Flowing from Words to Pixels: A Noise-Free Framework for Cross-Modality Evolution

Qihao Liu[1,2], Xi Yin[1], Alan Yuille[2], Andrew Brown[1], Mannat Singh[2]

[1] Meta GenAI [2] Johns Hopkins University

CVPR 2025 Highlight

Code and models

Motivation

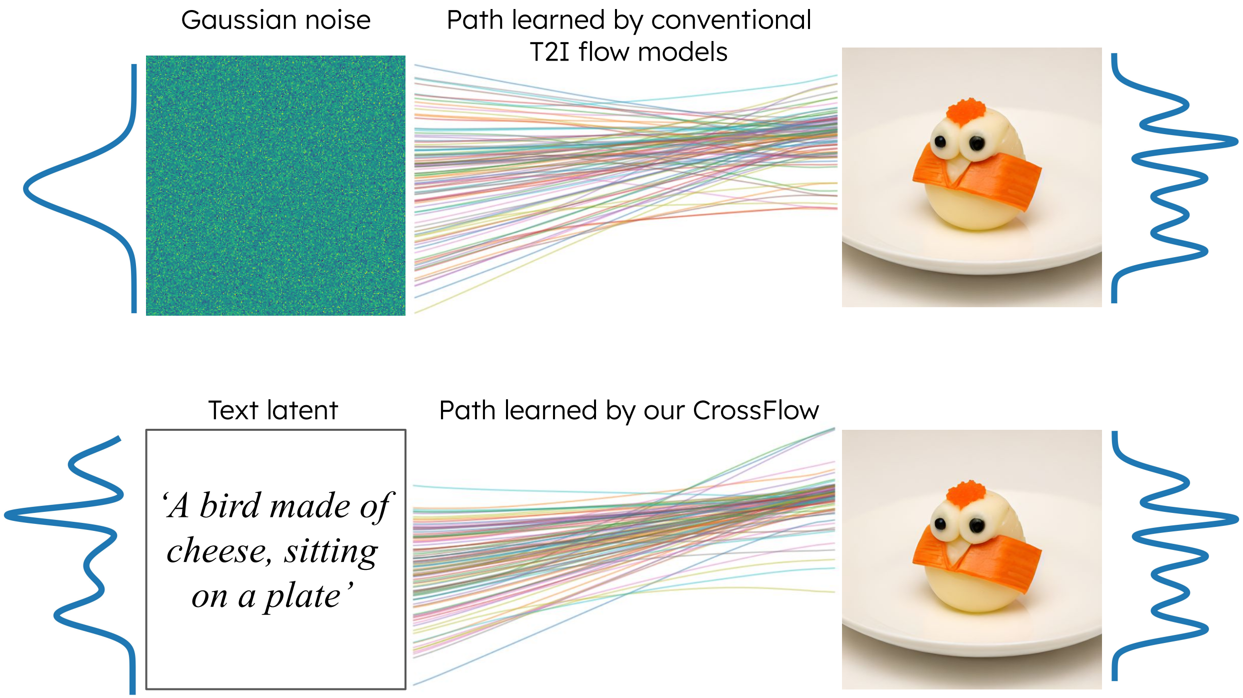
Diffusion models & Flow Matching models

Stable Diffusion 3

Movie Gen

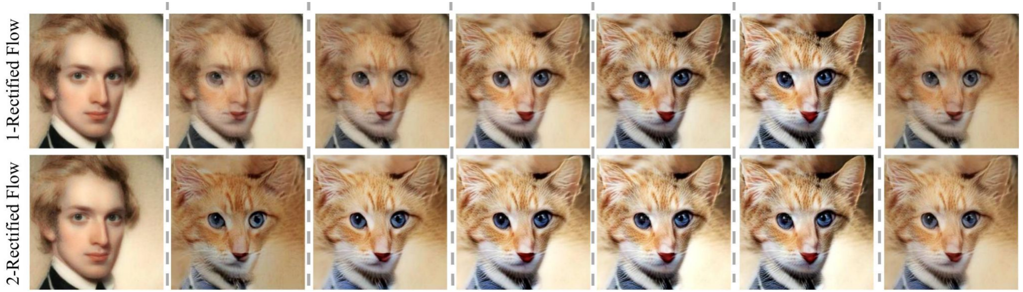
Motivation

Starting from Gaussian noise vs. from correlated distribution

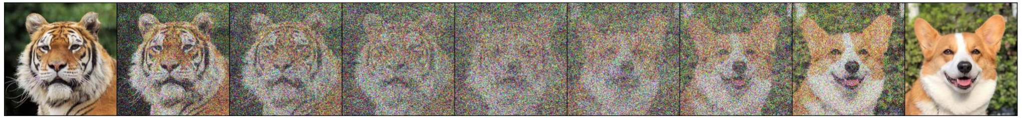


Related Work

Mapping between two similar intra-modal distributions (e.g., face to face)



