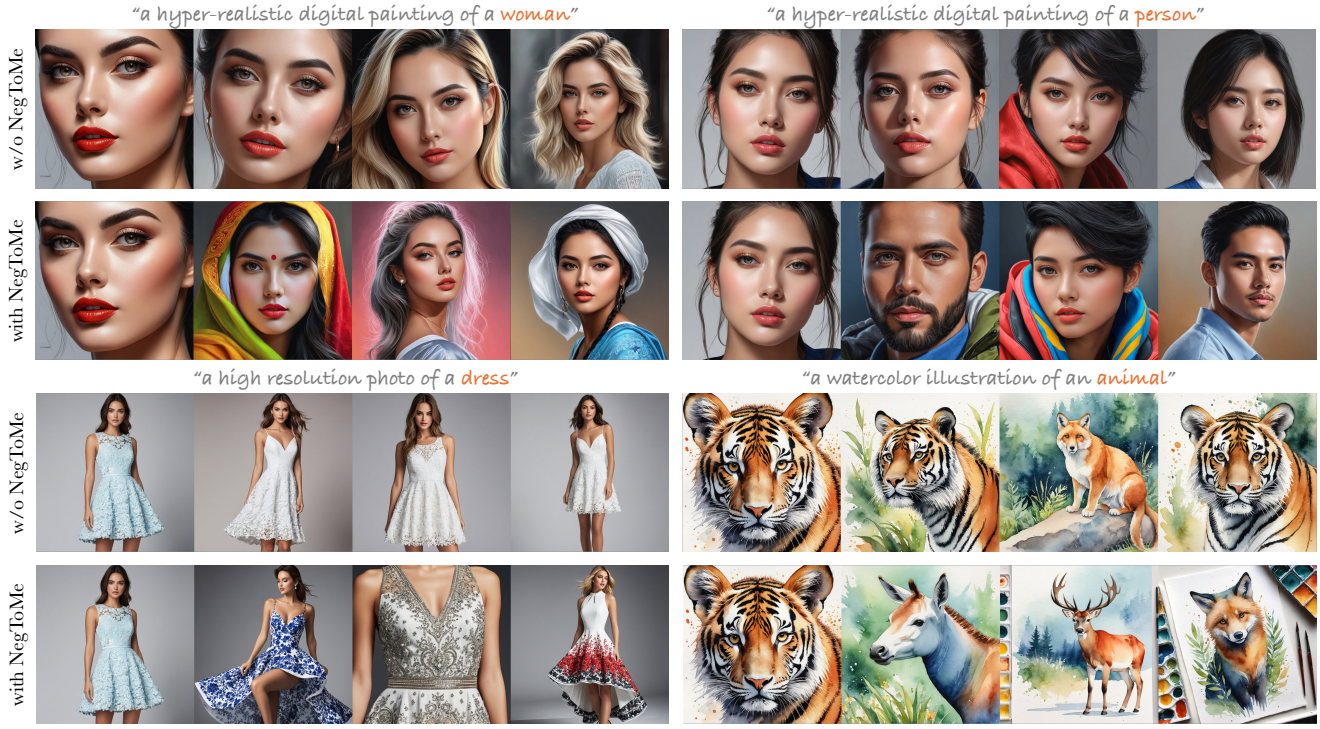
Negative Token Merging: Image-based Adversarial Feature Guidance

Jaskirat Singh[α]* Lindsey Li[β]* Weijia Shi[β]* Ranjay Krishna[βχ] Yejin Choi[β]
Pang Wei Koh[βχ] Michael F. Cohen[β] Stephen Gould[α] Liang Zheng[α] Luke Zettlemoyer[β]
[β]University of Washington [α]Australian National University [χ]Allen Institute for AI



(a) **Adversarial Guidance across Different Outputs:** State-of-the-art diffusion models are observed to suffer from limited diversity (*e.g.*, ethnic, racial, gender *etc.*). NegToMe can be used to improve output diversity by adversarially guiding each image away from each other during reverse diffusion process.

(b) **Adversarial Guidance with Copyrighted Content:** Diffusion models can generate copyrighted content. Moreover, using negative prompt for avoiding this is often insufficient. NegToMe helps better reduce similarity to copyrighted characters, by guiding diffusion features away from copyrighted images.

Figure 1. We introduce NegToMe, a training-free approach for adversarial guidance directly using images instead of a negative prompt. Above we show its applications for a) improving output diversity (visual, gender, racial) by guiding each image away from others, b) reducing visual similarity to copyrighted characters, by guiding outputs away from copyrighted images. (refer Sec. 4 for further applications).

Abstract

Text-based adversarial guidance using a negative prompt has emerged as a widely adopted approach to push the output features away from undesired concepts. While useful, performing adversarial guidance using text alone can be insufficient to capture complex visual concepts and avoid undesired visual elements like copyrighted characters. In this paper, for the first time we explore an alternate modality in this direction by performing adversarial guidance directly using visual features from a reference im-

*age or other images in a batch. In particular, we introduce **neg**ative **to**ken **me**rging (NegToMe), a simple but effective training-free approach which performs adversarial guidance by selectively pushing apart matching semantic features (between reference and output generation) during the reverse diffusion process. When used w.r.t. other images in the same batch, we observe that NegToMe significantly increases output diversity (racial, gender, visual) without sacrificing output image quality. Similarly, when used w.r.t. a reference copyrighted asset, NegToMe helps reduce visual similarity with copyrighted content by 34.57%. NegToMe is simple to implement using just few-lines of code, uses only marginally higher ($< 4\%$) inference times and generalizes to different diffusion architectures like Flux, which do not natively support the use of a separate negative prompt. Code is available at* https://negtome.github.io.

1. Introduction

Large-scale text-to-image (T2I) diffusion models [10, 32, 37, 38, 41, 49] have made unparalleled progress and allow for generation of powerful imagery. Despite these advances, guiding the generation process adversarially to avoid generation of undesired concepts [4] remains a challenging problem. Such guidance is advantageous for several applications such as improving image quality (by guiding away from low-quality features), improving output diversity (by guiding each image away from each other), avoiding undesired concepts such as copyrighted characters (Fig. 1, 2) [17] *etc*.

Existing methods in this direction predominantly rely on the use of negative prompt [3, 21] for adversarial guidance. However, use of negative-prompt *alone* for adversarial guidance suffers from some limitations; capturing complex visual concepts using text-alone can be hard (trying to capture every detail: pose, action, background *etc*. for *child in a park* in Fig. 2). The use of negative-prompt *alone* might be insufficient to remove undesirable visual features (*e.g*., copyrighted characters in Fig. 1). Furthermore, using a separate negative prompt itself may not be feasible when using *state-of-the-art* guidance distilled models like Flux [6].

