Bayesian Power Steering

An Effective Approach for Domain Adaptation of Diffusion Models

Ding Huang, Ting Li*, Jian Huang*

July 2024

DEPARTMENT OF APPLIED MATHEMATICS
應 用 數 學 系

Roadmap

1 Introduction

- **Diffusion Models**

  The advent of diffusion models (Ho et al., 2020) and their extensions has enabled effective learning of intricate probability measures for image data.

1 Introduction

- **Pre-train large-scale diffusion models: Stable Diffusion**, which is trained utilizing large text-image datasets, for example, the LAION-5B dataset.

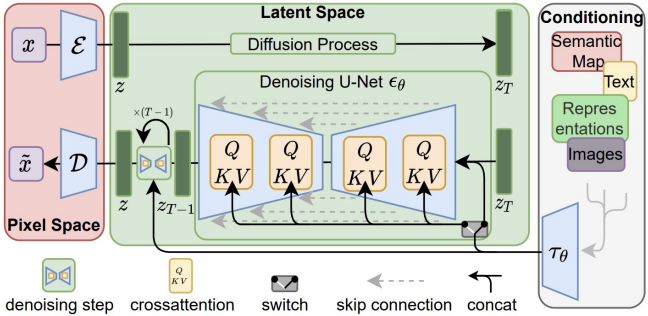


Figura: The overview of Stable Diffusion.

1 Introduction

Stable Diffusion (SD, Rombach et al. (2022)) first map the image data into the latent probability space, $Z_0 \in \mathcal{Z} \subseteq \mathbb{R}^d$. In the latent space,

- the diffusion process put forward the data to the Gaussian distribution,

$$Z_t = \sqrt{\bar{\alpha}_t} Z_0 + \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\eta}, \ \boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d); \tag{1}$$

1 Introduction

Stable Diffusion (SD, Rombach et al. (2022)) first map the image data into the latent probability space, $Z_0 \in \mathcal{Z} \subseteq \mathbb{R}^d$. In the latent space,

- the diffusion process put forward the data to the Gaussian distribution,

$$Z_t = \sqrt{\bar{\alpha}_t} Z_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\eta}, \ \boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d); \tag{1}$$

- The backward process $\{\tilde{Z}_t^{C_{\text{text}}}\}_{t=1}^T$ is used for generation, starting from $\tilde{Z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ with the following iteration:

$$\tilde{Z}_{t-1}^{C_{\text{text}}} = \frac{1}{1-\beta_t}\left(\tilde{Z}_t^{C_{\text{text}}} - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}^*(\tilde{Z}_t^{C_{\text{text}}}, t, C_{\text{text}})\right) + \sigma_t \boldsymbol{\eta}, \tag{2}$$

where $\boldsymbol{\epsilon}^*$ is the denoise function and $\beta_t := 1 - \bar{\alpha}_t/\bar{\alpha}_{t-1}$, $\sigma_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$.

1 Introduction

- According to Tweedie's formula (Efron, 2011), the denoise model $\boldsymbol{\epsilon}^*(\mathbf{z}_t, t, \mathbf{c}_{\text{text}})$ can be expressed as

$$\begin{aligned}\boldsymbol{\epsilon}^*(\mathbf{z}_t, t, \mathbf{c}_{\text{text}}) :&= -\sqrt{1-\bar{\alpha}_t}\nabla \log p(Z_t = \mathbf{z}_t \mid C_{\text{text}} = \mathbf{c}_{\text{text}}) \\ &= \mathbb{E}[\boldsymbol{\eta} \mid Z_t = \mathbf{z}_t, C_{\text{text}} = \mathbf{c}_{\text{text}}]. \end{aligned} \tag{3}$$

