CoSimGen: Controllable diffusion model for simultaneous image and segmentation mask generation

Anonymous Author(s)
Affiliation
Address
email

Abstract

Generating paired images and segmentation masks remains a core bottleneck in data-scarce domains such as medical imaging and remote sensing, where manual annotation is expensive, expertise-dependent, and ethically constrained. Existing generative approaches typically handle image or mask generation in isolation and offer limited control over spatial and semantic outputs. We introduce CoSimGen, a diffusion-based framework for controllable simultaneous generation of images and segmentation masks. CoSimGen integrates multi-level conditioning via (1) class-grounded textual prompts enabling hot-swapping of input control, (2) spatial embeddings for contextual coherence, and (3) spectral timestep embeddings for denoising control. To enforce alignment and generation fidelity, we combine contrastive triplet loss between text and class embeddings with diffusion and adversarial objectives. Low-resolution outputs (128 × 128) are super-resolved to 512×512, ensuring high-fidelity synthesis. Evaluated across five diverse datasets, CoSimGen achieves state-of-the-art performance in FID, KID, LPIPS, and Semantic-FID, with KID as low as 0.11 and LPIPS of 0.53. Our method enables scalable, controllable dataset generation and advances multimodal generative modeling in structured prediction tasks.

1 Introduction

Creating large-scale paired datasets of images and segmentation masks is a major bottleneck in domains like medical imaging [1], geospatial analysis [2], autonomous driving [9], and surgical AI [22]. Manual annotation is costly, domain-specific, and often ethically constrained. While generative models such as VAEs [18], GANs [11], and diffusion models [32, 14] have advanced image synthesis, most methods generate either images [25, 17] or masks [19, 7], not both. Simultaneous image-mask generation remains underexplored, especially with flexible, controllable conditioning—critical for simulation, data augmentation, and rare-case modeling.

We present **CoSimGen**, a diffusion-based framework for **Co**ntrollable **S**imultaneous **i**mage and segmentation **m**ask **Gen**eration. CoSimGen supports conditioning on either class labels or natural language prompts and unifies multimodal control in a single generation process (Figure 1). Built upon a Conditional Denoising Diffusion Probabilistic Model (DDPM) with a U-Net backbone [29], CoSimGen introduces two novel components: *(1) Spectron*: a spatio-spectral embedding fusion module that injects class and timestep embeddings into the network. Here, class features are fused spatially to guide object placement and structure; timestep embeddings are fused along channels to model denoising dynamics; and *(2) Textron*: a text-grounded semantic conditioning module that aligns class embeddings with language embeddings, enabling text prompts to be "hot-swapped" in place of class labels during inference. Contrastive learning aligns the embedding spaces to achieve this, as inspired by CLIP [24].

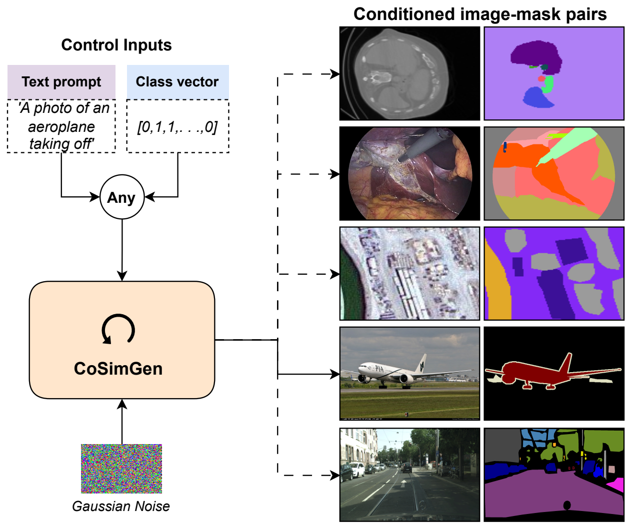

Figure 1: CoSimGen generates paired image and segmentation mask from a class or text prompt. Outputs are high-resolution, semantically aligned, and spatially coherent.

The training objective combines three losses: diffusion loss for denoising, contrastive triplet loss for text-label alignment, and adversarial loss for realism. We first generate low-resolution outputs (128×128) and super-resolve them to (512×512), maintaining fidelity and semantic alignment between image and mask. CoSimGen is evaluated on five diverse datasets—CholecSeg8k [15], BTCV [10], Cityscapes [6], PASCAL VOC [8], and MBRSC [16]—spanning surgical, medical, urban, natural, and satellite imagery domains. We use Fréchet Inception Distance (FID) [13], Kernel Inception Distance (KID) [4], Inception VGG Distance (VGG-D), Learned Perceptual Image Patch Similarity (LPIPS) [34], and Semantic FID (sFID) [3] to assess image quality and semantic realism. For mask-image alignment fidelity, we use sFID and Positive Predictive Value (PPV). CoSimGen outperforms strong baselines across all metrics and datasets. It enables controllable, high-resolution, semantically consistent generation of annotated data, offering a scalable alternative to manual labeling. Our approach is especially valuable in high-stakes domains like surgical training and medical diagnosis, where precise region-level control is essential. Furthermore, CoSimGen can serve as a generative pretraining tool, supporting domain adaptation and low-resource learning setups.

