TASKFORCE: LANGUAGE-GUIDED TOKENIZATION FOR IMAGE SUPER RESOLUTION

Adithya Narayan* **Kaustav Mukherjee*** **Santhoshini Gongidi*** **Tanya Choudhary***
{anaraya2, kaustavm, sgongidi, tchouda}@andrew.cmu.edu

ABSTRACT

Image super-resolution is a common problem with a wide use-case, where a model takes a low-resolution image input and outputs an upscaled image of higher resolution. The main challenge in super-resolution lies in recovering the finer details of high resolution image. Recent models have exhibited difficulties in reconstructing fine-grained details, understanding complex scenes and objects from low-resolution images, and require significant compute and time for inference. In this project, we introduce TokSR, an image super-resolution model that leverages multimodal, text-supported image tokenizers that replace the variational autoencoders often used for image super-resolution, on top of a diffusion transformer backbone. We show that this architecture has the capability to improve image quality and reduce compute and time requirements when compared to current image super-resolution models. We run experiments that demonstrate that our model offers improvements in reconstruction quality in key metrics like FID, while also significantly reducing the inference time and compute needed for image super-resolution.

1 INTRODUCTION AND PROBLEM DEFINITION

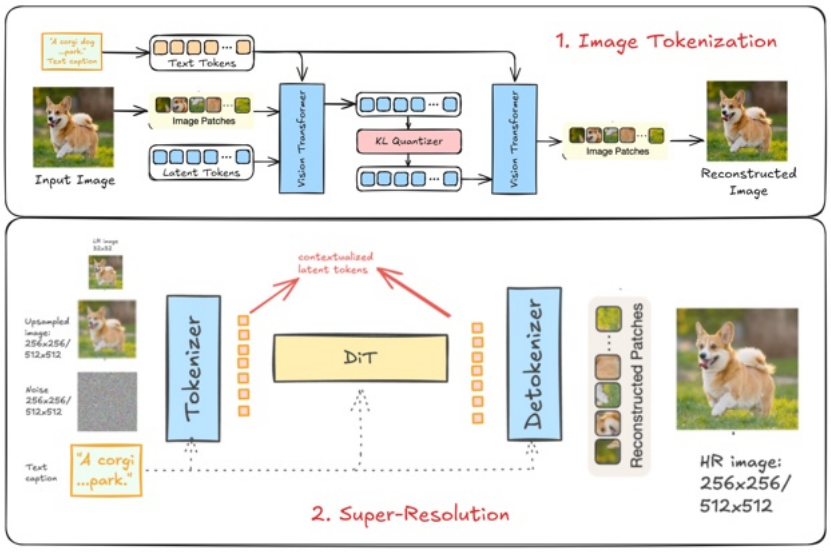


Figure 1: Overview of Proposed Image Tokenization and Super-Resolution Architectures

The goal of image super-resolution (SR) is to generate a high resolution (HR) output from a low resolution (LR) image input. It is a highly ill-posed problem, with an infinite number of possible HR outputs that can map to a single LR input. Therefore, these models are generative in nature, but must be controlled to ensure high-quality generations and consistency with the input image.

* Everyone Contributed Equally – Alphabetical order

Some of the first models, such as SRGAN (Ledig et al. (2017)) and ESRGAN (Wang et al. (2018)), used generative adversarial networks (GANs) to perform image super-resolution. However, since the introduction of diffusion models (Ho et al. (2020b)), and then diffusion transformers (DiTs) (Peebles & Xie (2023)), most modern super-resolution models have adopted these new architectures to create super-resolution images that are much higher in quality and and adhere well to the input images. However, these diffusion-based architectures are much more computationally expensive, meaning that modern models are time and resource inefficient to run. In addition, these models typically have difficulty with reconstructing high-definition regions of the image, such as faces, hands, and other fine-details. Furthermore, in complex scenes, they are often also unable to capture sufficient semantic information from the LR images, therefore resulting in incorrect SR outputs.

Simultaneously, work is being done to improve the performance of image encoders, which are key elements of any super-resolution model. While older models leveraged variational autoencoders (VAEs) to directly encode images into a latent space, newer techniques use both discrete and continuous VAEs to create sequences of tokens that can reconstruct the input image while having a much higher compression ratio (Kingma & Welling (2022)). Further improvements have been made in this field recently, allowing for further image compression using vision transformers and by incorporating textual information to improve reconstructions.

Given the intersection between these fields, we propose TokenSR, a DiT-based super-resolution model that leverages text-guided image tokenization to improve the inference speed and compute requirements while retaining, or even improving output image quality. This will be done by concatenating the low-resolution image input to a noise vector to create the model input, and incorporating text and image embeddings in multiple cross-attention blocks to refine the super-resolution output with additional semantic and fine-grained information. Evaluation results show that our model is able to significantly reduce compute and inference time, while retaining nearly all the performance of a VAE-based model. Additionally, we show that adding text inputs allows for sharper, more accurate super-resolution results.

2 Related Work and Background

There is a plethora of previous work that has been done in the fields of image super-resolution, image tokenizers, and token-based image generation models.

Super-Resolution A seminal work in image super-resolution is SR-GAN, a generative adversarial network-based model capable of natural image super-resolution of up to 4 times the resolution (Ledig et al. (2017)). It does this by incorporating both an adversarial loss, as well as a content loss, which is calculated using perceptual similarity. This concept is further explored in papers such as ESRGAN (Wang et al. (2018)) and Real-ESRGAN (Wang et al. (2021)), which iterate on the perceptual loss by using features before the final activation, and improve both generator and discriminator architectures. Since its advent, diffusion architectures have also been pivotal in the development of new super-resolution models. Introduced in the paper Denoising Diffusion Probabilistic Models by Ho et al. (2020b), these models iteratively denoise a noise sample to generate a high-fidelity output image. This was then expanded to latent diffusion models (Rombach et al. (2022)), which operate in the latent space instead of pixel space, using a variational auto encoder to encode and decode images into latents. This is coupled with a series of cross-attention blocks within the denoising U-Net that allow for conditioning inputs such as text, images, and class labels.

Techniques like ControlNet, introduced by Zhang et al. (2023), enabled low-resolution inputs to serve as conditions for latent diffusion models, effectively allowing image super-resolution. These would be further expanded upon with the advent of the diffusion transformer (DiT), which completely replaced the U-Net in diffusion blocks with transformers, further improving performance. (Peebles & Xie (2023)). Many super-resolution models capitalize on this DiT architecture, such as controlnets with newer latent diffusion models (Esser et al. (2024)), and Inf-DiT by Yang et al. (2024). Inf-DiT, in particular, introduces multimodal information within the diffusion process by using a CLIP image encoder (Radford et al. (2021)) and incorporating those with a cross-attention block, expanding on the cross-attention layers from latent diffusion models.

Image Tokenizers The ability to encode images is paramount for creating both unimodal models, and multimodal models that incorporate visual information. The first prevailing architecture used for this purpose is the variational auto-encoder (VAE) by Kingma & Welling (2022), which proposes a model that encodes an image into a gaussian distribution, then samples it and recreates the input image using a decoder. VAEs have been used as the encoders and decoders for both latent diffusion and DiT models extensively. These have been built upon to create tokenizers that have higher compression ratios, thus allowing more computationally efficient models to be built with them. One of the first works to do this was DALL-E by Ramesh et al. (2021), which uses a discrete VAE to convert images into tokenized representations. This creates a visual vocabulary, that can then be used by transformer models to autoregressively predict image tokens when given a text prompt. VQ-VAE is another discrete VAE architecture that uses a codebook of a fixed size instead of a continuous gaussian as its latent space (van den Oord et al. (2018)), which has been used in many tokenizers since. OmniTokenizer by Wang et al. (2025) uses a VQ-VAE to generate tokens, and uses a transformer to capture global dependencies. The discrete COSMOS Tokenizer (NVIDIA et al. (2025)) also draws from this, using a modified version of VQ-VAEs called Finite-Scalar-Quantization (FSQ) by Mentzer et al. (2023) to encode images into discrete latent codes, while the continuous COSMOS Tokenizer reverts back to a VAE backbone to generate continuous latent tokens, specifically for latent diffusion models like Stable Diffusion.

A more recent development is the Vision Transformer (ViT) by Dosovitskiy et al. (2021), which encodes an image by first patchifying it, adding positional encodings, and using a transformer-only encoder to generate an output latent. This ViT architecture was capitalized upon by Yu et al. (2024) to create a TiTok, a 1D image tokenizer, allowing for a much higher compression ratio than the 2D-token outputs of all the aforementioned models. These could then allow for multimodal tokenization, such as their successive work, Text-Aware TiTok (Kim et al. (2025a)). This model uses a CLIP text encoder to pass additional tokens into the detokenizer, which helps improve semantic alignment with the input image and text, as such details can often be hard to see or lost in LR images. Another model, TexTok (Zha et al. (2024)), also implements a very similar architecture, but adds text tokens to both the tokenizer and detokenizer. A newer model incorporating both text and image tokens is FlowTok He et al. (2025), which combines text and image tokens into a single token space, allowing for flow matching between the two modalities. Some other recent works on tokenization include ElasticTok, another ViT-based model which can tokenize images and videos by considering prior frames and adaptively assigning different numbers of tokens to different frames Yan et al. (2025). Another model, FlexTok, expands on this concept for images by representing them with a variable number of tokens and encoding information hierarchically, with a fewer-token representation containing only higher-level details Bachmann et al. (2025).

