decoders. Integrating Flow Matching models between a standard latent Diffusion model and the convolutional decoder enables a significant

reduction in the computational cost of the generation process by letting the expensive Diffusion model operate at a lower resolution and up-scaling its outputs using an efficient Flow Matching model. Our Flow Matching model efficiently enhances the resolution of the latent space without compromising quality. Our approach complements DMs with their advancements and is orthogonal to their recent enhancements such as sampling acceleration and distillation techniques e.g., LCM [43]. This allows for mutual benefits between different approaches and ensures the smooth integration of our method into existing frameworks.

## References

- [1] LAION-Aesthetics | <https://laion.ai/blog/laion-aesthetics>. 17
- [2] Unsplash | <https://unsplash.com/data>. 6
- [3] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *CVPR*, 2017. 15, 16
- [4] Michael S Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *ICLR*, 2023. 2, 3
- [5] Michael S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv*, 2023. 2
- [6] Michael S Albergo, Mark Goldstein, Nicholas M Boffi, Ramesh Ranganath, and Eric Vanden-Eijnden. Stochastic interpolants with data-dependent couplings. *arXiv*, 2023. 2, 3, 4
- [7] Sepehr Sameni Aram Davtyan and Paolo Favaro. Efficient video prediction via sparsely conditioned flow matching. In *ICCV*, 2023. 3
- [8] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv*, 2022. 2
- [9] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, 2023. 2, 3
- [10] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *ICCV*, 2019. 15, 16
- [11] Lucy Chai, Michael Gharbi, Eli Shechtman, Phillip Isola, and Richard Zhang. Any-resolution training for high-resolution image synthesis. In *ECCV*, 2022. 6
- [12] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv*, 2023. 2
- [13] Quan Dao, Hao Phung, Binh Nguyen, and Anh Tran. Flow matching in latent space. *arXiv*, 2023. 3, 5, 7
- [14] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 6
- [15] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7346–7356, 2023. 3
- [16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2
- [17] 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. *NeurIPS*, 2017. 6
- [18] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 17
- [19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 2, 3
- [20] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *JMLR*, 2022. 2, 5
- [21] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *arXiv*, 2022. 2, 3
- [22] Tao Hu, David W Zhang, Pascal Mettes, Meng Tang, Deli Zhao, and Cees G.M. Snoek. Latent space editing in transformer-based flow matching. In *AAAI*, 2024. 3
- [23] Zhiyu Jin, Xuli Shen, Bin Li, and Xiangyang Xue. Training-free Diffusion Model Adaptation for Variable-Sized Text-to-Image Synthesis, 2023. 6, 8, 17
- [24] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *CVPR*, 2023. 2, 3
- [25] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv*, 2017. 6
- [26] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 2, 6
- [27] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *NeurIPS*, 2022. 2
- [28] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. In *NeurIPS*, 2021. 2
- [29] Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashed Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, et al. Voicebox: Text-guided multilingual universal speech generation at scale. In *arXiv*, 2023. 3
- [30] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photorealistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017. 3
- [31] Sangyun Lee, Beomsu Kim, and Jong Chul Ye. Minimizing trajectory curvature of ode-based generative models. *arXiv*, 2023. 2
- [32] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 2022. 3
- [33] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *CVPR*, 2023. 3
- [34] Jie Liang, Hui Zeng, and Lei Zhang. Efficient and degradation-adaptive network for real-world image super-resolution. In *ECCV*, 2022. 16
- [35] 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 *ECCV*, 2014. 8

