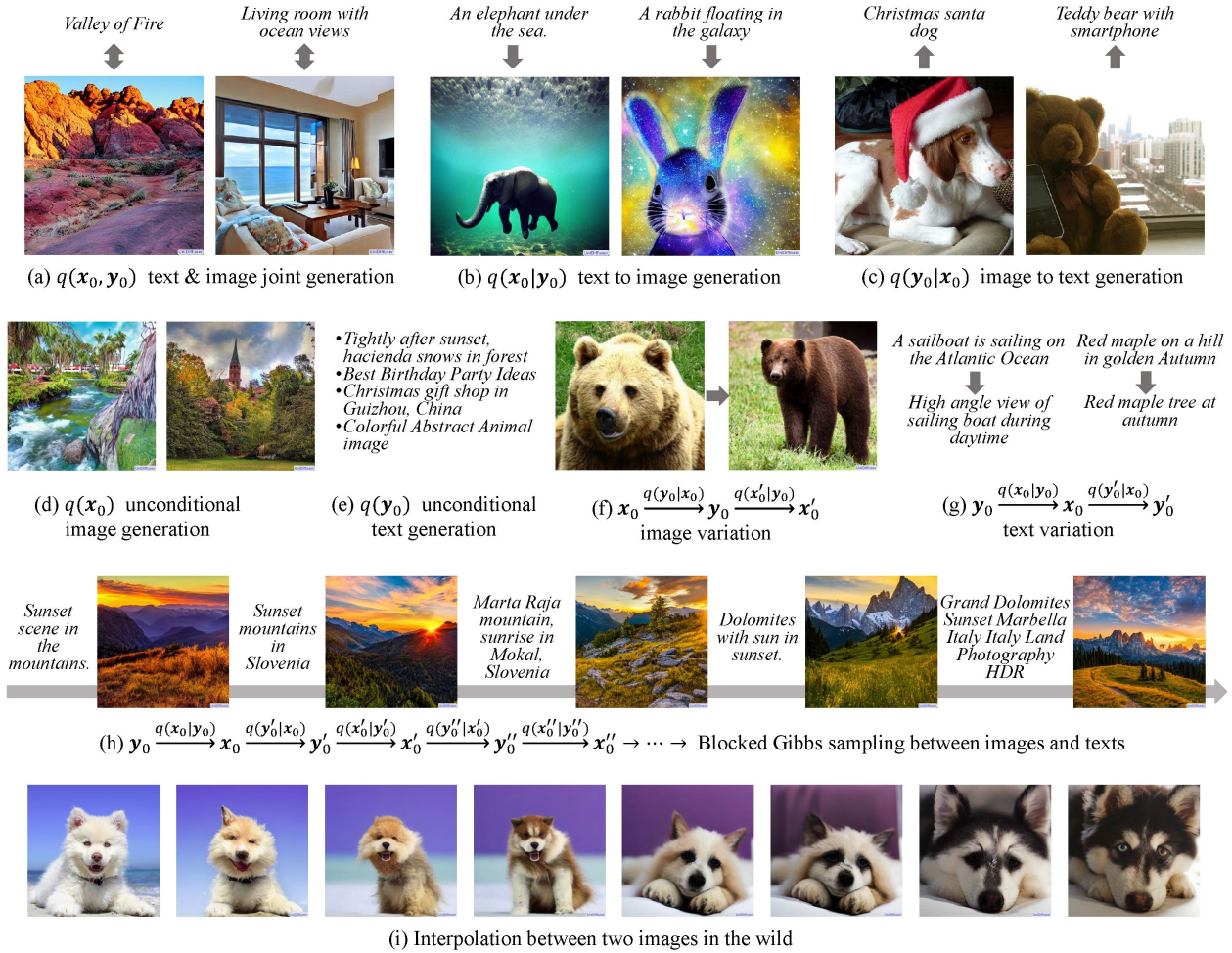
One Transformer Fits All Distributions in Multi-Modal Diffusion at Scale