Related Datasets For a dataset to be well-suited to this multimodal image SR task, it must have image-caption pairs of common, everyday objects that tokenizers and encoders will have priors on. One such dataset is Common Objects in Context (COCO) by Lin et al. (2015), which contains 83000 images, each with 5 different captions, and available in a variety of sizes. These can be downscaled to form LR-SR image pairs for super-resolution. Another, much larger and less curated dataset is LAION-400M (Schuhmann et al. (2021)), which has 400 million images-text pairs, with over 20 million with a size above 1024 by 1024. This type of dataset can be useful for training larger models with more data, or for a model trying to achieve super-resolution of higher-resolution input images. Another commonly used, large dataset is DataComp-1B, which achieved state-of-the-art results when used to train CLIP (Gadre et al. (2023)). This dataset also has an extremely large number of images, which will be sufficient to train super-resolution models of any size and architecture.

3 Task Setup and Data

3.1 Hypotheses

As previously highlighted, there are three key issues that current super-resolution models suffer from:

1. Inability to correctly construct fine-grained details such as faces and hands.
2. Difficulty capturing semantic information from LR images of complex scenes.
3. Large inference compute and time requirements due to complex architectures.

With respect to these issues, we make some primary hypotheses.

1. **Latent Image Tokens Provide Superior Representations:** We hypothesize that latent image tokens, learned through visual tokenizers, offer a more effective representation than traditional image patch tokens. By capturing high-level features instead of raw spatial patches, latent tokens may lead to improved reconstruction and super-resolution performance. We can test this by training a visual tokenizer and the fine-tuning a tokenizer-based transformer model for image super-resolution.
2. **Latent Image Tokens Allow for Performance Increases** We also propose that using latent image tokens can improve performance by compressing images into a smaller latent, therefore requiring less compute and less inference time from the DiT. Depending on the number and size of tokens, we expect minimal degradation of performance, and even an increase in output quality, despite a significantly smaller latent. We can test this by training an image super-resolution model with both traditional image encoding methods, and a latent image tokenizer, and compare the performance.
3. **Language Guidance Enhances the Quality of Latent Image Tokens** We further hypothesize that incorporating language signals during visual tokenization can enhance the quality of latent image tokens by introducing semantic awareness. By leveraging textual descriptions alongside visual information, tokenization may become more structured, leading to more meaningful and context-aware representations. To evaluate this, we can modify a visual tokenizer architecture to integrate language guidance and compare its effectiveness with a unimodal visual tokenizer.
4. **Textually Guided Tokens Improve Robustness to Low-Quality Inputs** Building on the previous hypothesis, we propose that text-guided tokenization can improve resilience to low-resolution or degraded images. We hypothesize that language-conditioned tokens provide additional semantic constraints, making reconstruction more robust in cases of blurring, noise, or missing details. To validate this, we can train both latent-only and text-conditioned visual tokenizers on clean images and test their ability to recover information from low-quality, corrupted inputs.

3.2 Data

A suitable dataset needs to be selected for the training of these models and the evaluation of the aforementioned hypotheses. In addition, the size of the models and input and output images for super-resolution is limited by the compute resources available to us for this project. Therefore, two different datasets were selected.

3.2.1 COCO

The COCO 2017 train, test, and val datasets were primarily used for training and evaluation of the super-resolution models. With over 82000 images and 414000 captions in the train set alone, this would be a sufficient number to train the DiT models. The important aspects of this dataset included the available image resolutions, as well as the caption length and quality.

From Figure 2, it can be seen that the images are available in a variety of sizes. However, there is a trend for one of the image dimensions to either by 500 or 640, which are ideal sizes for training a super-resolution model with the compute resources available for this project. The images are

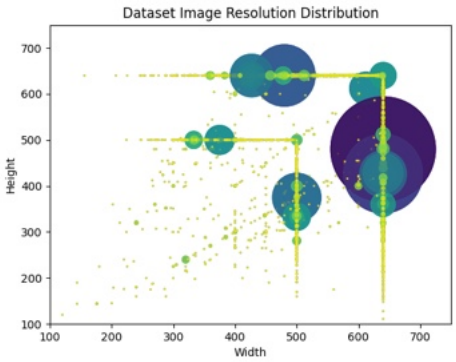


Figure 2: Image Resolutions

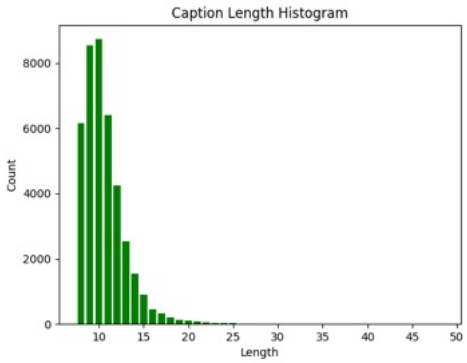


Figure 3: Caption Length Histogram

processed to squares to fit the model dimensions, then LR-SR pairs are created by simply downsampling the images.

Furthermore, Figure 3 indicates that most captions for COCO are fairly small, with a mean caption length of 10.62 words. Importantly, the captions are curated by hand, ensuring accuracy with regards to the input image, and multiple captions are available for each image. A lexical analysis of the captions indicate that the dataset has a Measure of Textual Lexical Diversity (MTLD) of 49.8, which can allow the model to learn a fair variety of input prompts, when compared to the average human MTLD of 95.72.

3.2.2 IMAGENET

For training the image tokenizer, a larger dataset than COCO would help provide better results. Therefore, ImageNet-1K-vl-Enriched from Visual Layer is selected, as it has over 1 million images with corresponding captions, which are generated using the BLIP2 captioning model. With 1000 different classes, this dataset offers a much greater variety in images than COCO, which will be necessary for training tokenizers with sufficient priors to address all the use-cases of the super-resolution model.

Additionally, the ImageNet dataset has a median caption length of 37 to 51 characters, which is similar to COCO, allowing for the trained tokenizer to fit well with the super-resolution model. Furthermore, the majority of the images fall in the same resolution range as the COCO images, meaning they can be transformed similarly for both tasks, and the tokenizer can be trained to best accomodate the COCO input images for the super-resolution model.

4 BASELINES

To establish effective baselines for our task, we can divide our experiments into two categories: super-resolution baselines, and image tokenizer baselines. These can be tested separately to determine the optimal architecture for our model.

4.1 IMAGE SUPER-RESOLUTION BASELINES

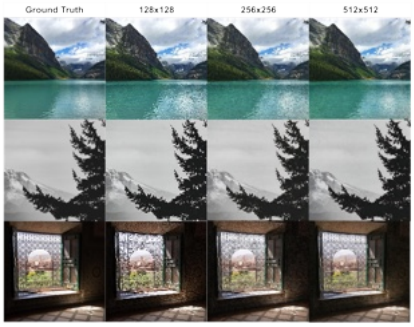
Two key baselines were selected for super-resolution: Inf-DiT and Stable Diffusion, each representing different architectures for comparison.

Inf-DiT is a unimodal image super-resolution model that can upscale low-resolution images by factors of 2 or 4. Built on a DiT-based architecture, it is augmented with a CLIP Encoder to incorporate additional information into the diffusion process. Unlike our LDM-based approach, Inf-DiT operates directly in the pixel space and is not trained explicitly to accept or model CLIP text embeddings. However, we test multimodal versions of this model by incorporating text embeddings from both "quality text" prompts (e.g., "HD, high resolution, clear") and negative prompts (e.g.,

"noisy, blurry, low resolution"), as well as short and long captions derived from ground truth and VLM-generated captions.

Stable Diffusion is a multi-modal Latent Diffusion Model (LDM), where the latent space is represented by a 2D image grid constructed using the VAE as described earlier. Stable Diffusion uses controlnets to condition the generation of super-resolution outputs from a low-resolution input image. We compare the performance of SD3 and SD3.5, both of which have been tested using the L2 loss metric for optimization. As shown in Table 6, SD3.5 achieves the lowest L2 loss, followed by Inf-DiT's unimodal and quality text models, with SD3 showing the highest loss. The superior performance of SD3.5 can be attributed to its larger model size, suggesting that a larger model can yield better results. The comparison underscores the importance of quality text conditioning, with CLIP text embeddings proving valuable for super-resolution tasks.

(a) Inf-DiT Baseline Results (Short Captions)



(b) SD3.5 Baseline Results

Figure 4: Comparison of text-conditioned image super-resolution baselines.

4.2 Image Tokenizer Baselines

Four state-of-the-art (SOTA) image tokenizers are selected as baselines, as they are publicly available and cover a wide span of compression ratios and tokenizer mechanisms. They are:

1. **Dall-E**, which contains a tokenizer that uses a discrete VAE to convert images into tokenized representations. This discretization creates a visual vocabulary that allows transformer models to predict image tokens autoregressively when conditioned on text. As one of the first successful image tokenization schemes, the discrete VAE pioneered the treatment of images as sequences of discrete tokens similar to words in text.
2. **OmniTokenizer**, which combines a VQ-VAE architecture with a transformer for image tokenization. The image is first encoded into discrete latent tokens using VQ-VAE, and these tokens are processed by a transformer to capture global dependencies. This joint approach enables efficient image representation and modeling for tasks like generation and classification.
3. **Discrete Cosmos tokenizer**, which encodes images into discrete latent codes, mapping them into quantized indices, as seen in autoregressive transformers such as VideoPoet 2024). This discrete representation is necessary for models such as GPT that are trained with the cross-entropy loss. It employs a Finite-Scalar-Quantization (FSQ) (Mentzer et al. (2023)) which is modified version of Vector Quantized Variational Autoencoder (VQ-VAE), for creating compressed discrete tokens.
4. **Continuous Cosmos tokenizer**, which also uses a VAE architecture that encodes images into continuous latent embeddings, as in latent diffusion models like Stable Diffusion. These embeddings are suitable for models that generate data by sampling from continuous distributions. Cosmos tokenizers are trained on high resolution images thus making them a good candidate for super-resolution tasks. The continuous latent space also aids in capturing high frequency details in an image, thus enabling high quality reconstructions.