- [36] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *ICLR*, 2023. 2, 3, 4, 6, 7, 10
- [37] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. In *ICML*, 2023. 2
- [38] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. In *ICLR*, 2022. 2
- [39] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2023. 2, 3
- [40] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *NeurIPS*, 2022. 2, 3
- [41] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv*, 2022. 2, 3
- [42] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv*, 2023. 2
- [43] Simian Luo et al. Lcm-lora: A universal stable-diffusion acceleration module. *arXiv preprint*, 2023. 1, 2, 6, 7, 11
- [44] Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *CVPR*, 2023. 2, 3
- [45] Kirill Neklyudov, Rob Brekelmans, Daniel Severo, and Alireza Makhzani. Action matching: Learning stochastic dynamics from samples. In *ICML*, 2023. 2
- [46] 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*, 2021. 3
- [47] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, 2021. 2, 3, 5
- [48] Pablo Pernias, Dominic Rampas, Mats L. Richter, Christopher J. Pal, and Marc Aubreville. Wuerstchen: An efficient architecture for large-scale text-to-image diffusion models, 2023. 6, 8
- [49] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv*, 2023. 1, 2, 3, 6, 8
- [50] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *CVPR*, 2022. 2
- [51] Markus N Rabe and Charles Staats. Self-attention does not need  $o(n^2)$  memory. *arXiv*, 2021. 2
- [52] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv*, 2022. 2, 3
- [53] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 2, 3, 5, 7, 11, 16, 17
- [54] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 4
- [55] 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. *NeurIPS*, 2022. 2, 3
- [56] Chitwan Saharia et al. Image super-resolution via iterative refinement. *TPAMI*, 2022. 3, 6, 10, 11
- [57] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *ICLR*, 2022. 2
- [58] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 2022. 6
- [59] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv*, 2022. 3
- [60] Ivan Skorokhodov, Grigorii Sotnikov, and Mohamed Elhoseiny. Aligning latent and image spaces to connect the unconnectable. In *ICCV*, 2021. 6
- [61] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 2
- [62] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 2, 3, 11
- [63] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021. 2
- [64] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *ICML*, 2023. 2, 6
- [65] Alexander Tong et al. Improving and generalizing flow-based generative models with minibatch optimal transport. In *ICML Workshop*, 2023. 2, 3, 5
- [66] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *arXiv*, 2023. 15, 16
- [67] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *ECCV Workshop*, 2018. 3, 8
- [68] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *ICCV*, 2021. 16
- [69] Xintao Wang et al. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *ICCV*, 2021. 15
- [70] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 2004. 6

- [71] Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion gans. *arXiv*, 2021. [2](#)
- [72] Zeyue Xue, Guanglu Song, Qiushan Guo, Boxiao Liu, Zhuofan Zong, Yu Liu, and Ping Luo. Raphael: Text-to-image generation via large mixture of diffusion paths. *arXiv*, 2023. [3](#)
- [73] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *NeurIPS*, 2024. [3](#), [16](#)
- [74] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *ICCV*, 2021. [3](#), [16](#)
- [75] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. [3](#)
- [76] Wendi Zheng, Jiayan Teng, Zhuoyi Yang, Weihang Wang, Jidong Chen, Xiaotao Gu, Yuxiao Dong, Ming Ding, and Jie Tang. Cogview3: Finer and faster text-to-image generation via relay diffusion. *arXiv*, 2024. [6](#), [8](#)

## A. Appendix for Boosting Latent Diffusion with Flow Matching

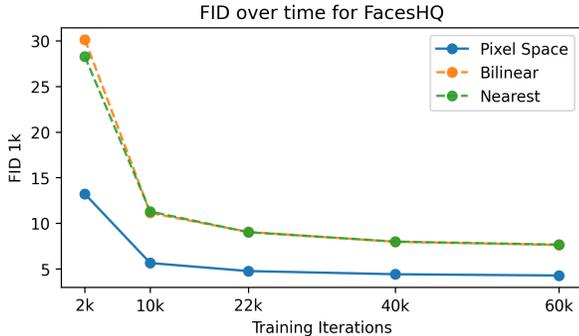


Figure 11. FID over training iterations for bilinear, nearest, and pixel-space upsampling of the low-resolution latent code on FacesHQ.

### A.1. Degraded Image Super-Resolution

Originally, our model was designed to boost the resolution of existing diffusion models at a reduced cost. This differs from traditional super-resolution (SR) methods in two ways. First, they usually perform SR at lower resolutions, and second, they apply image degradation methods, *e.g.*, compression artifacts, to obtain low-resolution images. Our model is not explicitly trained to be invariant to these degradations, but we can generalize to them. To support this claim, we additionally fine-tune our model on the Unsplash dataset, including image degradations following [69] on a  $4\times$  upsampling task from  $128^2$  to  $512^2$  pixels. We then perform *zero-shot* inference on the two common SR benchmark datasets DIV2K [3] and RealSR [10]. Tab. 5 quantitatively shows the comparison to other state-of-the-art SR methods, where our method achieves comparable or even superior results in FID and CLIP-IQA. Fig. 12 further compares the FID of our method against StableSR [66] in the low number of function evaluations (NFEs) regime. Our model excels particularly at low NFEs due to its straighter trajectories compared to the diffusion-based counterpart. Fig. 13 shows additional qualitative super-resolution results for mapping degraded low-resolution images to high-resolution images with our proposed CFM method.

