INFORMATION THEORETIC TEXT-TO-IMAGE ALIGNMENT

Chao Wang[1,2], Giulio Franzese[2], Alessandro Finamore[1], Massimo Gallo[1], Pietro Michiardi[2]
EURECOM[1], Huawei Technologies SASU, France[2]
{chao.wang, giulio.franzeze, pietro.michiardi}@eurecom.fr
{wang.chao3, alessandro.finamore, massimo.gallo}@huawei.com

ABSTRACT

Diffusion models for Text-to-Image (T2I) conditional generation have recently achieved tremendous success. Yet, aligning these models with user's intentions still involves a laborious trial-and-error process, and this challenging alignment problem has attracted considerable attention from the research community. In this work, instead of relying on fine-grained linguistic analyses of prompts, human annotation, or auxiliary vision-language models, we use Mutual Information (MI) to guide model alignment. In brief, our method uses self-supervised fine-tuning and relies on a point-wise MI estimation between prompts and images to create a synthetic fine-tuning set for improving model alignment. Our analysis indicates that our method is superior to the state-of-the-art, yet it only requires the pre-trained denoising network of the T2I model itself to estimate MI, and a simple fine-tuning strategy that improves alignment while maintaining image quality.

1 INTRODUCTION

Generative models used for Text-to-Image (T2I) conditional generation (Rombach et al., 2022; Ramesh et al., 2022; Saharia et al., 2022; Balaji et al., 2022b; Gafni et al., 2022; Podell et al., 2024) have reached impressive performance. In particular, diffusion models (Song & Ermon, 2019b; Ho et al., 2020; Kingma et al., 2021; Song & Ermon, 2020; Song et al., 2021; Dhariwal & Nichol, 2021) generate extremely high-quality images by specifying a natural text prompt that acts as a guiding signal (Ho & Salimans, 2022; Nichol et al., 2022; Rombach et al., 2022). Yet, accurately translating prompts into images with the intended semantics is still complex (Conwell & Ullman, 2022; Feng et al., 2023a; Wang et al., 2023a). Issues include catastrophic neglecting (i.e., prompt elements are not generated), incorrect attribute binding (i.e., elements attributes such as color, shape, and texture are missing or wrongly assigned), incorrect spatial layout (i.e., elements are not correctly positioned), and a general difficulty in handling complex prompts (Wu et al., 2024).

On the one hand, quantifying *model alignment* is not trivial. Various works (Hu et al., 2023; Gordon et al., 2023; Grimal et al., 2024) propose different metrics, most of which use complementary Visual Question Answering (VQA) models or Large Language Models (LLMs) to create scores measuring and explaining alignment. Moreover, a recent work (Huang et al., 2023) introduces a comprehensive benchmark suite to ease comparison among different metrics and modeling techniques via "categories", i.e., a pre-defined set of attribute binding, spatial-related, and other tasks.

On the other hand, addressing T2I model alignment is even more challenging than measuring it. Broadly, we can group the related literature into two main families: *inference-time* and *fine-tuning* methods. For inference-time methods, the key intuition is that the generative process can be optimized by modifying the reverse path of the latent variables. Some works (Chefer et al., 2023a; Li et al., 2023b; Rassin et al., 2023) mitigate failures by refining the cross-attention units (Tang et al., 2023) of the denoising network of Stable Diffusion (SD) (Rombach et al., 2022) on-the-fly, ensuring they attend to all subject tokens in the prompt (typically directly specified as a complementary prompt-specific input for the alignment process) and strengthen their activations. Other inference-time methods (Agarwal et al. (2023); Liu et al. (2022); Kang et al. (2023); Dahary et al. (2024); Meral et al. (2024); Feng et al. (2023b); Kim et al. (2023); Wu et al. (2023a); Zhang et al. (2024a;b)), focus on individual failure cases. These approaches (i) require a linguistic analysis of prompts, leading to specialized solutions that rely on auxiliary models for prompt understanding, and (ii) result in considerably longer image generation time due to extra optimization costs during sampling.

Considering fine-tuning methods, some works (Wu et al., 2023d; Lee et al., 2023) require human annotations to prepare a fine-tuning set, while others (Fan et al., 2023; Wallace et al., 2023; Clark et al., 2024) rely on Reinforcement Learning (RL), Direct Preference Optimization (DPO), or a differentiable reward function to steer model behavior. Recent methods use self-playing (Yuan et al., 2024; Xu et al., 2023; Sun et al., 2023; Wang et al., 2023b; Ma et al., 2023), auxiliary models such as VQA (Li et al., 2023a; Jiang et al., 2024) or segmentation maps (Kirillov et al., 2023) in a semi-supervised fine-tuning setting. While these methods do not introduce extra inference time costs, they still require human annotation (which is subjective, costly, and does not scale well) and/or auxiliary models to guide the fine-tuning.

Complementary to both families are *heuristic*-based methods that rely on a variety of "tricks", such as prompt engineering (Witteveen & Andrews, 2022; Liu & Chilton, 2022; Wang et al., 2023a), negative prompting (HuggingFace, 2022; Mahajan et al., 2023; Ogezi & Shi, 2024), prompt rewriting (Mañas et al., 2024) or brute force an appropriate seed selection (Samuel et al., 2024; Karthik et al., 2023). While these methods can be beneficial in specific cases, they fundamentally shift the alignment problem to users.

Overall, current approaches require extra information (human input, auxiliary models, and additional data). To the best of our knowledge, no previous work investigates *self-supervised* approaches for T2I alignment, i.e., the use of a pre-trained model to generate images given a specific set of prompts, and select the most aligned ones to prepare a fine-tuning set, without using auxiliary models. In this work, we investigate this strategy from an information theoretic perspective, by using MI to quantify the non-linear prompt-image relationship. In particular, we focus on the estimation of *point-wise* MI using neural estimators (Belghazi et al., 2018; Song & Ermon, 2019a; Brekelmans et al., 2022; Franzese et al., 2024; Kong et al., 2024), and study if and how MI can be used as a meaningful signal to improve T2I alignment, without relying on linguistic analysis of prompts, nor auxiliary models or heuristics. Our method unfolds as follows. We build upon the work in (Franzese et al., 2024) and extend it to compute point-wise MI. We then proceed with a self-supervised fine-tuning approach, whereby we use point-wise MI to construct a fine-tuning set using synthetic data generated by the T2I model itself. We then use the recent adapter presented in (Liu et al., 2024) to fine-tune a small fraction of weights injected in the T2I model denoising network. In summary, our work presents the following contributions:

(1) We define a point-wise MI estimator suitable for a discrete-time setting (§ 2). We empirically study whether MI between natural prompts and corresponding images considering both qualitative and quantitative approaches. Specifically, we show that MI provides a meaningful indication of alignment with respect to both alignment metrics (BLIP-VQA and HPS) as well as a users study (§ 3.1).