1 Introduction

- According to Tweedie's formula (Efron, 2011), the denoise model $\boldsymbol{\epsilon}^*(\mathbf{z}_t, t, \mathbf{c}_{\text{text}})$ can be expressed as

$$\begin{aligned}\boldsymbol{\epsilon}^*(\mathbf{z}_t, t, \mathbf{c}_{\text{text}}) :&= -\sqrt{1-\bar{\alpha}_t}\nabla \log p(Z_t = \mathbf{z}_t \mid C_{\text{text}} = \mathbf{c}_{\text{text}}) \\ &= \mathbb{E}[\boldsymbol{\eta} \mid Z_t = \mathbf{z}_t, C_{\text{text}} = \mathbf{c}_{\text{text}}]. \end{aligned} \tag{3}$$

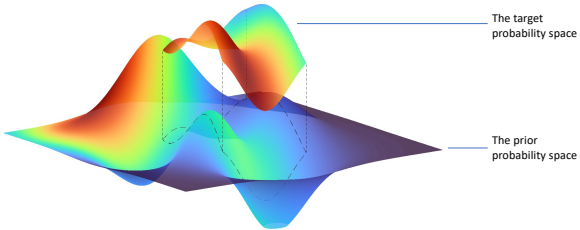
The diversity and abundance of data contributes to the exceptional generative capabilities of large-scale models $\hat{\boldsymbol{\epsilon}}$.

1 Introduction

- **Domain Adaptation**

  The aim is to transfer a broader knowledge base acquired from publicly available data to the task-specific distribution.

  - Privacy protection.
  - Sensitive medical data.
  - User-customized task.

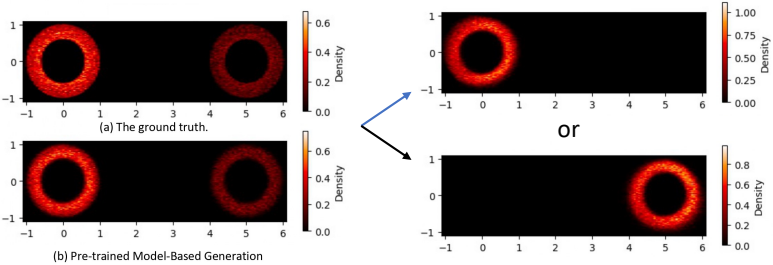


1 Introduction

- **In addition to private data, can public data provide more information for the distribution of private data?**

1 Introduction

- **Alternatively, can the deployment of such pre-trained large-scale models, in conjunction with the private dataset, effectively facilitate the transition from a large probability space to a small probability space?**



Roadmap

2 Formulation

- **Problem Formulation**
  - The latent probability space $\mathcal{Z} := (\Omega, \mathfrak{F}, \mathcal{P})$ be the generative target of the pre-trained model.

2 Formulation

- **Problem Formulation**
  - The latent probability space $\mathcal{Z} := (\Omega, \mathfrak{F}, \mathcal{P})$ be the generative target of the pre-trained model.
  - Our focus lies on a small latent probability space $\mathcal{Z}_\Delta := (\Delta, \Delta \cap \mathfrak{F}, \mathcal{P}_\Delta)$, where $\Delta \in \mathfrak{F}$ and $\mathcal{P}_\Delta(E) := \mathcal{P}(E)/\mathcal{P}(\Delta)$, for all $E \in \Delta \cap \mathfrak{F}$.

