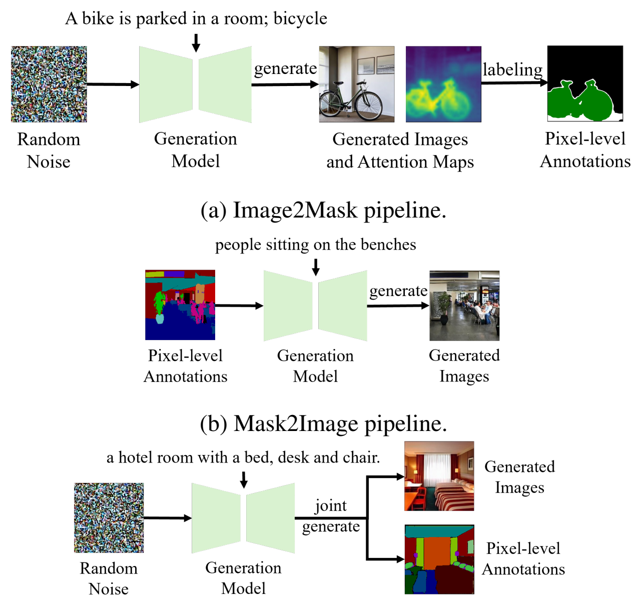
JoDiffusion: Jointly Diffusing Image with Pixel-Level Annotations for Semantic Segmentation Promotion

Haoyu Wang[1], **Lei Zhang**[1 †], **Wenrui Liu**[1], **Dengyang Jiang**[1], **Wei Wei**[1], **Chen Ding**[2]

[1]School of Computer Science, Northwestern Polytechnical University
[2]School of Computer Science & Technology, Xi'an University of Posts & Telecommunications
wanghaoyunwpu@mail.nwpu.edu.cn, nwpuzhanglei@nwpu.edu.cn

Abstract

Given the inherently costly and time-intensive nature of pixel-level annotation, the generation of synthetic datasets comprising sufficiently diverse synthetic images paired with ground-truth pixel-level annotations has garnered increasing attention recently for training high-performance semantic segmentation models. However, existing methods necessitate to either predict pseudo annotations after image generation or generate images conditioned on manual annotation masks, which incurs image-annotation semantic inconsistency or scalability problem. To migrate both problems with one stone, we present a novel dataset generative diffusion framework for semantic segmentation, termed JoDiffusion. Firstly, given a standard latent diffusion model, JoDiffusion incorporates an independent annotation variational auto-encoder (VAE) network to map annotation masks into the latent space shared by images. Then, the diffusion model is tailored to capture the joint distribution of each image and its annotation mask conditioned on a text prompt. By doing these, JoDiffusion enables simultaneously generating paired images and semantically consistent annotation masks solely conditioned on text prompts, thereby demonstrating superior scalability. Additionally, a mask optimization strategy is developed to mitigate the annotation noise produced during generation. Experiments on Pascal VOC, COCO, and ADE20K datasets show that the annotated dataset generated by JoDiffusion yields substantial performance improvements in semantic segmentation compared to existing methods.



(a) Image2Mask pipeline.

(b) Mask2Image pipeline.

(c) JoDiffusion pipeline.

Figure 1: Comparison of the proposed method with Image2Mask and Mask2Image pipelines. Compared with the other two step-by-step methods, JoDiffusion can directly generate images and corresponding pixel-level annotations.

1 Introduction

Semantic segmentation plays a crucial role in computer vision, which aims to assign a semantic label to each pixel. It has shown promising potential in plenty of practical applications including autonomous driving (Feng et al. 2020), medical image analysis (Asgari Taghanaki et al. 2021) and robot navigation (Song et al. 2023) etc.. Although deep neural networks have made significant progress in this task (Mo et al. 2022), their pleasing performance highly depends on a high-quality training dataset comprising large-scale paired images and ground-truth pixel-level annotations. However, due to high spatial resolution and diverse visual content, pixel-level manual annotation on image data is prohibitively costly and time-consuming, particularly in complex scenarios characterized by multi-object interaction or dense small-object distribution. This bottleneck significantly limits the adoption and deployment of semantic segmentation networks in real scenarios.

Inspired by the great success of deep generative models in image synthesis (Rombach et al. 2022; Podell et al. 2023; Esser et al. 2024), a promising solution lies in generating a synthetic dataset comprising sufficiently diverse synthetic images paired with ground-truth pixel-level annotations. Different from image generation for classification task (He et al. 2022), the dataset generation for semantic segmentation involves generation for paired image and pixel-level an-

†Corresponding author.