By introducing a general-purpose, controllable framework for paired data generation, CoSimGen addresses fundamental challenges in structured prediction and multimodal generative modeling—core areas of interest in multi-modal generative AI integrating vision and language modeling.

2 Related Work

Image generation from single to multimodal synthesis. Generative modeling has progressed from VAEs [18] and GANs [12] to diffusion models [14, 28], which now dominate high-fidelity image synthesis. Conditional GANs like Pix2Pix [17] and StyleGAN variants introduced structure via paired inputs or style codes. Diffusion-based methods, such as DALL·E [25], Imagen [30], and Surgical Imagen [23], enabled text-to-image synthesis with improved realism, but focus primarily on single-modality outputs.

Segmentation and data synthesis. To overcome annotation bottlenecks in segmentation tasks, works like Text4Seg [19] and SegGen [7] generate segmentation masks from text. Yet, they decouple image and mask generation or lack generalizability. Medical approaches such as HVAE [5] generate paired data but offer limited control and domain scope.

Paired image-mask generation. SimGen [3] and DiffuMask [33] explore joint generation of image and mask, but are either domain-restricted or conditionally limited. OVDiff [21] relies on fixed vocabularies for segmentation post-generation. Prior work lacks a unified, controllable model for simultaneous image-mask synthesis across domains. Our work addresses this gap with fine-grained conditionality, enabling scalable, multimodal generation from text or class vectors.

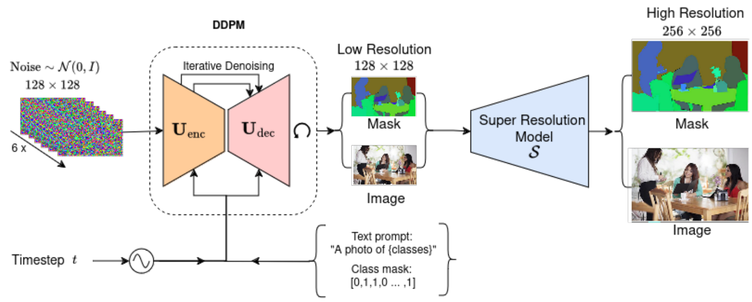

Figure 2: Architecture of CoSimGen for conditional generation of paired image-mask.

3 Methods

Our goal is to generate paired images and semantic segmentation masks, guided by user input prompts, for synthetic data creation and educational purposes. To achieve this, we propose a Controllable Simultaneous Image-Mask Generator (CoSimGen), a diffusion-based framework that utilizes contrastive learning to seamlessly condition generation on text or class labels, ensuring precise alignment between images and masks in a unified process.

3.1 Task Formalization

Let $\mathcal{D} = \{(\mathbf{X}_i, \mathbf{y}_i)\}_{i=1}^{n}$, where $\mathbf{X}_i \in \mathbb{R}^{C\times H\times W}$ represents an image, and $\mathbf{y}_i \in \{0,1\}^{H\times W}$ is the corresponding segmentation mask. The mask $\mathbf{y}_i$ contains the segmented objects, which are associated with class labels from $\mathcal{C} = \{c_1, c_2, \ldots, c_k\}$. The conditioning vector $\mathbf{c}_i$ is derived from the mask $\mathbf{y}_i$ and encodes the present classes. Alternatively, a text prompt $\mathbf{t}_i$ can be provided as a caption of the image, limited to describing only the objects present in the segmentation mask. The task is to train a model $\mathcal{M}$ that generates image-mask pairs $(\hat{\mathbf{X}}, \hat{\mathbf{y}})$ simultaneously, conditioned on the class labels $\hat{\mathbf{c}}$ and/or the text prompt $\hat{\mathbf{t}}$, such that the generated segmentation mask $\hat{\mathbf{y}}$ aligns with the generated image $\hat{\mathbf{X}}$ and both are similar to real samples from $\mathcal{D}$. The goal is to maximize the likelihood of generating paired data that closely resembles real samples, conditioned on the class labels or the textual description of the segmented objects.

3.2 Model Architecture

The proposed *CoSimGen* is built on a denoising diffusion process that iteratively refines noisy inputs into clean, coherent, paired image-mask samples. As illustrated in Fig. 2, the architecture comprises three main modules: (a) a **low-resolution (LR) generator** that establishes semantic alignment between the image, segmentation mask, and conditioning input prompt, (b) a **conditioning mechanism** that integrates text/class and timestep embeddings to enable flexible, user-controllable generation, and (c) a **super-resolution (SR) module** that upscales the LR coarse outputs into high-resolution (HR) spatial dimensions while preserving alignment and fidelity.

3.2.1 Low-Resolution Image-Mask Generation

CoSimGen employs a U-Net [29] diffusion backbone that leverages a Conditional Denoising Diffusion Probabilistic Model (DDPM) [14]. The U-Net consists of an encoder $\mathbf{U}_{\text{enc}}$ and decoder $\mathbf{U}_{\text{dec}}$ connected via two residual skip connections at each level and trained to denoise a noisy input $\mathbf{X}_t$ over multiple timesteps t. The U-Net is conditioned on semantic (class or text) and timestep embeddings via the mechanisms defined in Sec. 3.2.2, enabling the model to adaptively align its features with both context and noise-level information, guiding the network to produce class-consistent outputs during denoising. This backbone takes as input:

1. A noisy input $\mathbf{X}_t$, representing a corrupted image-mask pair at diffusion step t
2. A binary class mask $\mathbf{M} \in \{0,1\}^c$, indicating the queried classes
3. A text prompt $\mathbf{Z}_q$, e.g., *"A photo of {class}"*
4. The timestep value t, representing the current noise level.