Rectified Flow

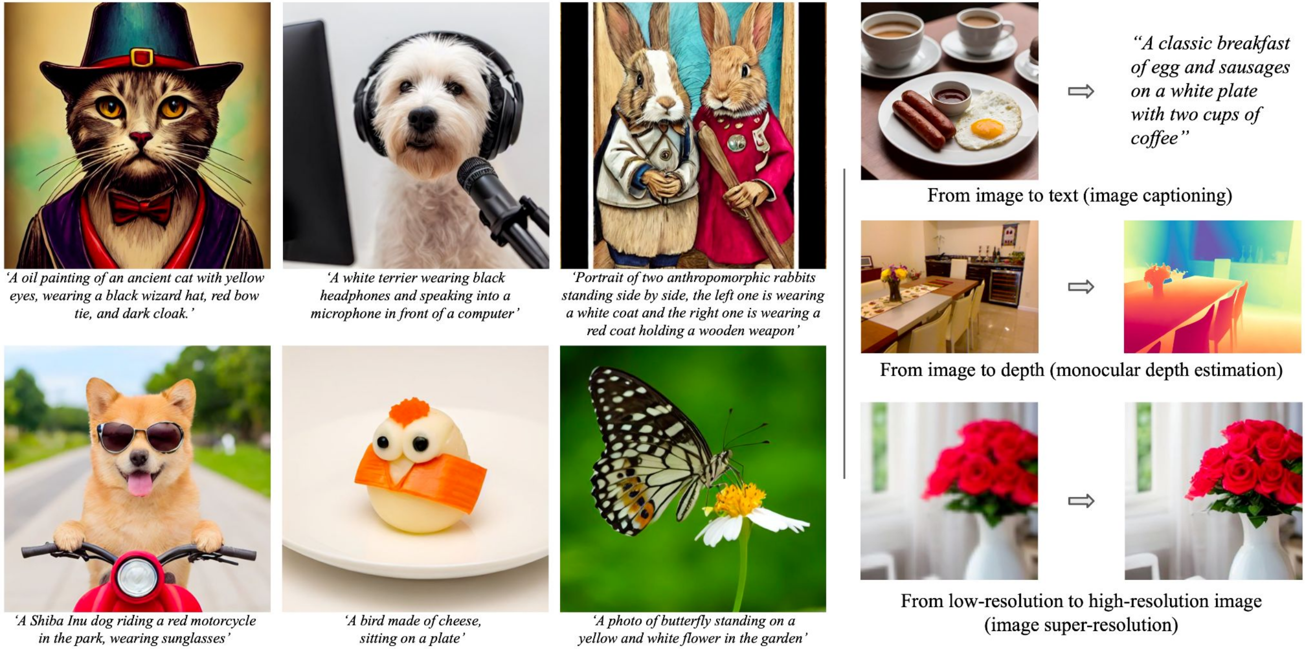


Denoising Diffusion Bridge Models

Generalized Schrödinger Bridge Matching

CrossFlow

Learning a direct mapping between modalities with flow matching



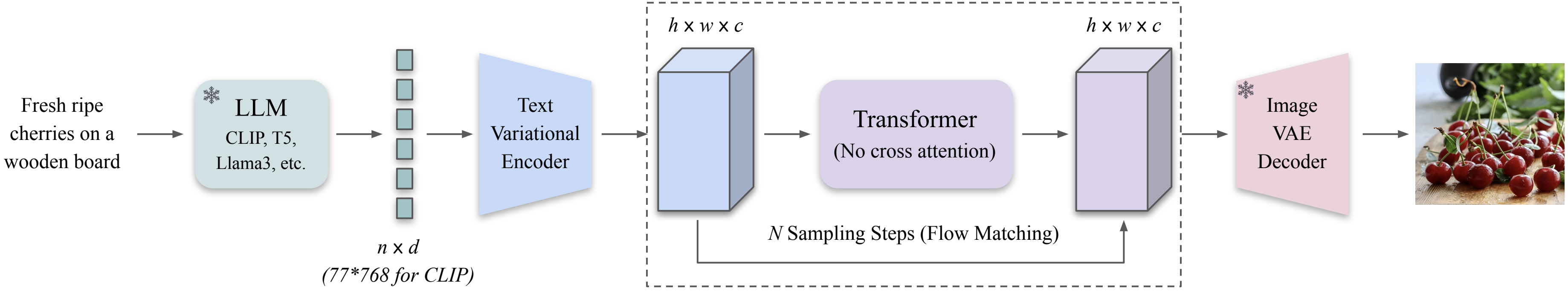
(a) Directly evolving text into images for Text-to-Image generation