(2) We design a self-supervised fine-tuning approach, called MI-TUNE (§ 3.2), that uses a small number of fine-tuning samples to align a pre-trained T2I model without extra auxiliary models or inference overhead.

(3) We perform an extensive experimental campaign using a recent T2I benchmark suite (Huang et al., 2023) and SD-2.1-base as base model obtaining sizable improvement compared to six alternative methods (§ 4). Those benefits hold also when considering more complex tasks (based on DiffusionDB (Wang et al., 2022)) and alternative base models (namely, SDXL (Podell et al., 2024)). Moreover we study the trade-off between T2I alignment and image quality that has been overlooked in the literature. Specifically, while the well-known FID, FD-DINO and CMMD metrics suggest a modest image quality/variety deterioration as a consequence of alignment objectives, optimizing the Classifier Free Guidance (CFG) hyper-parameter of the fine-tuned model at generation time, enables finding a "sweet spot" between T2I alignment and image quality.

2 PRELIMINARIES

Diffusion models. Denoising diffusion models (Ho et al., 2020; Sohl-Dickstein et al., 2015) are generative models characterized by a forward process, that is fixed to a Markov chain that gradually adds Gaussian noise to the data according to a carefully selected variance schedule β_t, and a corresponding discrete-time reverse process, that has a Markov structure as well. Intuitively, diffusion models rely on the principle of iterative denoising: starting from a simple distribution $\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, samples are generated by iterative applications of a denoising network $\boldsymbol{\epsilon_\theta}$, that removes noise over T denoising steps. A simple way to learn the denoising network $\boldsymbol{\epsilon_\theta}$ is to consider a re-weighted variational lower bound of the marginal likelihood:

$$\mathcal{L}_{\text{simple}}(\boldsymbol{\theta}) = \mathbb{E}_{t\sim U(0,T),\boldsymbol{x}_0\sim p_{\text{data}},\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})} \left[||\boldsymbol{\epsilon} - \boldsymbol{\epsilon_\theta}(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + (\sqrt{1-\bar{\alpha}_t})\boldsymbol{\epsilon}, t)||^2 \right], \tag{1}$$

where $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. For sampling, we let $\sigma_t^2 = \beta_t$. A similar variational objective can be obtained by switching perspective from discrete to continuous time (Song et al., 2021), whereby the denoising network approximates a score function of the data distribution. For image data, the denoising network is typically parameterized by a UNET (Ronneberger et al., 2015; Rombach et al., 2022).

This simple formulation has been extended to conditional generation (Ho & Salimans, 2021), whereby a conditioning signal $\boldsymbol{p}$ injects "external information" in the iterative denoising process. This requires a simple extension to the denoising network such that it can accept the conditioning signal: $\boldsymbol{\epsilon_\theta}(\boldsymbol{x}_t, \boldsymbol{p}, t)$. Then, during training, a randomized approach allows to learn both the conditional and unconditional variants of the denoising network, for example by assigning a null value to the conditioning signal. At sampling time, a weighted linear combination of the conditional and unconditional networks, such as $\tilde{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, \boldsymbol{p}, t) = \boldsymbol{\epsilon_\theta}(\boldsymbol{x}_t, \emptyset, t) + \gamma(\boldsymbol{\epsilon_\theta}(\boldsymbol{x}_t, \boldsymbol{p}, t) - \boldsymbol{\epsilon_\theta}(\boldsymbol{x}_t, \emptyset, t))$ can be used.

In this work, we use pre-trained latent diffusion models operating on a learned projection of the input data $\boldsymbol{x}_0$ into a corresponding latent variable $\boldsymbol{z}_0$ which is lower-dimensional compared to the original data. Moreover, the conditioning signal $\boldsymbol{p}$ is obtained by a text encoder such as CLIP (Radford et al., 2021).

MI estimation. MI is a central measure to study the non-linear dependence between random variables (Shannon, 1948; MacKay, 2003), and has been extensively used in machine learning for representation learning (Bell & Sejnowski, 1995; Stratos, 2019; Belghazi et al., 2018; Oord et al., 2018; Hjelm et al., 2019), and for both training (Alemi et al., 2016; Chen et al., 2016; Zhao et al., 2018) and evaluating generative models (Alemi & Fischer, 2018; Huang et al., 2020).

For many problems of interest, precise computation of MI is not trivial (McAllester & Stratos, 2020; Paninski, 2003). Consequently, a wide range of techniques for MI estimation have flourished. In this work, we focus on realistic and high-dimensional data, which calls for recent advances in MI estimation (Papamakarios et al., 2017; Belghazi et al., 2018; Oord et al., 2018; Song & Ermon, 2019a; Rhodes et al., 2020; Letizia & Tonello, 2022; Brekelmans et al., 2022; Kong et al., 2024). In particular, we capitalize on a recent method (Franzese et al., 2024), that relies on the theory behind continuous-time diffusion processes (Song et al., 2021) and uses the Girsanov Theorem (Øksendal, 2003) to show that score functions can be used to compute the Kullback-Leibler (KL) divergence between two distributions. In what follows, we use a simplified notation and gloss over several mathematical details to favor intuition over rigor. Here we consider discrete-time diffusion models, which are equivalent to the continuous-time counterpart under the variational formulation, up to constants and discretization errors (Song et al., 2021).

We begin by considering the two arbitrary random variables $\boldsymbol{z}$ and $\boldsymbol{p}$ which are sampled from the joint distribution $p_{\text{latent,prompt}}$, where the former corresponds to the distribution of the projections in a latent space of the image distribution, and the latter to the distribution of prompts used for conditional generation. Then, following the approach in (Franzese et al., 2024), with the necessary adaptation to the discrete domain (see Appendix A for details), point-wise MI estimation can be obtained as follows:

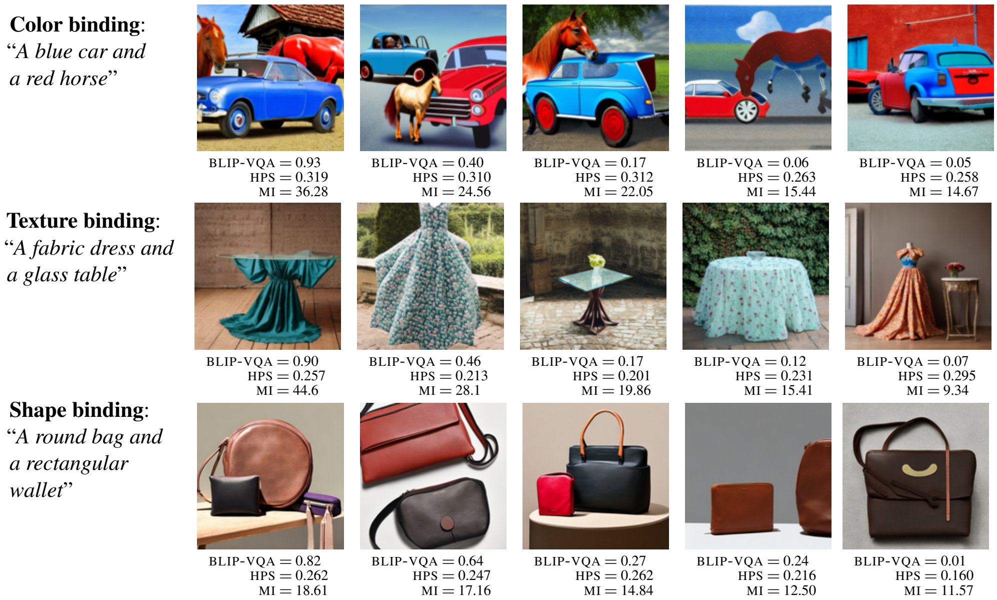

Figure 1: Qualitative analysis of MI as an alignment measure (all metrics decrease from left to right). See also Appendix I.