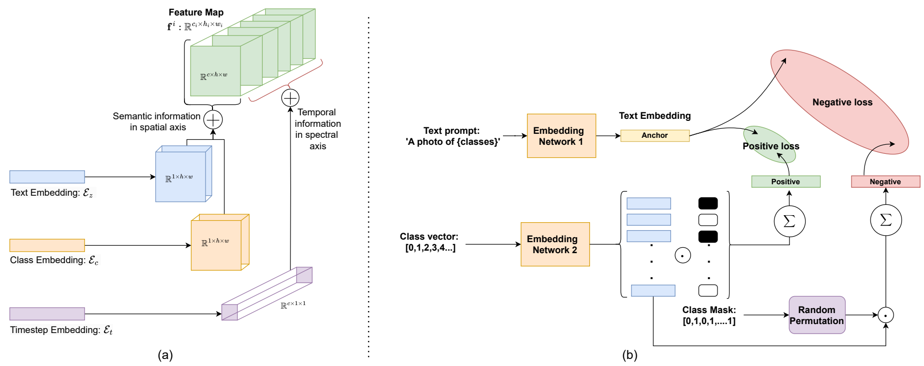

Figure 3: Conditioning mechanisms used in CoSimGen showing: (a) Spectron: spatio-spectral embedding fusion for semantic and temporal conditioning, and (b) Textron: contrastive alignment for interchangeable class/text conditioning.

3.2.2 Conditioning Mechanisms

Text encoder. The text encoder $\mathcal{E}_z$ processes input text $\mathbf{Z}_q$ using a frozen sentence transformer [27], followed by projection layers that embed it into a D-dimensional space: $\mathbf{Z}_{\text{emb}} \in \mathbb{R}^{1\times D} = \mathcal{E}_z(\mathbf{Z}_q, \theta_T)$

Class encoder. Given a binary class mask $\mathbf{M} \in \{0,1\}^c$, the class encoder $\mathcal{E}_c$ multiplies $\mathbf{M}$ with a learnable weight matrix $\mathbf{W}_c \in \mathbb{R}^{c\times d}$ effectively selecting the active class embeddings and sums across classes: $\mathbf{C}_{\text{feat}} = \sum_{i=1}^{c} \mathbf{M}_i \cdot \mathbf{W}_c[i,:]$. The $\mathbf{C}_{\text{feat}}$ passes through a series of linear transformations projecting it into a D-dimensional class embedding: $\mathbf{C}_{\text{emb}} \in \mathbb{R}^{1\times D} = \mathcal{E}_c(\mathbf{M}; \theta_{\mathcal{E}_c})$

Timestep encoder. The timestep encoder $\mathcal{E}_t$ maps scalar timestep t into a D-dimensional embedding via sinusoidal positional encoding followed by MLP projections: $\mathbf{T}_{\text{emb}} \in \mathbb{R}^{1\times D} = \mathcal{E}_t(t, \theta_{\mathcal{E}_t})$

These conditional embeddings ($\mathbf{Z}_{\text{emb}}, \mathbf{C}_{\text{emb}}, \mathbf{T}_{\text{emb}}$) are intuitively injected into the model via two approaches which we proposed in this work:

(a) **Spatio-spectral embedding fusion (Spectron):** In traditional generative models, conditioning feedback is applied by direct concatenation along the latent space or by adding conditions to feature maps along the channel axis. Diffusion models often follow a similar approach, where the conditional embeddings and timestep embeddings are introduced by adding them to the channel axis of the feature maps. While effective, this approach does not fully exploit the semantic richness of the conditional embeddings.

To bridge this gap, we introduce *Spectron* (Fig. 3a), a strategy that injects conditions into feature maps at all resolutions, allowing for a more intuitive conditioning process. Recognizing that class conditions, such as the class embedding $\mathbf{C}_{\text{emb}}$, represent a semantic understanding of the image and mask, we propose spatially embedding this information. This semantic representation governs the shape, outline, and textures within the generated image and mask. Therefore, it is intuitively powerful to apply the semantic conditional vectors along the spatial dimensions, thereby spatially conditioning the features $\mathbf{f}$ at each resolution i:

$$\mathbf{f}_{\text{cond}}^{i,\text{spatial}} = \mathbf{f}^i + \mathbf{C}_{\text{emb}}^i \tag{1}$$

where $\mathbf{f}^i : \mathbb{R}^{c_i\times h_i\times w_i}$ and $\mathbf{C}_{\text{emb}}^i : \mathbb{R}^{1\times h_i\times w_i}$ thus adding the conditional embedding in the spatial dimension.