(b) CrossFlow for various tasks

Method

Key: encoding the source modality data into a **regularized distribution**

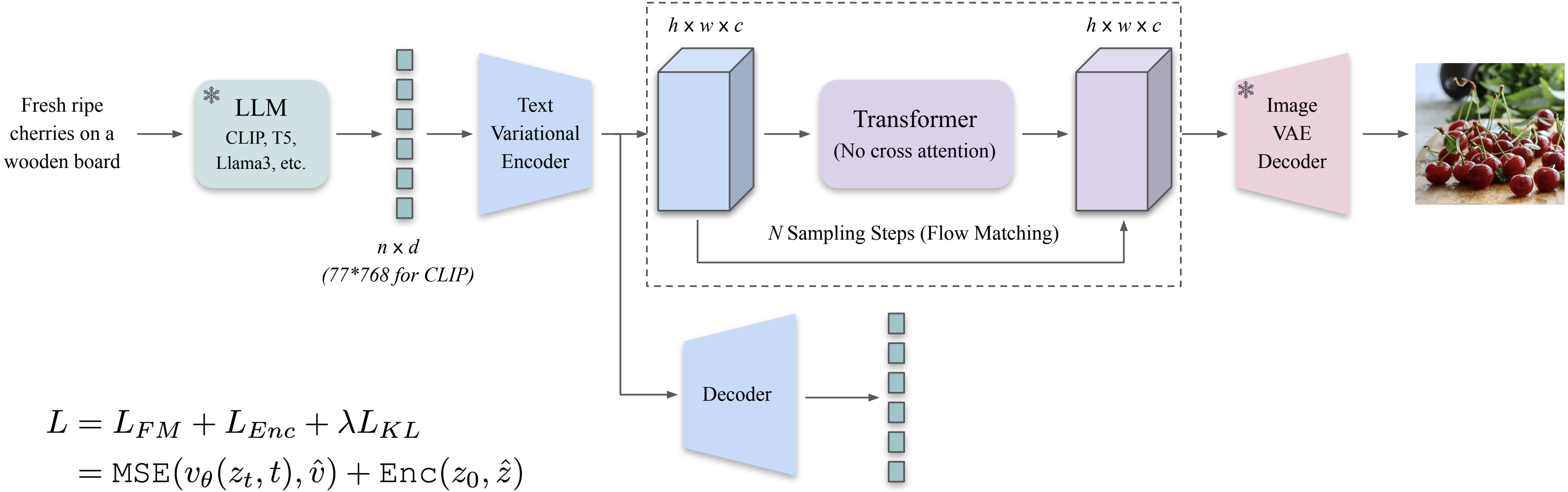
- Variational Encoder



Method

Key: encoding the source modality data into a **regularized distribution**

- Joint training with VE loss



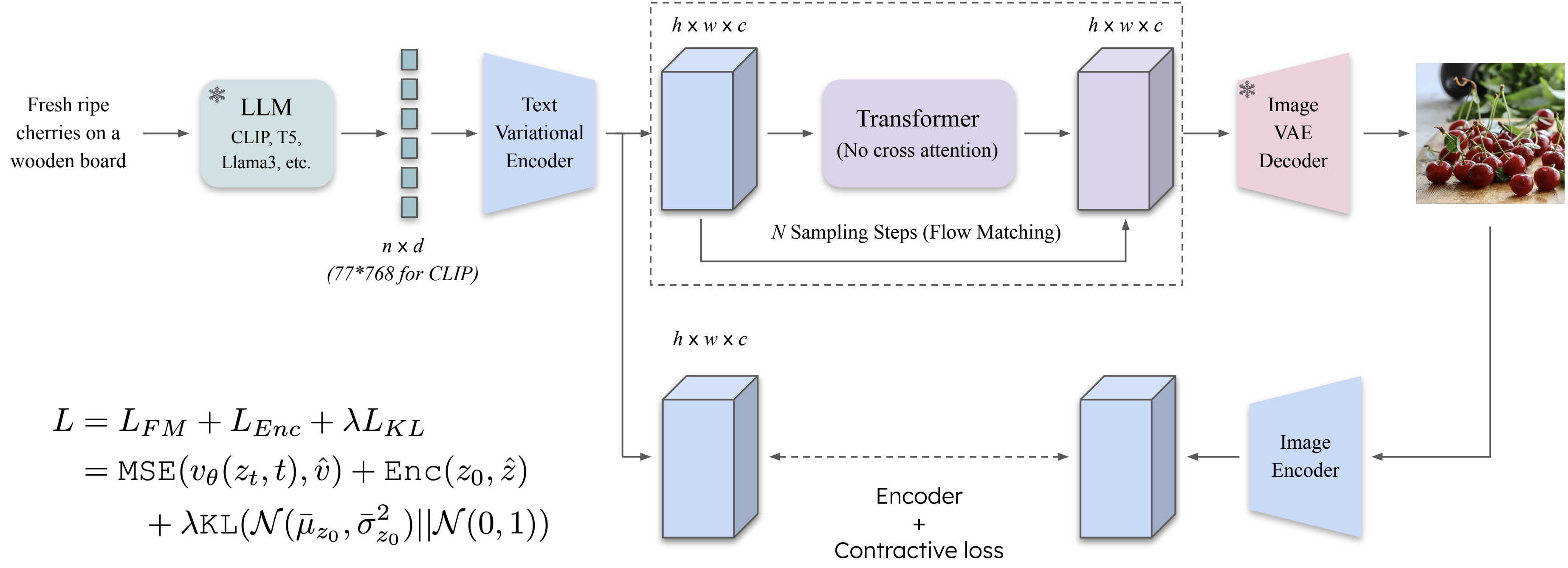
$$L = L_{FM} + L_{Enc} + \lambda L_{KL}$$
$$= \mathtt{MSE}(v_\theta(z_t, t), \hat{v}) + \mathtt{Enc}(z_0, \hat{z})$$
$$+ \lambda \mathtt{KL}(\mathcal{N}(\bar{\mu}_{z_0}, \bar{\sigma}^2_{z_0}) || \mathcal{N}(0, 1))$$