*Code is available at https://github.com/00why00/JoDiffusion.

notations. To this end, two lines of research have been investigated, including the Image2Mask (Wu et al. 2023; Nguyen et al. 2024; Tang et al. 2025) and Mask2Image (Yang et al. 2024; Ye et al. 2024). As shown in Fig. 1a, the Image2Mask framework initially employs a text-to-image diffusion model to generate synthetic images, and then a cross-attention-based pseudo-annotation scheme is applied to predict pixel-level pseudo annotations by leveraging text-image similarity computed in a latent feature space. Although this framework enables direct generation of synthetic semantic segmentation datasets conditioned solely on text prompts, the quality of pixel-level pseudo annotations remains suboptimal. Specifically, semantic inconsistencies between generated images and predicted pseudo annotations arise due to inevitable text-image similarity calculation errors and limited spatial resolution of feature maps compared with original image. Training models on such datasets impose ambiguous semantic information, ultimately leading to suboptimal generalization performance during inference. In contrast, the Mask2Image framework employs a dedicated diffusion model to generate synthetic images conditioned on both manual pixel-level annotation masks and text prompts, as shown in Fig. 1b. While the introduced high-quality annotation masks ensure semantic consistency with the generated images, the limited availability of manual annotations inherently restricts image content diversity beyond the scope of provided masks, resulting in suboptimal scalability.

To mitigate both limitations of existing methods, we present a novel semantic segmentation dataset generation framework, termed JoDiffusion. As illustrated in Fig. 1c, JoDiffusion differs fundamentally from existing frameworks by enabling simultaneous generation of paired images and pixel-level annotation masks through a joint diffusion model conditioned solely on text prompts. This framework not only guarantees semantic consistency between generated images and annotation masks but also achieves good scalability. To achieve this goal, we first establish a baseline framework leveraging a standard latent text-image diffusion model and integrate an annotation-specific variational auto-encoder (VAE) network to model the latent distribution of pixel-level annotations. This architecture enables paired images and pixel-level annotation masks to be mapped into a unified latent space, thereby facilitating the maintenance of semantic consistency during the generation process. Then, the diffusion model is tailored to jointly diffuse and denoise the input text prompts, images, and pixel-level annotation masks in the latent space. More importantly, the text prompts with random noise is forced to jointly recover the latent representation of each paired image and annotation mask during training. By doing these, the diffusion model can capture the joint distribution of paired images and annotation masks. This enables the simultaneous generation of semantically consistent paired images and annotation masks, relying solely on text prompts. Moreover, during the inference phase, without the requirement of additional manual annotation masks as the Mask2Image framework, the diffusion model can flexibly generalize beyond the limited set of manually annotated masks. In addition, we further develop a mask optimization strategy to mitigate the inevitable annotation noise produced during generation. With the generated high-quality synthetic dataset, we can train an effective segmentation model with better generalization performance. To testify this, we evaluate JoDiffusion onto three benchmark datasets including Pascal VOC (Everingham et al. 2015), MS COCO (Lin et al. 2014), and ADE20K (Zhou et al. 2017). The experimental results demonstrate that, compared to several state-of-the-art competitors, training the same semantic segmentation model with the synthetic dataset generated by JoDiffusion leads to substantially better generalization performance.

In summary, the primary contributions of this work can be succinctly articulated as follows:

- we propose a novel synthetic dataset generation framework for semantic segmentation. To the best of our knowledge, this is the first attempt to achieve simultaneous generation of semantically consistent paired images and pixel-level annotation masks conditioned solely on text promts.
- We also develop a mask optimization strategy to effectively mitigate the annotation noise produced during generation.
- We achieve new SOTA semantic segmentation performance when training the model using the generated synthetic dataset.

2 Related Work

Text-to-image Diffusion Models

Diffusion models have made breakthrough progress in the field of image generation in recent years. Early diffusion models (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020) achieved high-fidelity image generation through simple forward denoising and reverse denoising processes. Subsequently, models such as DALL·E (Ramesh et al. 2021, 2022; Betker et al. 2023) and Imagen (Saharia et al. 2022; Baldridge et al. 2024) adopted cross-modal conditional generation methods to apply diffusion models to text-to-image generation tasks, and surpassed GANs (Goodfellow et al. 2020) in terms of image clarity and semantic consistency. In order to improve inference efficiency, latent diffusion models (Rombach et al. 2022; Podell et al. 2023; Esser et al. 2024) perform diffusion modeling in the latent space, greatly reducing the computational complexity while maintaining high-quality generation capabilities, which has promoted the popularity of diffusion models. Subsequently, conditional image generation methods (Zhang, Rao, and Agrawala 2023; Ye et al. 2023; Zhao et al. 2024) introduced additional control in the diffusion model to make the generation process more controllable. Multimodal generation methods (Xu et al. 2023; Bao et al. 2023b) jointly model the joint distribution of different modalities, allowing information such as text and images to interact with each other during the diffusion process, thereby achieving bidirectional control generation.

Semantic Dataset Generation

Early attempts at semantic segmentation dataset generation leveraged GAN-based models (Zhang et al. 2021; Li et al.