### A.2. Additional Visualizations

#### A.2.1 Image Synthesis at 1024 and 2048 Resolutions

We present additional samples generated at resolutions of 1k and 2k from our pipeline in Figs. 6 and 15 to 17. Our method produces high-resolution images with remarkable fidelity while maintaining fast inference times.

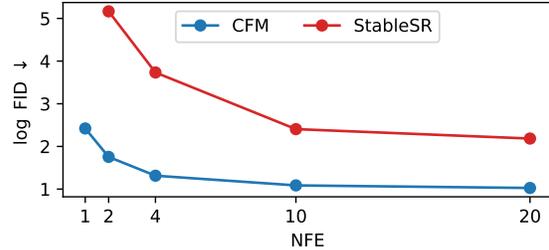


Figure 12. In the low NFE regime we outperform StableSR [66] on FID at the  $128^2 \rightarrow 512^2$  px super-resolution task.

### A.2.2 Super-Resolution

Fig. 13 shows qualitative results of our method applied to the image super-resolution task. Low-resolution images include degradations following [69] and we increase the resolution from  $128^2$  to  $512^2$  px. These results correspond to Tab. 5 from the main paper and showcase the generalizability of our model for super-resolution tasks.

### A.3. Additional Ablations

#### A.3.1 Up-Sampling Method

Our method requires matching resolutions between the source and target latent codes. We ablate three different options. The first two options involve directly upsampling the latent codes in the *latent space*, either via bilinear or nearest neighbor upsampling. The third method pixel space upsampling (PSU) first decodes the image into *pixel space*, then bilinearly upsamples the image, and encodes it back to latent space. This introduces some additional cost through the decoding and encoding operation, however, we found it to be negligible small.

Tab. 6 shows that pixel space upsampling (PSU) performs particularly well. Besides performing better overall, PSU also mitigates the problem of artifacts for low-resolution images. Fig. 18 and Fig. 19 show that the autoencoder introduces artifacts when encoding and decoding low-resolution images, *e.g.* of resolution  $128^2$  and  $256^2$  px. Increasing the resolution in pixel space and then encoding and decoding the image, avoids this issue. Since our primary goal is to speed up inference for image synthesis, using pixel space upsampling maintains most of the information present in the low-resolution latent code, which in turn ensures a well defined starting point for our Coupling Flow Matching module. This also makes the overall training more efficient as shown in

#### A.3.2 Number of Function Evaluations

Tab. 8 shows the performance of our CFM model for different number of function evaluations. We evaluate the results on a subset of 1k samples of FacesHQ and present the total inference time on a single NVIDIA A100 GPU. We can

Model	DIV2K [3]				RealSR [10]			
	SSIM↑	PSNR↑	FID↓	CLIP-IQA↑	SSIM↑	PSNR↑	FID↓	CLIP-IQA↑
BSRGAN	0.63	<u>24.60</u>	44.22	0.52	<u>0.77</u>	23.69	141.28	0.50
RealERSGAN	<b>0.64</b>	24.33	37.64	0.52	0.76	25.69	135.18	0.44
DASR	<u>0.63</u>	24.50	49.16	0.50	<b>0.77</b>	<b>27.02</b>	132.63	0.31
LDM	0.58	23.36	36.21	0.62	0.72	25.49	132.32	0.59
StableSR	0.57	23.31	<b>24.67</b>	<u>0.67</u>	0.70	24.69	<u>127.20</u>	<u>0.62</u>
ResShift	0.62	<b>24.69</b>	36.01	0.61	0.74	<u>26.31</u>	142.81	0.55
CFM ( <i>ours</i> )	0.52	21.63	<u>29.76</u>	<b>0.73</b>	0.65	23.31	<b>125.88</b>	<b>0.67</b>

Table 5. Quantitative comparison to other State-of-the-Art Super-Resolution models on two benchmark datasets. Comparison models are BSRGAN [74], RealERSGAN [68], DASR [34], LDM [53], StableSR [66], and ResShift [73].