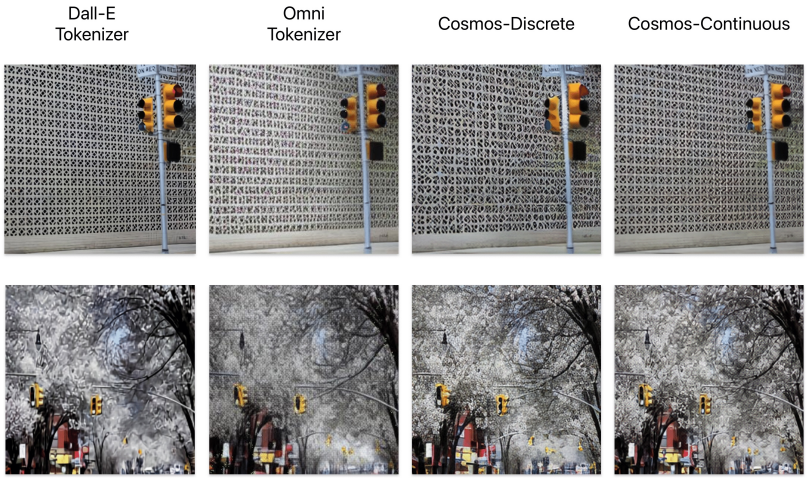


Figure 5: Tokenizer Baseline Qualitative Results

5 Proposed Model

5.1 Text-aware 1D Image Tokenizer-Detokenizer

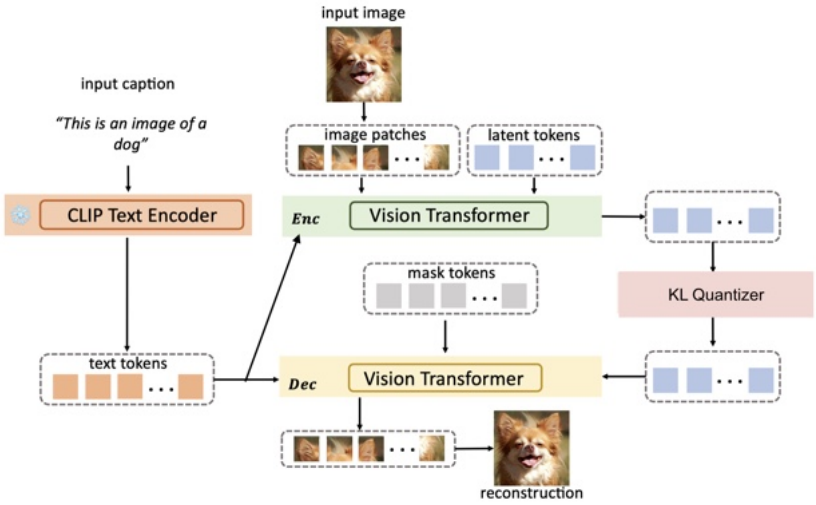


Figure 6: TexTok-VAE: Text-aware 1D Tokenizer Architecture.

Our tokenizer builds on the TA-TiTok (Kim et al. (2025b)) and TexTok (Zha et al. (2024)) frameworks. We implement a symmetric transformer-based encoder-decoder design, as shown in Figure 6. Unlike TA-TiTok, which introduces text conditioning only during detokenization, our model is fully text-aware: both the tokenizer and detokenizer attend to text inputs. Unlike Textok framework, we model the latent tokens Z_{1D} as a Gaussian distribution and apply KL-divergence regularization, yielding a continuous 1D VAE-style representation. This avoids the information loss of hard quantization and improves reconstruction quality.

The tokenizer takes projected image patch tokens, CLIP-encoded text tokens, and learnable latent tokens Z_{1D} as input. Through self-attention, the latent tokens attend to the image and text inputs, forming a semantically aligned representation. The detokenizer receives sampled latent vectors

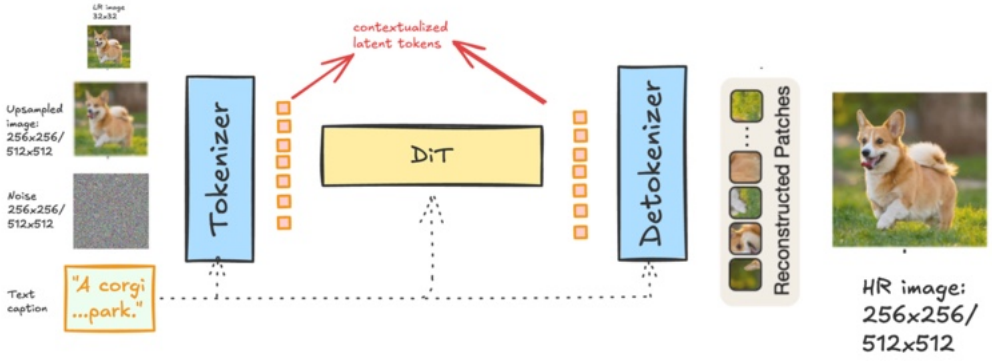


Figure 7: Overall achitecture of our approach

Z_{1D} and the text tokens to reconstruct the image patch tokens via transformer layers. These are then projected and reshaped into the final image.

We use a ViT-Base backbone (12 transformer layers with 768-dimensional embeddings) for both the tokenizer and detokenizer. Input images of size 256×256 are divided into 1024 non-overlapping 16×16 patches, following the standard ViT setup. Text inputs are encoded using a frozen CLIP text encoder, yielding 77 tokens per sample. The latent space consists of 32 learnable tokens, each with an embedding dimension of 16. We refer to this architecture as *TexTok-VAE* in this report.

This KL-regularized continuous latent space integrates seamlessly with diffusion models as a replacement for 2D VAEs. Training optimizations are detailed in Section 5.6.

5.2 LDM FOR IMAGE SUPER-RESOLUTION

We also tried using the LR image in training in two ways; first as a conditioning signal that cross-attends with the DiT and second by concatenating it to the noisy latent as done in the original LDM paper (Rombach et al. (2022)). Overall we find that the concatenation approach works best. Specifically, we begin by tokenizing both the HR image and LR image (upscaled with bi-cubic interpolation). Then we concatenate the noisy tokens of the HR image directly to the LR image tokens generating a final input of dims $[16, 32+32, 16]$. The simplified L2 is then computed between the first 32 output tokens and the original image tokens.

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{x_0,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(x_t, t)[:, :32, :]\|_2^2\right]$$

We also have to be careful when handling the positional embedding in this case. We consider two design options for incorporating positional embeddings into our model. The first is to use separate positional embedding layers for the low-resolution (LR) and high-resolution (HR) tokens, allowing the model to independently learn the spatial relationships within each domain. The second approach, which we adopt in this work, is to share the same positional embeddings across both LR and HR tokens. We do this by repeating the indexes fed into the positional embedding layer.

By using shared positional embeddings, we encourage the model to form joint spatial relationships between corresponding patches in the LR and HR images. In other words, this setup explicitly guides the model to learn associations between tokens occupying the same spatial location across different resolutions, promoting alignment and consistency in learned representations.

5.3 LOSS FUNCTIONS

To train the **TexTok-VAE** tokenizer effectively, we combine multiple loss functions that balance reconstruction quality, perceptual fidelity, semantic alignment, and adversarial realism. These loss functions form the standard set of loss functions used while training image tokenizers. While the reconstruction and perceptual losses ensure that the generated images match the input both pixel-wise

and visually, the KL divergence regularizes the latent space for smooth sampling. The adversarial and LeCam losses introduce realism by encouraging the model to produce sharper and more lifelike outputs through GAN-based feedback.

The total training loss is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \lambda_{\text{percep}}\mathcal{L}_{\text{percep}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}} + \lambda_{\text{disc}}\mathcal{L}_{\text{adv}} + \lambda_{\text{lecam}}\mathcal{L}_{\text{lecam}}$$

Each component of the total loss contributes as follows:

- $\mathcal{L}_{\text{rec}}$: L2 loss between original and reconstructed images, promoting pixel-wise accuracy.
- $\mathcal{L}_{\text{percep}}$: Perceptual similarity loss based on LPIPS with ConvNeXt-S, encouraging visual fidelity.
- $\mathcal{L}_{\text{KL}}$: KL divergence regularization on latent variables, enforcing smoothness in the learned latent space.
- $\mathcal{L}_{\text{adv}}$: Adversarial loss from a discriminator network, pushing reconstructions to be visually realistic.
- $\mathcal{L}_{\text{lecam}}$: LeCam regularization to stabilize discriminator training using exponential moving averages of logits.

The corresponding weights are: $\lambda_{\text{percep}} = 1.1$, $\lambda_{\text{KL}} = 10^{-6}$, $\lambda_{\text{disc}} = 0.1$, and $\lambda_{\text{lecam}} = 0.001$. The discrimator loss contributes to the model optimizations starting at 200k iterations.