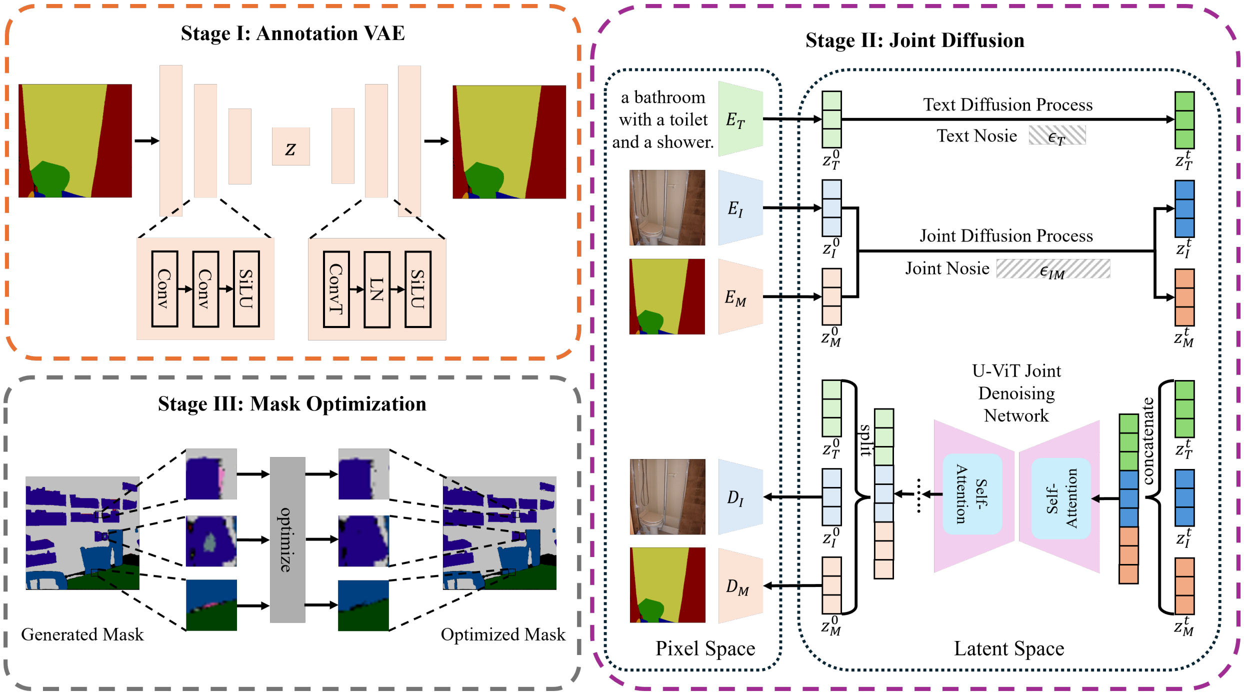


Figure 2: Three stages of JoDiffusion. We first train an annotation VAE to efficiently encode sparse and discrete category maps into a compact latent space for diffusion model alignment. Next, we jointly model the relationship between text, images, and pixel-level annotation masks to enhance the semantic consistency of the generated results. Finally, we optimize the generated annotation masks to improve the quality of semantic segmentation results.

2022), where semantic information was extracted from the latent space, and pixel-level annotations were inferred using additional decoders. With the emergence of diffusion models (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020) demonstrating superior image synthesis quality, recent efforts have shifted toward diffusion-based dataset generation. Existing approaches can be categorized into two main pipelines: Image2Mask and Mask2Image. The Image2Mask pipeline first generates images using a diffusion model, and infers the corresponding pixel-level annotation masks by parsing the features or attention maps in the generation process. For example, DiffuMask (Wu et al. 2023) extracts category-related salient areas from the image generation process by analyzing the cross-attention mechanism of the diffusion model, and further infers annotation masks using Affinity Net. Dataset Diffusion (Nguyen et al. 2024) is optimized on this basis, combining a large language model (Achiam et al. 2023) to generate more diverse text descriptions, and using self-attention maps to improve the quality of semantic masks. SDS (Tang et al. 2025) further introduces perturbation-based CLIP similarity and class-balance annotation similarity to filter the generated images to reduce data noise and improve the effectiveness of the dataset. In contrast, the Mask2Image method generates the corresponding images through the diffusion model based on the semantic masks. For example, FreeMask (Yang et al. 2024) uses the mask-to-image generation method FreestyleNet (Xue et al. 2023), and designs a series of filtering strategies to suppress erroneously synthesized areas to ensure the quality of generated data. SegGen (Ye et al. 2024) train an additional text-to-mask model to make the generated semantic masks more diverse, thereby improving the generalization ability of the semantic segmentation model.

3 Method

Problem Setup

Our goal is to learn a joint generative model $\mathcal{G}_\theta(I, M|T)$ that synthesizes images and corresponding annotation masks from text captions T, using a real-world semantic segmentation dataset $\mathcal{D}_{real} = \{(I_i, M_i)\}_{i=1}^{N_{real}}$ as supervision. The generated synthetic dataset $\mathcal{D}_{syn} = \{(I_i, M_i)\}_{i=1}^{N_{syn}}$ should align with $\mathcal{D}_{real}$ in terms of category distribution, object structures, and visual characteristics while introducing greater diversity to enhance the generalization of semantic segmentation models. Here, θ represents the parameters of the generative model, and I_i, M_i denote the RGB image and its corresponding annotation mask, respectively. Finally, we evaluate our approach by training semantic segmentation models on $\mathcal{D}_S$ and $\mathcal{D}_R \cup \mathcal{D}_S$.

