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

Motivation

- In theory, flow matching should work on any two distributions. However, prior works only works with matching similar distributions, or set the source distribution to gaussian.
- Cross-modality generation *guides* gaussian to target distribution mapping using conditioning mechanisms. However, directly mapping one modality to another without the need for noise should be easier and more efficient.

Preliminary: Flow Matching

Recall that a *flow matching* is a mapping from a source distribution p_0 to a target distribution p_1 via the prescribed ODE. Given an ODE, or a forward process

$$z_t = tz_1 + (1 - (1 - \sigma_{\mathsf{min}})t)z_0,$$

where $z_0 \sim p_0$, $z_1 \sim p_1$, the velocity is derived as

$$\hat{v}_t = \frac{dz_t}{dt} = z_1 - (1 - \sigma_{\mathsf{min}})z_0.$$

Then flow model $v_\theta(z_t, t)$ tries to learn the flow matching by approximating the ground truth velocity $\hat{v}_t$.

Framework

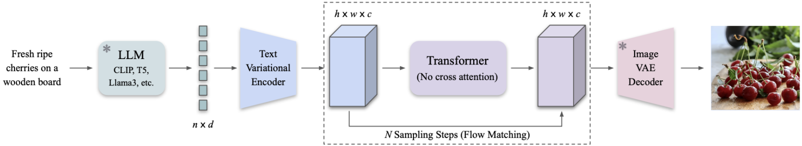

Figure: Framework of CrossFlow

Training loss:

$$\mathcal{L} = \mathcal{L}_{\text{VE}} + \mathcal{L}_{\text{FM}},$$

- $\mathcal{L}_{\text{VE}}$: Variational encoder loss
- $\mathcal{L}_{\text{FM}}$: Flow matching loss e.g. $\text{MSE}(v_\theta, \hat{v})$

Variational Encoder

Flow matching requires the shape of the source and target distributions to be the same. Hence, it is necessary to *convert* the input x from the source distribution to the shape z form the target distribution *without* losing information. Intuitively, one can use an encoder $\mathcal{E}$:

- Deterministic encoder: $z_0 = \mathcal{E}(x)$
- Deterministic encoder + noise: $z_0 = \mathcal{E}(x) + n$ with $n \sim \mathcal{N}(0, \sigma)$
- Variation encoder: $z_0 \sim \mathcal{N}(\mu_x, \sigma_x)$, where $\mu_x, \sigma_x = \mathcal{E}(x)$

Empirically, the authors have reported that the variational encoder yields the best performance. My intuition is that

- Gaussian to p_1 works well
- In diffusion model context, it's been reported that there exist *golden noises* for each text prompt
- Latent space of images consists of sparse disjoint clusters
- Robust to generalization (and composition)

VE loss

$$\mathcal{L}_{\mathsf{VE}} = \mathcal{L}_{\mathsf{enc}} + \lambda \mathcal{L}_{\mathsf{KL}}$$

- $\mathcal{L}_{\mathsf{KL}} = \mathsf{KL}\left(\mathcal{N}(\mu_x, \sigma_x) || \mathcal{N}(0, I)\right)$
  - Controls the noisyness, what about only controlling the variance?
  - Match the framework for image latent from Image VAE
  - Not so different from gaussian to target..?
- $\mathcal{L}_{\mathsf{enc}}$: encoding loss
  (1) Reconstruction loss
  (2) intra-modality contrastive loss
  (3) cross-modality contrastive loss
  (-) Empirically, (1) $<<<$ (2) $<$ (3)

Classifier-Free Guidance

CrossFlow utilizes two learnable tokens g_{c} and g_{uc} for conditional and unconditional generations, respectively. Then we have an analoguous framework as with the conventional CFG:

- $v_\theta(z_t, c) \leftrightarrow v_\theta\left(\mathsf{concat}(z_t, g_{\mathsf{c}})\right)$
- $v_\theta(z_t, \emptyset) \leftrightarrow v_\theta\left(\mathsf{concat}(z_t, g_{\mathsf{uc}})\right)$.

Then one can perform CFG with

$$v_\theta(z_t) = \omega \cdot v_\theta\left(\mathsf{concat}(z_t, g_{\mathsf{c}})\right) + (1-\omega) \cdot v_\theta\left(\mathsf{concat}(z_t, g_{\mathsf{uc}})\right).$$

Authors reported that as with other generative models, CrossFlow yields better performance with CFG.

Experiment: T2I generation

Method	#Params.	FID-30K ↓ *zero-shot*	GenEval ↑ score
DALL·E [68]	12.0B	27.50	-
GLIDE [59]	5.0B	12.24	-
LDM [73]	1.4B	12.63	-
DALL·E 2 [69]	6.5B	10.39	0.52
LDMv1.5 [73]	0.9B	9.62	0.43
Imagen [74]	3.0B	7.27	-
RAPHAEL [88]	3.0B	6.61	-
PixArt-α [10]	0.6B	7.32	0.48
LDMv3 (512^2) [22]	8.0B	-	0.68
CrossFlow	0.95B	9.63	0.55

Figure: Comparison with T2I models

Experiment: Arithmetic on the input latent space

Figure: Arithmetic on the input latent space

Experiment: Various tasks

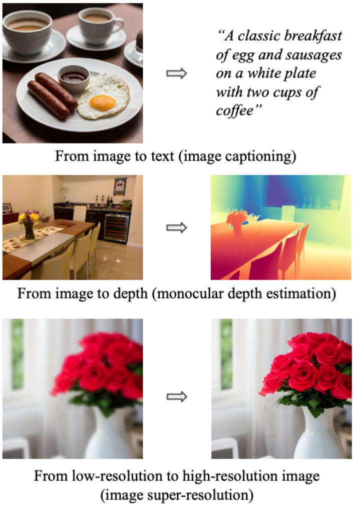

Figure: Various tasks: note for each task a separate model is trained.