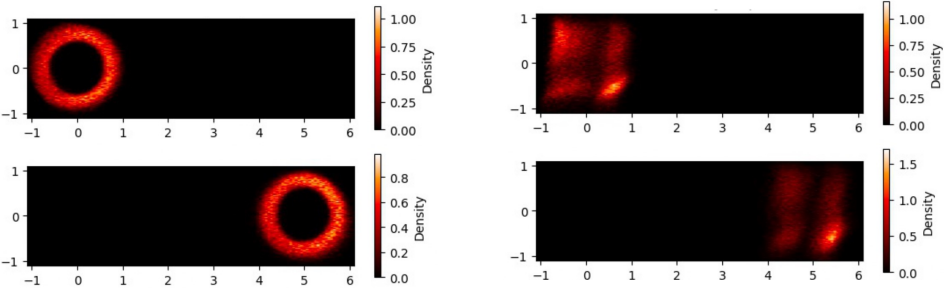
2 Formulation

- **Problem Formulation**
  - The latent probability space $\mathcal{Z} := (\Omega, \mathfrak{F}, \mathcal{P})$ be the generative target of the pre-trained model.
  - Our focus lies on a small latent probability space $\mathcal{Z}_\Delta := (\Delta, \Delta \cap \mathfrak{F}, \mathcal{P}_\Delta)$, where $\Delta \in \mathfrak{F}$ and $\mathcal{P}_\Delta(E) := \mathcal{P}(E)/\mathcal{P}(\Delta)$, for all $E \in \Delta \cap \mathfrak{F}$.

    Lemma 1. (Chung, 2001) If $\Delta \in \mathfrak{F}$, then there exists some measurable function $\boldsymbol{\psi}(\cdot)$ such that $\Delta(\omega) = \boldsymbol{\psi}(Z_0)$ for any $\Delta(\omega) := \omega \in \Delta \cap \mathfrak{F}$.

2 Formulation

Consequently, when the data $\mathbf{z}_0 \in \Delta$ aligns with suitable conditions $\mathbf{c}$, **the task of learning a "small distribution"is formulated as learning a probability measure of $Z_0 \mid C$**, where C is a random variable defined in $\mathcal{C} := (\boldsymbol{\psi}(\Delta), \boldsymbol{\psi}(\mathfrak{F}), \mathcal{P}') \subseteq \mathbb{R}^k$.



2 Formulation

- **Bayesian Formulation**

  Suppose condition $(\mathbf{c}_{\text{text}}, \mathbf{c}_{\text{add}}) \in \mathcal{C}$ provides a detailed characterization of the target domain

  - our primary objective is to learn the integrated denoise function, denoted as

$$\bar{\boldsymbol{\epsilon}}^*(\mathbf{z}_t, t, \mathbf{c}_{\text{text}}, \mathbf{c}_{\text{add}}) := \mathbb{E}[\boldsymbol{\eta} \mid \mathbf{z}_t, \mathbf{c}_{\text{text}}, \mathbf{c}_{\text{add}}].$$

2 Formulation

- **Bayesian Formulation**

  Suppose condition $(\mathbf{c}_{\text{text}}, \mathbf{c}_{\text{add}}) \in \mathcal{C}$ provides a detailed characterization of the target domain

  - our primary objective is to learn the integrated denoise function, denoted as

$$\bar{\boldsymbol{\epsilon}}^*(\mathbf{z}_t, t, \mathbf{c}_{\text{text}}, \mathbf{c}_{\text{add}}) := \mathbb{E}[\boldsymbol{\eta} \mid \mathbf{z}_t, \mathbf{c}_{\text{text}}, \mathbf{c}_{\text{add}}].$$

  - By Bayes' theorem, we have

$$p(\mathbf{z}_t \mid \mathbf{c}_{\text{text}}, \mathbf{c}_{\text{add}}) = \frac{p(\mathbf{c}_{\text{add}} \mid \mathbf{z}_t, \mathbf{c}_{\text{text}})}{p(\mathbf{c}_{\text{add}} \mid \mathbf{c}_{\text{text}})} p(\mathbf{z}_t \mid \mathbf{c}_{\text{text}}).$$

2 Formulation

- **Bayesian Formulation**

  Suppose condition $(\mathbf{c}_{\text{text}}, \mathbf{c}_{\text{add}}) \in \mathcal{C}$ provides a detailed characterization of the target domain

  - our primary objective is to learn the integrated denoise function, denoted as

$$\bar{\boldsymbol{\epsilon}}^*(\mathbf{z}_t, t, \mathbf{c}_{\text{text}}, \mathbf{c}_{\text{add}}) := \mathbb{E}[\boldsymbol{\eta} \mid \mathbf{z}_t, \mathbf{c}_{\text{text}}, \mathbf{c}_{\text{add}}].$$