Upsampling	FacesHQ					LHQ				
	PSNR↑	SSIM↑	MSE↓	FID↓	p-FID↓	PSNR↑	SSIM↑	MSE↓	FID↓	p-FID↓
Bilinear	25.68	0.71	0.012	3.30	3.96	23.07	0.60	0.036	3.81	4.13
Nearest	25.55	0.71	0.013	3.32	3.77	22.81	0.59	0.038	4.03	4.67
Pixel Space	<b>30.40</b>	<b>0.82</b>	<b>0.004</b>	<b>1.35</b>	<b>1.61</b>	<b>25.49</b>	<b>0.68</b>	<b>0.022</b>	<b>2.32</b>	<b>2.70</b>

Table 6. Ablation of upsampling methods for our Coupling Flow Matching model in the latent space. The *pixel space upsampling* (PSU) method exhibits constantly better results than the latent space upsampling methods.

Image Size	LAION-10k		
	SSIM ↑	PSNR ↑	FID ↓
256	0.82	26.42	2.47
512	0.86	28.65	1.28
1024	0.88	30.72	0.84
2048	0.88	32.20	0.60

Table 7. Evaluation of the pre-trained autoencoder for different image resolutions.

	FacesHQ	LHQ	Unsplash
Model size	113M	306M	306M
Channels	128	128	128
Depth	3	3	3
Channel multiplier	1, 2, 3, 4	1, 2, 4, 8	1, 2, 4, 8
Attention resolutions	16	16	16
Head channels	64	64	64
Number of heads	4	4	4
Batch size	96	128	768

Table 9. Hyperparameters and number of parameters for our Coupling Flow Matching module.

NFE	FacesHQ-1k			
	SSIM ↑	PSNR ↑	FID ↓	Time (ms/img) ↓
1	0.81	27.41	66.24	395
2	0.81	27.86	33.89	399
4	0.79	27.71	17.69	413
10	0.78	27.10	11.53	465
50	0.75	26.50	10.31	718
100	0.75	26.50	10.30	1,046
dopri5	0.75	26.31	10.34	996

Table 8. We ablate the CFM-400 model and compare different numbers of function evaluations (NFE) during inference for the *Euler* method, as well as the adaptive step-size Dormand-Prince solver.

clearly observe that our model achieves good performance already with as few as 10 euler steps, indicating straight ODE trajectories.

### A.3.3 Autoencoder Resolutions

Even though the autoencoder from LDM [53] was trained on a fixed resolution, we find it to generalize well to images at different scales. This is shown quantitatively in Tab. 7, as well as qualitatively in Fig. 19.

## A.4. Training Details

The model architecture details for different datasets are provided in Tab. 9, and we employ a learning rate of  $5 \times 10^{-5}$ . We precompute the image latents to enhance computational efficiency. For fairness, we train the diffusion model counterpart for upsampling using the same architecture, utilizing 1000 diffusion steps and a cosine diffusion schedule.

#### A.4.1 LDM $32^2$

In the main paper we use a Latent Diffusion Model that operates on a lower-dimensional latent space of  $32^2 \times 4$ . The standard SD1.5 [53] was trained to work on latents with a dimensionality of  $64^2$  pixels. After decoding the latent, this results in generated images with a resolution of  $512^2$  pixels. Sampling latents of a different dimensionality is also possible but results in degraded performance, which is particularly pronounced for sampling lower-resolution latents. This deterioration can be partially mitigated by changing the scale of the self-attention layers from the standard  $\sqrt{1/d}$  to  $\sqrt{\log_T N/d}$  where  $d$  is the inner attention dimensionality and  $T$  and  $N$  the number tokens during the training and inference phase respectively [23]. We use this to rescale the attention layers of SD-1.5 for  $32^2$  latents. Additionally, we finetune it on images of LAION-Aesthetics V2 6+ [1] rescaled to  $256^2$  pixels for one epoch, a learning rate of  $1e-5$ , and batch size of 256.