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

Given a pre-trained diffusion model, we compute an expectation (over diffusion times t) of the scaled squared norm of the difference between the conditional $\boldsymbol{\epsilon_\theta}(\boldsymbol{z}_t, \boldsymbol{p}, t)$ and unconditional networks $\boldsymbol{\epsilon_\theta}(\boldsymbol{z}_t, \emptyset, t)$, which corresponds to an estimate of the point-wise MI between an image and a prompt. Intuitively, the difference between these scores quantifies how much extra knowledge of the prompt helps in denoising the perturbed images. This is both a key ingredient and a competitive advantage of our method, as it enables a self-contained approach to alignment based on the T2I model alone without auxiliary models or human feedback.

3 Our Method: MI-Tune

The T2I alignment problem arises when user's intentions, as expressed through natural text prompts, fail to materialize in the generated image. Our novel approach aims to address alignment using a theoretically grounded MI estimation, that applies across various contexts. To improve model alignment, we introduce a self-supervised fine-tuning method. Leveraging the T2I model itself, we estimate MI and generate an information-theoretic enhanced fine-tuning dataset. While our focus in this work is on T2I alignment, our framework remains extensible to other modalities.

3.1 Is Mutual Information meaningful for alignment?

To the best of our knowledge, MI has never been evaluated as a *meaningful signal* for T2I alignment. As such, in this section we perform both qualitative and quantitative analyses to investigate this aspect.

Qualitative analysis. Starting with a qualitative analysis, we select a set of simple prompts to probe color, texture, and shape attribute binding from T2I-CompBench (Huang et al., 2023) using SD (Rombach et al., 2022) (specifically SD-2.1-base) to generate the corresponding images. We then measure the well-known BLIP-VQA (Huang et al., 2023) and Human Preference Score (HPS) (Wu et al., 2023b) alignment metrics as well as point-wise MI estimates. BLIP-VQA uses a large vision-language model to compute an alignment score, by casting questions against an image to verify that the prompt used to generate it is well represented. HPS is an elaborate metric that uses an auxiliary pre-trained model, blending alignment with aesthetics according to human perception, which are factors that can sometimes be in conflict. Figure 1 collects some examples and related metric scores revealing a substantial agreement among all measures: all metrics decrease from left to right in the figure, as prompt-image alignment deteriorates.

Quantitative analysis. To quantitatively measure the agreement between MI and well-established alignment metrics, we use all 700 prompts from T2I-CompBench and use SD (again, SD-2.1-base) to generate 50 images per prompt. We use point-wise MI to rank such images and select the 1st, 25th, and 50th. For these three representative images, we compute BLIP-VQA and HPS scores and re-rank them according to both metrics. Last, we measure agreement between the three rankings using Kendall's τ method (Kendall, 1938), and average results across all prompts. Results indicate good agreement between MI and BLIP-VQA ($\tau = 0.4$), and a strong agreement between MI and HPS ($\tau = 0.68$).

To strengthen our analysis, we also perform a users study eliciting human preference (see Appendix B.1 for details). Given a randomly selected prompt from T2I-CompBench that users can read, we present the top-ranked generated image (among the 50) according to MI, BLIP-VQA and HPS, in a randomized order. Users can select one or more images to indicate their preference regarding alignment and aesthetics, for a total of 10 random prompts per user. From the 102 surveys from 46 users, we find that human preference for prompt-image pairs goes to MI for 69.1%, BLIP-VQA for 73.5% and HPS for 52.2% of the cases, respectively.

Relevant literature. Overall, our analyses support our intuition by which MI *is a meaningful signal for alignment* (and possibly aesthetics too), setting the stage for our T2I alignment method. Our intuition is also supported by recent studies investigating the information flow in the generative process of diffusion models. Specifically, Kong et al. (2024) estimates pixel-wise mutual information between natural prompts and the images generated at each time-step of a backward diffusion process. They compare such "information maps" to cross-attention maps (Tang et al., 2023) in an experiment involving prompt manipulation – modifications of the initial prompt during reverse diffusion – and conclude that MI is much more sensitive to information flow from prompt to images. In a similar vein, Franzese et al. (2024) compute MI between prompt and images at different stages of the reverse process of image generation. Experimental evidence indicates that MI can be used to analyze various reverse diffusion phases: noise, semantic, and denoising stages (Balaji et al., 2022a). While previous studies do not explicitly focus on alignment, they *indirectly* support our intuition that MI estimated using a diffusion model gauges the amount of information a text prompt conveys about an image (and vice-versa) which is key for T2I alignment.

3.2 Self-supervised fine-tuning with MI-TUNE

In summary, given a pre-trained diffusion model such as SD (Rombach et al., 2022) or any variant, such as SDXL (Podell et al., 2024), we leverage our point-wise MI estimation method to select a small fine-tuning dataset set of information-theoretic aligned examples.

Our self-supervised alignment method **relies on the pre-trained model only** to produce a given amount of fine-tuning data, which is then filtered to retain prompt-image pairs with a high degree of alignment, according to pair-wise MI **estimates obtained using only the pre-trained model**. We begin with a set of fine-tuning prompts $\mathcal{P}$, which can be either manually crafted, or borrowed from available prompt collections (Wang et al., 2023a; Huang et al., 2023). Ideally, fine-tuning prompts should be conceived to stress the pre-trained model with challenging attribute and spatial bindings, or complex rendering tasks.