For **super-resolution model** we use the mean squared error loss for the diffusion process. Loss is applied against the noise generate in the latent space during forward diffusion process, and the noise predicted by the diffusion model in the backward pass. This loss was originally proposed by the DDPMHo et al. (2020a) in pixel space and the adapted to latent space by LDMRombach et al. (2022).

$$\mathcal{L}_{\text{MSE}} = \mathbb{E}_{x_0,\epsilon,t}\left[\|\epsilon_t - \hat{\epsilon}_t\|_2^2\right]$$

Where:

- $\mathbf{z}_t$ is the noisy latent variable at timestep t,
- $\mathbf{x}$ is the original image (or data),
- ϵ_t is the true noise added to the latent variable $\mathbf{z}_t$,
- $\hat{\epsilon}_t$ is the predicted noise by the model,

5.4 Changes to training data

The **TexTok-VAE** model is trained and validated on a curated subset of the ImageNet dataset. Specifically, we use 600K images along with their corresponding captions sourced from the `imagenet-1k-vl-enriched` dataset[1]. These automatically generated captions provide language supervision for our text-aware tokenizer. To enhance data diversity and improve generalization, we apply standard augmentations during training, including center cropping and horizontal flipping.

Given the limited scale of the COCO dataset compared to ImageNet, we opt to train our combined 24-layer ViT-based tokenizer and detokenizer on the ImageNet dataset. Although we also prepared the CC12M dataset for experimentation, computational constraints led us to restrict our training to ImageNet.

For **super-resolution model**, we train on the COCO2017 train dataset and eval on its val set, as proposed originally in the first report. We resize the images to 256x256 pixels, which serve as our super-resolution images. For generating low-resolution images, we use bicubic-interpolation to scale the images down to 64x64 pixels.

We also pre-tokenize the images using tokenizers and captions using CLIP before training our super-resolution model. We perfrom this for our training as well as validation set. This reduces our

[1] https://huggingface.co/datasets/visual-layer/imagenet-1k-vl-enriched

Figure 8: Conditioning via cross-attention and self attention. First column represents the groundtruth images, columns 2 to 4 are generated with cross-attention and 5 to 7 are with self attention conditioning. For both set of methods we vary the guidance scale in range as 4, 6 and 10

compute cost during training as we do not need the text and image encoders in GPU memory, as well as speeds up the training by allowing for greater batch sizes.

5.5 Conditioning Methods

One of the major design aspect we tested was conditioning method for our latent diffusion model. In LDM paper they describe two main conditioning mechanisms, cross-attention and self-attention. For the first one, conditioning is provided via cross-attention blocks within the diffusion transformer blocks. In our model we experimented with conditioning as combination of text caption encoded using CLIP, and low-resolution image encoded using DINOv2 Oquab et al. (2023). DINOv2 was chosen specifically as it is trained for downstream tasks in spatial grounding such as segmentation and depth estimation. For CLIP we use the pooled token and for DINO we use the CLS token as well mean of the remaining tokens as a method to capture spatial information. These two modalities are then combined by projection and concatenation, forming our conditioning signal.

The low-resolution conditioning via self-attention approach, we upscale the low-resolution image to same size as our desired super-resolution size of 256x256. Then we tokenize this upscaled image using TA-Titok tokenizer and concatenate these low-res up-scaled tokens with the noise latents as input to diffusion model. The model also is conditioned with clip embeddings of text captions via cross attention.

In our experimentation for super-resolution task we found that conditioning via self-attention worked much better than cross-attention for aligning the super-resolution output with low-resolution image. In case of cross-attention, model will hallucinate the details of the image entirely with minimal correlation with the low-resolution images in form of color and some structure. The self-attention on the other hand was able to achieve alignment well retaining the details of low resolution image and filling in with semantically aligned details.

5.6 Hyperparameters and their effects

During the development of the **TexTok-VAE** tokenizer, several training strategies and architectural decisions were explored to improve reconstruction quality, convergence stability, and training efficiency.

To represent the textual modality, we opted for CLIP-based text token embeddings over more compute-intensive alternatives like T5. While models such as T5 offer richer language representations, their inference cost and memory footprint are impractical for large-scale tokenizer training. CLIP embeddings, in contrast, provide a compact yet effective linguistic prior, yielding 77 tokens per sample with minimal computational overhead.

We initially attempted to train a discrete tokenizer using vector quantization in the latent space. However, this setup led to blurry reconstructions, even after 50k iterations, and failed to capture the finer structures of natural images. This limitation prompted a shift to a continuous latent repre-

Figure 9: Comparison of reconstructions from TexTok-VAE and TexTok-VQ after 50k training iterations. *The TexTok-VAE model, which employs a continuous latent representation with KL regularization, produces significantly sharper and more detailed reconstructions. In contrast, the TexTok-VQ model struggles with blurriness and fails to reconstruct higher-level object structures that are clearly visible in the TexTok-VAE model.*

sentation, where the latent tokens Z_{1D} are modeled as Gaussian distributions and trained with KL divergence regularization. The comparison shown in Figure 9 highlights the superior reconstruction quality achieved with the continuous variant over the discrete counterpart.

In another experiment, we leveraged pretrained weights from the TA-TiTok model, which publicly provides its image tokenizer and text-aware detokenizer. Our initial approach involved freezing the detokenizer and training only the text-aware image tokenizer. However, this led to training divergence and suboptimal performance. As a result, we adopted an end-to-end training strategy for both modules. To assist early convergence, we initialized the encoder with TA-TiTok weights, although further experiments are needed to quantify the true impact of this initialization on final performance.

Similarly, we also experimented with the positional embedding to concatenate the LR image tokens with the HR image tokens. To do this, we worked with two sets of intuitions. For the first, we simply interpolated the positional embeddings we obtain for the HR tokens to allow it to work with the LR tokens. Similarly, following the intuition that corresponding token positions from both the HR and LR image must contain correlated information, we also experimented with simply repeating the positional embeddings. Overall, we found that the latter produced the most compelling results.

Finally, all models were trained using mixed-precision (FP16) training, which significantly improved training throughput and memory efficiency. This allowed for larger batch sizes and faster convergence without compromising model quality.

For **super-resolution** models, we evaluate the effect of adding classifier-free guidance for low resolution images in the conditioning via cross attention case. For our base case (w/o classifier-free guidance) we set the low resolution image conditioning to be provided to model at all times. Whereas for the classifier-free guidance case, we switch the conditioning signal"ON" for 85% time and "OFF" for 15% times when we set the text labels to be zero. From these we observe that loss decreases to a lower value without classifier-free guidance than with classifier-free guidance.

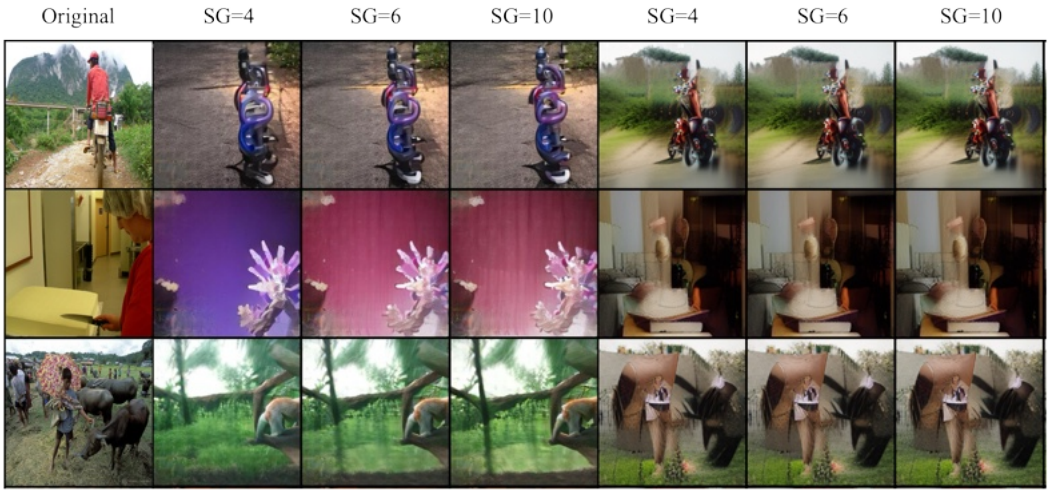


Figure 10: Results of super-resolution model trained with TiTok(VQ-VAE) and TA-Titok(VAE) with guidance scales 4, 6 and 10.

Guidance	L2 loss
w/o Classifier-free guidance	0.08
Classifier-free guidance	0.16

Table 1: Effect of low-resolution image conditioning on loss

With also try two different variants of tokenizers for super-resolution task. First is a VQ-VAE style TiTok, where images are encoded as discrete latents using a codebook. And second is VAE style Ta-Titok, which encodes to latents in a continuous space. We wanted to compare the effects of discrete vs continuous latent space for training a DiT model with text and image conditioning. The l2-loss for training with VQ-VAE (0.03) was lower than with VAE (0.06) encoded latents. However the generated results had more patterned artifacts with VQ-VAE. It was also more difficult to capture finer textures in generated images with VQ-VAE. Thus for our final model we decided to use VAE based TA-Titok tokenizer. Note that the low resolution image here was used as conditioning via cross attention along with text guidance, and thus fails to align with the low-res image.

