EURECOM Sophia Antipolis

HUAWEI

Check our paper & code:

Mail: chao.wang@eurecom.fr

Information Theoretic Text-to-Image Alignment

Chao Wang, Giulio Franzese, Alessandro Finamore, Massimo Gallo, Pietro Michiardi

ICLR 2025

Problem

For existing text-to-image (T2I) diffusion models, precise alignment between given texts and generated images is still challenging (e.g., attribute binding, missing object).

A round bag and a rectangular wallet

A man on the top of a turtle

Research question: Would it be possible to use Mutual Information (MI) between image and text to measure and guide the alignment?

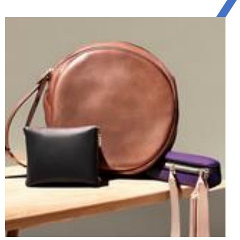
Prerequisites:

1) MI formula supported by discrete-time diffusion model
2) Is MI a meaningful signal for T2I alignment?

Prereq1: Point-wise MI estimation

Given 2 R.V. $\boldsymbol{z}, \boldsymbol{p}$ sampled from the joint distribution $p_{latent,prompt}$,

$$\mathrm{I}(\boldsymbol{z}, \boldsymbol{p}) = \mathbb{E}_{t, \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})} \left[\kappa_t || \boldsymbol{\epsilon_\theta}(\boldsymbol{z}_t, \boldsymbol{p}, t) - \boldsymbol{\epsilon_\theta}(\boldsymbol{z}_t, \emptyset, t) ||^2 \right], \quad \kappa_t = \frac{\beta_t T}{2\alpha_t (1 - \bar{\alpha}_t)}.$$



“A round bag and a rectangular wallet”

Discrete-time diffusion model:

- Forward process: $q(\boldsymbol{z}_{0:T}, \boldsymbol{p}) = q(\boldsymbol{z}_0, \boldsymbol{p}) \prod_{t=1}^{T} q(\boldsymbol{z}_t | \boldsymbol{z}_{t-1})$

  Hand-crafted transition kernel: $q(\boldsymbol{z}_t | \boldsymbol{z}_{t-1}, \boldsymbol{p}) = q(\boldsymbol{z}_t | \boldsymbol{z}_{t-1}) = \mathcal{N}(\boldsymbol{z}_t; \sqrt{1 - \beta_t} \boldsymbol{z}_{t-1}, \beta_t \boldsymbol{I})$

- Backward process: $p_\theta(\boldsymbol{z}_{0:T} | \boldsymbol{p}) = p(\boldsymbol{z}_T) \prod_{t=1}^{T} p_\theta(\boldsymbol{z}_{t-1} | \boldsymbol{z}_t, \boldsymbol{p})$

  Learnable transition kernel: $p_\theta(\boldsymbol{z}_{t-1} | \boldsymbol{z}_t, \boldsymbol{p}) = \mathcal{N}(\boldsymbol{z}_{t-1}; \boldsymbol{\mu}_\theta(\boldsymbol{z}_t, \boldsymbol{p}), \beta_t \boldsymbol{I})$, with $\boldsymbol{\mu}_\theta(\boldsymbol{z}_t, \boldsymbol{p}) = \frac{1}{\sqrt{\alpha_t}} \left(\boldsymbol{z}_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t, \boldsymbol{p}) \right)$

  The unguided version $p_\theta(\boldsymbol{z}_{t-1} | \boldsymbol{z}_t, \emptyset)$ was optimized at the same time to improve CFG sampling.