decoder + reconstruction loss

Method

Key: encoding the source modality data into a **regularized distribution**

- Joint training with VE loss

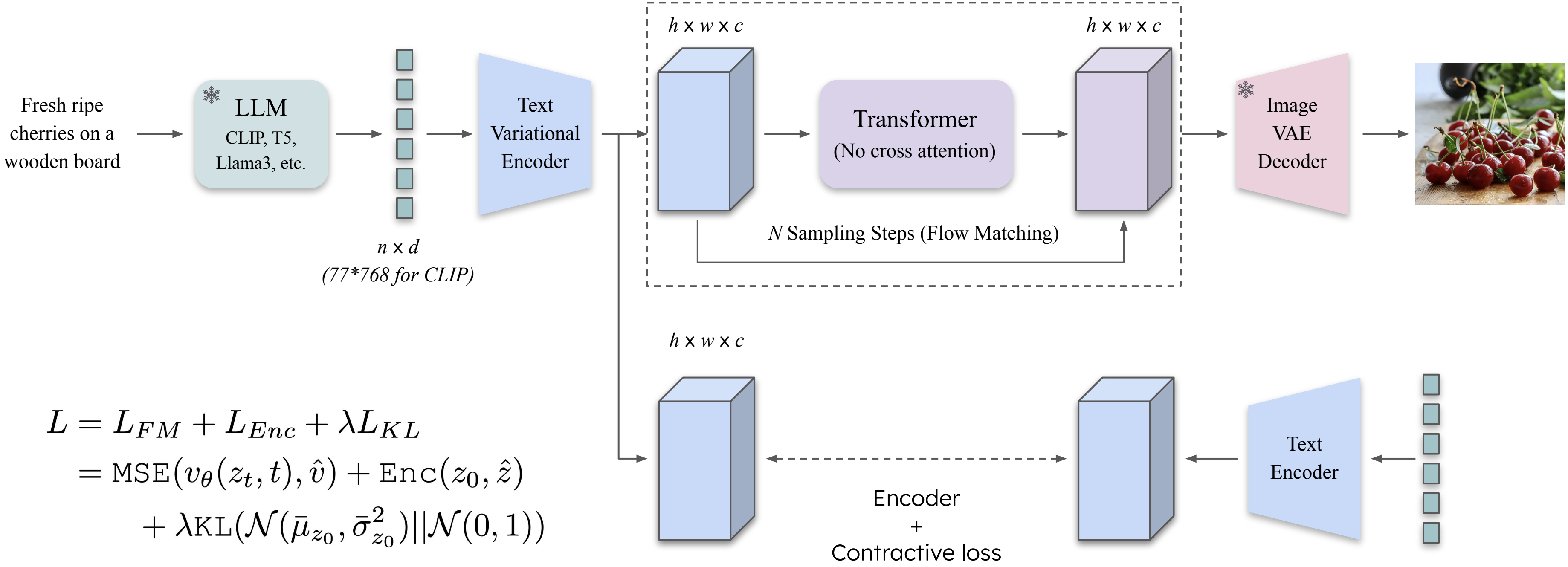


$$
\begin{aligned}
L &= L_{FM} + L_{Enc} + \lambda L_{KL} \\
&= \mathtt{MSE}(v_\theta(z_t, t), \hat{v}) + \mathtt{Enc}(z_0, \hat{z}) \\
&\quad + \lambda \mathtt{KL}(\mathcal{N}(\bar{\mu}_{z_0}, \bar{\sigma}^2_{z_0}) || \mathcal{N}(0, 1))
\end{aligned}
$$

Method

Key: encoding the source modality data into a **regularized distribution**

- Joint training with VE loss



$$\begin{aligned} L &= L_{FM} + L_{Enc} + \lambda L_{KL} \\ &= \mathtt{MSE}(v_\theta(z_t, t), \hat{v}) + \mathtt{Enc}(z_0, \hat{z}) \\ &\quad + \lambda \mathtt{KL}(\mathcal{N}(\bar{\mu}_{z_0}, \bar{\sigma}^2_{z_0}) || \mathcal{N}(0, 1)) \end{aligned}$$

Method

Classifier-free guidance

- Standard CFG can only be applied to methods with additional conditioning input *c*

$$\tilde{v}_\theta(z_t, c) = \omega v_\theta(z_t, c) + (1-\omega) v_\theta(z_t)$$