6 Results

6.1 Tokenizer Results

TexTok-VAE is trained in an end-to-end manner for 300K iterations on the ImageNet-1K Enriched dataset, with additional training details provided in Section 5.3 and 5.4. Reconstruction performance is evaluated using two primary metrics: reconstruction Fréchet Inception Distance (rFID) and Inception Score (IS), as reported in Table 2. The rFID measures the distributional distance between original and reconstructed images using features extracted from a pretrained Inception network, providing a perceptual evaluation of reconstruction quality where lower values indicate better fidelity. Inception Score, in contrast, evaluates the quality and diversity of generated images by assessing how confidently a pretrained Inception classifier assigns labels, with higher scores indicating both high visual fidelity and greater semantic diversity. All evaluations are conducted on the ImageNet validation set to facilitate comparison against TiTok variants,publicly available 1D tokenizer-VAE trained on the significantly larger and more diverse DataComp-1B dataset. Although the results for TiTok variants are reported in a zero-shot setting, the scale and diversity of DataComp present a considerable advantage over our comparatively smaller training corpus. Despite this disparity, TexTok-VAE achieves competitive reconstruction results, suggesting that given access

to larger-scale data and additional compute resources, the model can be scaled to match or exceed the performance of state-of-the-art open-source tokenizers.

Tokenizer	Text-Aware	Training Data	rFID ↓	IS ↑
TiTok	-	LAION-Aesthetics, LAION-Humans, OpenHuman (1B)	2.56	171.7
TA-TiTok	Detokenizer Only	DataComp (1B)	1.61	197.5
TexTok-VAE (Ours)	Tokenizer + Detokenizer	ImageNet (600K)	2.73	174.14

Table 2: Comparison of reconstruction metrics on ImageNet validation set. rFID evaluates perceptual fidelity (lower is better) and Inception Score (IS) reflects the quality and diversity of reconstructions (higher is better).

We also report zero-shot evaluation results of our model compared to baseline tokenizers introduced in Section 4, summarized in Table 3. The comparison primarily focuses on 2D versus 1D image tokenizers along two key axes: the number of tokens required for generation and detokenization in downstream generative tasks, and the reconstruction Fréchet Inception Distance (rFID) as a measure of reconstruction quality. Reducing the number of tokens is crucial because transformer-based models experience quadratic computational overhead with respect to the sequence length due to the self-attention mechanism. Thus, achieving a favorable trade-off between token count and reconstruction fidelity is critical for designing scalable and efficient generative models. A lower token count directly translates into faster inference times and reduced memory requirements, enabling practical deployment in high-resolution image generation tasks. Our results show that TexTok-VAE and TA-TiTok, despite being 1D tokenizers with a fixed token budget, achieve competitive SSIM and LPIPS scores compared to 2D tokenizers, while significantly reducing the token count required for generation. This efficiency underscores the potential of 1D tokenizers to scale large generative models without sacrificing output quality, suggesting promising directions for future multimodal and high-resolution generation systems. We suspect that the high FID observed for 1D tokenizers may be due to artifacts introduced during tokenization. Additionally, it is noteworthy that the FID differs between the ImageNet validation and COCO test sets, and we plan to investigate this discrepancy in future experiments.

Tokenizer	1D vs. 2D	VQ/VAE	# Tokens ↓	rFID ↓	SSIM ↑	LPIPS ↓
DALL-E	2D	VQ	256	0.18	0.64	0.14
OmniTokenizer	2D	VQ	256	0.29	0.70	0.18
CosmosTokenizer-1	2D	VQ	256	0.17	0.6938	0.0463
CosmosTokenizer-2	2D	VAE	1024	0.18	0.8433	0.00939
TA-TiTok	1D	VAE	32	**7.72**	**0.54**	**0.175**
TexTok-VAE (Ours)	1D	VAE	32	**9.71**	**0.43**	**0.21**

Table 3: Reconstruction metrics comparison for tokenizers, computed on the COCO test set . "# Tokens" indicates the number of effective image tokens that must be denoised by the super-resolution model during reconstruction.
Citations: DALL-E Ramesh et al. (2021), OmniTokenizer Wang et al. (2025), CosmosTokenizer Agarwal et al. (2025), TA-TiTok Kim et al. (2025b)

6.2 SuperResolution Results

As shown in Table 4, our tokenized latent diffusion based super-resolution method outperforms the baseline Inf-DiT which operates in pixel space in terms of FID and SSIM metrics, achieving a lower FID score (1.0574 vs. 1.272) and a higher SSIM value (0.4334 vs. 0.365), indicating that our model generates images that are both more realistic and structurally closer to the ground truth. While our approach results in a slightly lower PSNR (17.30 compared to 18.35) and a higher LPIPS score (0.300 vs. 0.175), suggesting a minor trade-off in pixel-wise fidelity and perceptual similarity,

the improvements in FID and SSIM demonstrate better overall perceptual quality and structural preservation, which are often more aligned with human visual preferences in super-resolution tasks.

	FID ↓	LPIPS ↓	SSIM ↑	PSNR ↑
Inf-DiT	1.272	0.175	0.365	18.35
Ours	1.0574	0.300	0.4334	17.30

Table 4: Comparison of our model with baseline on Super-resolution metrics

7 Analysis

7.1 Intrinsic Metrics

To assess the performance of image tokenizers, we conducted an evaluation measuring reconstruction quality under controlled random masking conditions. In this experiment, a fraction of the tokens was progressively replaced—using random codebook entries for discrete tokenizers and Gaussian noise for continuous tokenizers. The results, presented in Table 5, report both absolute reconstruction losses and their degradation rates as masking increases. The degradation slope serves as a proxy for codebook utilization efficiency, highlighting how effectively each tokenizer leverages its representational capacity.

OmniTokenizer demonstrated the least relative reconstruction decline, indicating that its codebook is highly utilized and information can be retrieved from multiple codebook vectors. In contrast, higher relative reconstruction declines suggest that the same information cannot be effectively retrieved from other vectors, indicating lower utilization of the latent space. Our model's intrinsic metrics indicate that it is trained to utilize all 32 tokens to encode the image effectively.

Tokenizer	**1D vs. 2D**	**VQ/VAE**	**Relative Reconstruction Decline**		
			Mask 0%	**Mask 30%**	**Mask 70%**
DALL-E	2D	VQ	0.0022	0.0085 (286.36%)	0.0512 (2227.27%)
OmniTokenizer	2D	VQ	0.254	0.268 (5.51%)	0.284 (11.81%)
CosmosTokenizer-1	2D	VQ	0.022	0.109 (395.45%)	0.318 (1345.45%)
CosmosTokenizer-2	2D	VAE	0.006	0.040 (566.67%)	0.040 (566.67%)
Textok-VAE (Ours)	1D	VAE	0.020	0.0415 (107.5%)	0.087 (335%)
TA-TiTok	1D	VAE	0.012	0.042 (250%)	0.090 (650%)

Table 5: Evaluating Tokenizer Efficiency through progressive randomization.
Results on 40,000 COCO test set images. Relative Reconstruction Decline (%) shows percentage increase in reconstruction loss compared to 0% masking. Each cell for Mask 30% and Mask 70% shows absolute loss followed by percentage decline.

As intrinsic metric for **super-resolution** we evaluate reconstruction loss between noise latents for DDIM Sampling. The primary intrinsic metric in analyzing diffusion-based super-resolution is L2 Loss, which is used as the main loss to optimize these models. Therefore, we report the L2 loss of the generated and initial images for our baselines and our model.

Model	Steps 1	5	10	25
SD3	0.3708	0.4887	0.3673	0.2472
SD3.5	0.3446	0.3446	0.2950	0.2724
Ours	0.5655	0.5385	0.5534	0.5767

Table 7: A table of L2 Loss for different denoising steps

Model	**L2 loss**
Stable Diffusion 3	0.3956
Stable Diffusion 3.5	0.2660
Inf-DiT (Default)	0.3163
Inf-DiT (with Quality Text)	0.3133
Ours	0.5769

Table 6: L2 reconstruction loss between super-resolution models

We also compare the performance of our model with Stable Diffusion 3 across different number steps for denoising. While for our baseline, more denoising steps led to lower amount of loss in latents. In the start our model follows a similar pattern as SD3 such that the L2 Loss between the latents increases at first, then decreases. But it starts to increase again as we notice on steps 10 and 25. To see if this behaviour continues we also experimented with denoising for upto 1000 timesteps. The loss for 1000 steps was 0.5517, which is lower then 25, thus is continues to decreases after increasing initially.

7.2 Qualitative Analysis and Examples

7.2.1 Tokenizer Qualitative Results and Analysis

The analysis of image tokenization reveals two primary findings: the varying performance of tokenization for simple versus complex images, and the significant impact of text guidance on image reconstruction. As shown in Figure 11, single-object reconstructions are much clearer than highly textured ones. Complex images tend to introduce noticeable artefacts, which are also seen in other open-source tokenizers like DALL-E and Cosmos. These artefacts are also present in our Textok-VAE model, indicating that current tokenization methods struggle with capturing intricate textures and fine details in multi-object images.

In Figure 12, the effect of text guidance on image tokenization is demonstrated. The first row shows that the instrument is reconstructed much more clearly with text guidance, while in the second row, the details of the watch's face are improved as well. This suggests that text guidance enables the model to capture the fine details necessary for accurate reconstruction by providing contextual information for the image tokens. The benefits of text guidance are evident across the rest of the rows, where the text helps reconstruct lower-level features that enhance overall fidelity.