  - By Bayes' theorem, we have

$$p(\mathbf{z}_t \mid \mathbf{c}_{\text{text}}, \mathbf{c}_{\text{add}}) = \frac{p(\mathbf{c}_{\text{add}} \mid \mathbf{z}_t, \mathbf{c}_{\text{text}})}{p(\mathbf{c}_{\text{add}} \mid \mathbf{c}_{\text{text}})} p(\mathbf{z}_t \mid \mathbf{c}_{\text{text}}).$$

  - To take advantage of the pretrained model $\boldsymbol{\epsilon}^*(\mathbf{z}_t, t, \mathbf{c}_{\text{text}}) := \mathbb{E}[\boldsymbol{\eta} \mid \mathbf{z}_t, \mathbf{c}_{\text{text}}]$, we have

$$\begin{aligned}\bar{\boldsymbol{\epsilon}}^*(\mathbf{z}_t, t, \mathbf{c}_{\text{text}}, \mathbf{c}_{\text{add}}) &= -\sqrt{1-\bar{\alpha}_t}[\nabla \log p(\mathbf{c}_{\text{add}} \mid \mathbf{z}_t, \mathbf{c}_{\text{text}}) + \nabla \log p(\mathbf{z}_t \mid \mathbf{c}_{\text{text}})] \\ &= -\sqrt{1-\bar{\alpha}_t} \nabla \log p(\mathbf{c}_{\text{add}} \mid \mathbf{z}_t, \mathbf{c}_{\text{text}}) + \boldsymbol{\epsilon}^*(\mathbf{z}_t, t, \mathbf{c}_{\text{text}}).\end{aligned}$$

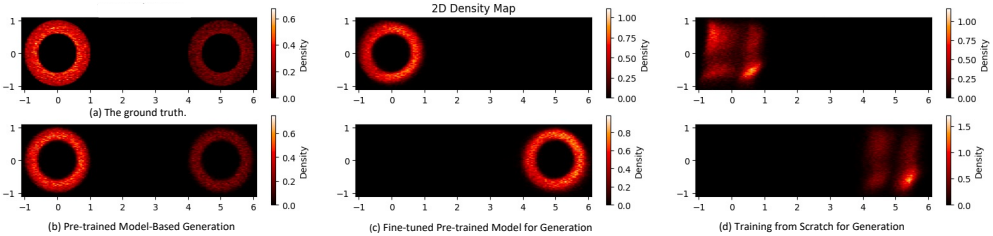
2 Formulation

- Alternatively, can the deployment of such pre-trained large-scale models, in conjunction with the private dataset, effectively facilitate the transition from a large probability space to a small probability space?

2 Formulation

- Alternatively, can the deployment of such pre-trained large-scale models, in conjunction with the private dataset, effectively facilitate the transition from a large probability space to a small probability space?

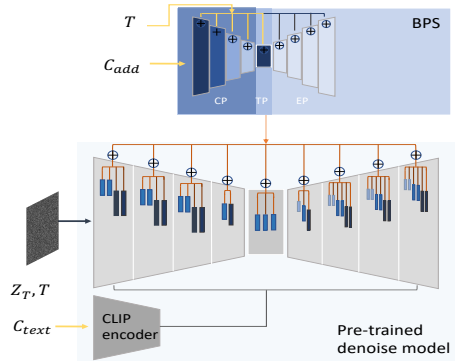
Yes! Through a Bayesian fine-tuning approach!