In this paper, we explore an alternate modality in this direction by performing adversarial guidance using images. Our key intuition is that even if describing the undesired concepts is not effective or feasible as text (*child in park* for Fig. 2), we can directly use the visual features from a reference image in order to adversarially guide the generation process. For instance in Fig. 2, instead of trying to exhaustively describe the *child's* attire, placement, pose, background *etc*., we wish to directly use the visual features from the reference image to guide the generation process. Similarly, in cases where a negative prompt alone is not sufficient (*e.g*., copyrighted characters in Fig. 1), we can better guide the generation away from undesired concepts by directly using the character images for adversarial guidance.

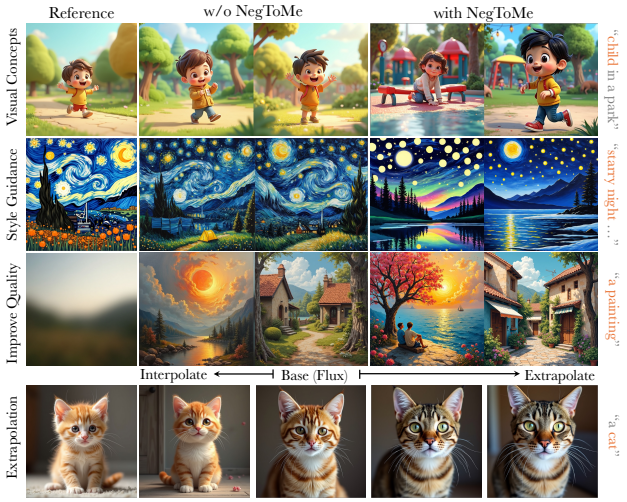


Figure 2. ***Image-based adversarial guidance.*** Simply adjusting the reference allows for a range of custom applications with NegToMe (*e.g*., improve output quality by using a blurry reference).

To this end, we propose negative token merging (NegToMe), a simple and training-free approach which achieves adversarial guidance by selectively pushing apart semantically matching features (between the reference image and generated outputs) during the reverse diffusion process. NegToMe is easy to implement with only a few lines of code and can be integrated into various diffusion architectures, enabling a range of custom applications (see Fig. 2); 1) adversarial guidance for visually complex concepts (*e.g*., *'child in a park'*), 2) style-guidance for excluding specific artistic elements, 3) enhancing output aesthetics using a blurry reference image, and, 4) object feature interpolation or extrapolation *e.g*., between a *kitten* and *young cat* (Fig. 2) to inter-extrapolate visual features for cat age, size *etc*.

In particular, to demonstrate the practical usefulness of the proposed approach, we identify two prominent use-cases of NegToMe; 1) improving output diversity (when performed *w.r.t* other images in same batch), and, 2) reducing visual similarity to copyrighted characters (when performed *w.r.t* a copyrighted RAG database). For instance, it has been empirically shown [29] that state-of-the-art diffusion models often struggle from the problem of limited output diversity (*e.g*., limited racial, gender diversity for a prompt for a *person* in Fig. 1). The use of NegToMe across a batch, inherently helps addresses this problem by pushing visual features of each image away from the each other during reverse diffusion process. Through both qualitative and quantitative results (refer Sec. 4.1) we show that NegToMe helps significantly improve output diversity (racial, gender, ethnic, visual) without requiring any training or finetuning.

Similarly, when performed *w.r.t* to a copyrighted RAG database (Sec. 4.2), we observe that NegToMe complements and improves over traditional negative-prompt based adversarial guidance [17] for reducing visual similarity with copyrighted characters. Experiments reveals that despite its

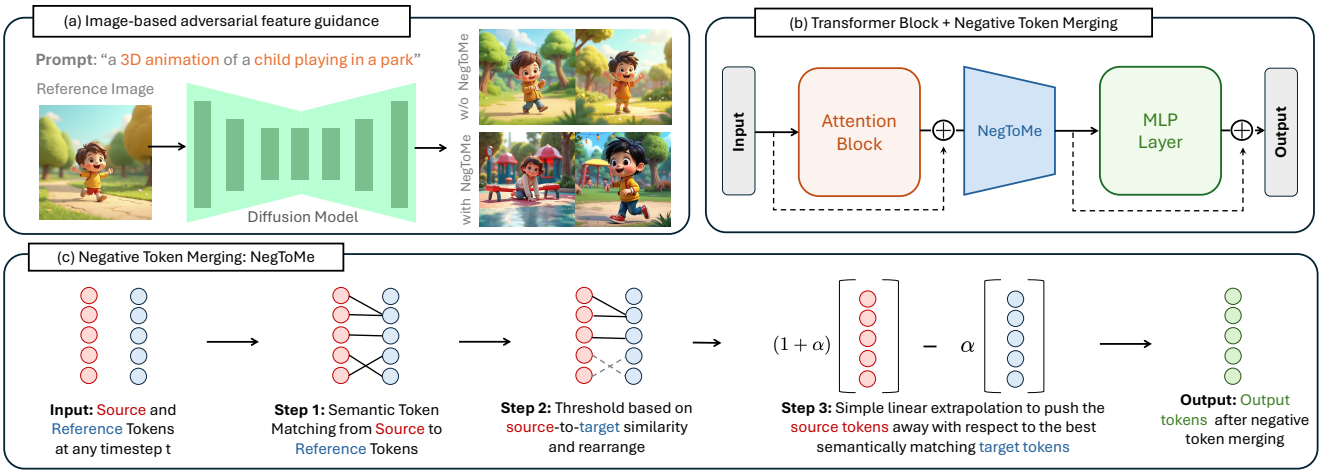


Figure 3. ***Method Overview.*** (a) The core idea of NegToMe is to perform adversarial guidance directly using visual features from a reference image (or other images in the same batch). (b) NegToMe is simple and can be applied in any transformer block. (c) A simple three step process for performing adversarial guidance using NegToMe (refer Sec. 3 and Alg. 1 for the detailed implementation).

simplicity, the proposed approach helps reduce visual similarity to copyrighted content by 34.57% while using only marginally higher ($< 4\%$) inference times (Sec. 4).

2. Related work

Adversarial feature guidance using a negative prompt has been widely explored for a range of applications [2, 4, 17, 21, 47]. While remarkable, as discussed in Sec. 1 we find that the use of a negative-prompt alone might not always be sufficient (Fig. 1, 2) or even sometimes feasible [6]. Our work thus explores a complementary modality by performing adversarial guidance directly using a reference image.