However, as shown in Figure 13, there are cases where text guidance fails to capture all the relevant details. In the first image, the captions fail to represent the variety of objects present, leading to incomplete reconstructions. For example, the cake and cabinet handle are poorly reconstructed and do not match the input image. In the second image, the reconstruction of a person fails to capture the gender-specific details, with the person now appearing as a man instead of a woman. These examples highlight the need for more accurate and detailed captions to fully utilize text guidance.

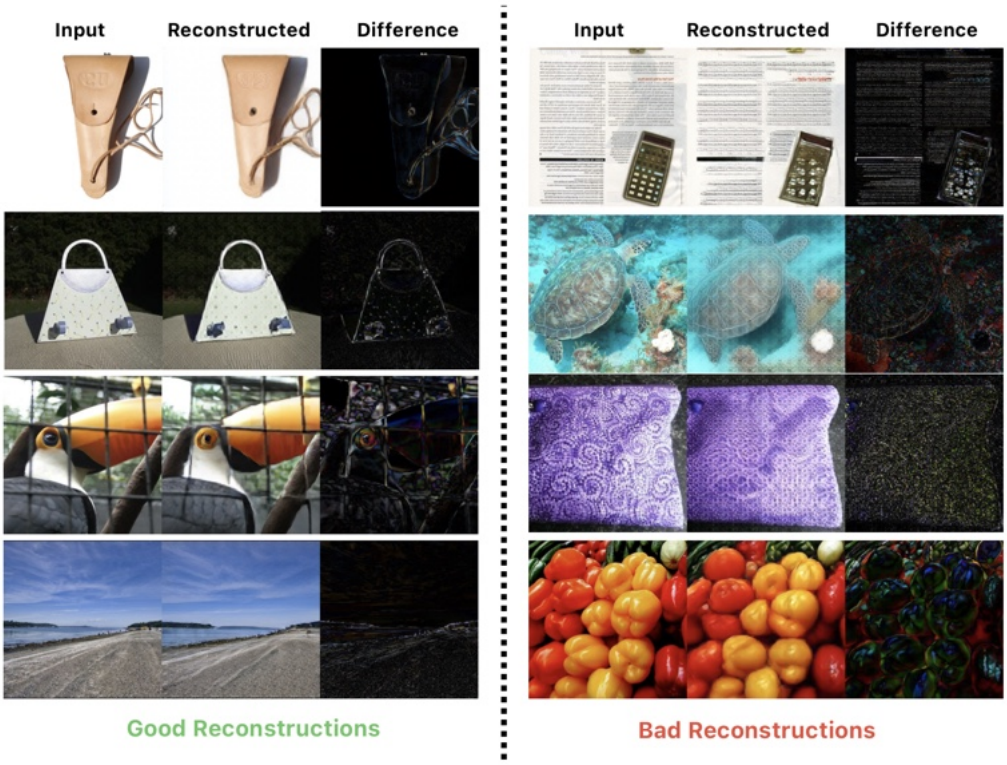


Figure 11: Reconstruction results from TexTok-VAE.
textitThis figure showcases both successful reconstructions and cases where the tokenizer struggles to encode the input into the compact 1D latent space of 32 tokens. Difference images are provided to highlight discrepancies between the reconstructed and original images, illustrating areas where the model's tokenization fails to capture fine details.

Figure 13: Limitation of current text guidiance on image tokenization. *Limited captioning of the image leads to mis-represented finer details in the detokenized iamges. These examples are discussed in more detail in section 7.2.1*

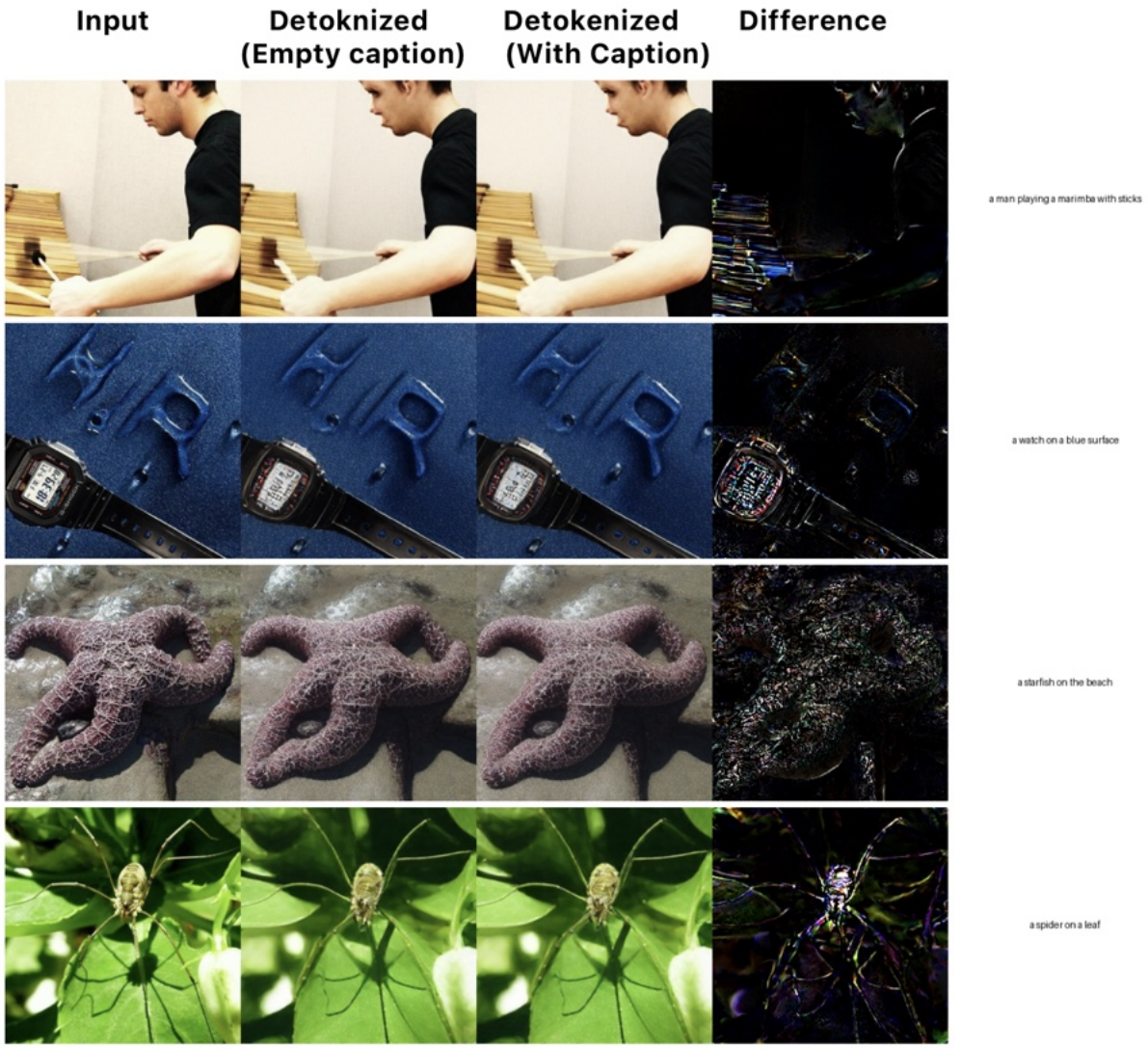


Figure 12: Effect of text guidance on image tokenization. *The "Detokenized" column displays reconstructed images with and without text guidance. The "Difference" columns highlight the texture details missed in the reconstruction when no caption is provided, emphasizing the improvements text guidance brings to capturing finer image details.*

In conclusion, while text guidance improves image tokenization by enabling the model to capture finer details, the quality of text inputs is crucial. Comprehensive and precise text descriptions can enhance the model's ability to reconstruct complex images, but current tokenization methods still face challenges in handling texture-rich and multi-object scenes. Further improvements in both the tokenization techniques and text alignment could enhance image fidelity, especially for more intricate and multi-faceted scenes.

7.2.2 Super-Resolution Qualitative Results and Analysis

Next, we can analyze the super-resolution results through a qualitative analysis of the outputs. As aforementioned, one key hyperparameter to optimize is the guidance scale, which controls how much the text guidance influences the final output. As seen in Figure 14, increased text guidance seems to worsen the model's performance. Hallucinations can be seen in the face in the third image, while a computer mouse seems to be generated in the fifth image despite the input image not having