The timestep embedding $\mathbf{T}_{\text{emb}}$, by contrast, encodes the noisiness of the input, and the noise level is assumed to affect all channels uniformly and equally. Hence, it becomes intuitive to apply the timestep conditioning along the channel dimension of the noisy feature maps, thereby spectrally conditioning the features $\mathbf{f}$ at each resolution i:

$$\mathbf{f}_{\text{cond}}^{i,\text{spectral}} = \mathbf{f}_{\text{cond}}^{i,\text{spatial}} + \mathbf{T}_{\text{emb}}^i \tag{2}$$

where $\mathbf{f}_{\text{enc}}^{i,spatial} : \mathbb{R}^{c_i\times h_i\times w_i}$ and $\mathbf{T}_{\text{emb}}^i : \mathbb{R}^{c_i\times 1\times 1}$. By combining these two perspectives, Spatio-Spectral Feature Mixing enables both semantic and temporal feedback, allowing $\mathbf{U}$ to generate features that are spatially aligned with the condition semantics and spectrally aligned with the temporal noise level. This dual conditioning mechanism ensures that the U-Net captures a deep alignment between the class condition and timestep information across all spatial and spectral dimensions, enhancing the model's generative capabilities.

(b) **Text-grounded class conditioning (Textron):** While class embeddings $\mathbf{C}_{\text{emb}}$ can condition image and mask generation independently, , they lack the flexibility to allow inference on text inputs directly. To address this, *Textron* aligns class embeddings with their corresponding text embeddings during training, enabling the model to accept either class or text embeddings interchangeably during inference. This is achieved by learning a shared embedding space where class and text representations are closely aligned.

Conventional generative models condition on text by learning a similarity metric between text and generated image features. While effective for evaluating alignment, this approach does not enable *direct substitution (or "hot-swapping")* of class embeddings with text embeddings during inference. Textron overcomes this limitation by introducing a contrastive triplet loss (Eq. 3) that explicitly aligns class and text embeddings. Given a text embedding $\mathbf{Z}_{\text{emb}}$ (anchor), the corresponding class embedding $\mathbf{C}_{\text{emb}}$ (positive), and a mismatched class embedding $\tilde{\mathbf{C}}_{\text{emb}}$ (negative), the loss is defined as:

$$\mathcal{L}_{\text{triplet}} = \max\left(0, \|\mathbf{Z}_{\text{emb}} - \mathbf{C}_{\text{emb}}\|^2 - \|\mathbf{Z}_{\text{emb}} - \tilde{\mathbf{C}}_{\text{emb}}\|^2 + \alpha\right) \tag{3}$$

This objective encourages the model to reduce the distance between matched text-class pairs while pushing apart mismatched ones, with margin α controlling the separation. As illustrated in Fig. 3(b), the text embedding $\mathbf{Z}_{\text{emb}}$ serves as the anchor; the corresponding class embedding $\mathbf{C}_{\text{emb}}$ is used as the positive, and a randomly selected class embedding $\tilde{\mathbf{C}}_{\text{emb}}$ forms the negative. By minimizing this loss, the model learns a unified embedding space where class and text representations are close, enabling the class encoder to be replaced with a text encoder during inference. This design allows the model to leverage the semantic richness of natural language, supporting flexible and efficient text-grounded generation.

3.2.3 Super-Resolution Module

Upscaling strategy. To enhance visual fidelity, the coarse outputs $\mathbf{X}_{\text{LR}} \in \mathbb{R}^{6\times128\times128}$ are passed to a super-resolution model ($\mathcal{SR}$). We adopt an Efficient Sub-Pixel CNN (ESPCNN) [31] as the $\mathcal{SR}$ with an upscale factor of 2.

Training objective. During training, Gaussian noise is added to the ground-truth low-resolution input:

$$\tilde{\mathbf{X}}^{\text{gt}}_{\text{LR}} = \mathbf{X}^{\text{gt}}_{\text{LR}} + \epsilon, \quad \mathcal{SR}(\tilde{\mathbf{X}}^{\text{gt}}_{\text{LR}}) \to \mathbf{X}^{\text{gt}}_{\text{HR}} \in \mathbb{R}^{6\times2h\times2w}$$

The SR model is trained progressively from $128 \times 128 \to 256 \times 256 \to 512 \times 512$, supporting multi-scale inference.

4 Experiments

Datasets. We evaluate on five diverse segmentation datasets: Cityscapes [6], PASCAL VOC [8], MBRSC [16], BTCV [10], and CholecSeg8k [15], covering general, remote sensing, radiology, and surgical domains. To enhance class separability, segmentation masks are mapped to a uniformly spaced Fibonacci RGB (F-RGB) space using a golden angle transformation.

Implementation. CoSimGen is trained at 128×128 resolution, with a residual U-Net backbone ($d = 64$, multipliers $1 : 8$). A separate super-resolution module ($\mathcal{SR}$) scales outputs to 256^2 and 512^2. We use Adam optimizer (lr=2×10^{-4}, batch size 24), PyTorch mixed-precision training, and NVIDIA H100 GPUs running for under 72 training hours.