Fan Bao , Shen Nie , Kaiwen Xue , Chongxuan Li , Shi Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su , Jun Zhu

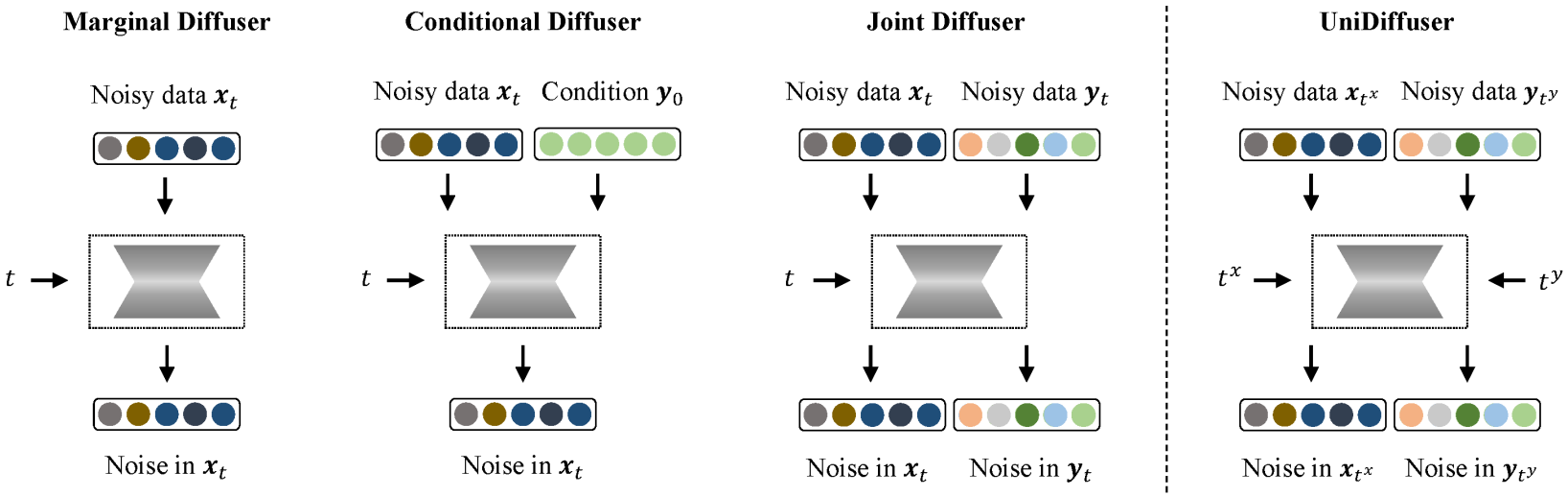
Introduction

Our paper presents a diffusion-based framework (dubbed UniDiffuser) that explicitly fits **all relevant distributions** in **one model without** introducing additional training or inference overhead.

Our key insight is – learning diffusion models for all distributions can be unified as predicting the noise in the perturbed data, where the perturbation levels (i.e. timesteps) can be different for different modalities.



Training loss

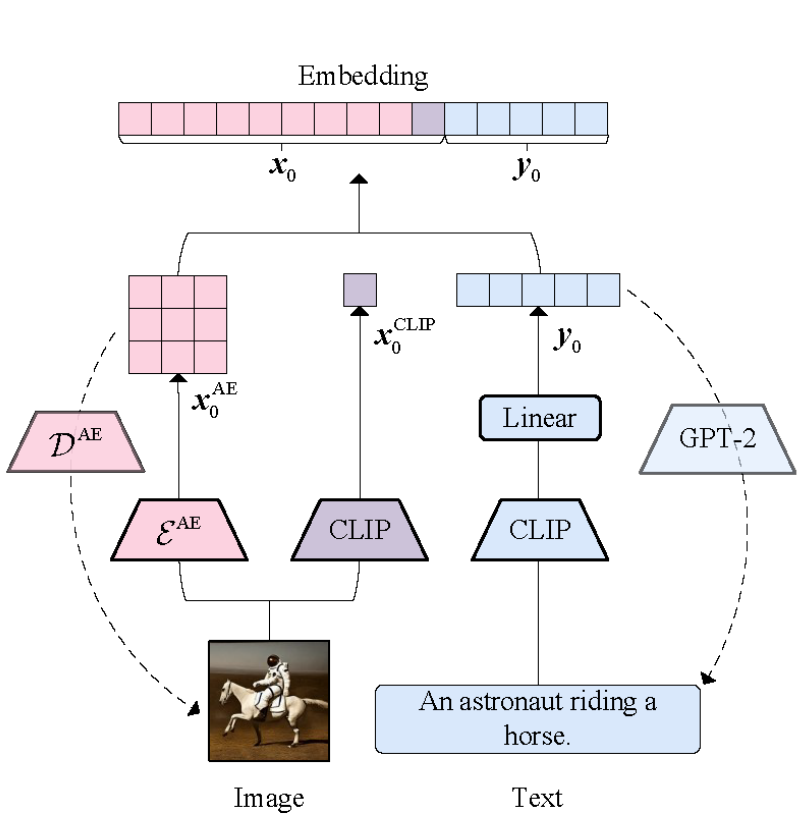


$$loss = \mathbb{E}_{x_0,y_0,\epsilon^x,\epsilon^y,t^x,t^y}||\epsilon_\theta(x_{t^x}, y_{t^y}, t^x, t^y) - [\epsilon^x, \epsilon^y]||_2^2$$

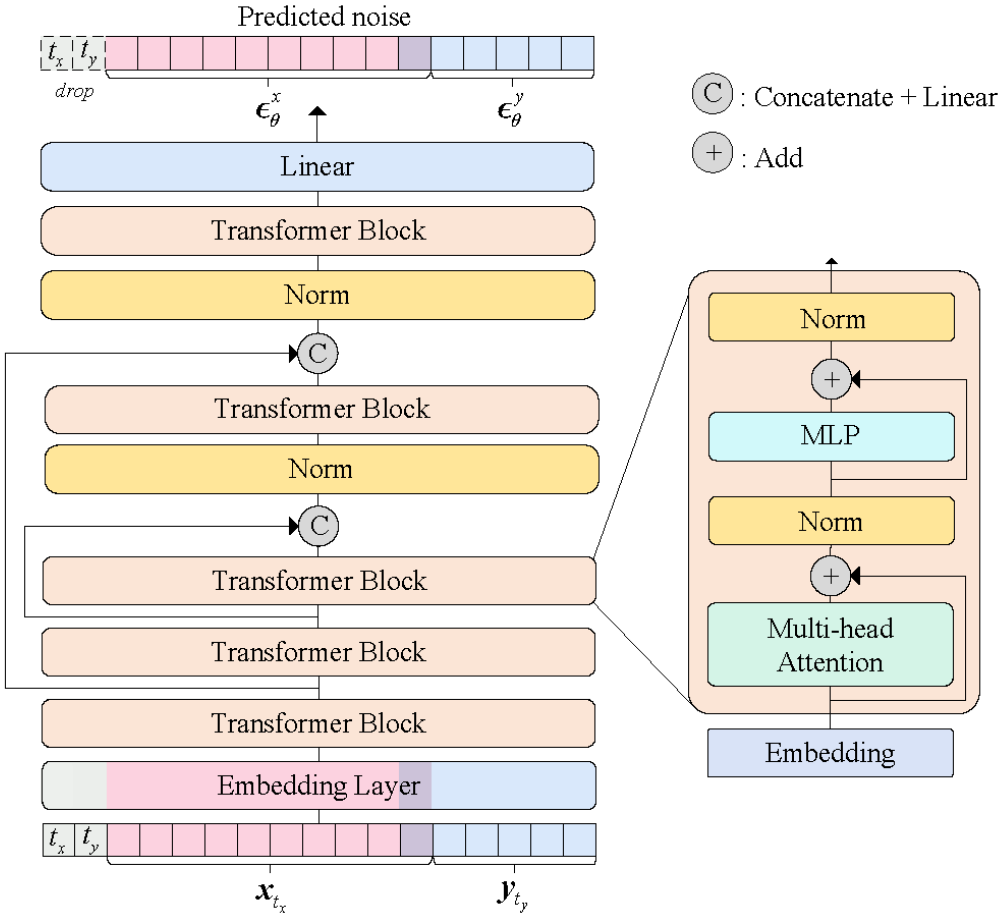
In comparison, learning a single joint distribution over multi-modal data requires additional procedures (e.g., Markov Chain Monte Carlo) to sample from the marginal or conditional distributions, which is unaffordable on large-scale multi-modal data

Notably, Classfier-Free Duidance can be directly applicable to the conditional and joint sampling of UniDiffuser without modifying the training process

Transformer-based network



(a) Encode images & texts into latent space



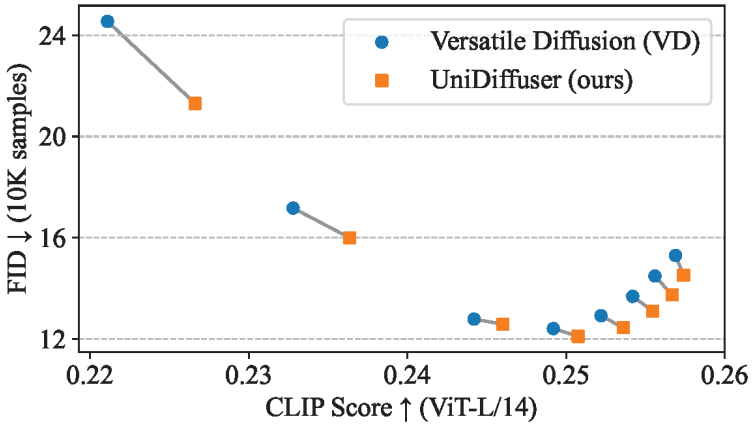
(b) The U-ViT backbone of the joint noise prediction network

Experiment

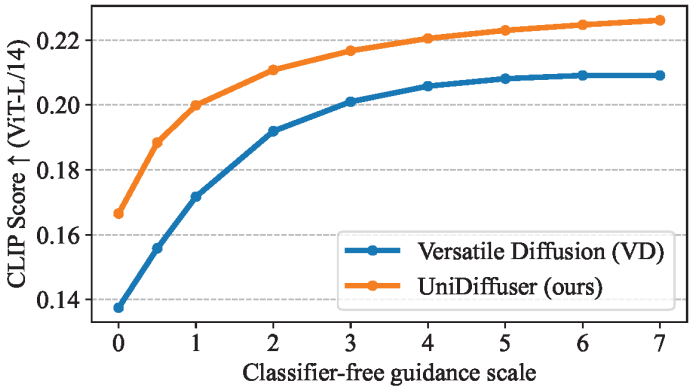
Zero-shot FID on the MS-COCO validation set.

Model	FID ↓
Bespoken models	
GLIDE (Nichol et al., 2022)	12.24
Make-A-Scene (Gafni et al., 2022)	11.84
DALL·E 2 (Ramesh et al., 2022)	10.39
Stable Diffusion† (Rombach et al., 2022)	8.59
Imagen (Saharia et al., 2022)	7.27
Parti (Yu et al., 2022)	**7.23**
General-purpose models	
Versatile Diffusion† (Xu et al., 2022)	10.09
UniDiffuser (**ours**)	**9.71**

Comparing UniDiffuser and Versatile Diffusion in text-to-image generation



Comparing UniDiffuser and Versatile Diffusion in image-to-text generation



Versatile Diffusion

Unidiffuser

A dog wearing a beret

Versatile Diffusion

Unidiffuser

An elephant under the sea

Thanks