- CFG with an indicator $1_c \in \{0, 1\}$

$$\tilde{v}_\theta(z_t) = \omega v_\theta(z_t, 1_c = 1) + (1-\omega) v_\theta(z_t, 1_c = 0)$$

'a cat playing chess'

Indicator 1c= 1

'a cat playing chess'

Indicator 1c= 0

Results

Text-to-image generation

Method	#Params (B)	#Steps (K)	FID ↓	CLIP ↑
Standard FM (Baseline)	1.04	300	10.79	0.29
CrossFlow (Ours)	0.95	300	10.13	0.29

Table 1. **Comparison between our CrossFlow and standard flow matching with cross-attention.**

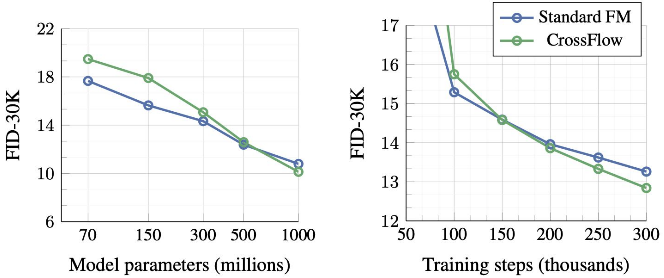


Figure 3. **Performance *vs*. Model Parameters and Iterations.**

Method	#Params.	FID-30K ↓ *zero-shot*	GenEval ↑ score
DALL·E [72]	12.0B	27.50	-
GLIDE [63]	5.0B	12.24	-
LDM [77]	1.4B	12.63	-
DALL·E 2 [73]	6.5B	10.39	0.52
LDMv1.5 [77]	0.9B	9.62	0.43
Imagen [78]	3.0B	7.27	-
RAPHAEL [92]	3.0B	6.61	-
PixArt-α [11]	0.6B	7.32	0.48
LDMv3 (512^2) [23]	8.0B	-	0.68
CrossFlow	0.95B	9.63	0.55
CrossFlow (Sin-Cos)	0.95B	8.95	0.57