Loss Functions. CoSimGen is optimized using a combination of noise prediction, alignment, adversarial, and perceptual objectives. The core diffusion model minimizes a conditional noise reconstruction loss $\mathcal{L}_{\text{diff}}$, guided by timestep t, text embedding $\mathbf{T}_{\text{emb}}$, and class embedding $\mathbf{C}_{\text{emb}}$. To enforce semantic alignment, we introduce a triplet loss $\mathcal{L}_{\text{triplet}}$ that attracts $\mathbf{T}_{\text{emb}}$ to $\mathbf{C}_{\text{emb}}$ and repels it from a negative class embedding. Realism of generated image-mask pairs is encouraged via an adversarial loss $\mathcal{L}_{\text{adv}}$ against a frozen discriminator. For high-resolution refinement, a super-resolution loss $\mathcal{L}_{\text{SR}}$ minimizes a weighted combination of MSE and perceptual loss. The full objective

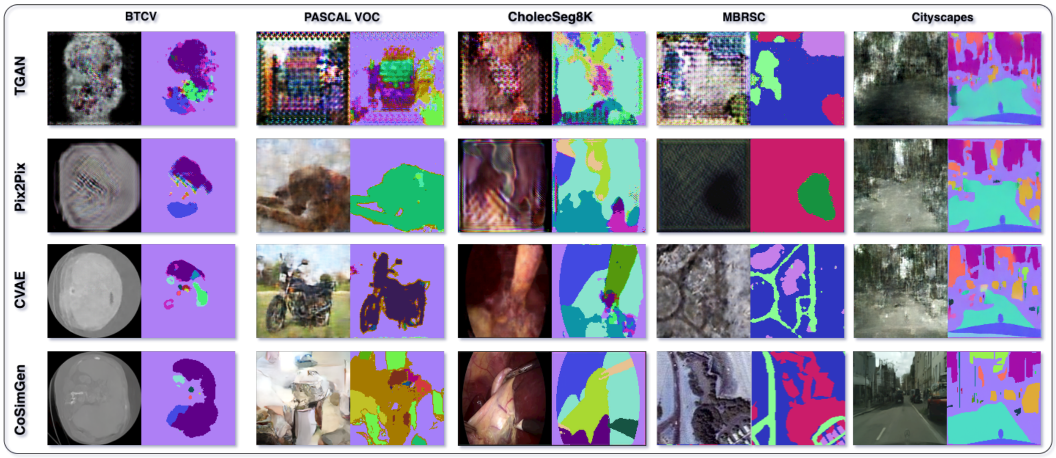

Figure 4: Qualitative comparisons of generated image-mask pairs (low-resolution outputs)

is: $\mathcal{L}_{\text{CoSimGen}} = \mathcal{L}_{\text{diff}} + \mathcal{L}_{\text{triplet}} + \beta \cdot \mathcal{L}_{\text{adv}} + \mathcal{L}_{\text{SR}}$. We provide detailed formulations for each loss component in Appendix.

Baselines. Given the novel challenge of entangled generation of image-mask pairs, we evaluate both regression-based and adversarial generative approaches. For adversarial methods, we adapt classical frameworks, including TGAN [26] and Pix2PixGAN [17], to handle dual outputs. As a regression-based baseline, we employ a conditional convolutional Variational Autoencoder (CVAE) [18] with joint image-mask reconstruction. All baselines are tuned for fairness across datasets. While recent generative models have shown strong results in single-modality settings (e.g., image-only or segmentation-only), directly applying them to the coupled image-mask generation task proves non-trivial. We observed that such models require extensive architectural modifications to jointly handle continuous and discrete modalities—often reducing them to their foundational counterparts. Thus, our baseline comparisons focus on principled, extensible variants of standard models, reflecting a fair and interpretable benchmark for CoSimGen.

Evaluation Protocols. We report Fréchet Inception Distance (FID) [13], Kernel Inception Distance (KID) [4], Learned Perceptual Image Patch Similarity (LPIPS) [34], and Inception VGG Distance (VGG-D) between real and generated images to assess visual realism. To evaluate mask-image alignment and regional generation fidelity, we compute *semantic FID* (sFID) [3], which extend the traditional FID to class-specific image regions guided by generated masks. These metrics quantify how well generated regions match target semantics. Further details on sFID are provided in Appendix. Conditioning quality is measured via Positive Predictive Value (PPV) of queried classes in the generated masks. For the super-resolution module $\mathcal{SR}$, we evaluate reconstruction quality by training on generated low- and high-resolution pairs and testing on low-resolution inputs.

5 Results

5.1 Image Quality

We evaluate the quality of generated image-mask pairs both qualitatively and quantitatively. Fig. 4 provides visual comparisons between CoSimGen and baseline models — CVAE, TGAN, and Pix2PixGAN — across diverse datasets. CoSimGen consistently produces sharper, crisper, more coherent images with structurally aligned masks, particularly excelling in datasets with ample training data. This fidelity is especially evident in complex domains like surgical scenes and urban layouts, where both visual detail and semantic alignment are critical.

While CVAE performs competitively on smaller datasets such as PASCAL VOC, reflecting its advantage in low-data regimes, CoSimGen significantly outperforms all baselines on larger-scale datasets, including Cityscapes, CholecSeg8k, MBRSC, and BTCV. This performance gap underscores the scalability of CoSimGen, driven by its design to handle high-resolution structures and complex spatial distributions.

Table 1: Evaluation of the fidelity of the generated images across 3 datasets in comparison with the baselines across four metrics: FID, KID, VGG distance, and LPIPS distance.