In Tab. 10 we provide metrics for our fine-tuned Latent Diffusion Model on the LAION dataset. We find that FID plateaus with a classifier-free guidance scale at 7.0. Consequently, we adopt this value for small resolution ( $32^2$  in the latent space), text-guided image synthesis.

CFG scale	FID ↓	CLIP ↑
1	41.95	0.157
3	17.13	0.214
5	14.03	0.231
7	13.47	0.239
9	13.50	0.243

Table 10. Results of our fine-tuned Latent Diffusion Model on the LAION dataset for different Classifier-free guidance (CFG) scales [18].

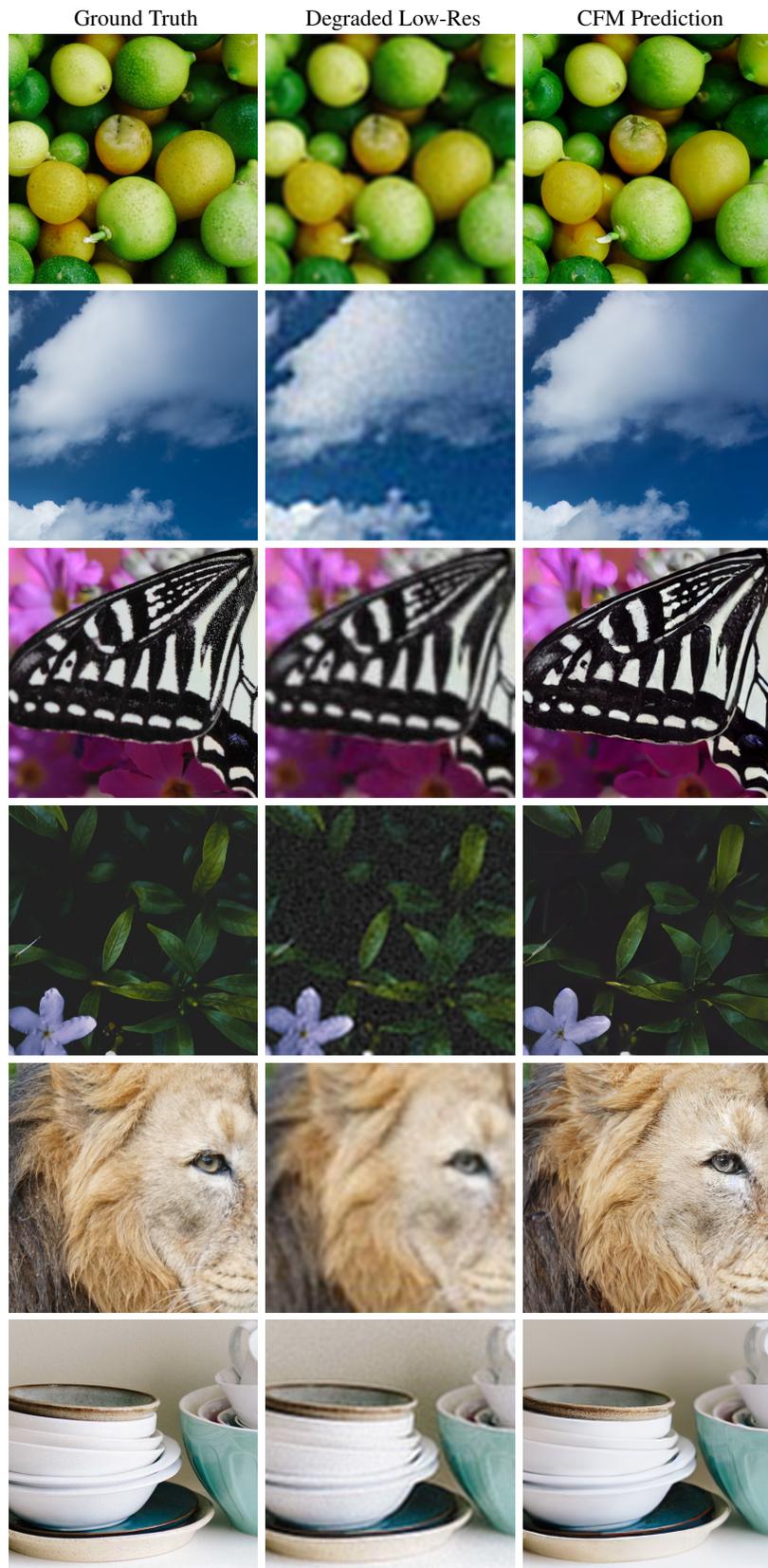


Figure 13. Qualitative results for image super-resolution on degraded images from  $128^2$  to  $512^2$  px with our CFM model.