Algorithm 1: MI-TUNE

Input : Pre-trained model: ϵ_θ, Prompt set: $\mathcal{P}$
Hyper par : Image pool size: M; Top MI-aligned samples: k
Output : Fine-tuned diffusion model ϵ_{θ^*}

// Fine-tuning set
$\mathcal{S} \leftarrow [\,]$
for $\boldsymbol{p}^{(i)}$ *in* $\mathcal{P}$ **do**
  for $j \in \{1, \cdots, M\}$ **do**
    // Generate and compute MI
    $\boldsymbol{z}^{(j)}, \mathrm{I}(\boldsymbol{z}^{(j)}, \boldsymbol{p}^{(i)})$ = `PointWiseMI` $(\boldsymbol{\epsilon}_\theta, \boldsymbol{p}^{(i)})$
    // Append samples and MI
    $\mathcal{S}[\boldsymbol{p}^{(i)}]$.append $(\boldsymbol{z}^{(j)}, \mathrm{I}(\boldsymbol{z}^{(j)}, \boldsymbol{p}^{(i)}))$
  end
  // Retain only Top-k elements
  $\mathcal{S}[\boldsymbol{p}^{(i)}]$ = Top-k $(\mathcal{S}[\boldsymbol{p}^{(i)}])$
end
return $\boldsymbol{\epsilon}_{\theta^*}$ = `FineTune` $(\boldsymbol{\epsilon}_\theta, \mathcal{S})$

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

As described in Algorithm 1, for each prompt $p^{(i)}$ in the fine-tuning set $\mathcal{P}$, we use the pre-trained model to generate a fixed number M of synthetic images. Given prompt-image pairs $(p^{(i)}; z^{(j)})$, $j \in [1, M]$, we estimate pair-wise MI and select the top k pairs, which will be part of the model fine-tuning dataset $\mathcal{S}$. Finally, we augment the pre-trained model with adapters (Hu et al., 2021; Liu et al., 2024), and proceed with fine-tuning. We study the impact of the adapter choice, and whether only the denoising network or both the denoising and text encoder networks should be fine-tuned (Appendix E.2). Moreover, we measure the impact of the number of fine-tuning rounds R to the pre-trained model, i.e., we renew the fine-tuning dataset $\mathcal{S}$ using the fine-tuned model, and re-fine-tune it using Algorithm 1 (§ 4.4). Our efficient implementation combines latent generation and point-wise MI computation as shown in Algorithm 2. Since MI estimation involves computing an expectation over diffusion times t, it is easy to combine generation and estimation in the same loop. Moreover, the function is easy to parallelize to significantly speed up the fine-tuning set $\mathcal{S}$ composition.

4 Experimental evaluation

4.1 Benchmark and metrics

Benchmark. We compare all techniques using T2I-CompBench (Huang et al., 2023), a benchmark composed of 700/300 (train/test) prompts across 6 *categories* including attribute binding (color, shape, and texture categories), object relationships (2D-spatial and non-spatial associations), and complex composition tasks. These prompts were generated with predefined rules or ChatGPT (OpenAI, 2024). We also assess MI-TUNE performance on more realistic prompts by sampling 5,000/1,250 (train/test) prompt-image pairs from DiffusionDB (Wang et al., 2022), a large-scale dataset composed of complex human-crafted prompts paired with the corresponding images generated from a SD model.

Alignment Metrics. Evaluating T2I alignment is difficult as it requires a detailed understanding of prompt-images pairs, and many metrics have been proposed, e.g., CLIP (Hessel et al., 2021; Radford et al., 2021), MINIGPT-4 (Zhu et al., 2024), and human evaluation. In our work we use BLIP-VQA (Huang et al., 2023), HPS (Wu et al., 2023c) and UniDet (Zhou et al., 2022). While BLIP-VQA computes a score with a questions-answers approach – a given prompt is decomposed and each part is transformed into a question for an auxiliary VQA model; then, answers are aggregated into a single score – only based on alignment, HPS includes both alignment and aesthetics – this is enabled by an auxiliary model pre-trained using human-annotated data. As in (Huang et al., 2023), the 2D-spatial category is evaluated using the UniDet object detection model.

We complement these metrics with a user study. We randomly select 100 prompts per category, and generate 10 pictures per prompt for each method we consider in our evaluation. Then, we run surveys composed of 12 rounds (2 for each category), each showing to the user a randomly selected prompt and a randomly selected image for each method, randomly arranged in a grid. At each round, users need to select zero or more images they consider aligned with the prompt. Overall, we collected 42 surveys from 5 users, from which we computed the total percentage of times each method was selected for each category (Appendix B.2).

Image quality metrics. Assessing performance only considering alignment metrics can hide undesired effects. Intuitively, a strong adherence to a given prompt reduces the generative process "degrees of freedom" and this trade-off might not be visible even to a trained eye. To investigate these dynamics we compute FID (Heusel et al., 2017), FD-DINO (Oquab et al., 2024) and CMMD (Jayasumana et al., 2024) scores – FID favors natural colors and textures but struggles to detect objects/shapes distortion, while FD-DINO and CMMD favor image content. Following (Imagen-Team et al., 2024), rather than using the T2I-CompBench test set, we compute the metrics using 30k samples of the MS-COCO-2014 (Lin et al., 2015) validation set.

4.2 MI-TUNE FINE-TUNING

Base models. We mainly run our benchmark using SD-2.1-base as base model, but we also report results of the application of MI-TUNE on SDXL to demonstrate its flexibility.

Fine-tuning sets. T2I-CompBench contains 700 training prompts for each category. When using MI-TUNE, we generate $M = 50$ images for each prompt using the pre-trained model, compute their point-wise MI, and select the top $k = 1$[1] (sensitivity to M and k in Appendix E.1). For the 2D-Spatial category, we also compose fine-tuning sets generating images from SPRIGHT (Chatterjee et al., 2024) – a model optimized for this (more challenging) category and fine-tuned from SD-2.1 (a higher resolution version of SD-2.1-base). Last, we also contrast MI-TUNE fine-tuning set composition against (i) using HPS rather than MI for image selection,[2] (ii) using both MI-selected and real-pictures and (iii) images from DiffusionDB.

Fine-tuning weights. In our work, fine-tuning corresponds to injecting DoRA (Liu et al., 2024) adapters (rank and scaling factor α are set to 32) only into the attention layers and fully connected layers of the denoising UNET network, whereas other layers are frozen.[3]

Other hyperparams search. We consider up to $R \in [1, 3]$ rounds of fine-tuning i.e., using as base model the one obtained from previous round and apply Algorithm 1, and Classifier Free Guidance (CFG) $\in [2.5, 7.5]$. For each fine-tuned model we then compute all alignment and image quality metrics. More fine-grained hyperparams details and computational costs considerations in Appendix C.

4.3 ALTERNATIVE METHODS

Inference-time methods. Pre-trained model alignment can be improved at inference by optimizing the latent variables z_t throughout the numerical integration used to generate the (latent) image. This process steers model alignment with an auxiliary loss based on attention maps and fine-grained linguistic analysis of the prompt (e.g., additional input is used to explicitly indicate which words to focus on). In this family, we consider 3 methods: Attend and Excite (A&E) (Chefer et al., 2023b), Structured Diffusion Guidance (SDG) (Feng et al., 2023b) and Semantic-aware Classifier-Free Guidance (SCG) (Shen et al., 2024).