Table 2. **Comparison with recent T2I models.**

Text encoder	FID ↓	CLIP ↑
Encoder	66.65	0.20
Encoder + noise	59.91	0.21
Variational Encoder	40.78	0.23

(a) **Variational Encoder** *

Loss	FID ↓	CLIP ↑
T-T Recon.	40.78	0.23
T-T Contrast.	34.67	0.24
I-T Contrast.	33.41	0.24

(b) **Text VE loss***

Method	FID ↓	CLIP ↑
No guidance	33.41	0.24
AG	26.36	0.25
CFG indicator	24.33	0.26

(c) **CFG with indicator**

Model	FID ↓	CLIP ↑
CLIP (0.4B)	24.33	0.26
T5-XXL (11B)	22.28	0.27
Llama3 (7B)	21.20	0.27

(d) **Language Model**

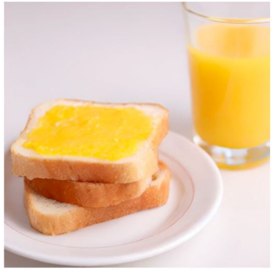
Train strategy	FID ↓	CLIP ↑
2-stage separate training	32.55	0.24
Joint training	24.33	0.26
2-stage w/ joint finetuning	23.79	0.26

(e) **Training strategy**

Table 3. **Ablation study** on Text Variational Encoder, training objective, CFG, language models, and training strategy. We conduct ablation study on our smallest model (70M), reporting *zero-shot* FID-10K and CLIP scores. Final settings used for CrossFlow are underlined. AG: Autoguidance. *: results without applying CFG.

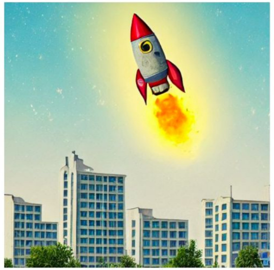
Results

Text-to-image generation



'a glass of orange juice to the right of a plate with buttered toast on it'

'a teddy bear on a skateboard in times square'



'a painting of a rocket lifting off from the city'

'a teddy bear sitting on a yellow toy pickup truck'

'a black dog is playing chess with a white dog'

'three birds standing on a wire stock'

'five frosted glass bottles'

'two cats doing research'

'a close-up of milk pouring into a white bowl against a black background'

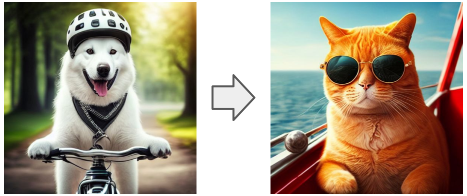
'a close-up of the eyes of an owl'

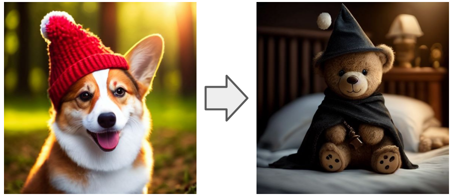
'a black and white landscape photograph of a black tree'

'a cute illustration of a horned owl with a graduation cap and diploma'

Results

Linear interpolation in latent space (z0)





Results

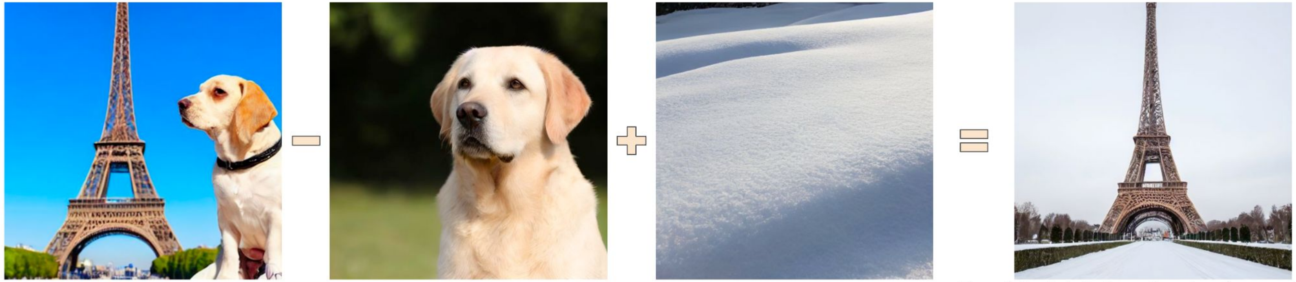
Arithmetic operations in latent space (z0)