Token merging [7, 8] proposes to increase the throughput of existing ViT [11] models by gradually merging redundant tokens in the transformer blocks. Recent works [27, 48] apply the idea of token merging for video editing in order to better maintain temporal coherence of the edited video. In contrast, we explore cross-frame negative token merging as a mechanism for providing adversarial feature guidance.

Increasing output diversity has been explored to address mode-collapse with diffusion models [5, 29, 34, 50]. The reduced diversity occurs due a several factors including training-data imbalance, classifier-free guidance [21], preference-optimization finetuning [26, 36] *etc*., and is hard to eliminate at pretraining stage. Prior works on addressing this often require costly retraining / finetuning [29]. In contrast, we propose a simple training-free approach which inherently improves semantic diversity of the output features.

Copyright mitigation. The growing concern over copyright infringement by generative models, has attracted significant attention in recent literature [18, 25, 31, 39, 40, 44]. A particularly pressing issue is generation of copyrighted characters by diffusion models [15, 17, 18, 24, 46]. Prior works for addressing these risks typically require expensive finetuning and unlearning [9, 16, 51] to remove copyrighted information from model weights. Our work (NegToMe) thus provides a simple approach for reducing visual similarity to copyrighted content in a training-free manner.

3. Negative Token Merging

Our goal is to insert a negative token merging module into an existing diffusion model [6, 33], in order to perform adversarial feature guidance *w.r.t* other images in a batch (Sec. 4.1) or a real-image input (Sec. 4.2). The core idea of our approach is to perform adversarial guidance by pushing each output token (*source*) away from its best matching token (*target*) in the reference image. The negative token merging module is applied between the attention and MLP branches of each transformer block (Fig. 3). In particular, given the output of the attention block, we first perform cross-image token matching to find the best matching *target* token for each output *source* token. We then apply simple linear extrapolation pushing each *source* token away from its best matching *target* token. Fig. 3 provides an overview.

Semantic Token Matching. A key idea behind NegToMe is to push each output token (*source*) away from its semantically closest token (*target*) in the reference image during the reverse diffusion process. This requires accurate computation of semantic token-to-token correspondences (*e.g*., head of child in Fig. 3) between the generated tokens and the tokens in the reference image. Luckily, we can leverage the rich semantic structure of the intermediate diffusion features [45, 52] itself to compute cross-image token-token similarities using noisy latent features itself.

In particular, given the output $\mathrm{O}_{src} \in \mathbb{R}^{B\times N\times D}$ of an attention block (B is the batch size and N is the number of image tokens), we first compute similarity *w.r.t* the reference image tokens $\mathrm{O}_{ref} \in \mathbb{R}^{1\times N\times D}$ as follows,

$$\mathcal{S}(\mathrm{O}_{src}, \mathrm{O}_{ref}) = \tilde{\mathrm{O}}_{src} \cdot \tilde{\mathrm{O}}_{ref}^{T}; \quad \mathcal{S} \in \mathbb{R}^{B.N\times N}, \tag{1}$$

where $\tilde{\mathrm{O}}_{src}$, $\tilde{\mathrm{O}}_{ref}$ refer to the frame-level normalized source and reference image tokens, respectively. We next

Algorithm 1 NegToMe: Negative Token Merging

```
def NegToMe(x_src, x_ref, alpha, threshold):
    """
    x_src: [B, N, D]
    x_ref: [N, D]
    alpha: float
    threshold: float
    """
    # 1) Normalization
    x_src_norm = F.normalize(x_src, dim=-1)
    x_ref_norm = F.normalize(x_ref, dim=-1)

    # 2) Cosine similarity
    cosine_similarity = x_src_norm @ x_ref_norm.T

    # 3) Find source-to-target match and rearrange
    max_similarity, argmax_indices =
         cosine_similarity.max(dim=-1)
    x_target = x_ref[argmax_indices]

    # 4) Threshold and merge
    threshold_mask = max_similarity > threshold
    x_merge = torch.where(
        threshold_mask.unsqueeze(-1),
        (1 + alpha) * x_src - alpha * x_target,
        x_src)
    return x_merge
```

use the similarity matrix $\mathcal{S} \in \mathbb{R}^{B.N \times N}$ in order to compute the best matching target token for each source token as,

$$\mathrm{O}_{target} = \mathrm{O}_{ref}\,[\mathrm{argmax}\,\{\mathcal{S}(\mathrm{O}_{src}, \mathrm{O}_{ref})\}] \quad (2)$$

$$\mathrm{O}_{target} = \mathrm{H} \odot \mathrm{O}_{target} + (1 - \mathrm{H}) \odot \mathrm{O}_{src}, \quad (3)$$

where $\mathrm{H} = \mathbb{1}\,[\max\{\{\mathcal{S}(\mathrm{O}_{src}, \mathrm{O}_{ref})\} > \tau]$ helps ensure that *source-tokens* with *source-to-target* token similarity below a threshold τ (*i.e.*, no good semantic match is available) are not modified during negative token merging.

Source-to-Target Token Extrapolation. Given the semantically-matched *target* token matrix for the *source* tokens O_{src}, we next perform a simple linear extrapolation between the *source* and *target* tokens as,

$$\mathrm{O}_{merge} = (1 + \alpha_t)\,\mathrm{O}_{src} - \alpha_t\,\mathrm{O}_{target}, \quad (4)$$

where α_t is a time-dependent affine-coefficient, which helps control the degree to which the *source* and *target* tokens are pushed apart during the reverse diffusion process.

Masked Adversarial Guidance. While performing adversarial guidance *w.r.t* the entire reference image is useful (*e.g.*, increasing diversity), we may also wish to perform adversarial guidance only *w.r.t* to certain parts of the provided reference. For instance, when performing copyright mitigation (Sec. 4.2), we may wish to reduce visual similarity only with respect to the copyrighted character without being affected by background noise or features. Similarly, masked guidance may also be useful in order to perform adversarial guidance with specific subparts (*e.g.*, red hat, mustache, pony tail) of the provided reference image.

In particular, given an additional mask binary input M_{ref} for the reference image, we can perform masked adversarial guidance by simply introducing a bias-term in the source-to-target similarity $\mathcal{S} \in \mathbb{R}^{B.N \times N}$ computation as,

$$\mathcal{S}(\mathrm{O}_{src}, \mathrm{O}_{ref}) = \tilde{\mathrm{O}}_{src} \cdot \tilde{\mathrm{O}}_{ref}^{T} + \log(\tilde{M}_{ref} + \epsilon), \quad (5)$$

where $\epsilon = 10^{-6}$ and $\tilde{M}_{ref} \in \mathbb{R}^{1 \times N}$ is original mask M_{ref} resized and flattened to match the corresponding sequence length N for attention block output $\mathrm{O}_{src} \in \mathbb{R}^{B \times N \times D}$.