Fine-tuning methods. Alternatively, a pre-trained model can be fine-tuned with adapters (Hu et al., 2021) optimized via a variety of RL or supervision methods. Specifically, we consider 3 approaches: Diffusion Policy Optimization with KL regularization (DPOK) (Fan et al., 2023), Generative mOdel finetuning with

[1]We remark that, albeit in a different context, this selection resembles an image retrieval task (Krojer et al., 2023)

[2]We exclude BLIP-VQA for the fine-tuning set composition to avoid biasing the evaluation (Huang et al., 2023).

[3]LoRA adapters (Hu et al., 2021) and fine-tuning also the CLIP-based text encoder do not provide performance improvements (Appendix E.2). Likewise, creating a multi-category model by "merging" different per-category models or using a fine-tuning set composed with images from all categories do not provide performance gains (Appendix E.3).

Table 1: Alignment results (%). Gray highlighted style when MI-TUNE outperforms all competitors; Grayed text for under-performing methods per-family; Green heatmaps show per-category absolute gains w.r.t. the base model.

	Method	BLIP-VQA 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)*
Infer.	A&E	61.43	47.39	64.10	16.18	66.21	51.69	*(51.17)*
Infer.	SDG	47.15	45.24	47.13	15.25	66.17	47.41	*(44.72)*
Infer.	SCG	49.82	43.28	50.16	16.31	66.60	51.07	*(46.21)*
FT	DPOK	53.28	45.63	52.84	17.19	66.95	51.97	*(47.98)*
FT	GORS	53.59	43.82	54.47	15.66	67.47	52.28	*(47.88)*
FT	HN-ITM	46.51	39.99	48.78	15.24	65.31	49.84	*(44.28)*
	MI-TUNE	65.04	50.08	65.82	†18.51	67.77	54.17	*(53.56)*
	best Infer. ⊟ base	11.78	4.68	14.11	0.54	0.37	1.16	*(5.44)*
	best FT ⊟ base	3.94	2.92	4.48	1.42	1.24	1.75	*(2.62)*
	MI-TUNE ⊟ base	15.39	7.37	15.83	2.74	1.54	3.64	*(7.75)*
	MI-TUNE ⊟ *best*⋆	3.61	2.69	1.72	1.32	0.30	1.89	(1.92)
	MI-TUNE % *best*⋆	5.88	5.68	2.68	7.68	0.44	3.62	(4.33)

	Method	HPS Color	Shape	Texture	2D-Sp.	Non-Sp.	Compl.	*(avg)*
	SD-2.1-base	27.64	24.56	24.99	27.50	26.66	25.70	*(26.17)*
Infer.	A&E	28.44	24.43	25.88	28.42	26.60	25.60	*(26.56)*
Infer.	SDG	27.25	24.40	24.71	27.10	26.12	25.83	*(25.90)*
Infer.	SCG	27.86	24.85	25.57	27.76	26.98	26.03	*(26.51)*
FT	DPOK	28.20	24.99	25.44	28.12	26.80	25.88	*(26.57)*
FT	GORS	28.15	24.79	25.56	27.90	26.88	26.07	*(26.56)*
FT	HN-ITM	26.90	24.33	24.63	27.15	25.40	25.22	*(25.60)*
	MI-TUNE	29.13	25.57	26.20	†28.50	27.15	26.70	*(27.21)*
	best Infer. ⊟ base	0.80	0.29	0.89	0.92	0.32	0.33	*(0.59)*
	best FT ⊟ base	0.56	0.43	0.57	0.62	0.22	0.37	*(0.46)*
	MI-TUNE ⊟ base	1.49	1.01	1.21	1.00	0.49	1.00	*(1.03)*
	MI-TUNE ⊟ *best*⋆	0.69	0.58	0.32	0.08	0.17	0.63	(0.41)
	MI-TUNE % *best*⋆	2.43	2.32	1.24	0.28	0.63	2.42	(1.55)

	Method	**Human** (*user study*) Color	Shape	Texture	2D-Sp.	Non-Sp.	Compl.	*(avg)*
	SD-2.1-base	29.76	11.90	40.48	35.71	66.67	29.76	*(35.71)*
Infer.	A&E	31.95	15.48	52.38	32.14	65.48	30.95	*(38.06)*
Infer.	SDG	26.19	15.48	38.10	38.10	61.90	29.76	*(34.92)*
Infer.	SCG	20.24	11.90	33.33	40.48	69.05	39.29	*(35.71)*
FT	DPOK	23.81	16.67	47.62	34.52	70.24	38.10	*(38.49)*
FT	GORS	34.52	14.29	48.81	36.90	65.48	30.95	*(38.49)*
FT	HN-ITM	23.81	19.05	30.95	20.24	47.62	23.81	*(27.58)*
	MI-TUNE	46.43	25.01	53.19	†45.24	73.81	46.43	*(48.35)*
	best Infer. ⊟ base	2.19	3.58	11.90	4.77	2.38	9.53	*(5.72)*
	best FT ⊟ base	4.76	7.15	8.33	1.19	3.57	8.34	*(5.56)*
	MI-TUNE ⊟ base	16.67	13.11	12.71	9.53	7.14	16.67	*(12.64)*
	MI-TUNE ⊟ *best*⋆	11.91	5.96	0.81	4.76	3.57	7.14	(5.69)
	MI-TUNE % *best*⋆	34.50	31.29	1.55	11.76	5.08	18.17	(17.06)

A ⊟ B indicates the absolute difference between A and B; A % B corresponds to the percentage difference (A - B) / B; †: Fine-tuning set obtained from SPRIGHT rather than SD-2.1-base; Human scores do not sum to 100 in each category as users can select multiple methods for each question.

Reward-driven Sample selection (GORS) (Huang et al., 2023) and Hard-Negatives Image-Text-Matching (HN-ITM) (Krojer et al., 2023). Notice that since results in the literature for both families do not necessarily refer to same base models, to guarantee a fair comparison, we adapted and evaluated all methods on SD-2.1-base.

4.4 RESULTS

Comparing methods. Table 1 reports the alignment results on T2I-CompBench. To simplify its reading, the bottom part of the table summarizes (i) the absolute gain with respect to the SD-2.1-base model for each of the best methods in each family and (ii) the percentage gains of MI-TUNE with respect to the alternative method for each category. We also summarize performance as averages across categories for each metric.

Despite performance varies, MI-TUNE achieves a new state of the art across all categories/metrics, often by a sizable margin. While this is more evident for BLIP-VQA and Human, the literature shows that HPS has natural small variations (see Appendix D), hence MI-TUNE gains are significant also for this metric.

Table 1 results are obtained generating fine-tuning sets from SD-2.1-base for all tasks but 2D-Spatial. For this category, we were able to obtain (at best) BLIP-VQA=15.93 and HPS=28.13. Conversely, generating the fine-tuning images from SPRIGHT resulted beneficial. We can link this result to the self-supervision nature of MI-TUNE. On the one hand, our methodology is not bounded to a specific model. On the other hand, the filtering operated via point-wise MI estimation can benefit from “pre-alignment” – MI-TUNE can strengthen existing alignment but might not be sufficient to “induce” it. Notice that all competitors suffer from this trade-off too as no single winner emerges. In particular, despite A&E and GORS are the most frequent best method in their family (winning in 10-out-of-18 scenarios), all competitors show less consistent performance across categories and metrics than MI-TUNE. For instance, for attribute binding (color, shape and texture), fine-tuning methods under-perform according to BLIP-VQA and Human, but the performance gaps are very close considering HPS. Yet, MI-TUNE achieves consistently higher performance across all categories, outperforming alternative fine-tuning methods by a large margin.