Roadmap

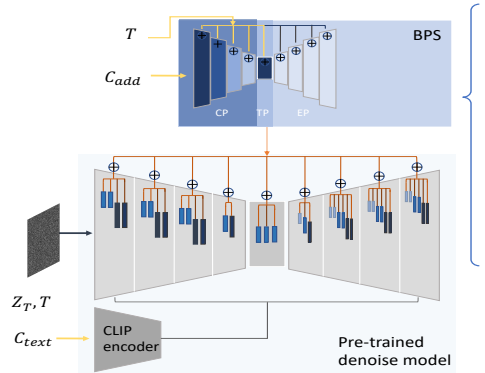
3 Methodology

- **Bayesian Power Steering (BPS)**



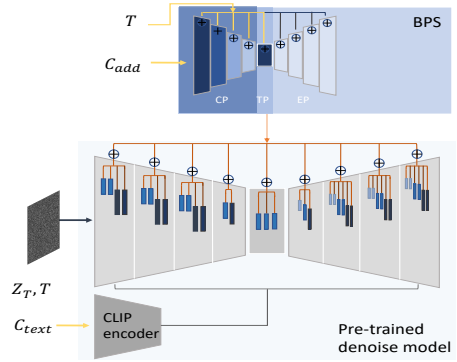
3 Methodology

- **Bayesian Power Steering (BPS)**



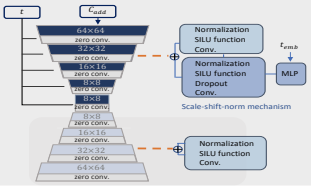
3 Methodology

- **Bayesian Power Steering (BPS)**



- **Architecture Design**

- Head-heavy and foot-light configuration.



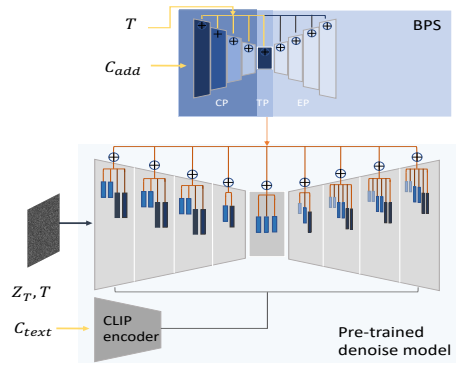
- Differentiated integration structure.

$$\boldsymbol{v}_{add} = \boldsymbol{B}_{\boldsymbol{\phi}}(t, \boldsymbol{c}_{add})$$

$$\widehat{\boldsymbol{h}}^{i,j} = \boldsymbol{h}^{i,j} + \omega_i \boldsymbol{v}_{add}^{i}, i = 1,2,\dots,21$$

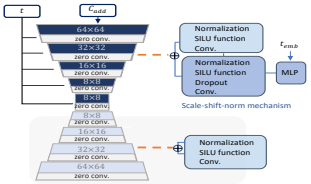
3 Methodology

- **Bayesian Power Steering (BPS)**



- **Architecture Design**

- Head-heavy and foot-light configuration.



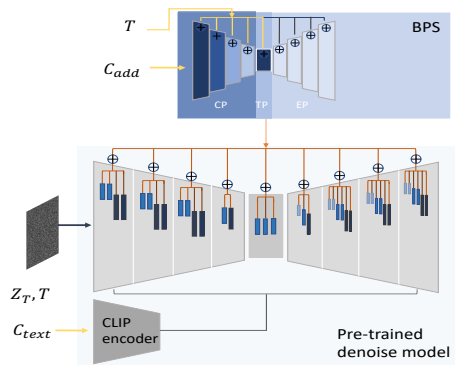
- Differentiated integration structure.

$$\boldsymbol{v}_{add} = \boldsymbol{B}_{\boldsymbol{\phi}}(t, \boldsymbol{c}_{add})$$

$$\widehat{\boldsymbol{h}}^{i,j} = \boldsymbol{h}^{i,j} + \omega_i \boldsymbol{v}_{add}^{i}, i = 1,2,\dots,21$$

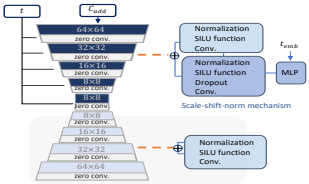
3 Methodology

- **Bayesian Power Steering (BPS)**



- **Architecture Design**

- Head-heavy and foot-light configuration.