Model	Pascal VOC				MBRSC				BTCV			
	FID	KID	VGG-D	LPIPS-D	FID	KID	VGG-D	LPIPS-D	FID	KID	VGG-D	LPIPS-D
TGAN	348.19	0.29	221.53	0.77	394.86	0.27	113.09	0.72	394.05	0.53	146.31	0.60
Pix2PixGAN	348.05	0.30	225.56	0.79	410.18	0.34	117.94	0.70	284.60	0.36	152.80	0.54
CVAE	337.41	0.35	**204.97**	0.76	326.16	0.27	**106.79**	0.70	192.21	0.19	144.84	0.45
CoSimGen (Ours)	**206.29**	**0.20**	227.64	**0.74**	**203.67**	**0.11**	110.43	**0.63**	**159.92**	**0.13**	**139.35**	**0.53**

Table 2: Comparative evaluation of the generated mask-image alignment in 4 datasets across two metrics: semantic fréchet inception distance (sFID) and positive predicted value (PPV).

Model	Cityscapes		Pascal VOC		MBRSC		BTCV	
	sFID	PPV	sFID	PPV	sFID	PPV	sFID	PPV
TGAN	345.29±39.21	0.68±0.08	348.21±54.32	**0.98±0.02**	435.12±20.87	0.43±0.02	405.07±23.52	0.50±0.09
Pix2PixGAN	128.12±12.09	0.92±0.03	**326.66±87.23**	0.81±0.06	462.74±19.15	0.84±0.06	323.26±28.25	0.43±0.04
CVAE	204.27±9.82	0.86±0.10	381.53±29.11	0.90±0.07	422.80±33.10	**0.91±0.08**	250.52±17.67	**0.56±0.09**
CoSimGen (Ours)	**54.08±8.93**	**0.98±0.01**	343.66±10.89	0.78±0.01	**294.66±12.20**	0.87±0.00	**198.74±5.68**	0.35±0.00

The visual analysis in Fig. 4 reveals key qualitative distinctions. TGAN and Pix2PixGAN, while visually plausible in isolated cases, suffer from mode collapse, texture artifacts, and misaligned masks. In contrast, CoSimGen preserves semantic boundaries with high fidelity, generating clinically plausible, spatially consistent outputs. Compared to CVAE, CoSimGen offers crisper structural detail and greater diversity in textures—especially visible in Cityscapes and surgical datasets.

Quantitative results in Table 1 reinforce these findings. CoSimGen achieves the lowest FID, KID, and LPIPS scores on four of the five datasets (MBRSC, BTCV, CholecSeg8k, Cityscapes), demonstrating superior realism and perceptual quality. Notably, our model sets a new benchmark on BTCV and MBRSC in both image quality and semantic fidelity. For radiology datasets, where accurate anatomical delineation is essential, CoSimGen maintains high mask-image coherence—outperforming adversarial models and matching or exceeding CVAE.

Overall, CoSimGen delivers state-of-the-art visual and semantic quality across a wide range of image domains, capturing both global realism and fine-grained structural alignment. Additional samples and visual comparisons are provided in the Appendix.

5.2 Image-Mask Alignment

Table 2 reports the alignment quality between generated images and segmentation masks across five datasets, using Semantic FID (sFID) and Positive Predictive Value (PPV). CoSimGen achieves the best sFID scores on Cityscapes, MBRSC, and BTCV, confirming its strong ability to preserve semantic consistency at the regional level. It also ranks second in PPV on MBRSC and BTCV, demonstrating reliable class-conditional generation. Performance on CholecSeg8k mirrors that on Cityscapes, reaffirming CoSimGen's alignment strength even in surgical domains with complex spatial priors.

On Pascal VOC, CoSimGen lags behind, which we attribute to the dataset's small size and high variance. Interestingly, CVAE performs better in this low-resource setting, suggesting VAE-style reconstruction is more stable when semantic structure is weakly represented. However, TGAN's deceptively high PPV stems from repetitive outputs—highlighting a trade-off between mask correctness and sample diversity, which CoSimGen manages more effectively.

Across datasets, results indicate that CoSimGen scales better with data complexity and maintains high semantic alignment under diverse conditions.

5.3 Input-Output Alignment

We assess alignment between input prompts and generated outputs using semantic FID (sFID), which measures class-wise fidelity by comparing semantic regions of generated images to real ones. Lower scores indicate both accurate class presence and precise spatial consistency. As reported in Table 2, CoSimGen achieves substantially lower sFID across all datasets, confirming reliable grounding of

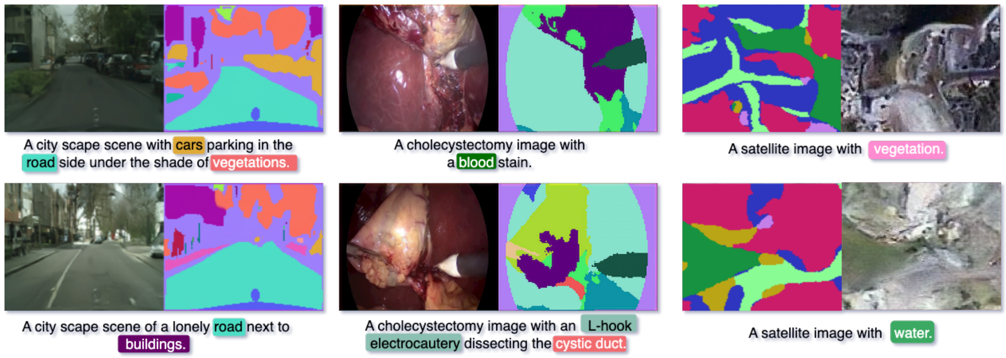

Figure 5: Qualitative results showing text(class)-conditioned image-mask generation

generation in the prompted classes. Visualizations in Fig. 5 further support this: masks not only include all requested classes, but their spatial layout reflects scene plausibility, which is also preserved in the corresponding images. Unlike baselines, which often miss prompted classes or distort their geometry, CoSimGen maintains both semantic and structural consistency. These results underscore the model's capacity to respect discrete mask structure while simultaneously generating continuous images—an essential requirement for controllable, high-quality surgical data synthesis.