Raw alignment performance apart, it is important to highlight MI-TUNE key differences compared to the alternative fine-tuning methods. DPOK uses RL with a reward model (pre-trained with human-labeled real images) to define a prompt-image alignment score to guide the fine-tuning, HN-ITM uses a contrastive learning approach based on an ad-hoc dataset with real positive (good alignment) and negative (poor alignment) prompt-image pairs, and GORS composes a fine-tuning set generating images from the diffusion model and selecting them based on BLIP-VQA. While GORS is very close in spirit to MI-TUNE, its performance is "biased" – the filtering criteria overlaps with the final evaluation strategy – as explicitly acknowledged by its authors (Huang et al., 2023). Overall, while both DPOK and GORS still require external assistance, MI-TUNE generates images *and* selects them using the target model itself, i.e., it is the first fully self-supervised model for T2I alignment to the best of our knowledge.

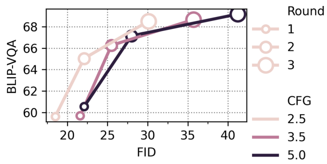

Figure 2: Hyper-params search.

Table 2: Comparing image quality/variety scores.

Metric	SD-2.1-base	MI-TUNE (R=2, CFG=2.5) Color	Shape	Texture	Spatial	Non-sp.	Comp.	(*avg*)	DALLE-3‡	IMAGEN-3‡	SDXL‡
FID(↓)	17.1	22.1	16.8	17.3	18.8	16.8	20.6	(*18.7*)	20.1	17.2	13.2
FD-DINO(↓)	229.1	279.0	236.9	250.4	251.7	231.9	255.6	(*250.9*)	284.4	213.9	185.6
CMMD(↓)	0.641	0.681	0.634	0.694	0.669	0.709	0.671	(*0.680*)	0.894	0.854	0.898

Results from 30k samples of MS-COCO-2014 *validation set; ‡ results from (Imagen-Team et al., 2024)*

Alignment/image quality-variety trade-offs. MI-TUNE results in Table 1 are obtained from a grid search across multiple fine-tuning rounds R and CFG values. In fact, we observe different trade-offs between alignment and image quality across different configurations. We exemplify this in Figure 2, for the Color category. The figure highlights two opposite dynamics: T2I alignment benefits from multiple fine-tuning rounds (higher BLIP-VQA) but can introduce image artifacts and reduce measured diversity (higher FID). While this trade-off is neither mentioned nor quantified in the literature of the considered methods, it is to be expected – strictly abiding to a prompt impacts the "generative pathways" at sampling time. Interestingly, lowering CFG (typically set to 7.5) counterbalances these dynamics and enables a "sweet spot" – as the model better aligns to a category thanks to fine-tuning, one can alleviate the guidance scale dependency at generation. Table 2 complements this analysis by showing FID, FD-DINO and CMMD scores for all categories, as well for SD-2.1-base and three state of the art models – while all metrics indeed suggest a possible reduction in image variety considering SD-2.1-base, MI-TUNE scores are comparable with other state-of-the-art models (see Figure 3 for example images).

Table 3: FT set selection.

Strategy	BLIP-VQA Color	BLIP-VQA Shape	HPS Color	HPS Shape
MI only	65.04	50.08	29.13	25.57
HPS only	59.43	46.87	n.a.	n.a.
MI+Real(0.25)	61.34	48.47	29.16	25.87
MI+Real(0.5)	61.63	49.50	29.38	25.92
MI+Real(0.9)	59.83	48.92	28.60	25.60

Table 4: Alignment (%) using SDXL.

Method	BLIP-VQA Color	Shape	Texture	2D-Sp.	Non-Sp.	Comp.	HPS Color	Shape	Texture	2D-Sp.	Non-Sp.	Comp.
(*ref*) SDXL	60.78	49.70	55.78	21.02	68.16	52.68	28.47	24.99	25.85	28.50	26.64	25.90
SD-2.1-base	49.65	42.71	49.99	15.77	66.23	50.53	27.64	24.56	24.99	27.50	26.66	25.70
MI-TUNE	69.66	55.86	66.74	22.18	72.17	57.74	29.03	25.90	27.15	29.57	27.56	26.70
MI-TUNE ⊟ (*ref*)	8.88	6.16	10.96	1.16	4.01	5.06	0.56	0.91	1.30	1.07	0.92	0.80
MI-TUNE % (*ref*)	14.61	12.39	19.65	5.52	5.88	9.61	1.97	3.64	5.03	3.75	3.45	3.09

Table 5: DiffusionDB.

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

Fine-tuning set composition. The strategy to select prompt-image pairs for the fine-tuning set has a large design space beyond the use of MI. In Table 3, we report (for two categories for brevity) alignment performance using two alternative strategies. Specifically, using HPS rather than MI degrades performance.[4] Results when composing the fine-tuning set by mixing MI-selected and real images selected from the CC2M dataset (Changpinyo et al., 2021) are instead inconsistent (BLIP-VQA steadily degrades but HPS signals an improvement in some scenarios).

SDXL and DiffusionDB. We complete our evaluation by presenting results obtained applying MI-TUNE on SDXL in Table 4, and considering an alternative scenario closer to real user application using DiffusionDB in Table 5 to complement the synthetic nature of T2I-CompBench. As expected, "vanilla" SDXL significantly outperforms SD-2.1-base, yet MI-TUNE enables sizable improvements on SDXL alignment (see Figure 4). For the realistic alignment use case in Table 5, we select prompt-images pairs from DiffusionDB and we contrast alignment when fine-tuning using the images already paired with prompts against MI-selected ones. We use SD-2.1-base as base model and report only HPS scores[5] in Table 5. Overall, fine-tuning with DiffusionDB images improves the base model, yet MI-TUNE enables superior performance (see Figure 5).

[4] We compute only BLIP-VQA to avoid evaluation bias (Huang et al., 2023).

[5] The higher prompt complexity does not well suit BLIP-VQA text decomposition (see Appendix H.1).