Overview

As shown in Fig. 2, our method consists of three key stages: 1) Annotation VAE training: we first train an annotation VAE network to encode the annotation masks to obtain a compact latent representation. 2) Joint diffusion modeling: we train the diffusion model based on text, images, and annotation masks to jointly model the relationship between the

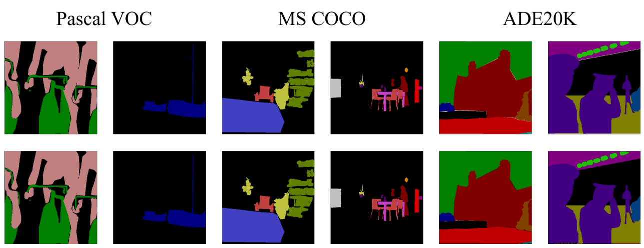


Figure 3: visualization of reconstructed pixel-level annotation masks on the validation sets. The first line is the input and the second line is the reconstruction result.

Dataset	mIoU ↑
Pascal VOC	99.50
MS COCO	98.85
ADE20K	98.74

Table 1: Reconstruction mIoU of pixel-level annotation masks on the validation sets.

image latent variable z_I and the mask latent variable z_M in the latent space, and guide them through the text condition T. In each denoising process, the model not only reconstructs the image features, but also ensures that the category of the annotation masks is consistent with the image content. 3) Mask optimization strategy: since the diffusion process may introduce label inconsistencies in small regions, we post-process them by using the majority class of its edge pixels, to correct the pixel-level annotations and optimize the final segmentation quality. The resulting dataset is used to train the semantic segmentation model.

Annotation VAE

To enable joint text-based generation of images and annotation masks, we adopt a latent diffusion model (Bao et al. 2023a,b), where the image encoder maps RGB images into a latent space. To maintain consistency with this setup, we introduce an Annotation VAE to encode annotation masks into a corresponding latent representation.

Annotation masks are typically stored as single-channel category indices. Directly normalizing and feeding it into VAE may result in adjacent category values being too close, making it challenging for the model to accurately differentiate them. In order to improve the category discrimination and reduce the computational overhead, we employ binary encoding as the input representation of the annotation VAE. Specifically, the category of each pixel $M(i,j)$ is converted into a binary representation M_{bin}.

Annotation VAE follows a lightweight architecture comprising of an encoder E_M and a decoder D_M, both utilizing a small number of convolutional and transposed convolutional layers. Compared to the image VAE used in the diffusion model, the annotation VAE not only significantly reduces the number of parameters ($\approx$ 50M $vs.$ 300M), as shown in Fig. 1, but also maintains high reconstruction quality.

Since annotation VAE serves purely as a compression tool rather than a generative model, we do not impose a standard normal prior on its latent variables. Consequently, KL divergence regularization is omitted, and the model is trained solely using cross-entropy loss, defined as:

$$\mathcal{L}_{\text{Annotation VAE}} = -\sum_{(i,j)} \sum_{c=0}^{N_C} M_{\text{one-hot},(i,j,c)} \log \bar{M}_{(i,j,c)}, \quad (1)$$

where $M_{\text{one-hot},(i,j,c)}$ represents the ground truth one-hot category at pixel (i,j), and $\bar{M}_{(i,j,c)}$ is the predicted probability obtained from the softmax output of the decoder. After training, given the latent representation z_M encoded by E_M, the reconstructed semantic mask is obtained by applying an argmax operation over the softmax output of the decoder: $\hat{M} = \arg\max(D_M(z_M))$.

Joint Diffusion

To ensure that the generated image and its corresponding pixel-level annotation masks remain semantically consistent, we adopt a joint diffusion process that models their shared distribution. Unlike Image2Mask pipeline, which first generates an image and infers its annotation masks, or Mask2Image pipeline, which generates annotation masks and then conditions the image generation, our approach diffuses and denoises images and annotation masks simultaneously. This bidirectional feature interaction allows for richer semantic alignment and improved scalability.

Our method builds upon Unidiffuser (Bao et al. 2023b). Compared to methods like SDXL (Podell et al. 2023), which rely on cross-attention to model text-image relationships, it concatenates text and image features and applies self-attention to model them, which offers greater flexibility for tuning. Specifically, given an image I, we first generate a descriptive caption T using BLIP-2 (Li et al. 2023). We then use the CLIP (Radford et al. 2021) text encoder $\mathcal{E}_T$, image encoder $\mathcal{E}_I$, and the image VAE E_I encode them into latent space:

$$z_T = \mathcal{E}_T(T), \quad z_I = [\mathcal{E}_I(I), E_I(I)]. \quad (2)$$

To integrate annotation masks M into this process, we leverage the Annotation VAE trained in the previous stage to obtain their latent representation: $z_M = E_M(M)$. To ensure consistency between images with annotation masks, we diffuse z_I and z_M jointly instead of treating them as independent diffusion processes. We achieve this by introducing a shared noise perturbation ϵ_{IM}, maintaining semantic alignment during diffusion.

The forward process progressively injects Gaussian noise into z_I^0 and z_M^0, simulating a degradation path that enables effective denoising:

$$q(z_I^t, z_M^t | z_I^0, z_M^0) = \mathcal{N}(\sqrt{\bar{\alpha}_t} \begin{bmatrix} z_I^0 \\ z_M^0 \end{bmatrix}, (1-\bar{\alpha}_t)I), \quad (3)$$

where $z_I^0 = z_I, z_M^0 = z_M$ and $\bar{\alpha}_t$ controls the noise schedule at timestep t. This formulation ensures that both the image and annotation masks share the same noise perturbation

Prompt: a street with cars parked on both sides and a mountain in the background.