Application to MM-DiT Models. A key advantage of the proposed approach is that it is also easily extendable to guidance-distilled MM-DiT architectures such as Flux [6], which do not natively support the use of a separate negative prompt. In particular, given an output $\mathrm{O}_{joint} \leftarrow (\mathrm{O}_{text}, \mathrm{O}_{img})$ of the joint-attention block [6, 13], NegToMe can be easily applied as follows,

$$\mathrm{O}_{joint} \leftarrow \{\mathrm{O}_{text} \oplus f_{neg}(\mathrm{O}_{img}, \mathrm{O}_{ref}, \alpha_t, \tau)\}, \quad (6)$$

where $f_{neg}(.)$ is NegToMe function from Alg. 1, and $\oplus$ is the matrix concatenation operation along sequence length.

4. Experiments

In this section, we demonstrate the practical usefulness of our approach for two prominent applications of NegToMe; increasing output diversity (Sec. 4.1) and reducing visual similarities to copyrighted characters (Sec. 4.2). We further showcase more general applications of NegToMe in Sec. 5.

4.1. Increasing Output Diversity

We evaluate the performance of our approach for increasing output diversity when performing negative token merging *w.r.t* to other images in the batch. To facilitate easy visual comparison, we perform negative token merging *w.r.t* the first image in each batch. Unless otherwise specified, all results are reported using the same text-to-image generation base-model [6, 33] on a single Nvidia-H100 GPU.

Dataset and Setup. We first construct an input prompt dataset comprising 20 general object categories (*e.g.*, animal, woman, bird, car *etc.*) across 7 different prompt templates (*e.g.*, "a photo of a", "a high-contrast image of a")[1]. For each category, we sample 280 images with 10 random seeds (4 per batch) both with and without NegToMe. The real images for FID [20] calculation are sourced from LAION-Aesthetics-v2 6+ dataset [43], where we use CLIP [35] to retrieve the top-1K images for each category.

Evaluation metrics. Following [21], we report the results for output-quality using 1) FID [20] and 2) Inception Score (IS) [42]. 3) Pairwise dreamsim-score [14]: is used to measure output feature diversity. 4) VQAScore [28] and 5) CLIPScore [19] are used to evaluate image-text alignment. Furthermore following [12], we also use 6) Entropy-Score which measures the degree to which outputs for a particular object category (*e.g.*, person) are spread across its subcategories (racial, gender, ethnic *etc.*). For human images, we use the FairFace classifier [22] to detect race, gender, and

[1] We adopt CLIP prompt templates [35] for CIFAR10 image classification, excluding ones that imply low-quality (*e.g.*, *"a blurry photo of a"*)

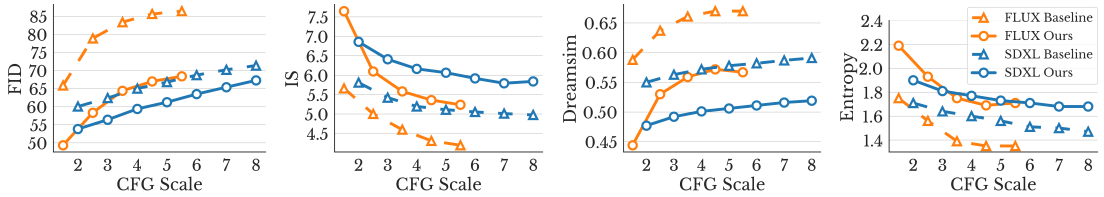


Figure 4. ***Quantitative Results for Output Diversity***. NegToMe (ours) helps improve output diversity (lower DreamSim score and higher Entropy) while preserving or improving quality (lower FID and higher IS) across different CFG scales for both SDXL and FLUX.

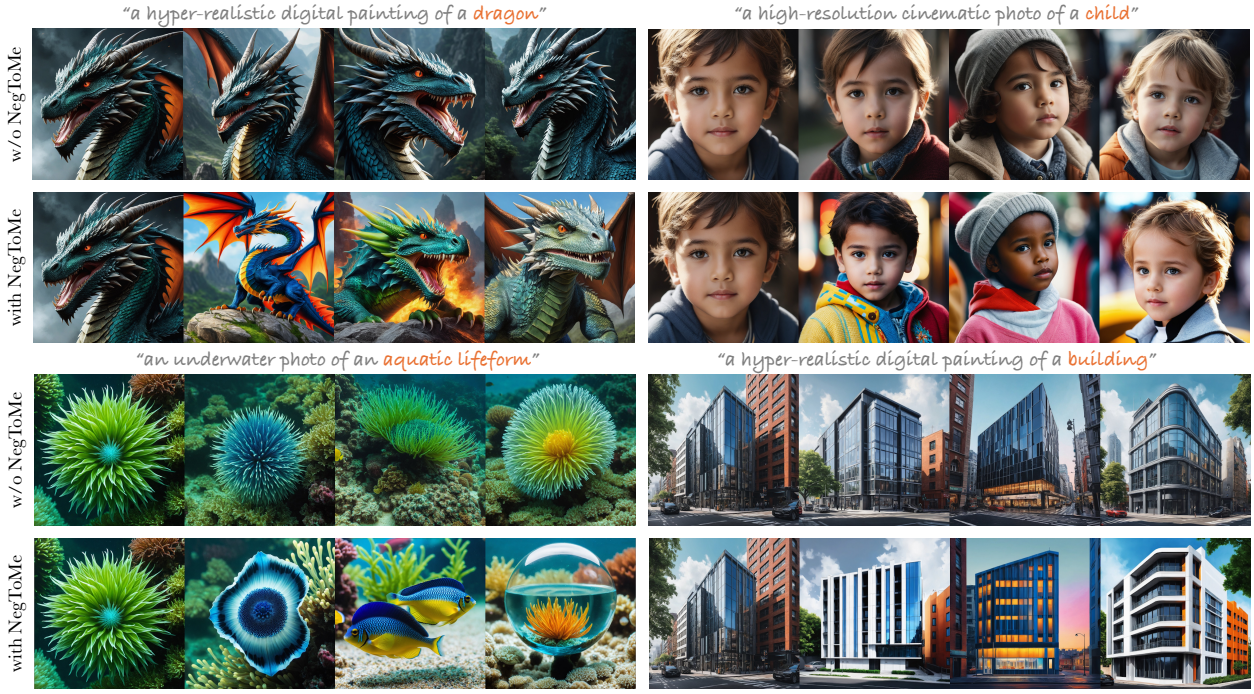


Figure 5. ***Increasing Output Diversity.*** We observe that when performed *w.r.t* to images in the same batch (the first image of each batch in above), NegToMe significantly improves output diversity (racial, ethnic, visual) while preserving output image quality.

age. For non-human categories (*e.g.*, bird), sub-categories are extracted via WordNet [30] and classified using CLIP.