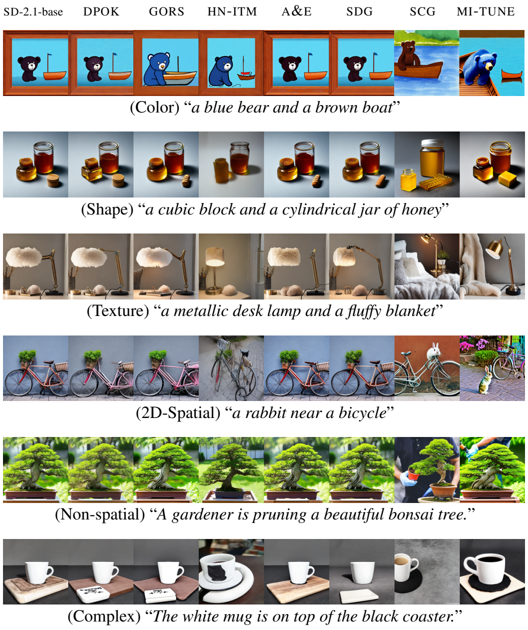

Figure 3: Qualitative examples from Table 1 (same seed used for a given prompt). More examples in Appendix F.

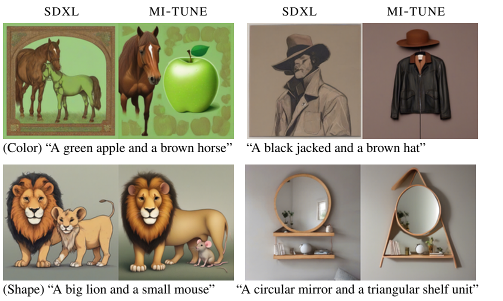

Figure 4: Qualitative examples from Table 4 (same seed used for a given prompt). More examples in Appendix G.

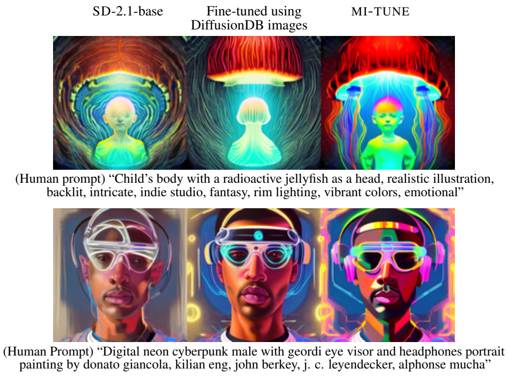

Figure 5: Qualitative examples from Table 5 (same seed used for a given prompt). More examples in Appendix H.

5 Conclusion

T2I alignment emerged as an important endeavor to steer image generation to follow the semantics and user intent expressed through a natural text prompt, as it can save considerable manual effort. In this work, we presented a novel approach to improve model alignment, that uses point-wise MI between prompt-image pairs as a meaningful signal to evaluate the amount of information "flowing" between natural text and images. We demonstrated, both qualitatively and quantitatively, that point-wise MI is coherent with existing alignment measures that either use auxiliary VQA models or elicit human intervention.

We presented MI-TUNE, a lightweight, self-supervised fine-tuning method that uses a pre-trained T2I model such as SD to estimate MI, and to generate a synthetic set of aligned prompt-image pairs, which is then used in a parameter-efficient fine-tuning stage, to align the T2I model. Our approach does not require human annotation, auxiliary VQA models, nor costly inference-time techniques, and achieves a new state-of-the-art across all categories/metrics explored in the literature, often by a sizable margin. These results carry on in more complex tasks, and for various base models, illustrating the flexibility of MI-TUNE.

REFERENCES

Aishwarya Agarwal, Srikrishna Karanam, K J Joseph, Apoorv Saxena, Koustava Goswami, and Balaji Vasan Srinivasan. A-star: Test-time attention segregation and retention for text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2283–2293, October 2023.

Alexander A Alemi and Ian Fischer. Gilbo: One metric to measure them all. *Advances in Neural Information Processing Systems*, 31, 2018.

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2016.

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. ediff-i: Text-to-image diffusion models with ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022a.

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. ediff-i: Text-to-image diffusion models with ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022b.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Anthony J Bell and Terrence J Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159, 1995.

Rob Brekelmans, Sicong Huang, Marzyeh Ghassemi, Greg Ver Steeg, Roger Baker Grosse, and Alireza Makhzani. Improving mutual information estimation with annealed and energy-based bounds. In *International Conference on Learning Representations*, 2022.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3558–3568, 2021.

Agneet Chatterjee, Gabriela Ben Melech Stan, Estelle Aflalo, Sayak Paul, Dhruba Ghosh, Tejas Gokhale, Ludwig Schmidt, Hannaneh Hajishirzi, Vasudev Lal, Chitta Baral, and Yezhou Yang. Getting it right: Improving spatial consistency in text-to-image models, 2024.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Trans. Graph.*, 42(4), 2023a.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models, 2023b.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, 29, 2016.

Kevin Clark, Paul Vicol, Kevin Swersky, and David J. Fleet. Directly fine-tuning diffusion models on differentiable rewards. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=1vmSEVL19f.

Colin Conwell and Tomer Ullman. Testing relational understanding in text-guided image generation, 2022.

Omer Dahary, Or Patashnik, Kfir Aberman, and Daniel Cohen-Or. Be yourself: Bounded attention for multi-subject text-to-image generation, 2024.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794. Curran Associates, Inc., 2021.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Reinforcement learning for fine-tuning text-to-image diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=8OTPepXzeh.

Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Reddy Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. In *The Eleventh International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=PUIqjT4rzq7.

Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Reddy Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. In *The Eleventh International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=PUIqjT4rzq7.

Giulio Franzese, Mustapha Bounoua, and Pietro Michiardi. MINDE: Mutual information neural diffusion estimation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=0kWd8SJq8d.

Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors, 2022.

Brian Gordon, Yonatan Bitton, Yonatan Shafir, Roopal Garg, Xi Chen, Dani Lischinski, Daniel Cohen-Or, and Idan Szpektor. Mismatch quest: Visual and textual feedback for image-text misalignment, 2023.

Paul Grimal, Hervé Le Borgne, Olivier Ferret, and Julien Tourille. Tiam - a metric for evaluating alignment in text-to-image generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2890–2899, January 2024.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2019.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. URL https://openreview.net/forum?id=qw8AKxfYbI.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20406–20417, October 2023.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=weHBzTLXpH.

Sicong Huang, Alireza Makhzani, Yanshuai Cao, and Roger Grosse. Evaluating lossy compression rates of deep generative models. In *International Conference on Machine Learning*. PMLR, 2020.

HuggingFace. Negative prompts. https://huggingface.co/spaces/stabilityai/stable-diffusion/discussions/7857, 2022.

Imagen-Team et al. Imagen 3, 2024.

Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9307–9315, June 2024.

Dongzhi Jiang, Guanglu Song, Xiaoshi Wu, Renrui Zhang, Dazhong Shen, Zhuofan Zong, Yu Liu, and Hongsheng Li. Comat: Aligning text-to-image diffusion model with image-to-text concept matching, 2024. URL https://arxiv.org/abs/2404.03653.

Wonjun Kang, Kevin Galim, and Hyung Il Koo. Counting guidance for high fidelity text-to-image synthesis, 2023.

Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. If at first you don't succeed, try, try again: Faithful diffusion-based text-to-image generation by selection. *arXiv preprint arXiv:2305.13308*, 2023.

M. G. Kendall. A Ner Measure of Rank Correlation. *Biometrika*, 30(1-2):81–93, 06 1938.

Yunji Kim, Jiyoung Lee, Jin-Hwa Kim, Jung-Woo Ha, and Jun-Yan Zhu. Dense text-to-image generation with attention modulation. In *ICCV*, 2023.

Diederik P Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=2LdBqxc1Yv.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, October 2023.

Xianghao Kong, Ollie Liu, Han Li, Dani Yogatama, and Greg Ver Steeg. Interpretable diffusion via information decomposition. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=X6tNkN6ate.

Benno Krojer, Elinor Poole-Dayan, Vikram Voleti, Christopher Pal, and Siva Reddy. Are diffusion models vision-and-language reasoners? *arXiv preprint arXiv:2305.16397*, 2023.

Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback, 2023.

Nunzio A Letizia and Andrea M Tonello. Copula density neural estimation. *arXiv preprint arXiv:2211.15353*, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, 2023a.

Yumeng Li, Margret Keuper, Dan Zhang, and Anna Khoreva. Divide & bind your attention for improved generative semantic nursing. In *34th British Machine Vision Conference 2023, BMVC 2023*, 2023b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. URL https://arxiv.org/abs/1405.0312.

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B. Tenenbaum. Compositional visual generation with composable diffusion models. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*, pp. 423–439, 2022.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation, 2024.

Vivian Liu and Lydia B Chilton. Design guidelines for prompt engineering text-to-image generative models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022.

Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-diffusion:open domain personalized text-to-image generation without test-time fine-tuning, 2023.

David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

Shweta Mahajan, Tanzila Rahman, Kwang Moo Yi, and Leonid Sigal. Prompting hard or hardly prompting: Prompt inversion for text-to-image diffusion models, 2023.

Oscar Mañas, Pietro Astolfi, Melissa Hall, Candace Ross, Jack Urbanek, Adina Williams, Aishwarya Agrawal, Adriana Romero-Soriano, and Michal Drozdzal. Improving text-to-image consistency via automatic prompt optimization, 2024.

David McAllester and Karl Stratos. Formal limitations on the measurement of mutual information. In *International Conference on Artificial Intelligence and Statistics*, 2020.

Tuna Han Salih Meral, Enis Simsar, Federico Tombari, and Pinar Yanardag. Conform: Contrast is all you need for high-fidelity text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models, 2022.

Michael Ogezi and Ning Shi. Optimizing negative prompts for enhanced aesthetics and fidelity in text-to-image generation, 2024.

Bernt Øksendal. *Stochastic differential equations*. Springer, 2003.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *Advances in neural information processing systems*, 2018.

OpenAI. Chatgpt (gpt-4) [large language model], 2024. URL https://chat.openai.com/chat.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt.

Liam Paninski. Estimation of entropy and mutual information. *Neural computation*, 15(6):1191–1253, 2003.

George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. *Advances in neural information processing systems*, 30, 2017.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=di52zR8xgf.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=AOKU4nRw1W.

Benjamin Rhodes, Kai Xu, and Michael U Gutmann. Telescoping density-ratio estimation. *Advances in neural information processing systems*, 2020.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241, 2015.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, volume 35, pp. 36479–36494, 2022.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models, 2022.

Dvir Samuel, Rami Ben-Ari, Simon Raviv, Nir Darshan, and Gal Chechik. Generating images of rare concepts using pre-trained diffusion models. In *AAAI*, 2024.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Dazhong Shen, Guanglu Song, Zeyue Xue, Fu-Yun Wang, and Yu Liu. Rethinking the spatial inconsistency in classifier-free diffusion guidance, 2024.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pp. 2256–2265, 2015.

Jiaming Song and Stefano Ermon. Understanding the limitations of variational mutual information estimators. In *International Conference on Learning Representations*, 2019a.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019b.

Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12438–12448. Curran Associates, Inc., 2020.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Karl Stratos. Mutual information maximization for simple and accurate part-of-speech induction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019.

Jiao Sun, Deqing Fu, Yushi Hu, Su Wang, Royi Rassin, Da-Cheng Juan, Dana Alon, Charles Herrmann, Sjoerd van Steenkiste, Ranjay Krishna, and Cyrus Rashtchian. Dreamsync: Aligning text-to-image generation with image understanding feedback, 2023.

Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the DAAM: Interpreting stable diffusion using cross attention. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5644–5659, July 2023.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization, 2023.

Zijie J. Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv:2210.14896 [cs]*, 2022. URL https://arxiv.org/abs/2210.14896.

Zijie J. Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023a.

Zirui Wang, Zhizhou Sha, Zheng Ding, Yilin Wang, and Zhuowen Tu. Tokencompose: Grounding diffusion with token-level supervision, 2023b.

Sam Witteveen and Martin Andrews. Investigating prompt engineering in diffusion models, 2022.

Qiucheng Wu, Yujian Liu, Handong Zhao, Trung Bui, Zhe Lin, Yang Zhang, and Shiyu Chang. Harnessing the spatial-temporal attention of diffusion models for high-fidelity text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7766–7776, October 2023a.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis, 2023b.

Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference, 2023c.

Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference, 2023d.

Xindi Wu, Dingli Yu, Yangsibo Huang, Olga Russakovsky, and Sanjeev Arora. Conceptmix: A compositional image generation benchmark with controllable difficulty, 2024. URL https://arxiv.org/abs/2408.14339.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=JVzeOYEx6d.

Huizhuo Yuan, Zixiang Chen, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning of diffusion models for text-to-image generation, 2024.

Xinchen Zhang, Ling Yang, Yaqi Cai, Zhaochen Yu, Kai-Ni Wang, Jiake Xie, Ye Tian, Minkai Xu, Yong Tang, Yujiu Yang, and Bin Cui. Realcompo: Balancing realism and compositionality improves text-to-image diffusion models, 2024a. URL https://arxiv.org/abs/2402.12908.

Yang Zhang, Rui Zhang, Xuecheng Nie, Haochen Li, Jikun Chen, Yifan Hao, Xin Zhang, Luoqi Liu, and Ling Li. Spdiffusion: Semantic protection diffusion for multi-concept text-to-image generation, 2024b. URL https://arxiv.org/abs/2409.01327.

Lirui Zhao, Yue Yang, Kaipeng Zhang, Wenqi Shao, Yuxin Zhang, Yu Qiao, Ping Luo, and Rongrong Ji. Diffagent: Fast and accurate text-to-image api selection with large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6390–6399, June 2024.

Shengjia Zhao, Jiaming Song, and Stefano Ermon. A lagrangian perspective on latent variable generative models. In *Proc. 34th Conference on Uncertainty in Artificial Intelligence*, 2018.

Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Simple multi-dataset detection. In *CVPR*, 2022.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=1tZbq88f27.