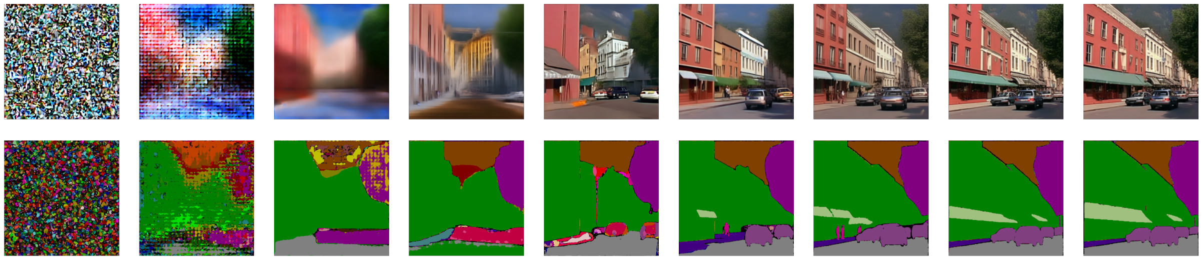


Figure 4: Visualization of joint generation result at different timesteps. A color map is applied for better visualization.

$\epsilon_{IM} \sim \mathcal{N}(0, I)$, maintaining structural consistency during training.

To recover the original image and pixel-level annotation masks pair from the noisy latent variables (z_I^t, z_M^t), we model the joint denoising distribution:

$$p_\theta(z_I^{t-1}, z_M^{t-1} | z_I^t, z_M^t, z_T) = \mathcal{N}(\mu_\theta(z_I^t, z_M^t, z_T, t), \sigma_t^2 I), \quad (4)$$

where σ_t^2 is determined by the predefined noise schedule and controls the level of randomness at each denoising step. The denoised mean μ_θ captures the underlying relationship between the image and pixel-level annotation masks:

$$\mu_\theta(z_I^t, z_M^t, z_T, t) = \frac{1}{\sqrt{\alpha_t}} \left(\begin{bmatrix} z_I^t \\ z_M^t \end{bmatrix} - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(z_I^t, z_M^t, z_T, t) \right) \quad (5)$$

where $\epsilon_\theta(z_I^t, z_M^t, z_T, t)$ is the denoising network, which predicts the noise added during the forward diffusion process. Instead of estimating independent noise components, the network learns a joint representation, leveraging shared information between the image and annotation masks.

The model is trained using the standard mean squared error loss, where the image and annotation masks part is:

$$\mathcal{L}_{\text{denoising}} = \mathbb{E}_{t, z_I^0, z_M^0, \epsilon} \left[\| \epsilon_\theta(z_I^t, z_M^t, z_T, t) - \epsilon_{IM} \|^2 \right], \quad (6)$$

where ϵ_{IM} is the noise that was added during the forward diffusion process. By minimizing it, the model effectively denoises latent representations while preserving semantic integrity between the image and annotation masks. This reinforces semantic alignment in generated pairs, leading to improved performance in downstream segmentation tasks.

Mask Optimization

While the joint diffusion process ensures semantic consistency between images and annotation masks, the pixel-level annotations generated by the model may still contain noise, especially around small target areas and object boundaries. This noise can manifest as speckle or label inconsistencies, which often lead to local deviations in labels and degrade the performance of downstream segmentation tasks. To address this, we propose a boundary mode-based mask optimization strategy. This method analyzes label distribution of boundary pixels and corrects small regions by replacing their labels with the most frequent category in that region, thus enhancing label consistency and suppressing noise.

Let $R \subset \{1, ..., H\} \times \{1, ..., W\}$ denote a small target region in the annotation mask, satisfying $|R| < \tau$, where $|R|$ being the number of pixels in R, and τ is a dataset-dependent threshold, typically set to identify small objects or noise regions. Small regions are particularly prone to noise, necessitating targeted refinement. To correct the labels in R, we first define its boundary pixel set as $\hat{R}$ and compute the mode of the label values among these boundary pixels:

$$c^* = \arg\max_c \sum_{(i,j) \in \hat{R}} \mathbb{I}(x_{i,j} = c), \quad (7)$$

where $x_{i,j}$ is the label at pixel (i, j), and $\mathbb{I}(\cdot)$ is an indicator function that counts occurrences of category in the boundary pixels. The calculated mode c^* represents the most frequent category in $\hat{R}$, which is then used to reassign all pixels in R:

$$\forall (i, j) \in R, \quad x_{i,j} \leftarrow c^*. \quad (8)$$

The effectiveness of this correction method is grounded in statistical estimation principles. Given a small target region R, its true category label may be ambiguous due to noise introduced in the diffusion process. However, the boundary pixels $\hat{R}$ are more likely to retain correct labels due to the inherent continuity of semantic regions in natural images. This assumption is supported by two key observations: adjacent pixels in real-world images typically belong to the same category, and errors introduced by the diffusion process tend to be randomly distributed in small isolated regions rather than along structured object boundaries.