Quantitative Results. Results are shown in Figure 4. We observe that NegToMe helps improve output diversity (*i.e.* lower Dreamsim scores and higher Entropy) while preserving or even improving quality (*i.e.* lower FID and higher IS) across different classifier-free guidance (cfg) scales [21] for both SDXL and FLUX. We also perform human evaluation evaluating diversity, quality and prompt alignment of the proposed approach using actual human users (Fig. 8). Similar to automated metric evaluation, we observe that NegToMe helps improve output diversity without sacrificing output image quality and prompt alignment performance.

Qualitative Results. We visualize the outputs with and without NegToMe for the base model SDXL (Fig. 1 and 5) and FLUX (Fig. 9). To avoid cherrypicking, all visualizations are from a fixed seed 0. We observe that the base-model often suffers from the problem of mode-collapse with the generated images exhibiting limited visual feature diversity (*e.g.*, visually similar dragons/buildings in Fig. 5, cats in Fig. 9). Also the outputs might simply collapse into the

Method	Dreamsim ↓	CLIPScore ↑	IS ↑	Inf. Time ↓
Base Prompt	0.812	0.334	3.197	13.2 s
Base Prompt + Ours	0.780	**0.336**	3.355	13.7 s
PW (`gpt-4o`)	0.743	0.332	3.686	15.4 s
PW + Ours	**0.712**	0.333	**3.747**	15.9 s

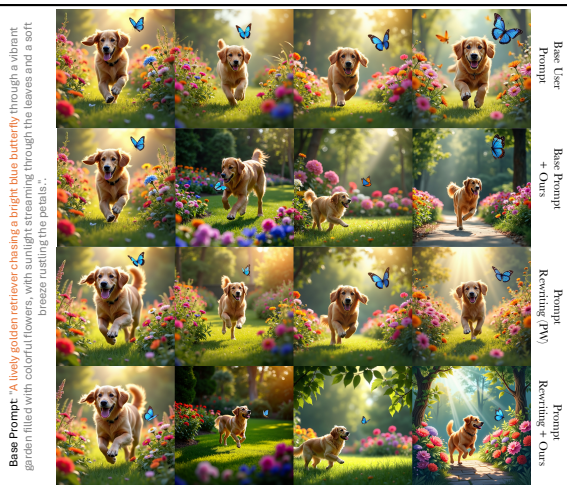


Figure 6. NegToMe helps improve output diversity both with (row-2) and without explicit prompt-rewriting (PW) (row-4).

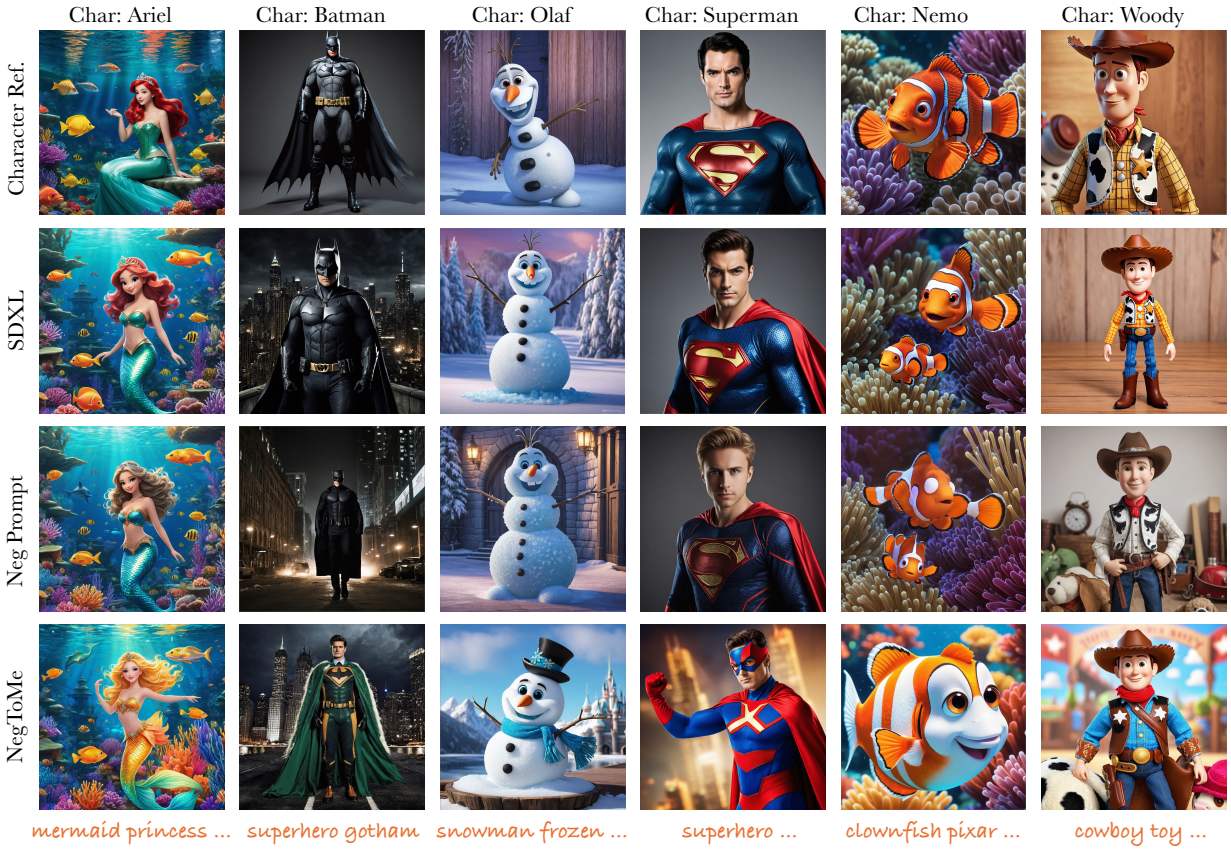


Figure 7. ***Copyright Mitigation.*** When used *w.r.t* a copyright RAG image dataset, NegToMe helps reduce visual similarities with copyrighted characters without sacrificing output image quality (Tab. 1). Complete prompts and further results are provided in the appendix.