z_0 = VE('*A white dog wearing a black hat*')

z_0 = VE('*Sunglasses*')

z_0 = VE('*A hat*')

z_0 = VE('*A white dog wearing a black hat*') + VE('*Sunglasses*') - VE('*A hat*')



z_0 = VE('*A labrador in front of Eiffel Tower*')

z_0 = VE('*A labrador*')

z_0 = VE('*snow*')

z_0 = VE('*A labrador in front of Eiffel Tower*') - VE('*A labrador*') + VE('*snow*')

Results

Arithmetic operations in latent space (z0)

Arithmetic Operation	Success Rate (%)
Addition	95.3
Subtraction	92.7
Combination	87.5
Overall	91.4

Table 8. **Success rate of arithmetic operation on 1,000 prompts from COCO-val.**

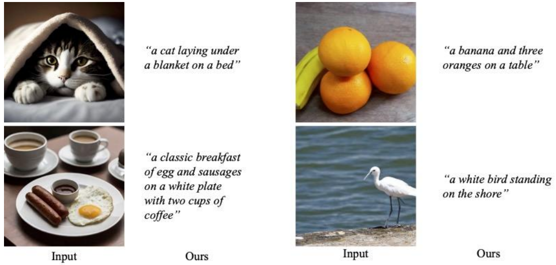
Results

CrossFlow for various tasks (without task specific architectures)

Image captioning

Method	B@4 ↑	M ↑	R ↑	C ↑	S ↑
MNIC [24]	30.9	27.5	55.6	108.1	21.0
MIR [43]	32.5	27.2	-	109.5	20.6
NAIC-CMAL [28]	35.3	27.3	56.9	115.5	20.8
SATIC [96]	32.9	27.0	-	111.0	20.5
SCD-Net [58]	37.3	28.1	58.0	118.0	21.6
CrossFlow (Ours)	36.4	27.8	57.1	116.2	20.4

Table 4. **Image captioning on COCO Karpathy split.**



Depth estimation

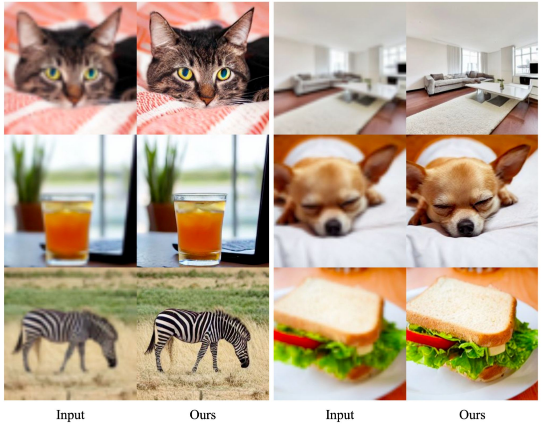
Method	KITTI		NYUv2	
	AbsRel (↓)	δ_1 (↑)	AbsRel (↓)	δ_1 (↑)
TransDepth [89]	0.064	0.956	0.106	0.900
AdaBins [6]	0.058	0.964	0.103	0.903
DepthFormer [45]	0.052	0.975	0.096	0.921
BinsFormer [46]	0.052	0.974	0.094	0.925
DiffusionDepth [18]	0.050	0.977	0.085	0.939
CrossFlow (Ours)	0.053	0.973	0.094	0.928

Table 5. **Monocular depth estimation on KITTI and NYUv2.**

Image super-resolution

Method	FID ↓	IS ↑	PSNR ↑	SSIM ↑
Reference	1.9	240.8	-	-
Regression	15.2	121.1	27.9	0.801
SR3 [75]	5.2	180.1	26.4	0.762
Flow Matching [50]	3.4	200.8	24.7	0.747
CrossFlow (Ours)	3.0	207.2	25.6	0.764

Table 6. **Image super-resolution on the ImageNet validation set.** Compared with standard SR method with flow matching, our



Thank you