Prereq2: MI is a meaningful signal for alignment

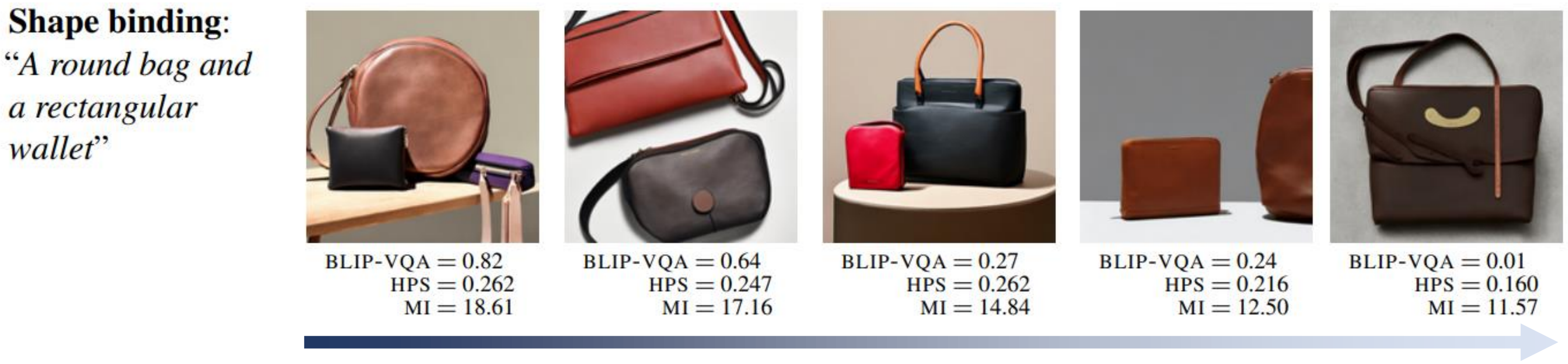
Comparison between MI and well-established alignment metrics (BLIP-VQA[1],HPS[2])

1. Kendall rank correlation coefficient

- good agreement between MI and BLIP-VQA (τ = 0.4)
- strong agreement between MI and HPS (τ = 0.68)

2. Human preference among the top-ranked images

- MI for 69.1%
- BLIP-VQA for 73.5%
- HPS for 52.2%



high scores = good alignment

low scores = poor alignment

[1] T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation, NeurIPS23

[2] Human Preference Score v2: A Solid Benchmark for Evaluating Human Preferences of Text-to-Image Synthesis

MI-TUNE: Self-supervised fine-tuning

Algorithm 2: Point-wise MI Estimation

Input : Pre-trained model: $\boldsymbol{\epsilon}_\theta$; Prompt: $\boldsymbol{p}$
Output : Generated latent: $\boldsymbol{z}$; Point-wise MI: $\mathrm{I}(\boldsymbol{z}, \boldsymbol{p})$

Function `PointWiseMI` $(\boldsymbol{\epsilon}_\theta, \boldsymbol{p})$:
  `// Initial latent sample`
  $\boldsymbol{z}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ **for** t *in* $T, \ldots, 0$ **do**
    `// MI estimation (Eq. (2))`
    $\mathrm{I}(\boldsymbol{z}_t, \boldsymbol{p}) \mathrel{+}= \left[\kappa_t \|\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{z}_t, \boldsymbol{p}, t) - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{z}_t, \emptyset, t)\|^2\right]$
    `// Noise sample`
    $\boldsymbol{w} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ if $t > 1$, else $\boldsymbol{w} = \mathbf{0}$
    `// Sampling step`
    $\boldsymbol{z}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{z}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{z}_t, \boldsymbol{p}, t)\right) + \sigma_t \boldsymbol{w}$
  end
  return $\boldsymbol{z}, \mathrm{I}(\boldsymbol{z}, \boldsymbol{p})$

Method:

1. For each prompt, generate 50 images
2. Filter the images with the highest I(image, text)
3. Finetune DoRA on these better aligned samples → improve T2I alignment

This process can be repeated for multiple rounds.

Advantage: no extra information needed

Experimental results

- **Benchmark on T2I-CompBench prompts: 6 categories**
- **MI-TUNE achieves new SOTA text-image alignment.**

Alignment results (%)