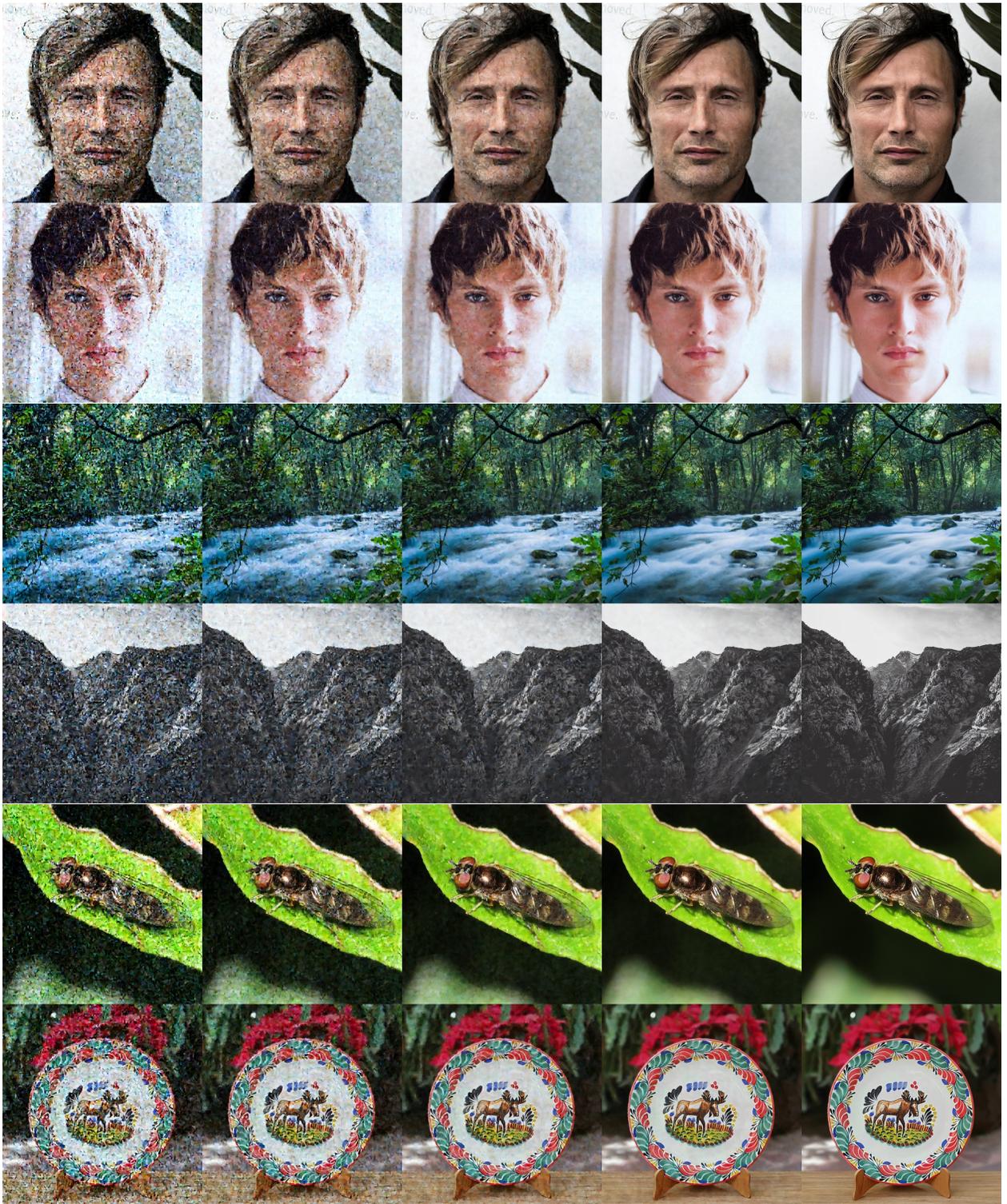


Figure 14. Samples decoded along the ODE trajectory for FacesHQ, LHQ, and LAION at  $t \in \{0, 0.25, 0.5, 0.75, 1\}$ , with a total number of function evaluations of 100. Best viewed when zoomed in.



Figure 15. Samples synthesized in  $1024^2$  px. The comparison is highlighted on the top-left corners between the samples generated solely from DM and those generated from the combination between DM and CFM.

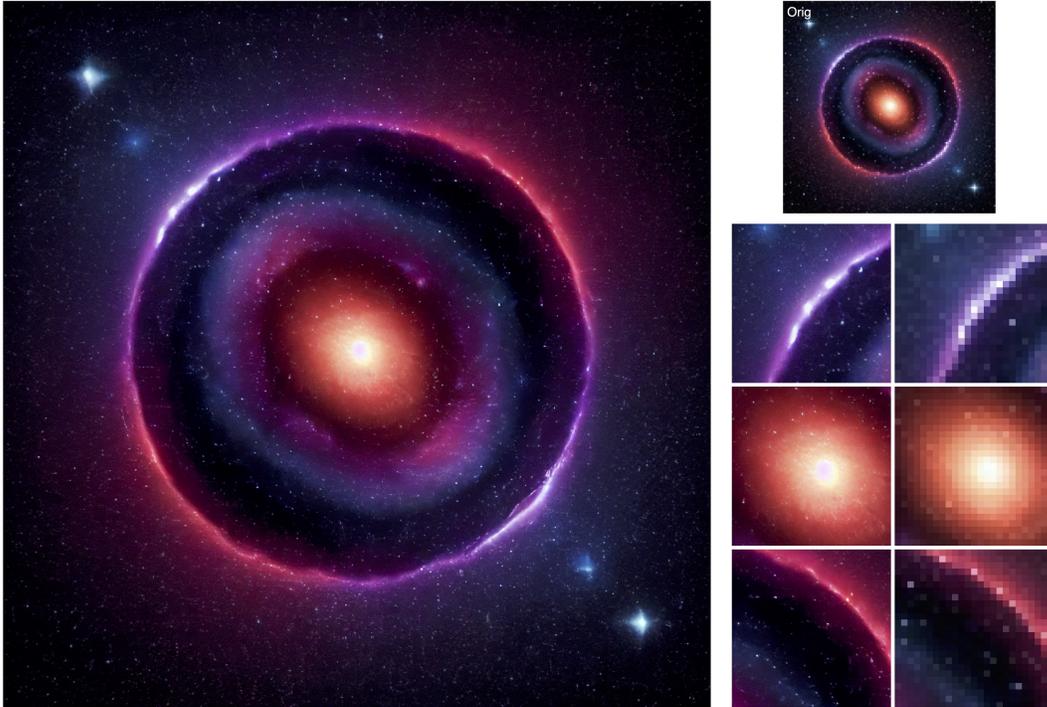


Figure 16. Chaining our models enables elevating the image resolution from  $128^2$  to  $2048^2$  px. The contrast before and after upsampling is presented in the right column, with the original low-resolution image positioned in the top-right corner for reference.