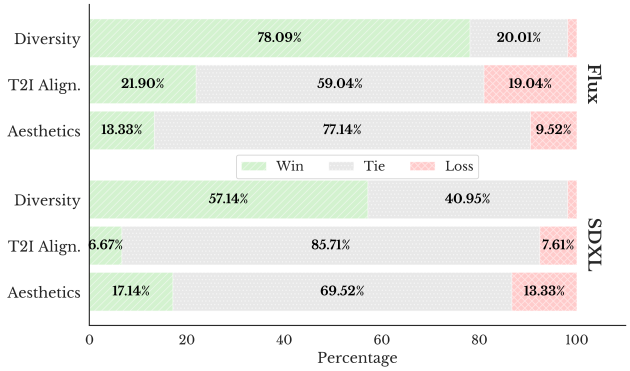


Figure 8. ***Human User Study.*** NegToMe helps improve output diversity while preserving text-to-image alignment performance.

same subcategory, even when the prompt is about a general category (*e.g.*, aquatic lifeform in Fig. 5). In contrast, despite its simplicity, NegToMe is able to better harness the underlying diffusion prior in order to generate diverse images in terms of demographics (*e.g.*, person in Fig. 1), subcategory (*e.g.*, aquatic lifeform in Fig. 5), viewpoints (*e.g.*, dragon in Fig. 5), image layout, pose, visual appearances of both foreground and background (*e.g.*, child in Fig. 9) *etc*.

Output diversity with explicit prompt-rewriting. A key advantage of NegToMe is that it helps improve diversity without the need for extensive prompt-rewriting, which also

Mitigation Strategy		Evaluation Metrics			
NegPrompt	NegToMe	Dreamsim ↓	VQAScore ↑	CLIPScore ↑	IS ↑
✗	✗	0.766	**0.913**	0.344	3.431
✓	✗	0.684	0.876	0.339	3.790
✗	✓	0.703	0.906	**0.346**	3.678
✓	✓	**0.638**	0.879	0.339	**3.864**

Table 1. ***Copyright Mitigation***. NegToMe reduces visual similarity to copyright characters while preserving T2I performance.

presents as a feasible yet expensive (time & memory) approach for improving diversity. This is particularly relevant when the user-prompts are quite detailed and long. To provide a fair comparison with prompt-rewriting setting, we therefore first curate a set of 20 detailed prompts across diverse settings (see appendix). For each prompt we then use a large-language model [1] in order to generate diverse variations of the original base prompt. The final images for both base-prompt and rewritten prompts are sampled across 10 random seeds. Results are shown in Fig. 6. While prompt-rewriting helps improve diversity, it comes at the cost of increased inference time. Furthermore, some of the generated outputs might still appear similar (*e.g.*, col-1 and col-3: Fig. 6). In contrast, NegToMe can adaptively improve the output diversity (with both base and rewritten prompts), while on average using only $< 4\%$ higher inferences times.

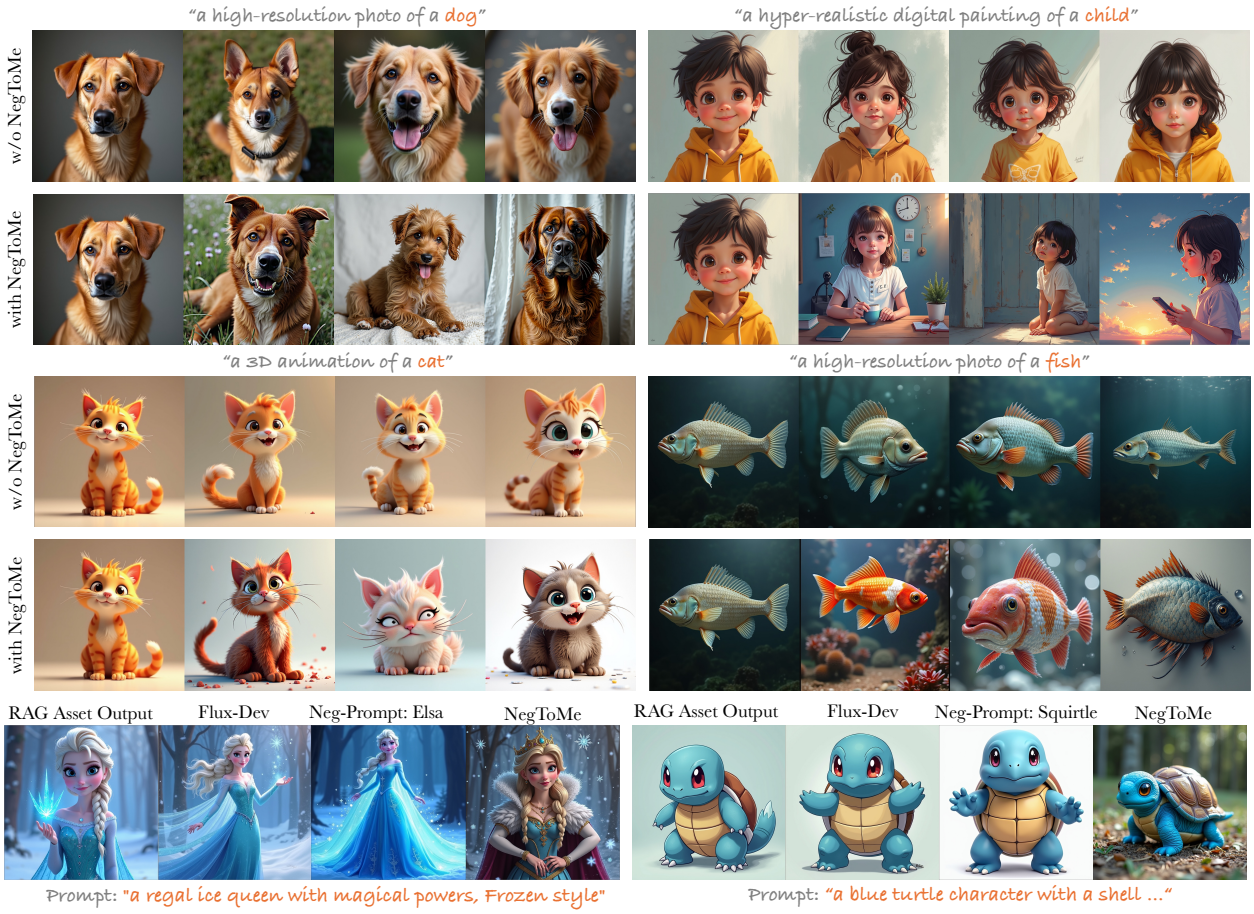


Figure 9. ***Application to MM-DiT models (Flux).*** NegToMe is model-agnostic and also applicable to MM-DiT models like Flux [6]. NegToMe significantly increases the output diversity (*top*), and helps reduce copyright violation (*bottom*).

4.2. Copyright Mitigation

We next show the efficacy of our approach for reducing visual similarities with copyrighted characters when performing NegToMe *w.r.t* a copyrighted image RAG database.