5.4 High-Resolution (HR) Outputs

The super-resolution (SR) images produced by ESPCNN [31], utilized in our proposed CoSimGen framework, are compared with baseline outputs from SRGAN [20] on CholecSeg8K [15] and BTCV [10] datasets. The results in Fig. 6 demonstrate that ESPCNN effectively captures high-frequency details that SRGAN fails to reproduce. This distinction is particularly evident in the sharper boundaries between textures, such as those of organs, bones, and blood vessels, highlighting ESPCNN's superior ability to preserve structural details. More results are provided in the Appendix.

Detailed ablation results analyzing the impact of CoSimGen's core contributions, triplet loss for text-grounded alignment and discriminator loss for fidelity regularization, along with their combined effect, are provided in the Appendix to demonstrate their individual and complementary benefits.

5.5 Discussion

Our experiments reveal that models optimized with regression objectives exhibit greater stability compared to those with adversarial objectives, particularly in tasks involving joint estimation of continuous and discrete distribution pairs. In practice, we observed that adversarial models frequently suffered from mode collapse, which undermines their reliability for such tasks. By contrast, regression-based approaches minimize prediction error in a smoother optimization landscape, which contributes to improved stability. We also found that incorporating adversarial loss as a regularizer in our model introduced oscillatory behavior in early training stages, where generation quality fluctuates. However, as training progresses, these oscillations diminishes, and generation quality stabilizes. This suggests that adversarial loss, while beneficial as a regularizer, may require careful tuning to balance stability with generation fidelity.

5.6 Limitations

CoSimGen, like most diffusion-based generative models, is highly data-dependent. It requires substantial amounts of annotated segmentation masks paired with class-specific text labels to perform effectively. Our results show a noticeable drop in performance on datasets with limited samples, such as PASCAL VOC, indicating that CoSimGen is optimized for high-fidelity generation in moderate to large-scale data settings. While data augmentation techniques (e.g., random rotations and flips) can partially mitigate this limitation in specific domains like satellite imagery, their applicability to natural or medical scenes is limited. This highlights a key challenge: the need for more effective augmentation strategies or lightweight adaptations of CoSimGen to make it viable for low-resource and few-shot scenarios.

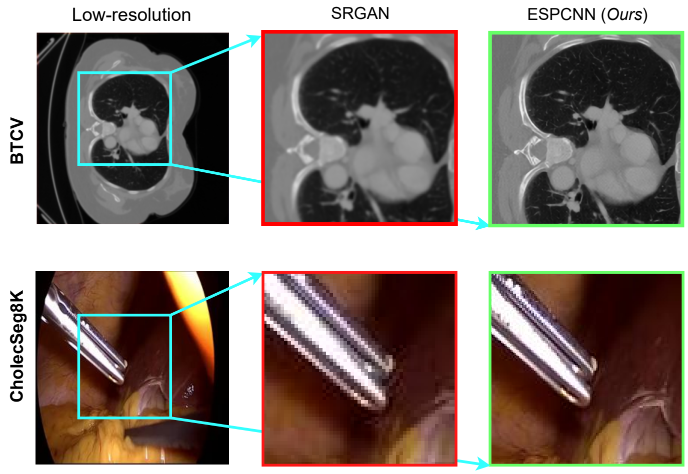

Figure 6: Comparison of super-resolution result of ESPCNN (in our model) and SRGAN (baseline).

Moreover, diffusion training is computationally intensive, which may restrict accessibility to users without high-end compute resources. Developing more efficient variants of CoSimGen or integrating it with faster approximation techniques could broaden its usability.

In terms of scope, all datasets used in this study lack identifiable human features. As such, we were unable to evaluate the model's behavior on data involving personally identifiable information. Users should exercise caution when deploying CoSimGen in privacy-sensitive settings, as its privacy-preserving capabilities remain unverified.

Although CoSimGen demonstrates strong input-output alignment and fidelity on curated benchmarks, its generalizability to out-of-distribution prompts or unseen object categories has not been systematically evaluated, leaving open questions about robustness in more diverse real-world applications.

6 Conclusion

This work introduces CoSimGen, a novel diffusion-based framework for controllable simultaneous image and segmentation mask generation. By addressing the critical challenges in existing generative models, CoSimGen provides a unified solution for producing high-quality paired datasets with precise control during generation. The model leverages text-grounded class conditioning, spatial-temporal embedding fusion, and multi-loss optimization, enabling robust performance across applications requiring spatial accuracy and flexibility. CoSimGen demonstrates state-of-the-art performance on diverse datasets, making it a versatile tool for augmenting datasets, simulating rare scenarios, and tackling domain-specific challenges. Its outputs offer a scalable alternative to manual annotation, significantly reducing the time and resources required for dataset creation. Moreover, the generated paired data serve as a ready source for pretraining models, given the framework's ability to produce an unlimited variety of high-fidelity, condition-adherent examples. Beyond its utility in dataset augmentation, CoSimGen establishes a foundation for future research in multi-modal, multi-class, and domain-adaptive generative frameworks. By bridging the gap between generative AI and real-world applications, the framework addresses critical bottlenecks in precision-driven and privacy-sensitive domains, advancing cross-domain AI research and deployment. CoSimGen represents a significant step forward in enabling scalable, controllable data generation, unlocking new possibilities for pretraining, robustness testing, and real-world impact.