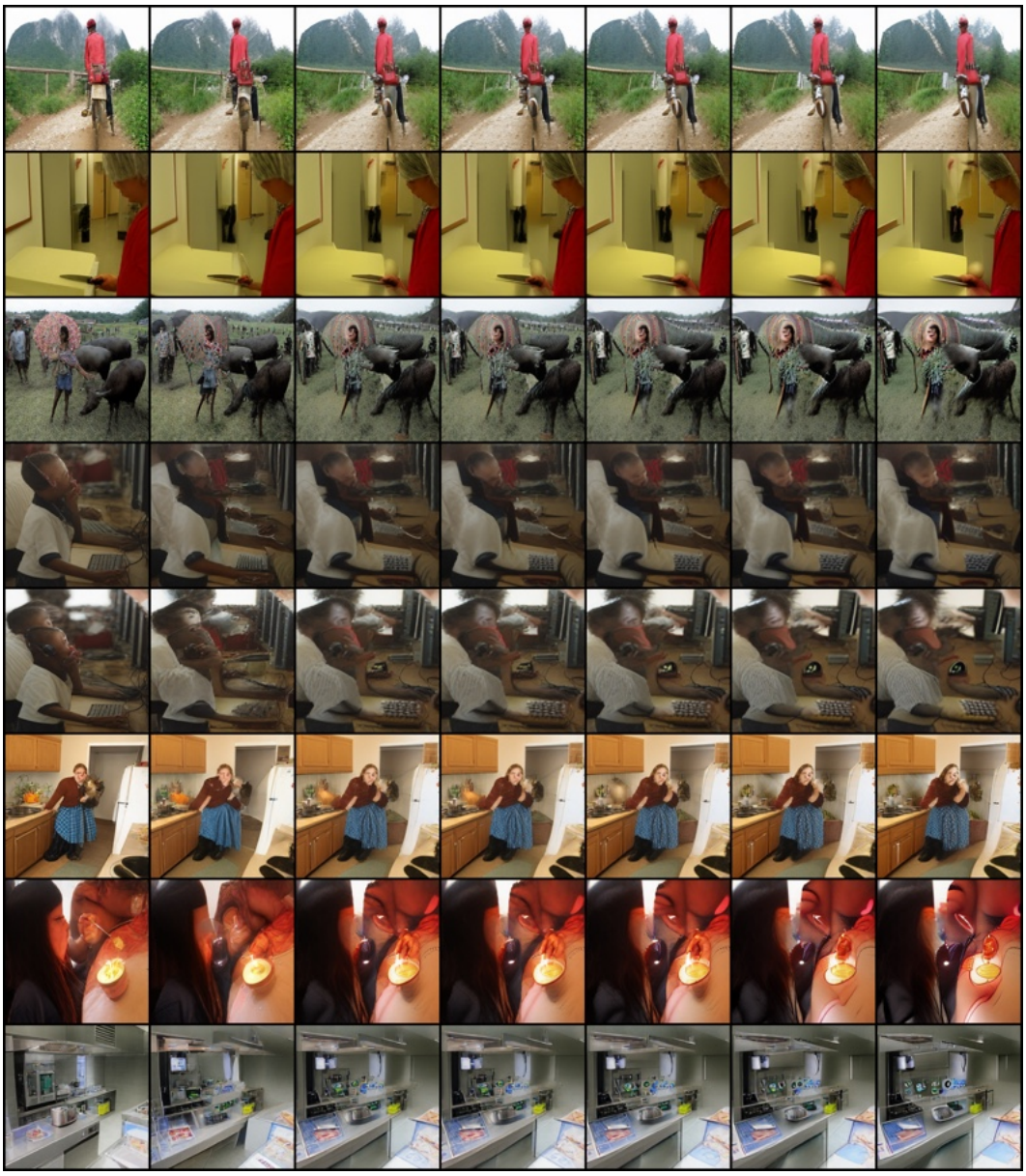


Figure 14: Evaluating the impact of guidance scale on super-resolution. **Column 1** is the output of a simple tokenization-detokenization process which we treat as the ground truth. The columns after show the output of the SR model at guidance scales $[0, 2, 4, 6, 8, 10]$.

one. The lighting is also different in the seventh image with a higher guidance scale. However, in the eight image, it is noticeable that some of the intermediate guidance scales results in improved outputs, where the items on the table and green section on the wall are not shown in the image with 0 guidance scale. This indicates that the text inputs are not very well aligned with the latent space of the images, which mostly results in hallucinations instead of reinforcing the concepts seen in the low-resolution input image. In practice, it is likely that the input text will not have a one-to-one alignment with the low-resolution input, meaning improvements will have to be made to the model to better ingest text.

Furthermore, the general qualitative outputs from Figure 15 indicates that there is a good general structural adhesion to the input image from super-resolution. When compared to the detokenized image, which indicate how much of the quality is lost to tokenization when compared to how much

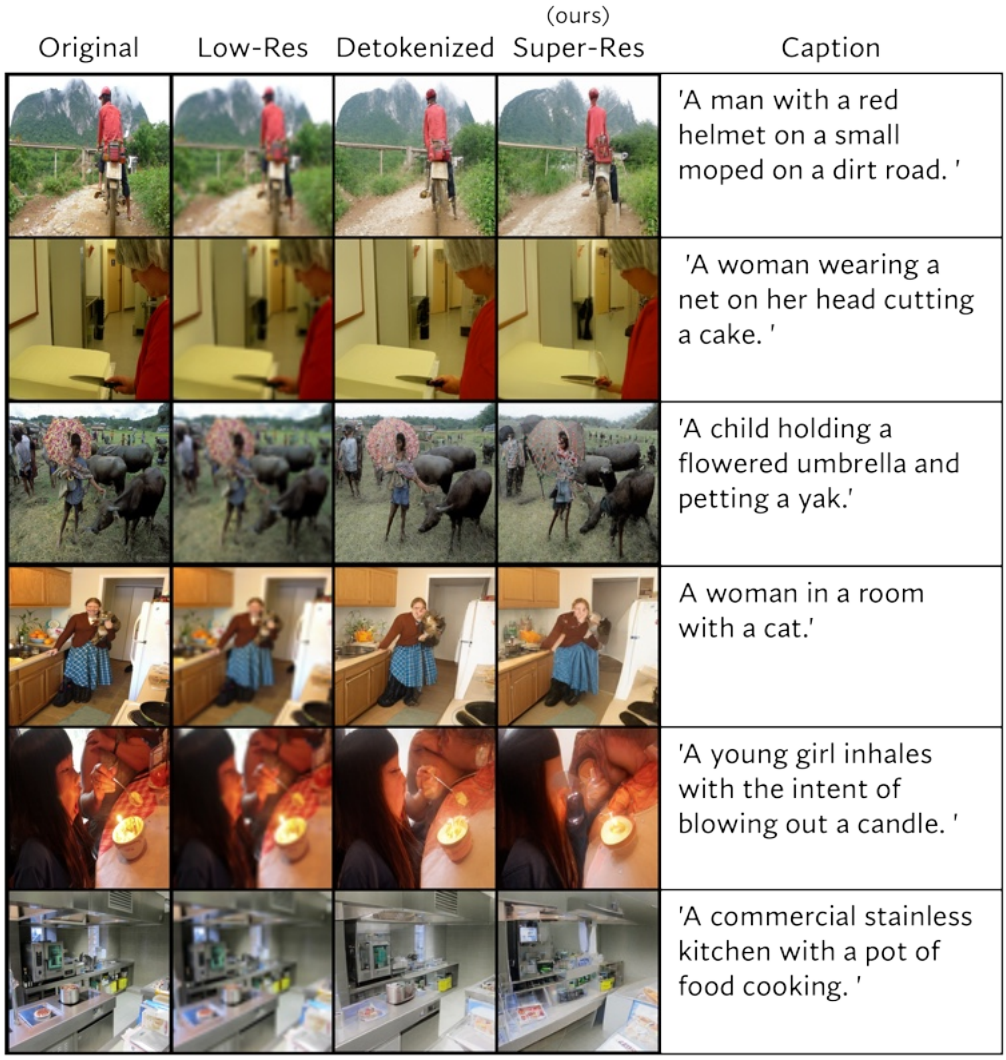


Figure 15: Qualitative results from our super-resolution models, compared to the original, low-resolution, and detokenized images, along with their respective captions.

is lost to the super-resolution itself, the super-resolution model manages to capture details about where human body parts such as hands and faces are, as well as other objects like vehicles and animals. However, there are significant issues with the quality of the generated faces and hands, and other details such as the appliances in the last image being completely changed through the super-resolution process. While increasing the guidance scale slightly can result in better results for some images, as was seen in Figure 14, there is still work to be done to ensure proper generation of fine-grained details. However, importantly, it should be noted that the qualitative results from the baselines also showed issues in detail recovery from the LR images and in the generation of complex objects and scenes, indicating that our model is able to achieve similar qualitative results while having a much higher compression ratio, and therefore compute and inference time improvements.

7.3 Attention Maps

In order to better visualize the effects of concatenating the low resolution image to the noisy latent, we visualized the self-attention maps for each layer of the model, with brighter colors indicating stronger activations between the specific token from the LR image and noisy latent. These were generated by taking the mean across attention heads for a mean of multiple noise samples.

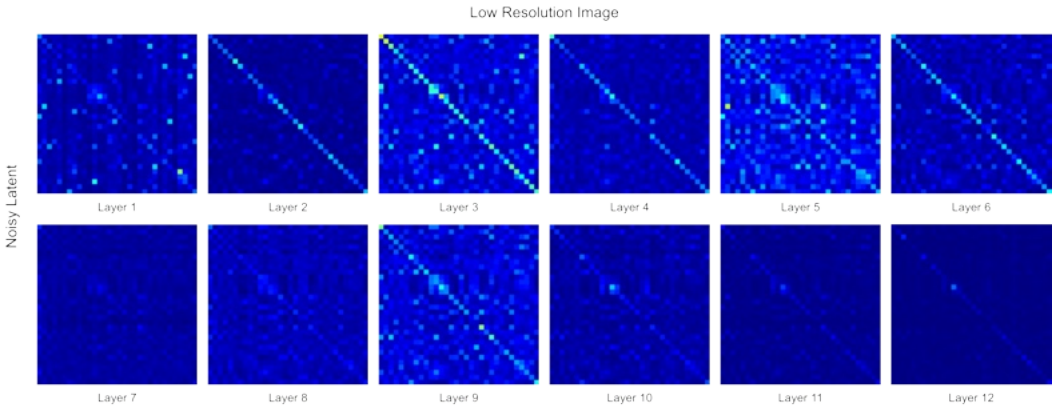


Figure 16: Self-Attention Maps for TaTiTok

As can be seen in Figure 16, the greatest activation is observed along the diagonals of most layers, which indicate that there is a significant amount of attention between the noisy latent tokens and their corresponding tokens in the concatenated LR image. However, some layers also show significant attention between different regions of the image, particularly layers 1, 3, 5, and 9, showing that these global connections enable better super-resolution performance across various semantic levels. There does not, however, seem to be a significant trend in how the attention changes from layer to layer, indicating that there is important information being gained at each layer, and that the chosen model depth is appropriate for the task.