Under these assumptions, the mode c^* of the boundary labels provides a reliable estimate of the true category of the target region. From a statistical perspective, this process can be viewed as a maximum likelihood estimation, where the most frequent category among the boundary pixels serves as the most probable label assignment for R. Formally, this can be expressed as:

$$c^* = \arg\max_c P(c | \hat{R}), \quad (9)$$

where $P(c|\hat{R})$ represents the empirical distribution of labels in the boundary region. Assuming an approximately uniform prior over categories, this estimation reduces to selecting the mode of the boundary labels. By replacing the labels in R with c^*, we effectively minimize the probability of incorrect category assignments while preserving structural coherence in the segmentation mask.

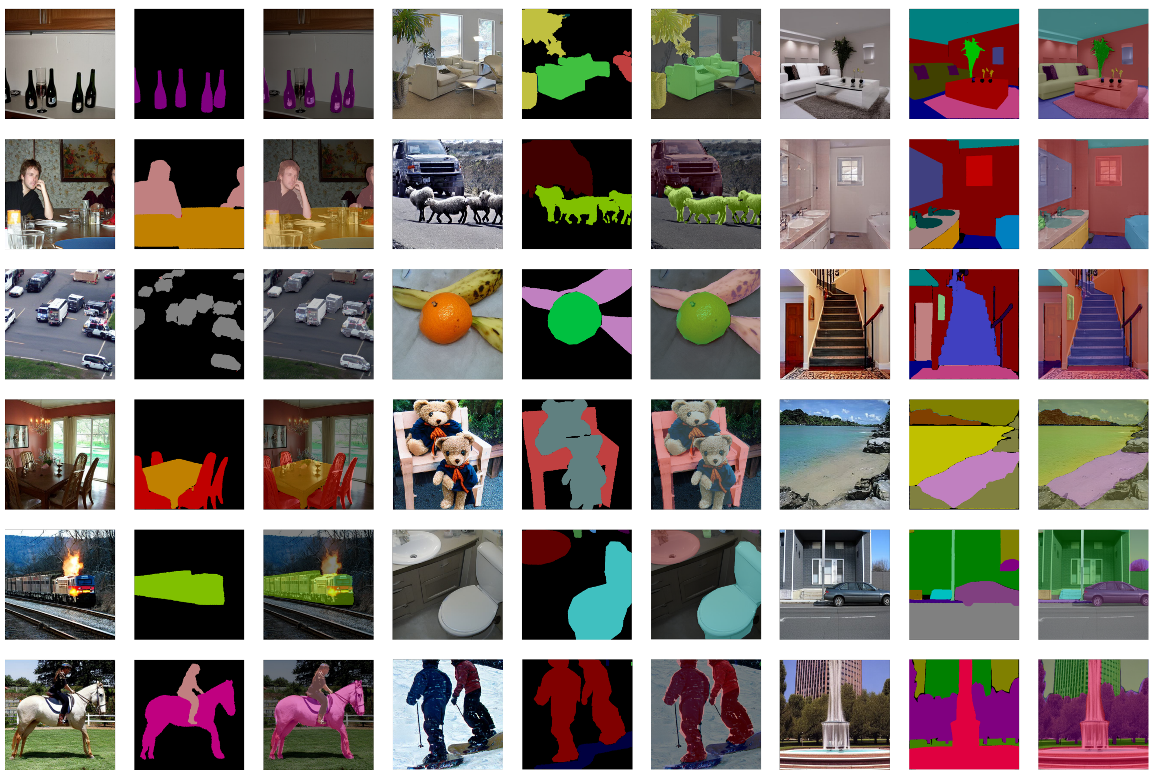


Figure 5: Visualization of joint generation result of three datasets. A color map is applied for better visualization.

4 Experiments

Datasets

Pascal VOC (Everingham et al. 2015) is a widely used benchmark for semantic segmentation, containing 20 object categories and 1 background category. Following previous work (Nguyen et al. 2024; Tang et al. 2025), we incorporate the Semantic Boundaries Dataset (Hariharan et al. 2011) to extend its training set, resulting in 10,582 training images and 1,449 validation images. **MS-COCO** (Lin et al. 2014) provides a more diverse and complex dataset with 80 object categories and one background category. It consists of 118,288 training images and 5,000 validation images, offering a challenging setting with high intra-class variance and occlusions. **ADE20K** (Zhou et al. 2017) is a scene parsing dataset containing 150 fine-grained semantic concepts, making it suitable for dense prediction tasks. It includes 20,210 training images and 2,000 validation images, covering a broad range of indoor and outdoor scenes.

Implementation Details

For all datasets, we resize images and annotation masks to 512×512 512 for training both VAE and diffusion models. We use the AdamW (Loshchilov and Hutter 2017) optimizer for all training stages, and apply random horizontal flipping as data augmentation. Additional architecture details and hyperparameters are provided in the supplementary material.

Comparison with State-of-the-Arts

We compare our method with the state-of-the-art Image2Mask methods, including DiffuMask (Wu et al. 2023), Dataset Diffusion (Nguyen et al. 2024) and SDS (Tang et al. 2025), as well as Mask2Image methods such as FreeMask (Yang et al. 2024) and SegGen (Ye et al. 2024). To ensure a fair comparison, we reproduced FreeMask using the same amount of data as ours and applied the same filtering strategy to our method.