		BLIP-VQA							**HPS**							**Human** (*user study*)						
	Method	Color	Shape	Texture	2D-Sp.	Non-Sp.	Compl.	(*avg*)	Color	Shape	Texture	2D-Sp.	Non-Sp.	Compl.	(*avg*)	Color	Shape	Texture	2D-Sp.	Non-Sp.	Compl.	(*avg*)
	SD-2.1-base	49.65	42.71	49.99	15.77	66.23	50.53	(*45.81*)	27.64	24.56	24.99	27.50	26.66	25.70	(*26.17*)	29.76	11.90	40.48	35.71	66.67	29.76	(*35.71*)
Infer.	A&E	61.43	47.39	64.10	16.18	66.21	51.69	(*51.17*)	28.44	24.43	25.88	28.42	26.60	25.60	(*26.56*)	31.95	15.48	52.38	32.14	65.48	30.95	(*38.06*)
	SDG	47.15	45.24	47.13	15.25	66.17	47.41	(*44.72*)	27.25	24.40	24.71	27.10	26.12	25.83	(*25.90*)	26.19	15.48	38.10	38.10	61.90	29.76	(*34.92*)
	SCG	49.82	43.28	50.16	16.31	66.60	51.07	(*46.21*)	27.86	24.85	25.57	27.76	26.98	26.03	(*26.51*)	20.24	11.90	33.33	40.48	69.05	39.29	(*35.71*)
FT	DPOK	53.28	45.63	52.84	17.19	66.95	51.97	(*47.98*)	28.20	24.99	25.44	28.12	26.80	25.88	(*26.57*)	23.81	16.67	47.62	34.52	70.24	38.10	(*38.49*)
	GORS	53.59	43.82	54.47	15.66	67.47	52.28	(*47.88*)	28.15	24.79	25.56	27.90	26.88	26.07	(*26.56*)	34.52	14.29	48.81	36.90	65.48	30.95	(*38.49*)
	HN-ITM	46.51	39.99	48.78	15.24	65.31	49.84	(*44.28*)	26.90	24.33	24.63	27.15	25.40	25.22	(*25.60*)	23.81	19.05	30.95	20.24	47.62	23.81	(*27.58*)
	MI-TUNE	65.04	50.08	65.82	†18.51	67.77	54.17	(*53.56*)	29.13	25.57	26.20	†28.50	27.15	26.70	(*27.21*)	46.43	25.01	53.19	†45.24	73.81	46.43	(*48.35*)

SD-2.1-base DPOK GORS HN-ITM A&E SDG SCG MI-TUNE

(Color) "*a blue bear and a brown boat*"

Experimental results

SD-XL

Method	BLIP-VQA					
	Color	Shape	Texture	2D-Sp.	Non-Sp.	Comp.
(ref) SDXL	60.78	49.70	55.78	21.02	68.16	52.68
SD-2.1-base	49.65	42.71	49.99	15.77	66.23	50.53
MI-TUNE	69.66	55.86	66.74	22.18	72.17	57.74
MI-TUNE ⊟ *(ref)*	8.88	6.16	10.96	1.16	4.01	5.06
MI-TUNE % *(ref)*	14.61	12.39	19.65	5.52	5.88	9.61

Human Prompts

Model	HPS
SD-2.1-base	23.99
DiffusionDB	24.35
MI-TUNE	25.32
MI-TUNE ⊟ *base*	1.33
MI-TUNE ⊟ *DiffusionDB*	0.97

SDXL MI-TUNE

(Color) “A green apple and a brown horse”

SDXL MI-TUNE

“A black jacked and a brown hat”

SD-2.1-base | Fine-tuned using DiffusionDB images | MI-TUNE

(Human prompt) “Child’s body with a radioactive jellyfish as a head, realistic illustration, backlit, intricate, indie studio, fantasy, rim lighting, vibrant colors, emotional”

What about Rectified Flows (e.g., SD3)?

Check out our ICLR 2025 Delta Workshop paper:

RFMI: Estimating Mutual Information on Rectified Flow for Text-to-Image Alignment

$$\mathrm{I}(X; y) = \int_0^1 \mathbb{E}_{X_t|Y=y}\left[\frac{t}{1-t} u_t(X_t|Y=y) \cdot \left(u_t(X_t|Y=y) - u_t(X_t)\right)\right] dt$$

SD3.5-M RFMI FT

(Shape) "a round bag and a square box"

Conclusion

- A point-wise MI estimator suitable for a discrete-time setting
- MI-TUNE : a self-supervised fine-tuning approach to align a pre-trained T2I model without extra auxiliary models or inference overhead
- Extensive experimental campaign on multiple prompts datasets and pre-trained models

Thank you!

Mail: chao.wang@eurecom.fr

Check our paper & code: