Intermediate Text Representation Guided Text-to-Image Generation for Enhancing One-and-Only Alignment

Soyoun Won[1], Aryan Yazdan Parast[1], Basim Azam[1], Jean Honorio[1], and Naveed Akhtar[1]

The University of Melbourne, Australia
{soyoun.won, aryan.yazdanparast}@student.unimelb.edu.au,
{basim.azam, jean.honorio, naveed.akhtar1}@unimelb.edu.au

Abstract. Text-to-image (T2I) diffusion models often fail to faithfully render explicit textual descriptions, instead defaulting to strongly learned visual priors due to a phenomenon referred to as *concept association bias*. We show that such bias is particularly strong for *one-and-only (OAO)* objects, entities that exist in a single canonical form, such as celestial bodies, landmarks, and artworks. The deeply ingrained visual identity for these concepts often resists modification through prompting alone. Addressing this challenge, we first identify through an information-theoretic analysis that the final text embedding discards concept-level information present in the intermediate-layer text representations, reducing the mutual information available to the subsequent denoising process. We then propose Intermediate Text Representation (IR)-guided diffusion, which injects intermediate hidden states of the text encoder into the conditioning signal during early denoising steps, recovering suppressed concepts without any additional training, optimization, or external models. To systematically evaluate the challenging task of aligning generative outputs with unusual prompts for OAO objects, we introduce OAO-AttackBench, a benchmark comprising counterfactual prompts that directly conflict with the core visual identity of OAO objects. Experiments on four benchmarks, including OAO-AttackBench, show that our method achieves up to a 19.1 percentage-point improvement in VQAScore while preserving generation fidelity and human preference. Project page: https://soyoun-won.github.io/one-and-only-ir-guidance/.

Keywords: Text-to-image generation · Text-to-image alignment · Diffusion models

1 Introduction

Recent text-to-image (T2I) diffusion models have demonstrated remarkable ability to generate photorealistic images from natural language prompts [10, 21, 22, 39]. Despite this progress, these models frequently fail to faithfully reflect the input prompt, instead defaulting to strongly learned visual priors [3, 5]. Several

Fig. 1: Concept association bias in one-and-only entities. SD3.0 defaults to its learned visual prior for the Earth, ignoring explicitly specified attributes (✗). SD3.0 fails to generate individual concepts, and where it does (yellow, square), it fails to appropriately associate them with the Earth. Our method faithfully reflects all specified concepts and their relationships (✓) without additional training or external models.

methods have been proposed to mitigate this issue [13, 32]; however, these approaches require prior knowledge of the specific prompt at inference time, which is difficult to obtain in practice.

Even when a text prompt is underspecified, state-of-the-art T2I models generate high-quality images by filling in the details from the internal knowledge acquired during the training process [32, 54]. While this ability produces plausible outputs in a typical setting, it becomes a liability when the prompt explicitly contradicts the model's learned associations [3,12,20]. We find that in such cases, T2I models tend to ignore the explicit textual instruction in favour of their ingrained priors.

We refer to this phenomenon as *concept association bias*: the tendency of a model to bind certain visual attributes so tightly to a concept that explicit textual overrides are suppressed. Importantly, we do not frame this *bias* as inherently harmful. It is precisely what enables T2I models to generate rich images from sparse descriptions. However, it fundamentally limits the realization of creative or counterfactual prompts. We observe that concept association bias is particularly strong for entities with little variation in their visual form and becomes most acute for what we call *one-and-only* entities, entities that exist in a single canonical instance, such as the Earth, the Mona Lisa, or the Eiffel Tower. We provide a motivating example in Figure 1 using the prompt, *"Earth is a {color} planet"*, where the target color attribute deviates from the canonical blue-green appearance of the Earth. Stable Diffusion 3.0 (SD3.0) consistently

fails to override its prior, either ignoring the color attribute entirely or failing to associate it with the Earth.

Our work is motivated by a key observation: simple concepts emerge in the early layers but some of these concepts are lost in the final embedding [45]. We hypothesize that the concepts ignored in images generated by T2I models are often encoded in the intermediate layers of the text encoder, yet fail to be materialized due to excessive abstraction. We formalize this intuition under an information-theoretic perspective, showing that intermediate hidden states carry greater mutual information with the input text than the final embedding, and that conditioning on these richer representations during the denoising stage of T2I generation can recover the suppressed attribute-level information.

Based on this analysis, we propose Intermediate Text Representation (IR)-guided diffusion, which injects intermediate representations of text encoders into the conditioning signal during the early denoising steps. IR-guidance operates entirely within the existing model at inference time, requiring no additional training, optimization, or external models. Unlike prior training-free approaches that require prior knowledge about text prompts [5], our method leverages an untapped internal resource of the text-to-image pipeline itself and employs it for any concept in the prompt without manual specification.

We are particularly interested in the alignment of generative outputs for one-and-only (OAO) objects. To rigorously evaluate this largely underexplored setting, where concept association bias is especially strong due to the fixed visual identity of OAO objects, we introduce a new benchmark, OAO-AttackBench. OAO-AttackBench is designed to challenge the core visual identity of OAO entities across three categories: celestial objects, landmarks, and artworks. We evaluate our method on OAO-AttackBench, as well as on existing datasets that challenge input alignment, including Whoops [3], Gecko(R), and Gecko(S) [52], to demonstrate generalization beyond OAO entities. Experimental results confirm that IR-guidance effectively steers the model toward the explicit text prompt, overriding deeply learned associations while preserving generation fidelity and human preference. Our contributions are summarized below:

- Through an information-theoretic analysis, we show that conditioning on intermediate text representations with greater mutual information improves text-to-image alignment for prompts with strong learned associations.
- We propose a training- and optimization-free method for improved T2I generation that leverages intermediate text representations during the denoising process to recover suppressed concepts without external models or predefined concept sets.
- We introduce OAO-AttackBench, a benchmark of counterfactual prompts targeting the core visual characteristics of one-and-only (OAO) entities across three categories, providing a systematic evaluation protocol for concept association bias.

2 Related Work

Our work draws on three lines of research: compositional T2I generation, bias in T2I diffusion models, and leveraging internal representations of the text encoders. We review each below and position our approach relative to prior methods.

Compositional Text-to-Image Generation. Faithfully composing multiple concepts specified in a text prompt remains a central challenge in text-to-image generation. Existing lines of work include manipulating the tokens of the text embedding [16, 40], modifying cross-attention layers [14, 24, 35], or incorporating large language models [17, 43]. However, these approaches often require prompt-specific information at inference time, introduce substantial computational overhead due to additional training or optimization procedures, or rely on external models.

Bias in Text-to-Image Diffusion Models. Previous works have revealed the biased nature of T2I diffusion models, with extensive work focusing on gender bias [1, 6, 11, 23, 26, 34, 49] and cultural bias [42, 48, 55] demonstrating that certain occupations are more strongly associated with particular genders or races. While widely studied, existing formulations of bias are often limited and require predefined sets. In a more general sense, T2I models inherit strong association biases between objects and their typical visual attributes from training data, often neglecting explicitly specified attributes in the prompt [19, 37, 46]. For instance, objects such as tomatoes and sunflowers resist recoloring because their typical attributes dominate the cross-attention maps [37]. Editing these associations through projection space has shown promise [32], but it requires prior knowledge of which associations to target. Our method does not require any predefined bias set and can address concept associations without prior specification of the target correlations.

Notably, these approaches often aim to remove certain features from generative outputs or introduce diversity associated with the concepts. In other words, they typically aim to address cases where the model associates certain concepts that the user did not explicitly request. Our objective is fundamentally different: we address the complementary problem, where the model fails to produce concepts that the user explicitly specified, due to strong association priors ingrained in the model's training data.

Internal Representations of Text Encoders. Recent work has shown that intermediate hidden states of text encoders carry information missing from the final text embedding [24, 27, 45]. Diffusion Lens [45] investigated the hidden states by using them as textual conditions for T2I models. The authors found that simple concepts emerge in the earlier layers, but some of these are absent from the generated images.

Kim et al. [24] found that the self-attention map of the text encoder encodes syntax information within the prompt and proposed a method that aligns the cross-attention maps of the denoising network to the self-attention maps. However, they require optimization, which adds computational complexity and only works in diffusion models with a single text encoder.

3 Proposed Method

We present IR-guided diffusion, a training-free method that leverages intermediate hidden states of the text encoder to recover concept information lost in the final embedding. We first formalize the concept omission problem through the lens of information theory (§3.1), then show that intermediate text representations provably carry richer information (§3.2), describe how this signal is injected into the denoising process in our method (§3.3), and introduce an adaptive scheduler that automatically determines the stopping point for injection (§3.4).

3.1 Problem Formulation

Consider a text prompt Y that describes an object A together with a specified attribute b (e.g., $A =$ "Earth" and $b =$ "purple" in the prompt "*Earth is a purple planet*"). We write $Y = \{A, b\}$, where both A and b may themselves consist of sub-concepts. Let x_t be the latent variable at diffusion timestep t and $p_\theta(x_t \mid \cdot)$ be the conditional distribution parameterized by θ. A concept omission arises when the model ignores the specified attribute. This is characterized by the scenario where images generated with and without b are distributionally similar. Formally,

$$D_{\mathrm{KL}}\big(p_\theta(x_t \mid A, b) || p_\theta(x_t \mid A)\big) \approx 0, \tag{1}$$

where D_{KL} denotes the Kullback–Leibler divergence. From an information-theoretic perspective, Eq. (1) is equivalent to

$$I_\theta(x_t;\, b \mid A) \approx 0, \tag{2}$$

where $I_\theta(\cdot;\cdot \mid \cdot)$ denotes the conditional mutual information under the model distribution parameterized by θ. This suggests that the generated latent contains negligible information about b once the object A is accounted for. By the chain rule of mutual information, the total information decomposes as

$$I(Y;\, x_t) = I(A;\, x_t) + I(b;\, x_t \mid A). \tag{3}$$

Our goal is to increase $I(b;\, x_t \mid A)$, the attribute-specific information in the generated image, while preserving $I(A;\, x_t)$. This formulation naturally handles the concept association bias in the generated image.

3.2 Theory: Intermediate Representations

A typical contemporary text encoder processes the text prompt Y through a sequence of transformer blocks. Let c_h denote the representation at an intermediate block of the underlying network, and c_f denote the representation of the final one. Without loss of generality, we assume a simple network in our analysis that has only these two blocks. Since each processing block is a deterministic function of its input, the encoder forms a Markov chain:

$$Y \;\to\; c_h \;\to\; c_f, \qquad c_f = h(c_h), \tag{4}$$

where h is a deterministic transformation function that maps c_h to c_f for our simple network, and denotes an arbitrary composition of processing blocks for a more complex network.

Lemma 1. *Since c_f is a deterministic function of c_h, the chain $Y \to c_h \to c_f$ satisfies the Markov property. By data processing inequality (DPI) [7],*

$$I(Y;\, c_h) \;\geq\; I(Y;\, c_f). \tag{5}$$

A formal proof of the lemma is provided in Supp. Mat. §A. Intuitively, a later abstraction block in a network can only transform the signal from an earlier block, which can either preserve or lose information. Eq. (5) establishes that intermediate hidden states carry at least as much mutual information with the text prompt as the final embeddings. Since the denoising network conditions on the text representation, we posit that richer conditioning can translate to richer latents for the model. Mathematically,

$$I(Y;\, x_t^h) \;\geq\; I(Y;\, x_t^f), \tag{6}$$

where x_t^h and x_t^f denote the latent at timestep t when conditioned on c_h and c_f, respectively. We validate this assumption in §4: images conditioned on c_h consistently recover the specified attribute that c_f fails to express.

Prior work has shown that object-level concepts emerge in early layers and are largely preserved through later layers [45], suggesting $I(A;\, x_t^h) \approx I(A;\, x_t^f)$. Combining this with Eq. (6) and the chain rule (Eq. (3)), we get:

$$I(b;\, x_t^h \mid A) \;\geq\; I(b;\, x_t^f \mid A). \tag{7}$$

This implies that conditioning on the intermediate representation can enhance attribute-specific mutual information during denoising.

3.3 IR-guided Diffusion

Motivated by the analysis above, we inject intermediate representations into the conditioning signal during denoising. The overview is shown in Figure 2. We treat c_h as a complementary guidance signal and construct an *IR-guided embedding*:

$$\tilde{c} = c_f + \lambda c_h, \tag{8}$$

where $\lambda \in [0, 1]$ controls the injection strength. Prior to injection, we apply the text encoder's final layer normalization to c_h to align its distribution with c_f; see Supp. Mat. §B for a distributional analysis. The modified score function is then

$$\nabla_{x_t} \log p_\theta(x_t \mid \tilde{c}). \tag{9}$$

Multi-Encoder Architectures. Diffusion architectures with a single text encoder work directly with Eq. (8). For architectures employing N text encoders whose outputs are concatenated before conditioning, we extract intermediate

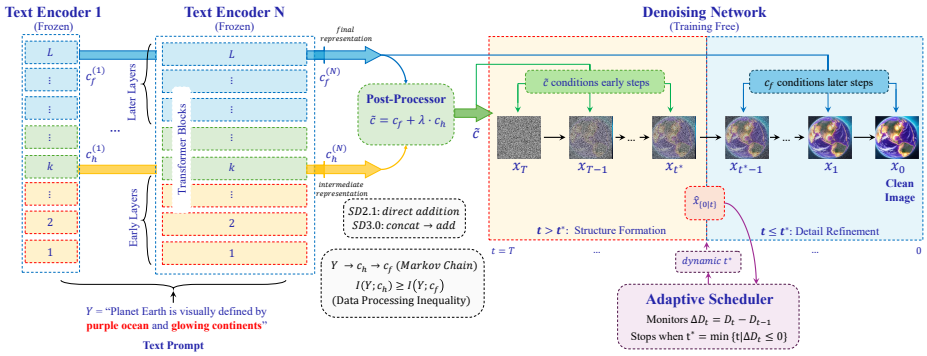

Fig. 2: (Left) The frozen text encoder produces the final text embedding c_f (layer L) and an intermediate hidden state c_h (layer k), which carries information suppressed in c_f. (Middle) The Post-Processor constructs an IR-guided embedding $\tilde{c} = c_f + \lambda c_h$, injecting suppressed concept information back into the conditioning signal. (Right) $\tilde{c}$ conditions the early denoising steps ($t > t^*$) where global structure is determined, while c_f conditions the later steps ($t \leq t^*$) for detail refinement. The Adaptive Scheduler dynamically determines t^* by monitoring structural acceleration ΔD_t. No training or optimization is required.

representations from each encoder independently, concatenate them following the native scheme, and add the result to the concatenated final embedding:

$$\tilde{c} = \left[c_f^{(1)} \parallel \cdots \parallel c_f^{(N)} \right] + \lambda \left[c_h^{(1)} \parallel \cdots \parallel c_h^{(N)} \right], \tag{10}$$

where $[\cdots \| \cdots]$ denotes the concatenation and superscripts index the encoders. This preserves the native conditioning interface for each architecture without requiring any architectural modification.

Post-Processing Strategies. To validate the effectiveness of additive injection as the post-processing strategy, we formulate four alternatives. Following [44], the final layer normalization is applied to the intermediate representation for all variants. *None* uses the intermediate representation directly $\tilde{c} = c_h$. *Mean-Std* normalizes c_h to match the samplewise statistics of c_f:

$$\tilde{c}_{\text{mean-std}} = \frac{c_h - \mu(c_h)}{\sigma(c_h)} \cdot \sigma(c_f) + \mu(c_f), \tag{11}$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ denote the mean and standard deviation, respectively. *Orthogonal projection* decomposes c_h into a component aligned with c_f and a residual. The projection matrix is

$$P = c_f \left(c_f^\top c_f\right)^{-1} c_f^\top, \tag{12}$$

yielding two variants: *Projection* ($\tilde{c}_{\text{proj}} = Pc_h$) and *Residual* ($\tilde{c}_{\text{res}} = c_h - Pc_h$). See Supp. Mat. §B for a comparative analysis.

3.4 Adaptive Scheduling

While intermediate representations carry richer information, conditioning on $\tilde{c}$ throughout the entire denoising process can introduce artifacts. Since diffusion models establish global structure in early denoising steps and refine fine-grained detail in later steps [5,14,50], we condition on $\tilde{c}$ only during the early, structure-forming phase and revert to the standard embedding c_f once the layout stabilizes.

To determine the transition point t^* dynamically, we monitor structural changes in the estimated clean image $\hat{x}_0$ via Tweedie's formula [9, 38] at each timestep:

$$\hat{x}_{0|t} = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\,\epsilon_\theta(x_t, t)}{\sqrt{\bar{\alpha}_t}}, \tag{13}$$

where $\bar{\alpha}_t$ is a set of monotonically decreasing parameters that control the noise level and ϵ_θ denotes the predicted noise parameterized by θ. Then, we measure the magnitude of structural updates as the squared difference between consecutive predictions

$$D_t = \left\| \hat{x}_{0|t} - \hat{x}_{0|t-1} \right\|_2^2, \tag{14}$$

and define the *structural acceleration* as its first-order difference:

$$\Delta D_t = D_t - D_{t-1}. \tag{15}$$

$\Delta D_t > 0$ indicates that the structural changes are intensifying, and $\Delta D_t \leq 0$ suggests that the layout is stabilizing. The transition point is the first timestep at which acceleration becomes non-positive:

$$t^* = \min\left\{ t \;\middle|\; \Delta D_t \leq 0 \right\}. \tag{16}$$

The complete denoising process is:

$$x_{t-1} = \begin{cases} \text{denoise}(x_t,\, \tilde{c}) & \text{if } t > t^*, \\ \text{denoise}(x_t,\, c_f) & \text{if } t \leq t^*. \end{cases} \tag{17}$$

This adaptive strategy avoids the need for a fixed hyperparameter to control the transition, and the overhead of computing D_t is negligible compared to the denoising step itself; see Supp. Mat. §K for computational cost analysis.

3.5 OAO-AttackBench

Models pick up biases from the datasets they are exposed to during training [4,41,54]. We observe from the previous section that such concept association bias is severe for one-and-only entities, objects that exist in a single canonical form. To address the challenging cases of concept association, we present **OAO-AttackBench**. OAO-AttackBench is designed to attack the implicit knowledge embedded in one-and-only entities. The term *attack* reflects the dataset's nature, which contradicts or violates the core visual identity of the one-and-only

entities. Unlike everyday objects, one-and-only entities are strongly tied to a single visual form, making generation from unusual descriptions more challenging. OAO-AttackBench consists of three categories: celestial objects, landmarks, and artworks, each of which is sub-categorized by shape and material, pattern and material, and genre, style and material, for a total of 504 prompts. Details on the dataset are provided in Supp. Mat. §C.

4 Experiments

We evaluate IR-guided diffusion on four benchmarks spanning counterfactual, commonsense-violating, and general-purpose compositional prompts across multiple diffusion backbones. Our experiments address three questions: (i) does IR-guidance improve text-to-image alignment for prompts that trigger concept association bias? (ii) does it preserve generation fidelity and human preference? (iii) does it generalize to standard prompts without degrading quality?

4.1 Experimental Setup

Datasets. We evaluate our method on four benchmarks: our proposed OAO-AttackBench, Whoops [3], Gecko(R), and Gecko(S) [52]. *OAO-AttackBench* is designed to evaluate a challenging form of concept association bias, where each subject is assumed to exist in only one canonical form. *Whoops* includes "weird" prompts that are designed to violate commonsense about everyday objects, testing whether models can override typical associations such as physical rules, social knowledge, and cultural norms. *Gecko* is a dataset containing 2K prompts, divided into Gecko(R) and Gecko(S). *Gecko(R)* contains texts resampled from existing datasets [1], and *Gecko(S)* contains LLM-generated texts that span categories covered less in Gecko(R).

Models. We implement IR-guided diffusion on Stable Diffusion (SD) 2.1 [39] and 3.0 [10]. SD2.1 employs a pretrained OpenCLIP ViT-H/14 text encoder, and SD3.0 has three pretrained text encoders: OpenCLIP ViT-G, CLIP ViT-L, and T5-XXL. Additional implementation details are provided in Supp. Mat. §D. To exploit word-level distinctiveness suppressed in the final embedding but retained in earlier-layer representations, we extract intermediate representations after the first transformer block of the text encoder ($k=1$). Analyses across encoder depth are provided in Supp. Mat. §E.

Metrics. We measure text-to-image alignment with two complementary metrics. *CLIPScore* [15] computes the cosine similarity between image and text embeddings from OpenAI CLIP ViT-L/14. *VQAScore* [29] evaluates compositional alignment by querying the CLIP-FlanT5 VQA model with the question "Does this figure show {text prompt}?" and averaging the probability of "yes". Because VQAScore probes the presence of fine-grained attributes, it is more sensitive to concept omission than CLIPScore, which captures global semantic similarity.

[1] TIFA [18], Stanford Paragraph [25], Localized Narratives [36], CountBench [33], VRD [31], DiffusionDB [51], MJ [47], PoseScript [8], Whoops [3], and DrawText-Creative [30]

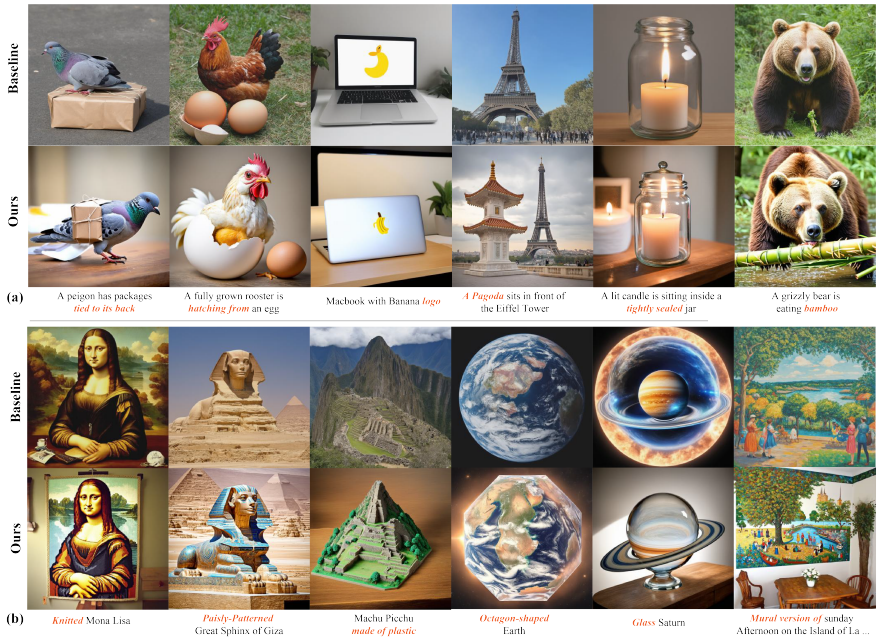

Fig. 3: *Qualitative comparison on SD3.0.* (a) Whoops prompts targeting everyday objects. The baseline generates the individual concepts but fails to express the specified attribute (orange text). (b) OAO-AttackBench prompts targeting one-and-only entities. The baseline preserves canonical appearances despite explicit contradictions in the prompt; the Mona Lisa remains a painting, the Earth remains spherical, and Saturn remains opaque. IR-guided diffusion successfully overrides these priors (*Knitted*, *Paisley Patterned*, *made of plastic*, *Octagon-shaped*, *Glass*) while preserving object identity.

To verify IR-guidance does not introduce artifacts, we report the *Kernel Inception Distance (KID)* [2] between the generated images and COCO [28] reference images. Human preference is quantified with *HPSv2* and *HPSv2.1* [53].

4.2 Results

Qualitative Comparison. Figure 3 presents side-by-side comparisons on SD3.0. On Whoops (top), the baseline generates both individual concepts but fails to express their specified relationships: the packages appear detached from the pigeon rather than *tied to its back*, and the candle sits in an open jar rather than *tightly sealed*. IR-guidance recovers these suppressed attributes and relationships. The contrast is more striking on OAO-AttackBench (bottom), where the baseline preserves canonical appearances despite explicit contradictions in the prompt. Saturn remains opaque despite "*Glass* Saturn"; Earth remains spherical despite "*Octagon-shaped* Earth"; the Mona Lisa retains its painted appearance despite "*Knitted* Mona Lisa." IR-guidance successfully overrides these deeply learned priors while preserving object identity. More qualitative results are provided in Supp. Mat. §F.

Table 1: Quantitative comparison on Whoops and OAO-AttackBench. **Bold** and italic values denote the best and the second-best scores, respectively.

			OAO-AttackBench							**Whoops**
			Celestial Objects		Landmarks		Artworks			
			Shape	Material	Pattern	Material	Genre	Style	Material	
VQA	SD3.0	Baseline	0.514	0.687	0.778	0.525	0.719	*0.733*	0.657	0.778
		None	0.509	0.716	0.773	0.630	0.750	0.693	0.706	0.773
		Projection	*0.563*	0.703	*0.792*	*0.658*	0.748	0.733	0.709	*0.792*
		Residual	0.516	0.694	0.779	0.630	**0.778**	0.700	*0.726*	0.779
		Mean-Std	0.544	*0.725*	0.760	0.617	0.734	0.701	0.716	0.760
		Ours	**0.565**	**0.726**	**0.802**	**0.716**	*0.753*	**0.736**	**0.756**	**0.802**
	SD2.1	Baseline	0.478	0.685	0.629	0.516	*0.688*	0.563	0.675	0.629
		None	0.501	0.700	*0.632*	*0.527*	0.658	0.581	0.702	*0.632*
		Projection	**0.522**	0.701	0.627	0.511	0.646	0.568	*0.702*	0.627
		Residual	*0.509*	**0.711**	0.617	0.510	0.640	*0.585*	0.692	0.617
		Mean-Std	0.465	0.680	0.617	**0.534**	0.668	0.570	0.674	0.617
		Ours	*0.509*	*0.703*	**0.641**	0.514	**0.693**	**0.622**	**0.715**	**0.641**
CLIP	SD3.0	Baseline	0.257	0.244	0.291	0.263	0.263	0.278	0.263	0.291
		None	0.246	0.240	0.293	0.278	0.264	0.267	0.270	0.293
		Projection	*0.258*	0.245	*0.293*	*0.284*	0.269	*0.279*	0.271	*0.293*
		Residual	0.249	0.244	0.291	0.276	*0.269*	0.276	*0.272*	0.291
		Mean-Std	0.252	**0.253**	0.291	0.279	0.260	0.274	0.265	0.291
		Ours	**0.259**	*0.253*	**0.296**	**0.289**	**0.274**	**0.281**	**0.276**	**0.296**
	SD2.1	Baseline	*0.246*	**0.241**	0.280	0.265	0.274	0.271	*0.282*	0.280
		None	0.243	0.239	**0.282**	*0.268*	*0.276*	*0.276*	0.281	**0.282**
		Projection	0.244	0.238	0.280	0.266	0.273	0.275	**0.284**	0.280
		Residual	**0.247**	*0.239*	0.279	0.265	0.273	0.276	0.282	0.279
		Mean-Std	0.245	0.232	0.278	**0.273**	**0.277**	0.275	0.281	0.278
		Ours	0.239	0.239	*0.281*	0.266	0.272	**0.277**	0.280	*0.281*

Text-image Alignment. Table 1 reports VQAScore and CLIPScore on OAO-AttackBench and Whoops. IR-guidance consistently outperforms baseline under VQAScore, achieving up to a 19.1 percentage-point improvement on OAO-AttackBench. The modest variation in CLIPScore suggests that the global semantic alignment is largely preserved, whereas the improvement in VQAScore indicates enhanced fine-grained text-to-image alignment. We also compare against the four alternative post-processing strategies defined in §3.3: None, Mean-Std, Projection, and Residual. As shown in Table 1, Ours (additive injection) consistently outperforms or matches all alternatives. Evaluation results on additional metrics are provided in Supp. Mat. §I.

We additionally report Kernel Inception Distance (KID) in Table 2. IR-guided diffusion achieves comparable or better scores than the baseline across all categories, confirming that the method preserves distributional fidelity. Experimental results suggest that IR-guidance not only improves compo-

Table 2: KID evaluation results on OAO-AttackBench (lower is better). Our method achieves comparable scores across categories.

	Celestial Objects	Landmarks	Artworks
SD2.1	0.0551	0.0559	0.0624
Ours$_{2.1}$	**0.0491**	**0.0557**	**0.0616**
SD3.0	0.0576	0.0583	**0.0382**
Ours$_{3.0}$	**0.0483**	**0.0500**	0.0397

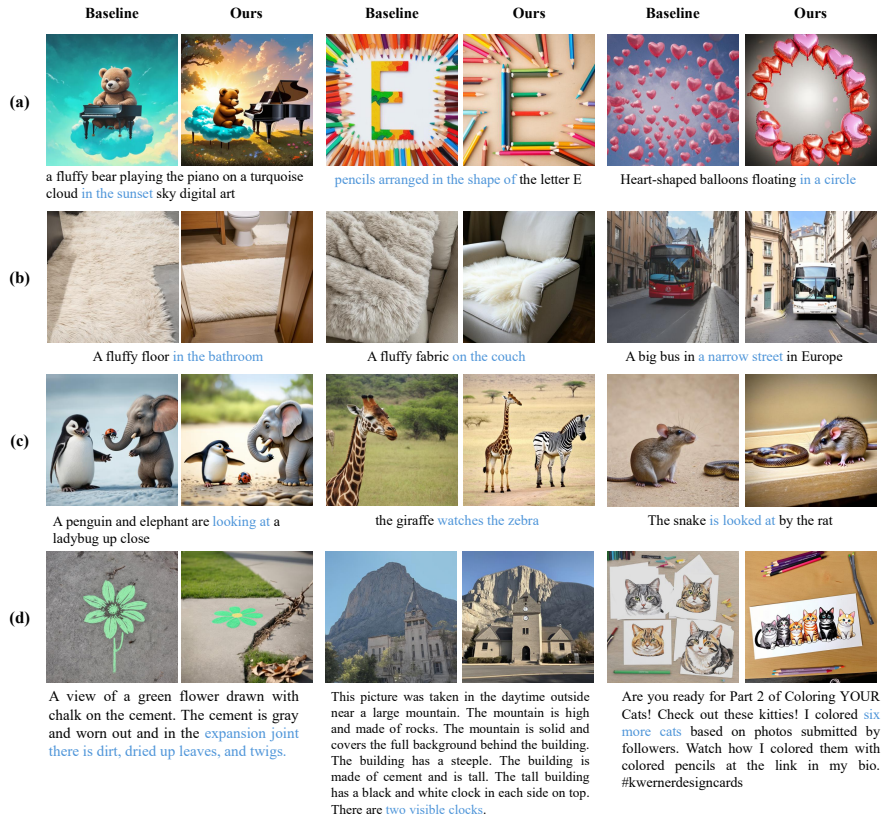

Fig. 4: Sample comparison on Gecko. Blue text denotes concepts ignored or misinterpreted in the baseline output. IR-guidance improves (a) general concept omission problem beyond one-and-only entities, (b) environmental context, (c) relational scenarios involving multiple objects, and (d) long-text prompts.

sitional accuracy but also maintains, and in most cases enhances, the overall distribution quality of the generated images.

Human Preference. Table 3 reports HPSv2 and HPSv2.1 on SD3.0. IR-guided diffusion improves human preference scores on both OAO-AttackBench and Whoops relative to the baseline. The improvement confirms that IR-guided images are not only more faithful to the text prompt but also more visually plausible to human observers.

Generation Beyond Unusual Prompts. So far, we have examined IR-guidance on the set of prompts that attack the core visual characteristics of one-and-only entities (OAO-AttackBench) and on prompts that violate commonsense (Whoops). Here, we further assess its compositional ability in more general settings. Table 4 summarizes the evaluation results on Gecko with SD3.0. As shown in the table, the IR-guided model achieves comparable performance to the base-

line. While improvements are modest, the results indicate that IR-guidance does not compromise standard generation quality on relatively general text prompts.

We visualize samples of IR-guided diffusion on Gecko in Figure 4. *Concept Omission*: For underrepresented concepts, Figure 4(a) demonstrates that IR-guidance effectively mitigates concept omission beyond one-and-only entities. *Contextual and Relational Alignment*: Figure 4(b-c) shows that the baseline generates plausible subjects but often misses the surrounding context or misinterprets interactions between objects. IR-guidance improves contextual cues and object relations. *Long Prompts*: As shown in Figure 4(d), IR-guidance preserves concepts specified in extended textual descriptions while also mitigating concept omission. Across these diverse settings, IR-guidance improves text-to-image alignment while preserving overall image quality.

4.3 Additional Analysis

Effect of IR-guidance Duration. Figure 5 illustrates how the generated images change as IR-guidance is applied over different ranges of timesteps on SD2.1. The first column shows the baseline without IR-guidance, and the subsequent columns progressively extend the range over which IR-guidance is applied. As IR-guidance is applied over more timesteps, the generated images increasingly exhibit the previously missing concepts.

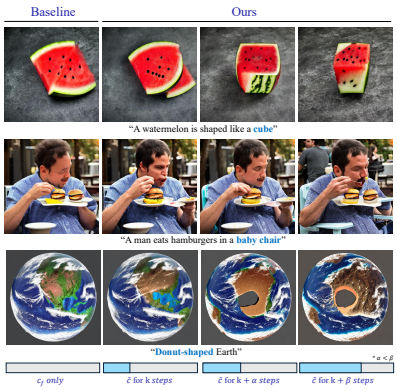

Fig. 5: Effect of IR-guidance duration. Each row shows the baseline (c_f only) and three variants where $\tilde{c}$ is used as conditioning for progressively more denoising steps. The timeline bars at the bottom indicate the conditioning schedule. As IR-guidance is extended to more timesteps, the specified attribute (blue text) is more faithfully expressed.

What does IR-guidance do? To further investigate the step-wise influence of IR-guidance, we calculate the mean squared difference between predicted samples $\hat{x}_0$ from two consecutive timesteps. Figure 6(a) shows the mean squared difference over the denoising process, and Figure 6(b) visualizes the predicted samples at various timesteps for SD2.1. A high mean squared difference indicates a substantial change in the predicted image, suggesting a shift in the overall structure of the generated image.

Figure 6(b) shows that the baseline's estimated output already closely resembles the final generated image at the early timestep $T-6$. In contrast, IR-guided estimation at the same timestep shows a noisier predicted image, where the overall structure is substantially different from the final generated image. IR-guidance incorporates intermediate text representations for the first few steps, then switches the textual conditioning to the final text embedding. Naturally, this introduces

Table 3: Evaluation results on human preference metrics.

Dataset	Model	HPSv2	HPSv2.1
OAO-AttackBench	Baseline	0.269	0.263
	Ours	**0.273**	**0.271**
Whoops	Baseline	0.291	0.295
	Ours	**0.293**	**0.299**

Table 4: Evaluation results on general datasets Gecko(R) and Gecko(S).

Dataset	Model	CLIP	VQA	HPSv2	HPSv2.1
Gecko(R)	Baseline	0.281	0.866	0.281	0.264
	Ours	**0.284**	**0.871**	**0.284**	**0.271**
Gecko(S)	Baseline	0.282	0.744	0.283	0.274
	Ours	**0.285**	**0.748**	**0.286**	**0.278**

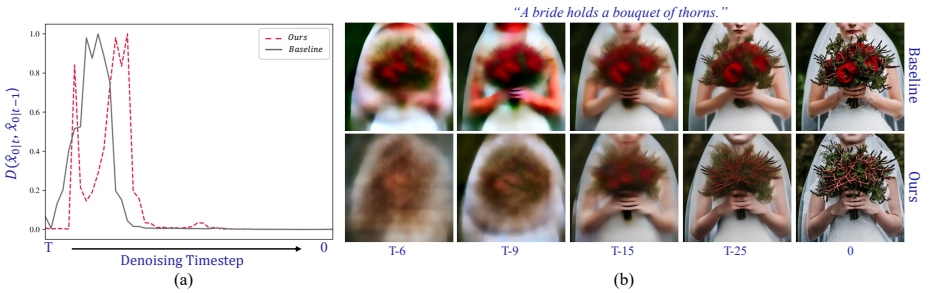

Fig. 6: (a) Mean squared difference of predicted images between two consecutive denoising timesteps. (b) Predicted generated image at each denoising timestep on the baseline (top) and IR-guided diffusion (bottom). IR-guided diffusion delays premature structural convergence, allowing concept exploration during early denoising steps.

a steep increase in the mean squared difference at the transition point. Interestingly, IR-guided diffusion (dotted red line in Figure 6(a)) shows two major local maxima after the transition, suggesting a shift in the denoising trajectory that prevents premature structural convergence.

Ablation Study. Figure 7 reports VQAScore and CLIPScore across different layers of the text encoder for extracting c_h on SD3.0. Earlier layers yield higher VQAScore on OAO-AttackBench, consistent with the observation that word-level information is present in early layers and is compressed in later layers (Supp. Mat. §E). On Whoops, the sensitivity is lower, reflecting the less rigid priors associated with everyday settings. Additional ablation studies are provided in Supp. Mat. §G.

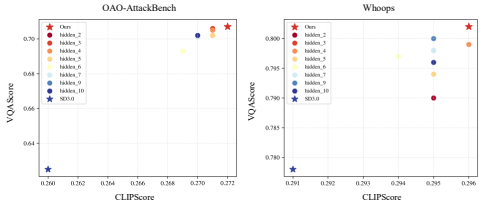

Fig. 7: VQAScore vs. CLIPScore of c_h from different transformer blocks. Earlier layers yield a higher VQAScore on OAO-AttackBench.

5 Conclusion

In this work, we introduce IR-guided diffusion, which leverages the internal states of the text encoder during the denoising process for prompt-faithful generation, accentuating underrepresented concepts. Our approach requires no predefined set

of biases, additional training, optimization, or external models. To validate its effectiveness, we introduce OAO-AttackBench, a challenging dataset of prompts that attack the canonical identity of one-and-only entities. Experimental results show that our approach is effective not only for one-and-only entities but also for everyday objects, while maintaining comparable performance on fidelity and human preference scores compared with the state-of-the-art baseline.

Acknowledgements

Naveed Akhtar is a recipient of the Australian Research Council Discovery Early Career Researcher Award (project # DE230101058) funded by the Australian Government. This research was supported by The University of Melbourne's Research Computing Services and the Petascale Campus Initiative.This work is also partially supported by the Google Research Scholar Program Award.

References

1. Azam, B., Akhtar, N.: Plug-and-play interpretable responsible text-to-image generation via dual-space multi-facet concept control. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2976–2985 (2025)
2. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying MMD GANs. In: International Conference on Learning Representations (2018)
3. Bitton-Guetta, N., Bitton, Y., Hessel, J., Schmidt, L., Elovici, Y., Stanovsky, G., Schwartz, R.: Breaking common sense: Whoops! a vision-and-language benchmark of synthetic and compositional images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2616–2627 (2023)
4. Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramer, F., Balle, B., Ippolito, D., Wallace, E.: Extracting training data from diffusion models. In: 32nd USENIX security symposium (USENIX Security 23). pp. 5253–5270 (2023)
5. Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., Cohen-Or, D.: Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. ACM transactions on Graphics (TOG) **42**(4), 1–10 (2023)
6. Chuang, C.Y., Jampani, V., Li, Y., Torralba, A., Jegelka, S.: Debiasing vision-language models via biased prompts. arXiv preprint arXiv:2302.00070 (2023)
7. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley-Interscience, 2 edn. (2006)
8. Delmas, G., Weinzaepfel, P., Lucas, T., Moreno-Noguer, F., Rogez, G.: Posescript: Linking 3d human poses and natural language. IEEE transactions on pattern analysis and machine intelligence (2024)
9. Efron, B.: Tweedie's formula and selection bias. Journal of the American Statistical Association **106**(496), 1602–1614 (2011)
10. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)
11. Friedrich, F., Brack, M., Struppek, L., Hintersdorf, D., Schramowski, P., Luccioni, S., Kersting, K.: Fair diffusion: Instructing text-to-image generation models on fairness. arXiv preprint arXiv:2302.10893 (2023)

12. Fu, X., He, M., Lu, Y., Wang, W.Y., Roth, D.: Commonsense-t2i challenge: Can text-to-image generation models understand commonsense? arXiv preprint arXiv:2406.07546 (2024)
13. He, F., Zhang, C., Zhao, Z.: Implicit priors editing in stable diffusion via targeted token adjustment. arXiv preprint arXiv:2412.03400 (2024)
14. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022)
15. Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. In: Proceedings of the 2021 conference on empirical methods in natural language processing. pp. 7514–7528 (2021)
16. Hu, T., Li, L., Van de Weijer, J., Gao, H., Shahbaz Khan, F., Yang, J., Cheng, M.M., Wang, K., Wang, Y.: Token merging for training-free semantic binding in text-to-image synthesis. Advances in Neural Information Processing Systems **37**, 137646–137672 (2024)
17. Hu, X., Wang, R., Fang, Y., Fu, B., Cheng, P., Yu, G.: Ella: Equip diffusion models with llm for enhanced semantic alignment. arXiv preprint arXiv:2403.05135 (2024)
18. Hu, Y., Liu, B., Kasai, J., Wang, Y., Ostendorf, M., Krishna, R., Smith, N.A.: Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20406–20417 (2023)
19. Huang, K., Sun, K., Xie, E., Li, Z., Liu, X.: T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. In: Advances in Neural Information Processing Systems. vol. 36 (2023)
20. Huberman, S., Patashnik, O., Dahary, O., Mokady, R., Cohen-Or, D.: Image generation from contextually-contradictory prompts. arXiv preprint arXiv:2506.01929 (2025)
21. Islam, K., AKHTAR, N.: Context-guided responsible data augmentation with diffusion models. In: ICLR 2025 Workshop on Navigating and Addressing Data Problems for Foundation Models
22. Islam, K., Zaheer, M.Z., Mahmood, A., Nandakumar, K., Akhtar, N.: Genmix: effective data augmentation with generative diffusion model image editing. Expert Systems with Applications p. 132273 (2026)
23. Kim, E., Kim, S., Park, M., Entezari, R., Yoon, S.: Rethinking training for de-biasing text-to-image generation: Unlocking the potential of stable diffusion. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 13361–13370 (2025)
24. Kim, J., Esmaeili, E., Qiu, Q.: Text embedding is not all you need: Attention control for text-to-image semantic alignment with text self-attention maps. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 8031–8040 (2025)
25. Krause, J., Johnson, J., Krishna, R., Fei-Fei, L.: A hierarchical approach for generating descriptive image paragraphs. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 317–325 (2017)
26. Li, J., Hu, L., Zhang, J., Zheng, T., Zhang, H., Wang, D.: Fair text-to-image diffusion via fair mapping. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 26256–26264 (2025)
27. Li, Y., Salehi, S., Ungar, L., Kording, K.P.: Does object binding naturally emerge in large pretrained vision transformers? arXiv preprint arXiv:2510.24709 (2025)

28. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
29. Lin, Z., Pathak, D., Li, B., Li, J., Xia, X., Neubig, G., Zhang, P., Ramanan, D.: Evaluating text-to-visual generation with image-to-text generation. In: European Conference on Computer Vision. pp. 366–384. Springer (2024)
30. Liu, R., Garrette, D., Saharia, C., Chan, W., Roberts, A., Narang, S., Blok, I., Mical, R., Norouzi, M., Constant, N.: Character-aware models improve visual text rendering. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 16270–16297 (2023)
31. Lu, C., Krishna, R., Bernstein, M., Fei-Fei, L.: Visual relationship detection with language priors. In: European conference on computer vision. pp. 852–869. Springer (2016)
32. Orgad, H., Kawar, B., Belinkov, Y.: Editing implicit assumptions in text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7030–7038 (2023)
33. Paiss, R., Ephrat, A., Tov, O., Zada, S., Mosseri, I., Irani, M., Dekel, T.: Teaching clip to count to ten. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3170–3180 (2023)
34. Parast, A.Y., Azam, B., Akhtar, N.: Ddb: Diffusion driven balancing to address spurious correlations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17526–17535 (2025)
35. Phung, Q., Ge, S., Huang, J.B.: Grounded text-to-image synthesis with attention refocusing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7932–7942 (2024)
36. Pont-Tuset, J., Uijlings, J., Changpinyo, S., Soricut, R., Ferrari, V.: Connecting vision and language with localized narratives. In: European conference on computer vision. pp. 647–664. Springer (2020)
37. Rassin, R., Hirsch, E., Glickman, D., Ravfogel, S., Goldberg, Y., Chechik, G.: Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. Advances in Neural Information Processing Systems **36** (2024)
38. Robbins, H.E.: An empirical bayes approach to statistics. In: Breakthroughs in Statistics: Foundations and basic theory, pp. 388–394. Springer (1992)
39. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)
40. Seo, H., Bang, J., Lee, H., Lee, J., Lee, B.H., Chun, S.Y.: Geometrical properties of text token embeddings for strong semantic binding in text-to-image generation. arXiv preprint arXiv:2503.23011 (2025)
41. Seshadri, P., Singh, S., Elazar, Y.: The bias amplification paradox in text-to-image generation. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 6367–6384 (2024)
42. Shen, X., Du, C., Pang, T., Lin, M., Wong, Y., Kankanhalli, M.: Finetuning text-to-image diffusion models for fairness. arXiv preprint arXiv:2311.07604 (2023)
43. Sun, J., Fu, D., Hu, Y., Wang, S., Rassin, R., Juan, D.C., Alon, D., Herrmann, C., Van Steenkiste, S., Krishna, R., et al.: Dreamsync: Aligning text-to-image generation with image understanding feedback. In: Proceedings of the 2025 Conference

of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 5920–5945 (2025)
44. Tang, Y., Yamada, Y., Zhang, Y., Yildirim, I.: When are lemons purple? the concept association bias of vision-language models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 14333–14348 (2023)
45. Toker, M., Orgad, H., Ventura, M., Arad, D., Belinkov, Y.: Diffusion lens: Interpreting text encoders in text-to-image pipelines. arXiv preprint arXiv:2403.05846 (2024)
46. Trusca, M.M., et al.: Object-attribute binding in text-to-image generation: Evaluation and control (2024), arXiv:2404.13766
47. Turc, I., Nemade, G.: Midjourney user prompts & generated images (250k) (2023)
48. Um, S., Ye, J.C.: Minority-focused text-to-image generation via prompt optimization. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 20926–20936 (2025)
49. Vice, J., Akhtar, N., Sigal, L., Hartley, R., Mian, A.: On the fairness, diversity and reliability of text-to-image generative models. Artificial Intelligence Review **59**(2), 57 (2026)
50. Wang, B., Vastola, J.J.: Diffusion models generate images like painters: an analytical theory of outline first, details later. arXiv preprint arXiv:2303.02490 (2023)
51. Wang, Z.J., Montoya, E., Munechika, D., Yang, H., Hoover, B., Chau, D.H.: 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). pp. 893–911 (2023)
52. Wiles, O., Zhang, C., Albuquerque, I., Kajić, I., Wang, S., Bugliarello, E., Onoe, Y., Papalampidi, P., Ktena, I., Knutsen, C., et al.: Revisiting text-to-image evaluation with gecko: On metrics, prompts, and human ratings. arXiv preprint arXiv:2404.16820 (2024)
53. Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., Li, H.: Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. arXiv preprint arXiv:2306.09341 (2023)
54. Yang, Y., Lin, Y., Liu, H., Shao, W., Chen, R., Shang, H., Wang, Y., Qiao, Y., Zhang, K., Luo, P.: Position: towards implicit prompt for text-to-image models. arXiv preprint arXiv:2403.02118 (2024)
55. Zhang, H., Guo, Y., Kankanhalli, M.: Joint vision-language social bias removal for clip. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 4246–4255 (2025)