Qualitative Results Fig. 3 illustrates the reconstruction performance of our annotation VAE. The first row presents the ground-truth pixel-level annotation masks, while the second row shows the reconstructed annotations after encoding and decoding. Our model effectively preserves the structural and categorical details of the original annotations, achieving high reconstruction fidelity with minimal information loss.

Fig. 4 visualizes the intermediate diffusion steps of our joint generation process. Starting from an initial noisy representation, our method progressively refines both the image and its corresponding annotation masks, improving structural and semantic coherence over time. As diffusion progresses, contours become sharper, and the generated annotations better align with object semantics, demonstrating the effectiveness of our joint generation strategy.

Fig. 5 presents the final synthesized image and annotations pairs on three datasets. For each dataset, the third column overlays the generated image and annotations for bet-

Segmenter	Backbone	Method	Pascal VOC			MS-COCO		
			Data Size	mIoU (Syn)	mIoU (Real+Syn)	Data Size	mIoU (Syn)	mIoU (Real+Syn)
DeepLabV3	ResNet50	Raw Dataset	$11.5k$	77.4		$118k$	48.9	
		SDS	$26k$	60.4	77.6	$50k$	31.0	50.3
		Dataset Diffusion	$40k$	61.6	77.6	$80k$	32.4	54.6
		JoDiffusion	$40k$	**72.5**	**78.3**	$80k$	**42.6**	**56.4**
	ResNet101	Raw Dataset	$11.5k$	79.9		$118k$	54.9	
		SDS	$26k$	59.1	79.8	$50k$	31.8	56.8
		Dataset Diffusion	$40k$	64.8	80.3	$80k$	34.2	57.4
		JoDiffusion	$40k$	**75.8**	**80.7**	$80k$	**44.9**	**59.1**
Mask2Former	ResNet50	Raw Dataset	$11.5k$	77.3		$118k$	57.8	
		DiffuMask	$60k$	57.4	77.5	-	-	-
		SDS	$26k$	59.8	78.1	50k	29.8	57.7
		Dataset Diffusion	$40k$	60.2	78.2	$80k$	31.0	57.8
		JoDiffusion	$40k$	**74.5**	**79.4**	$80k$	**44.6**	**58.5**

Table 2: Comparisons in mIoU with Image2Mask methods on Pascal VOC and MS-COCO dataset.

Backbone	Method	Pascal VOC		ADE20K	
		Data Size	mIoU	Data Size	mIoU
ResNet50	Raw Data	$11.5k$	77.3	$20k$	47.2
	SegGen	-	-	1M	**49.9**
	FreeMask	40k	<u>77.9</u>†	$40k$	48.2†
	JoDiffusion	40k	**79.4**	$40k$	<u>48.4</u>
Swin-S	Raw Data	$11.5k$	83.8	$20k$	51.6
	FreeMask	40k	<u>84.2</u>†	$40k$	<u>52.1</u>†
	JoDiffusion	40k	**85.1**	$40k$	**52.2**

Table 3: Comparisons in mIoU with Mask2Image methods on ADE20K dataset. † means our reproduced results.

ter visualization. The results indicate that our approach not only produces high-quality images across diverse datasets but also maintains strong semantic alignment between generated annotations and image content. Additional qualitative results can be found in the supplementary material.

Quantitative Results Tab. 1 reports the mIoU of our trained annotation VAE on three datasets. Our method achieves reconstruction accuracy exceeding 98%, demonstrating its effectiveness in compactly encoding annotation masks while preserving critical structural information.

Tab. 2 compares our method with Image2Mask approaches on Pascal VOC and MS-COCO datasets. Across multiple segmentation architectures and backbones, our approach significantly outperforms prior methods.

Tab. 3 presents the results on Pascal VOC and ADE20K datasets, where we follow the Mask2Image paradigm by training the Mask2Former segmenter with both real and synthetic data. Our approach consistently outperforms existing Mask2Image methods across multiple backbones. Additional results can be found in the supplementary material.

Discussion

Effectiveness of the mask optimization strategy. We analyze the effect of different regional thresholds τ on segmentation performance on Psacal VOC dataset. As shown in Tab. 4, applying mask optimization improves performance compared to the baseline without optimization.

τ	$\tau = 0$	$\tau = 20$	$\tau = 50$	$\tau = 100$
mIoU ↑	71.37	**72.47**	72.38	72.38

Table 4: Results on different mask optimization threshold τ.

Effectiveness of the generated data size. We investigate the impact of different amounts of generated training data on segmentation performance on Psacal VOC dataset. As shown in Tab. 5, increasing the dataset size consistently improves performance.

Data Size	$5k$	$10k$	$20k$	$40k$
mIoU ↑	68.54	70.02	70.97	**72.47**

Table 5: Results on different data sizes.