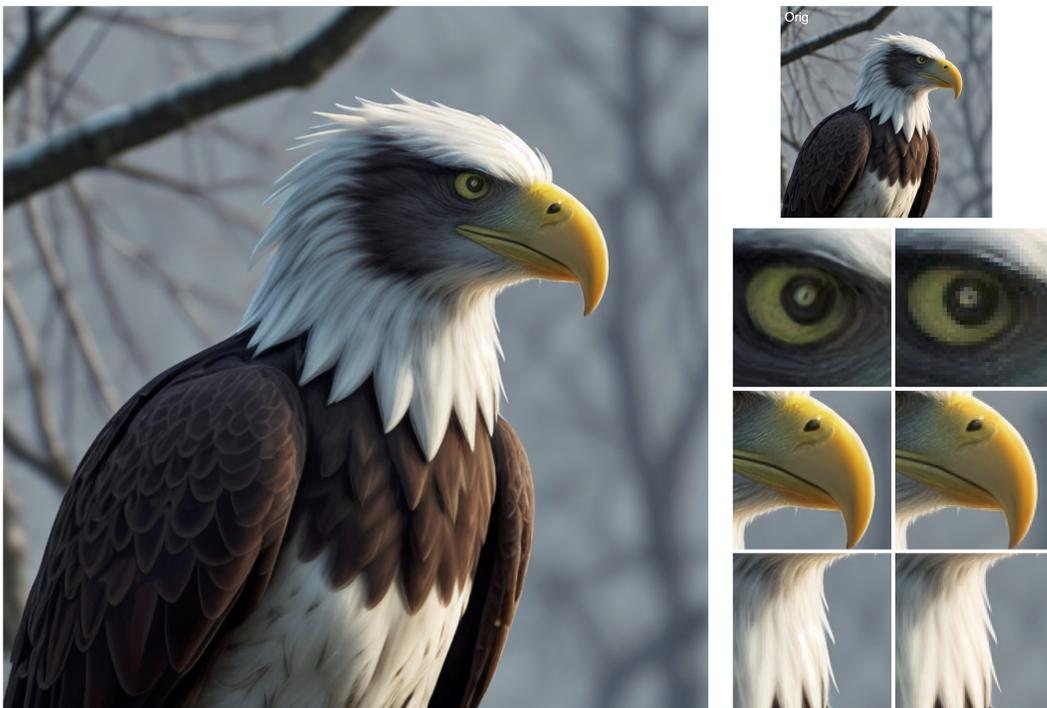


Figure 17. Upsampling results from  $512^2$  to  $2048^2$  px. The contrast before and after upsampling is presented in the right column, with the original low-resolution image positioned in the top-right corner for reference.

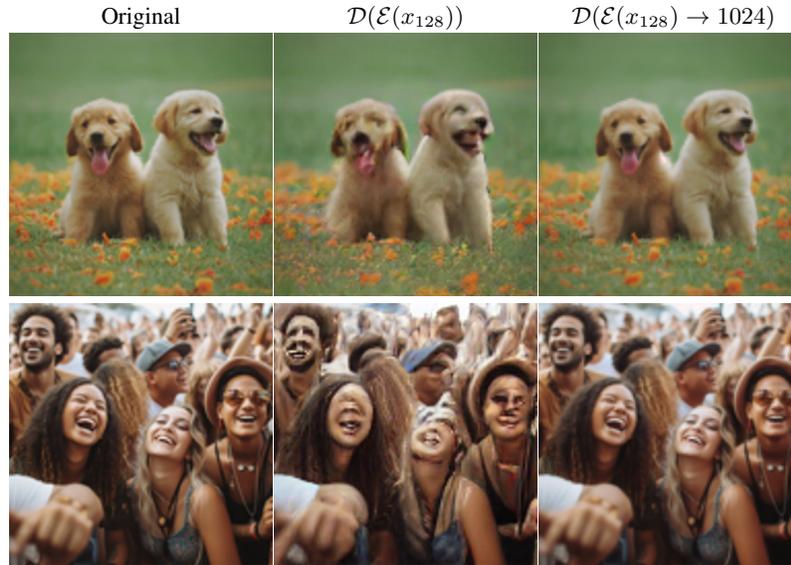


Figure 18. *Left*: Original image with  $128^2$  px resolution, bilinearly upsampled to  $1024^2$  px. *Middle*:  $128^2$  px image encoded and decoded with the autoencoder, clearly showing artifacts. *Right*:  $128^2$  px image bilinearly upsampled to  $1024^2$  px in pixel space and then encoded and decoded with the autoencoder. Besides being blurry, the decoded image shows no artifacts.



Figure 19. Comparison between reconstructed images and their high-resolution original input using the same pre-trained autoencoder. We can observe that the (i) pre-trained autoencoder can encode and decode images at different scales. (ii) It cannot reconstruct faces correctly in low resolution. (iii) The artifacts diminish with a higher resolution. Best viewed when zoomed in.