- Differentiated integration structure.

$$\boldsymbol{v}_{add} = \boldsymbol{B}_{\boldsymbol{\phi}}(t, \boldsymbol{c}_{add})$$

$$\widehat{\boldsymbol{h}}^{i,j} = \boldsymbol{h}^{i,j} + \omega_i \boldsymbol{v}_{add}^{i}, i = 1,2,\dots,21$$

- **Optimization Process**

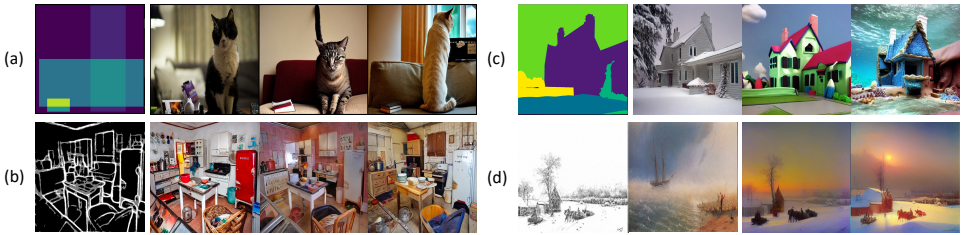
$$\mathcal{L}_{\phi} = \mathbb{E}_{z_0,t,\boldsymbol{\epsilon},\boldsymbol{c}_{add}}[\| \boldsymbol{\epsilon} - \bar{\boldsymbol{\epsilon}}_{\bar{\boldsymbol{\theta}},\boldsymbol{\phi}}(\boldsymbol{z}_t, t, \boldsymbol{c}_{text}, \boldsymbol{c}_{add}) \|_2^2]$$

$\boldsymbol{c}_{text}$
- i. a generalized overview
- ii. a object description
- iii. a detailed portrayal of the objects and their states

Roadmap

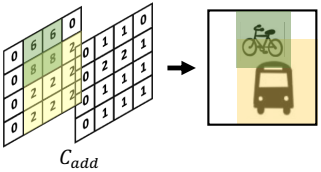
4 Experiments

- **Diverse applications**, including **(Figure (a))** layout-to-image, **(Figure (b))** sketch-to-image, **(Figure (c))** segmentation-to-image, and even **(Figure (d))** multi-conditional tasks (e.g., with style and line constraints).

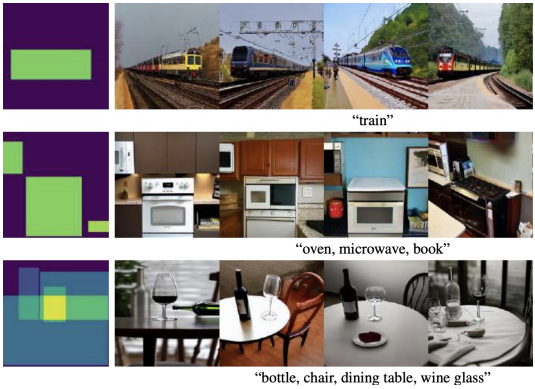


4 Experiments

- layout-to-image task.

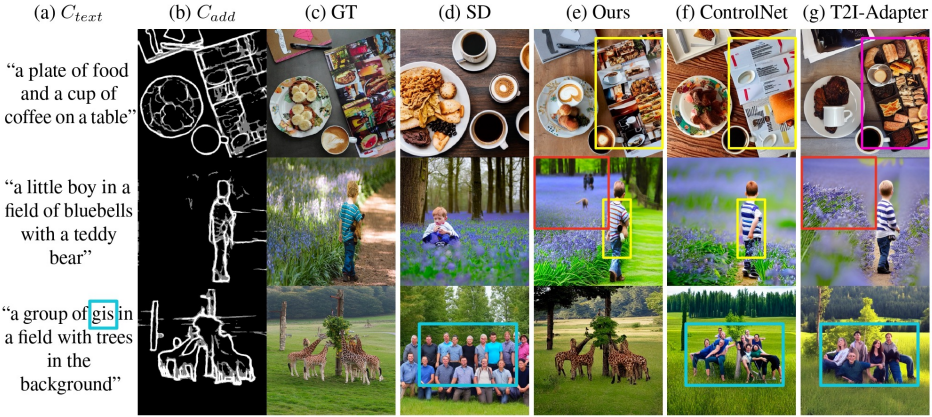


Class label
2: Bus
6: Bicycle



4 Experiments

- **Comparison**



4 Experiments

Tabela: Quantitative comparison: Stable Diffusion (Rombach et al., 2022), ControlNet (Zhang et al., 2023), T2I-Adapter (Mou et al., 2023) and our BPS.

	SD	**ControlNet**	**T2I-Adapter**	**Ours**
FID ↓	20.59	19.41	18.39	**10.49**
CLIP Score ↑	**0.2647**	0.2361	0.2642	0.2614
Quality↑	/	1.77	1.868	**2.38**
Fidelity↑	/	1.53	1.95	**2.52**

4 Experiments

- **Ablation Study**
  - **Architecture.**

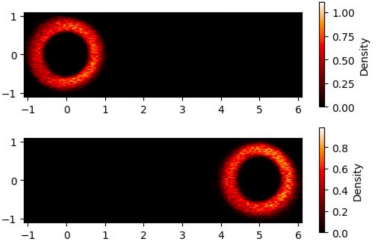
    Our study utilizes a pretrained model $\epsilon_{\hat{\theta}}$ based on the U-net architecture. This architecture consists of an encoder (E.), a middle block (MB.), a skip-connected decoder (D.), and skip-connections between the encoder and decoder (E-D.).

Tabela: Schemes for integrating residual structures across various hierarchical levels of the feature space.

Mode	E.	MB.	D.	E-D.
ALL	√	√	√	√
EMD	√	√	√	×
E	√	×	×	×
EM	√	√	×	×
D	×	×	√	×
MD	×	√	√	×
E-D	×	×	×	√
ME-D	×	√	×	√
M	×	√	×	×

4 Experiments

(a) Generation of multimodal 2D data.



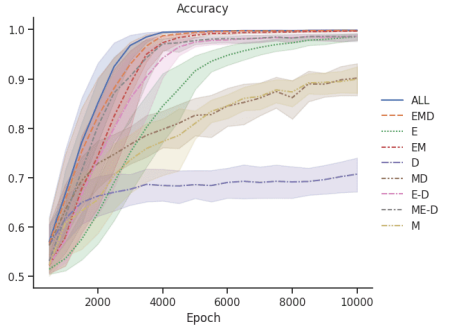
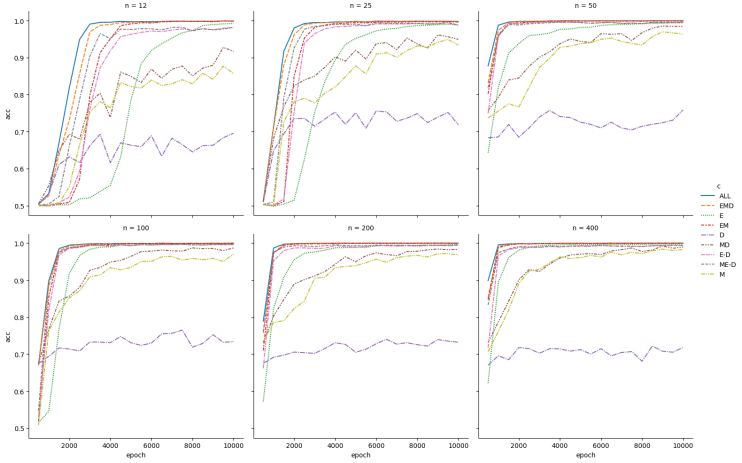


Figura: Classification accuracy of generated samples from distinct integration modes.