8 Future work and Limitations

8.1 Tokenizers

We list a few limitations of the models explored in this project below. While the proposed model architectures can certainly be improved by scaling training data and compute to match competitive open-source baselines, we focus here on non-training-related limitations:

Limitations of Tokenizers: Despite the advancements presented, our model exhibits several limitations:

- **Naive Text-Image Fusion:** Currently, text and image tokens are combined using basic self-attention mechanisms. Employing vision-aligned text embeddings, particularly those aligned with the image latent space, could improve multimodal integration and enhance reconstruction quality.
- **Visual Artifacts:** Image tokenizers can introduce artifacts in reconstructions. Integrating denoising or diffusion-based steps within the decoder may enhance image fidelity.
- **Text Dominance:** Lengthy text inputs can overshadow visual information, potentially biasing the model's outputs towards textual content.
- **Fixed Token Count:** Representing every image with a fixed number of tokens (e.g., 32) limits scalability across varying resolutions. Adaptive tokenization strategies may offer better flexibility.

8.2 Super Resolution

From the previous results and analysis, numerous limitations have been found with regards to the super-resolution model and process.

- **Hallucinations** As was seen in 14, hallucinations were present in the super-resolution outputs, especially with increased text guidance scale. Future works needs to address this issue, possibly through improving the alignment of text embeddings and low-resolution image latents, or an updated architecture that further prioritizes information from the low-resolution image.
- **Lack of Quality in Fine-Grained Details** The super-resolution outputs from 15 further indicate issues in generating fine-grained details such as human faces and hands. While the text modality was intended to address this issue, more work needs to be done to better integrate it. Additionally, tests with larger and more recent model architectures could improve these results.

9 Ethical Concerns and Considerations (unintentional, malicious, and dual-use)

There are numerous ethical concerns with the use of such image super-resolution models. It is important that users take care to consider them and the consequences of using such models in unethical manners.

- **Creation of Harmful Content** Image super-resolution models like this one come without rigorous controls to prevent the creation of content that can be classified as explicit, inappropriate or harmful to certain parties or individuals. Examples of this can include the unintentional creation of such content, such as in situations where the diffusion model generates content that does not fit the low-resolution input and may be inappropriate in nature, or the intentional super-resolution of harmful, inappropriate or violent content, with the intention to distribute such images.
- **Forgery or Manipulation** There is risk of image super-resolution models being used to intentionally produce high-quality images that can be used for the purposes of forgery or other manipulation. Examples include domains such as biomedical images, political or educational images or content, and official documents, all for the purpose of intentionally misleading its audience.
- **Deepfakes** One key issue in recent years has been the creation of deepfakes that can be used without authorization for purposes including disinformation and other misuse. Such image super-resolution models can be used for the creation of deepfakes, making them more convincing through effective super-resolution, and therefore raises privacy concerns.

A series of solutions are available to mitigate these concerns in practice:

- **Ethical and NSFW Filters** Crucial to the safe use of such models is the deployment of filters that can prevent users from inputting text prompts and low-resolution input images that can potentially be inappropriate or violate ethical use standards. Additionally, a filter must also be put into place to restrict the model from returning outputs that are inappropriate.
- **User Awareness** Also key in the prevention of abuse of super-resolution models is making users aware of the potential harmful uses, and especially users who intend to fine-tune or otherwise edit the model, as they can expose other users to these effects through their models. Furthermore, informing the general public about the potential of diffusion model misuse is equally important, ensuring that they aren't mislead by users who exploit these models for unethical purposes.
- **Watermarks** Another avenue is the intentional inclusion of watermarks that can indicate that the output image has been generated by an AI model, preventing its misuse for disinformation or misleading purposes.

References

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

Roman Bachmann, Jesse Allardice, David Mizrahi, Enrico Fini, Oğuzhan Fatih Kar, Elmira Amirloo, Alaaeldin El-Nouby, Amir Zamir, and Afshin Dehghan. Flextok: Resampling images into 1d token sequences of flexible length, 2025. URL https://arxiv.org/abs/2502.13967.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow

transformers for high-resolution image synthesis, 2024. URL https://arxiv.org/abs/2403.03206.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. Datacomp: In search of the next generation of multimodal datasets, 2023. URL https://arxiv.org/abs/2304.14108.

Ju He, Qihang Yu, Qihao Liu, and Liang-Chieh Chen. Flowtok: Flowing seamlessly across text and image tokens, 2025. URL https://arxiv.org/abs/2503.10772.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020b. URL https://arxiv.org/abs/2006.11239.

Dongwon Kim, Ju He, Qihang Yu, Chenglin Yang, Xiaohui Shen, Suha Kwak, and Liang-Chieh Chen. Democratizing text-to-image masked generative models with compact text-aware one-dimensional tokens, 2025a. URL https://arxiv.org/abs/2501.07730.

Dongwon Kim, Ju He, Qihang Yu Yu, Chenglin Yang, Xiaohui Shen, Suha Kwak, and Chen Liang-Chieh. Democratizing text-to-image masked generative models with compact text-aware one-dimensional tokens. *arXiv preprint arXiv:2501.07730*, 2025b.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022. URL https://arxiv.org/abs/1312.6114.

Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network, 2017. URL https://arxiv.org/abs/1609.04802.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. URL https://arxiv.org/abs/1405.0312.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: Vq-vae made simple, 2023. URL https://arxiv.org/abs/2309.15505.

NVIDIA, :, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, Jiaojiao Fan, Michele Fenzi, Francesco Ferroni, Sanja Fidler, Dieter Fox, Songwei Ge, Yunhao Ge, Jinwei Gu, Siddharth Gururani, Ethan He, Jiahui Huang, Jacob Huffman, Pooya Jannaty, Jingyi Jin, Seung Wook Kim, Gergely Klár, Grace Lam, Shiyi Lan, Laura Leal-Taixe, Anqi Li, Zhaoshuo Li, Chen-Hsuan Lin, Tsung-Yi Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Arsalan Mousavian, Seungjun Nah, Sriharsha Niverty, David Page, Despoina Paschalidou, Zeeshan Patel, Lindsey Pavao, Morteza Ramezanali, Fitsum Reda, Xiaowei Ren, Vasanth Rao Naik Sabavat, Ed Schmerling, Stella Shi, Bartosz Stefaniak, Shitao Tang, Lyne Tchapmi, Przemek Tredak, Wei-Cheng Tseng, Jibin Varghese, Hao Wang, Haoxiang Wang, Heng Wang, Ting-Chun Wang, Fangyin Wei, Xinyue Wei, Jay Zhangjie Wu, Jiashu Xu, Wei Yang, Lin Yen-Chen, Xiaohui Zeng, Yu Zeng, Jing Zhang, Qinsheng Zhang, Yuxuan Zhang, Qingqing Zhao, and Artur Zolkowski. Cosmos world foundation model platform for physical ai, 2025. URL https://arxiv.org/abs/2501.03575.

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.

William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023. URL https://arxiv.org/abs/2212.09748.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pp. 8821–8831. Pmlr, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs, 2021. URL https://arxiv.org/abs/2111.02114.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning, 2018. URL https://arxiv.org/abs/1711.00937.

Junke Wang, Yi Jiang, Zehuan Yuan, Bingyue Peng, Zuxuan Wu, and Yu-Gang Jiang. Omnitokenizer: A joint image-video tokenizer for visual generation. *Advances in Neural Information Processing Systems*, 37:28281–28295, 2025.

Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Chen Change Loy, Yu Qiao, and Xiaoou Tang. Esrgan: Enhanced super-resolution generative adversarial networks, 2018. URL https://arxiv.org/abs/1809.00219.

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data, 2021. URL https://arxiv.org/abs/2107.10833.

Wilson Yan, Volodymyr Mnih, Aleksandra Faust, Matei Zaharia, Pieter Abbeel, and Hao Liu. Elastictok: Adaptive tokenization for image and video, 2025. URL https://arxiv.org/abs/2410.08368.

Zhuoyi Yang, Heyang Jiang, Wenyi Hong, Jiayan Teng, Wendi Zheng, Yuxiao Dong, Ming Ding, and Jie Tang. Inf-dit: Upsampling any-resolution image with memory-efficient diffusion transformer. In *European Conference on Computer Vision*, pp. 141–156. Springer, 2024.

Qihang Yu, Mark Weber, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. An image is worth 32 tokens for reconstruction and generation, 2024. URL https://arxiv.org/abs/2406.07550.

Kaiwen Zha, Lijun Yu, Alireza Fathi, David A Ross, Cordelia Schmid, Dina Katabi, and Xiuye Gu. Language-guided image tokenization for generation. *arXiv preprint arXiv:2412.05796*, 2024.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. URL https://arxiv.org/abs/2302.05543.

A Appendix

You may include other additional sections here.