5 Conclusion

In this paper, we introduce JoDiffusion, a novel framework for joint image and annotation mask generation framework. Unlike traditional Image2Mask and Mask2Image approaches, our method directly models the joint distribution of images and their corresponding annotation masks. By incorporating an annotation VAE and an effective mask optimization strategy, our approach significantly outperforms prior methods in segmentation performance on Pascal VOC, MS-COCO, and ADE20K, demonstrating its efficacy in generating high-quality synthetic segmentation data.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China under Grand 62372379, Grant 62472359, and Grant 62472350; in part by the Xi'an's Key Industrial Chain Core Technology Breakthrough Project: AI Core Technology Breakthrough under Grand 23ZDCYJSGG0003-2023; in part by National Key Laboratory of Science and Technology on Space-Born Intelligent Information Processing fundation under Grant TJ-04-23-04; in part by Innovation Foundation for Doctor Dissertation of Northwestern Polytechnical University under Grant CX2025092.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Asgari Taghanaki, S.; Abhishek, K.; Cohen, J. P.; Cohen-Adad, J.; and Hamarneh, G. 2021. Deep semantic segmentation of natural and medical images: a review. *Artificial Intelligence Review*, 54: 137–178.

Baldridge, J.; Bauer, J.; Bhutani, M.; Brichtova, N.; Bunner, A.; Chan, K.; Chen, Y.; Dieleman, S.; Du, Y.; Eaton-Rosen, Z.; et al. 2024. Imagen 3. *arXiv preprint arXiv:2408.07009*.

Bao, F.; Nie, S.; Xue, K.; Cao, Y.; Li, C.; Su, H.; and Zhu, J. 2023a. All are worth words: A vit backbone for diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 22669–22679.

Bao, F.; Nie, S.; Xue, K.; Li, C.; Pu, S.; Wang, Y.; Yue, G.; Cao, Y.; Su, H.; and Zhu, J. 2023b. One transformer fits all distributions in multi-modal diffusion at scale. In *International Conference on Machine Learning*, 1692–1717. PMLR.

Betker, J.; Goh, G.; Jing, L.; Brooks, T.; Wang, J.; Li, L.; Ouyang, L.; Zhuang, J.; Lee, J.; Guo, Y.; et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3): 8.

Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*.

Everingham, M.; Eslami, S. A.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2015. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111: 98–136.

Feng, D.; Haase-Schütz, C.; Rosenbaum, L.; Hertlein, H.; Glaeser, C.; Timm, F.; Wiesbeck, W.; and Dietmayer, K. 2020. 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.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Hariharan, B.; Arbeláez, P.; Bourdev, L.; Maji, S.; and Malik, J. 2011. Semantic contours from inverse detectors. In *2011 international conference on computer vision*, 991–998. IEEE.

He, R.; Sun, S.; Yu, X.; Xue, C.; Zhang, W.; Torr, P.; Bai, S.; and Qi, X. 2022. Is synthetic data from generative models ready for image recognition? *arXiv preprint arXiv:2210.07574*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Li, D.; Ling, H.; Kim, S. W.; Kreis, K.; Fidler, S.; and Torralba, A. 2022. Bigdatasetgan: Synthesizing imagenet with pixel-wise annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21330–21340.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Mo, Y.; Wu, Y.; Yang, X.; Liu, F.; and Liao, Y. 2022. Review the state-of-the-art technologies of semantic segmentation based on deep learning. *Neurocomputing*, 493: 626–646.

Nguyen, Q.; Vu, T.; Tran, A.; and Nguyen, K. 2024. Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation. *Advances in Neural Information Processing Systems*, 36.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, 8821–8831. Pmlr.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Song, Y.; Xu, F.; Yao, Q.; Liu, J.; and Yang, S. 2023. Navigation algorithm based on semantic segmentation in wheat fields using an RGB-D camera. *Information Processing in Agriculture*, 10(4): 475–490.

Tang, H.; Yu, S.; Pang, J.; and Zhang, B. 2025. A Training-free Synthetic Data Selection Method for Semantic Segmentation. *arXiv preprint arXiv:2501.15201*.

Wu, W.; Zhao, Y.; Shou, M. Z.; Zhou, H.; and Shen, C. 2023. Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1206–1217.

Xu, X.; Wang, Z.; Zhang, G.; Wang, K.; and Shi, H. 2023. Versatile diffusion: Text, images and variations all in one diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7754–7765.

Xue, H.; Huang, Z.; Sun, Q.; Song, L.; and Zhang, W. 2023. Freestyle layout-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14256–14266.

Yang, L.; Xu, X.; Kang, B.; Shi, Y.; and Zhao, H. 2024. Freemask: Synthetic images with dense annotations make stronger segmentation models. *Advances in Neural Information Processing Systems*, 36.

Ye, H.; Kuen, J.; Liu, Q.; Lin, Z.; Price, B.; and Xu, D. 2024. Seggen: Supercharging segmentation models with text2mask and mask2img synthesis. In *European Conference on Computer Vision*, 352–370. Springer.

Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zhang, Y.; Ling, H.; Gao, J.; Yin, K.; Lafleche, J.-F.; Barriuso, A.; Torralba, A.; and Fidler, S. 2021. Datasetgan: Efficient labeled data factory with minimal human effort. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10145–10155.

Zhao, S.; Chen, D.; Chen, Y.-C.; Bao, J.; Hao, S.; Yuan, L.; and Wong, K.-Y. K. 2024. Uni-controlnet: All-in-one control to text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36.

Zhou, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; and Torralba, A. 2017. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 633–641.