Dataset and Setup. We first construct a dataset of 50 copyrighted characters (*e.g.*, Mario, Elsa, Batman), and curate input prompts which to trigger these characters without explicitly mentioning their names (see appendix). For each character, we compile a reference dataset of approximately 30 high-quality images depicting the character in diverse settings. Masked negative token merging is then performed for each prompt, using the best-matching RAG asset (asset with highest Dreamsim score) from the reference dataset. The mask for each asset is computed using HQ-SAM [23].

Qualitative Results. Results are shown in Fig. 7 (SDXL) and Fig. 9 (Flux). We observe that the base model still generates copyrighted characters, even when the corresponding character name is not mentioned in the input prompt. Using the character name as the negative prompt alone often is not sufficient, as the output images still show high visual similarity to the copyrighted character. In contrast, by applying adversarial guidance directly using character visual features, NegToMe reduces similarity to copyrighted characters while maintaining text-to-image alignment.

Method	AES ↑	VQAScore ↑	CLIPScore ↑	Human Pref. ↑
Flux-Dev	6.428	**0.866**	0.320	22.5 %
Flux-Dev + Ours	**6.604**	0.861	**0.322**	**77.5** %

Figure 10. ***Improving aesthetics.*** Using a blurry reference with NegToMe improves output aesthetics without any training [36].

Quantitative Results. We evaluate our approach using the base model SDXL, comparing it to pure negative-prompt copyright mitigation strategies. For each prompt, we sample 50 images with 50 different random seeds. For evaluation, we use the 1) maximum DreamSim score [14] across all RAG assets (excluding the reference used for NegToMe) for measuring visual similarity to copyrighted characters. 2) VQAScore [28], CLIPScore [19] for text-to-image align-

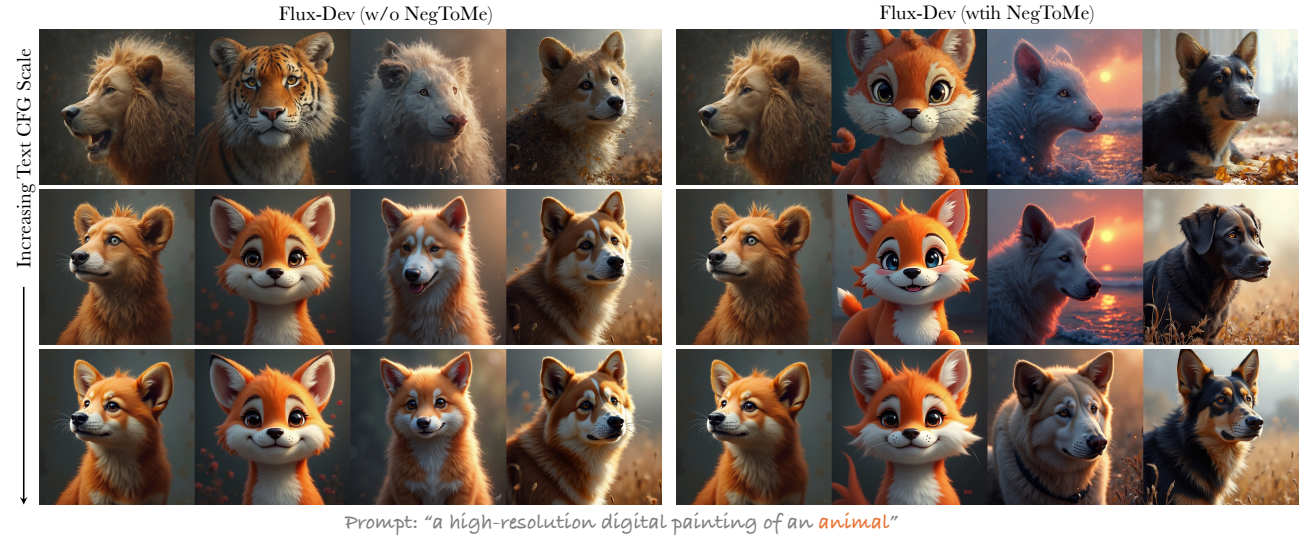


Figure 11. ***Variation with cfg scale*** leads to improved output quality at the cost reduced diversity (*left*). NegToMe not only improves output quality at lower *cfg* values (by guiding away from poor-quality features, see Fig. 10) but also helps improve output diversity for higher *cfg*.

Figure 12. ***Variation with merging alpha.*** Increasing the value of α (refer to Sec. 3) for NegToMe gradually increases output diversity in terms of gender, race, ethnicity, lighting, style *etc*.

ment, 3) IS [20] for image quality. Results are shown in Tab. 1. We observe that NegToMe reduces visual similarity to copyrighted characters without sacrificing text-to-image alignment and image quality. Furthermore, NegToMe is complementary to negative prompting, with the best performance achieved when combining both methods.

5. Method Analysis and Applications

Improving output aesthetics. As noted in Fig. 2, we note that NegToMe allows for a range of custom applications by appropriately adjusting the reference inputs. Notably, we find that when using a poor quality image as reference, NegToMe helps improve output aesthetics and image quality without requiring any training / finetuning [36] (Fig. 10).

Variation across text-guidance scale. Results are shown in Fig. 11. We observe that traditional text-based classifier-free guidance [21] suffers from a tradeoff between output diversity and image quality. In contrast, we find that NegToMe is able to improve output diversity while preserving image quality across different scales of classifier-free guidance. Interestingly, we also observe that the increase in output diversity with NegToMe is often accompanied by an increase in output image quality especially at lower *cfg* values. This happens due to the use of a poor quality reference image (*e.g.*, *lion*: Fig. 11) at lower *cfg* scales, which in addition to improving diversity also tends to improve output image aesthetics and details (*lamb, dog* in row-1: Fig. 11).

Variation with merging alpha. Results are shown in Fig. 12. We observe that NegToMe provides an easy to use mechanism for controlling output image diversity. As seen in Fig. 12, we observe that gradually increasing the value of α (Sec. 3) helps the user easily control output diversity in terms of race, gender, ethnicity, lighting, style *etc*.

6. Conclusion

In this paper we introduce NegToMe, a simple training-free approach which complements traditional text-based negative-prompt guidance, by performing adversarial guidance directly using visual features of a reference image. NegToMe is simple, training-free and can be incorporated with most *state-of-art* diffusion models using just few lines of code (Alg. 1). By simply varying the reference image, NegToMe enables a range of custom applications such as increasing output diversity (Sec. 4.1), reducing simi-

larity with copyrighted images (Sec. 4.2), improving output aesthetics (Fig. 10) *etc*., while on average using only marginally $< 4\%$ higher inference times. We excitedly hope that our research helps users better leverage *state-of-the-art* diffusion models for diverse creative applications.