4 Experiments



Robustness analysis of the integrated Models of distinct modes with respect to sample size and number of iterations. The horizontal coordinate is the number of epochs and the vertical coordinate is the accuracy.

(b) Segmentation-to-image.

For this task, the performance is assessed using the ADE20K dataset (Zhou et al., 2017). The conditioning fidelity is evaluated through Mean Intersection-over-Union (mIoU).

Tabela: Evaluation of semantic segmentation label reconstruction with mIoU ↑.

ALL	EMD	EM	ME-D	MD	BPS
0.351	0.351	0.350	0.163	0.240	**0.366**

4 Experiments

— **Fusion with prompt information.**

(a) C_{add}

(b) “Nestled amidst lush foliage, the country house unveils its enchanting beauty as summer cascades upon it.”

(c) “Amidst the fall’s magical embrace, a country house stands adorned with a vibrant carpet of fallen leaves.”

(d) “On a serene winter day, the country house stands gracefully amidst a pristine blanket of snow. ”

(e) “A wooden house built in the desert.”

(f) “a house is built under the sea.”

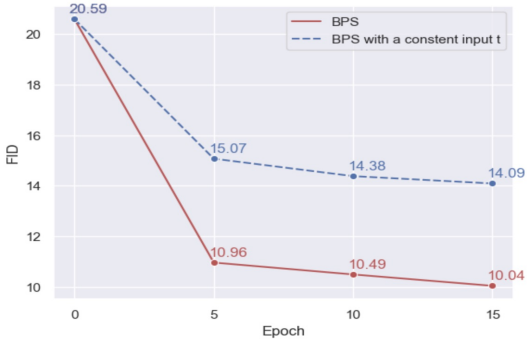
(g) “” a country house is constructed with toy bricks.”

(h) “a House made of toy plasticine.”

4 Experiments

— **Implication of time-step information and the training efficiency.**

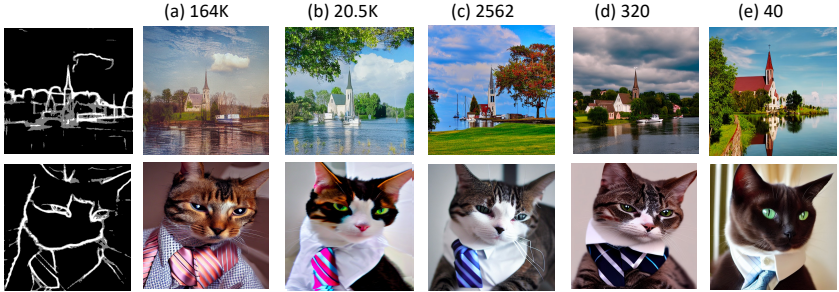
In the sketch-to-image task, we conduct a controlled experiment by setting the time input of BPS to a constant value. The figure below shows the convergence of models.



4 Experiments

— **Data scarcity.**

To assess the model's generalization ability in scenarios with limited data, we randomly select subsets of 40, 320, 2562, 20500, and 164K images from the training and validation sets for fine-tuning.



4 Experiments

— **Intervention ability.**

Explore the significant influence of the intervention weights (w_i, where $1 \leq i \leq 21$) within the BPS framework on the outcomes of the interventions. Specifically, decreasing these weights contributes to an increase in diversity among the generated results.

References I

Chung, K. L. (2001). *A course in probability theory*. Academic press.

Efron, B. (2011). Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614.

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

Mou, C., Wang, X., Xie, L., Zhang, J., Qi, Z., Shan, Y., and Qie, X. (2023). T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*.

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*, pages 10684–10695.

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*, pages 3836–3847.

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*, pages 633–641.

Thank you!

Paper