References

[1] Preeti Aggarwal, Renu Vig, Sonali Bhadoria, and CG Dethe. Role of segmentation in medical imaging: A comparative study. *International Journal of Computer Applications*, 29(1):54–61, 2011.

[2] H Gökhan Akçay and Selim Aksoy. Automatic detection of geospatial objects using multiple hierarchical segmentations. *IEEE transactions on Geoscience and Remote Sensing*, 46(7): 2097–2111, 2008.

[3] Aditya Bhat, Rupak Bose, Chinedu Innocent Nwoye, and Nicolas Padoy. Simgen: A diffusion-based framework for simultaneous surgical image and segmentation mask generation. *arXiv preprint arXiv:2501.09008*, 2025.

[4] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

[5] Alice Brown and Bob Green. End-to-end autoencoding architecture for simultaneous image and mask generation in medical imaging. *Medical Image Analysis*, 78:102345, 2023.

[6] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[7] John Doe and Jane Smith. Seggen: Supercharging segmentation models with text2mask and mask2img generation. *arXiv preprint arXiv:2301.67890*, 2023.

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html, 2012.

[9] Di Feng, Christian Haase-Schütz, Lars Rosenbaum, Heinz Hertlein, Claudius Glaeser, Fabian Timm, Werner Wiesbeck, and Klaus Dietmayer. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Transactions on Intelligent Transportation Systems*, 22(3):1341–1360, 2020.

[10] Eli Gibson, Francesco Giganti, Yipeng Hu, Ester Bonmati, Steve Bandula, Kurinchi Gurusamy, Brian Davidson, Stephen P. Pereira, Matthew J. Clarkson, and Dean C. Barratt. Multi-organ abdominal ct reference standard segmentations, February 2018. URL `https://doi.org/10.5281/zenodo.1169361`.

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. In *Communications of the ACM*, volume 63, pages 139–144. ACM New York, NY, USA, 2020.

[13] 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, 2017.

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020.

[15] W-Y Hong, C-L Kao, Y-H Kuo, J-R Wang, W-L Chang, and C-S Shih. Cholecseg8k: a semantic segmentation dataset for laparoscopic cholecystectomy based on cholec80. *arXiv preprint arXiv:2012.12453*, 2020.

[16] Humans in the Loop. Semantic segmentation dataset, 2024. URL `https://humansintheloop.org/resources/datasets/semantic-segmentation-dataset-2/`. Accessed: 2024-10-11.

[17] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

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

[19] Mengchen Lan et al. Text4seg: Reimagining image segmentation as text generation. *arXiv preprint arXiv:2401.12345*, 2024.

[20] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[21] Pablo Marcos-Manchón, Roberto Alcover-Couso, Juan C SanMiguel, and Jose M Martínez. Open-vocabulary attention maps with token optimization for semantic segmentation in diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9242–9252, 2024.

[22] Chinedu Innocent Nwoye. *Deep learning methods for the detection and recognition of surgical tools and activities in laparoscopic videos*. PhD thesis, Université de Strasbourg, 2021.

[23] Chinedu Innocent Nwoye, Rupak Bose, Kareem Elgohary, Lorenzo Arboit, Giorgio Carlino, Joël L Lavanchy, Pietro Mascagni, and Nicolas Padoy. Surgical text-to-image generation. *Pattern Recognition Letters*, 2025.

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

[25] 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*, pages 8821–8831. Pmlr, 2021.

[26] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR, 2016.

[27] N Reimers. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

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

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

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

[31] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network, 2016. URL `https://arxiv.org/abs/1609.05158`.

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

[33] Weijia Wu, Yuzhong Zhao, Mike Zheng Shou, Hong Zhou, and Chunhua Shen. Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1206–1217, 2023.

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

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The contributions that we mentioned in the abstract and introduction are in the paper's analysis and experimental sections.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discussed the limitations of the work in Section 5.5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper contains no theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Our proposed method is fully detailed in Section 3 while all hyperparmeters used and training strategies are provided in Section 4 and Appendix

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Datasets use are publicly available, baseline models explored already open-sourced codebases and the remainder of the code will be released on GitHub shortly after this submission.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Some of these information are stated in the experiments section 4, and the rest are provided in the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: This is a generation task, we generated large number of samples which is individually stochastic and randomly sample a subset in multiple rounds and report error bars in Table 2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide information on the computer resources in Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have ensured that our research conforms to the code of ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Societal impact is discussed in the Appendix.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: We do not use controversial dataset and we positioned our model utility for educational purpose.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite all external sources of assets and we will include our asset license permit in the README.md of the code.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Our released code constitutes a new assets and will be well documented on GitHub to complement the documentation provided by this paper.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We do not conduct crowdsourcing experiments.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper did not invovle crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: Development of the method and research do not involve LLMs.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.