Acknowledgments

We would like to thank Ishan Misra for helpful discussions and feedback on experiment design and quantitative evaluations. We are also thankful to Jonas Kohler and Junshen Chen for early discussions on negative token merging.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 6

[2] Andrew. How to use negative prompts?, 2023. 3

[3] Mohammadreza Armandpour, Ali Sadeghian, Huangjie Zheng, Amir Sadeghian, and Mingyuan Zhou. Re-imagine the negative prompt algorithm: Transform 2d diffusion into 3d, alleviate janus problem and beyond. *arXiv preprint arXiv:2304.04968*, 2023. 2

[4] Yuanhao Ban, Ruochen Wang, Tianyi Zhou, Minhao Cheng, Boqing Gong, and Cho-Jui Hsieh. Understanding the impact of negative prompts: When and how do they take effect? *arXiv preprint arXiv:2406.02965*, 2024. 2, 3

[5] Hritik Bansal, Da Yin, Masoud Monajatipoor, and Kai-Wei Chang. How well can text-to-image generative models understand ethical natural language interventions? *arXiv preprint arXiv:2210.15230*, 2022. 3

[6] Andreas Blattmann, Axel Sauer, Dominik Lorenz, Dustin Podell, Frederic Boesel, Harry Saini, Jonas Müller, Kyle Lacey, Patrick Esser, Robin Rombach, Sumith Kulal, Tim Dockhorn, Yam Levi, and Zion English. Scaling rectified flow transformers for high-resolution image synthesis. https://github.com/black-forest-labs/flux, 2024. Accessed: 2024-09-12. 2, 3, 4, 7

[7] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022. 3

[8] Daniel Bolya and Judy Hoffman. Token merging for fast stable diffusion. *CVPR Workshop on Efficient Deep Learning for Computer Vision*, 2023. 3

[9] 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 Transactions on Graphics (TOG)*, 42(4):1–10, 2023. 3

[10] Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, et al. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv preprint arXiv:2309.15807*, 2023. 2

[11] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[12] Ahmed Elgammal. Can: Creative adversarial networks, generating "art" by learning about styles and deviating from style norms. *arXiv preprint arXiv:1706.07068*, 6:2017, 2017. 4

[13] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 4

[14] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023. 4, 7

[15] Aditya Golatkar, Alessandro Achille, Luca Zancato, Yu-Xiang Wang, Ashwin Swaminathan, and Stefano Soatto. CPR: Retrieval Augmented Generation for Copyright Protection. In *CVPR*, 2024. 3

[16] Chao Gong, Kai Chen, Zhipeng Wei, Jingjing Chen, and Yu-Gang Jiang. Reliable and Efficient Concept Erasure of Text-to-Image Diffusion Models, 2024. 3

[17] Luxi He, Yangsibo Huang, Weijia Shi, Tinghao Xie, Haotian Liu, Yue Wang, Luke Zettlemoyer, Chiyuan Zhang, Danqi Chen, and Peter Henderson. Fantastic copyrighted beasts and how (not) to generate them. *arXiv preprint arXiv:2406.14526*, 2024. 2, 3

[18] Peter Henderson, Xuechen Li, Dan Jurafsky, Tatsunori Hashimoto, Mark A. Lemley, and Percy Liang. Foundation Models and Fair Use. *ArXiv*, abs/2303.15715, 2023. 3

[19] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 4, 7

[20] 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. *Advances in neural information processing systems*, 30, 2017. 4, 8

[21] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2, 3, 4, 5, 8

[22] Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age. *ArXiv*, abs/1908.04913, 2019. 4

[23] Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan Liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Segment anything in high quality. In *NeurIPS*, 2023. 7

[24] Katherine Lee, A Feder Cooper, and James Grimmelmann. Talkin"bout ai generation: Copyright and the generative-ai supply chain. *arXiv preprint arXiv:2309.08133*, 2023. 3

[25] Katherine Lee, A. Feder Cooper, and James Grimmelmann. Talkin' 'Bout AI Generation: Copyright and the Generative-AI Supply Chain, 2024. 3

[26] 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. *arXiv preprint arXiv:2302.12192*, 2023. 3

[27] Xirui Li, Chao Ma, Xiaokang Yang, and Ming-Hsuan Yang. Vidtome: Video token merging for zero-shot video editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7486–7495, 2024. 3

[28] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. *arXiv preprint arXiv:2404.01291*, 2024. 4, 7

[29] Zichen Miao, Jiang Wang, Ze Wang, Zhengyuan Yang, Lijuan Wang, Qiang Qiu, and Zicheng Liu. Training diffusion models towards diverse image generation with reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10844–10853, 2024. 2, 3

[30] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 5

[31] Sewon Min, Suchin Gururangan, Eric Wallace, Weijia Shi, Hannaneh Hajishirzi, Noah A Smith, and Luke Zettlemoyer. SILO Language Models: Isolating Legal Risk In a Nonparametric Datastore. In *ICLR*, 2023. 3

[32] 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. *arXiv preprint arXiv:2112.10741*, 2021. 2

[33] 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. *arXiv preprint arXiv:2307.01952*, 2023. 3, 4

[34] Yiming Qin, Huangjie Zheng, Jiangchao Yao, Mingyuan Zhou, and Ya Zhang. Class-balancing diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18434–18443, 2023. 3

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4

[36] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024. 3, 7, 8

[37] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 2

[39] Matthew Sag. The new legal landscape for text mining and machine learning. *J. Copyright Soc'y USA*, 66:291, 2018. 3

[40] Matthew Sag. Copyright safety for generative ai. *Forthcoming in the Houston Law Review*, 2023. 3

[41] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 2

[42] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 4

[43] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 4

[44] Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting Pretraining Data from Large Language Models. In *The Twelfth International Conference on Learning Representations*, 2024. 3

[45] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems*, 36:1363–1389, 2023. 3

[46] Boyi Wei, Weijia Shi, Yangsibo Huang, Noah A. Smith, Chiyuan Zhang, Luke Zettlemoyer, Kai Li, and Peter Henderson. Evaluating copyright takedown methods for language models, 2024. 3

[47] Max Woolf. Stable diffusion 2.0 and the importance of negative prompts for good results, 2023. Accessed: [insert date]. 3

[48] Bichen Wu, Ching-Yao Chuang, Xiaoyan Wang, Yichen Jia, Kapil Krishnakumar, Tong Xiao, Feng Liang, Licheng Yu, and Peter Vajda. Fairy: Fast parallelized instruction-guided video-to-video synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8261–8270, 2024. 3

[49] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022. 2

[50] Cheng Zhang, Xuanbai Chen, Siqi Chai, Chen Henry Wu, Dmitry Lagun, Thabo Beeler, and Fernando De la Torre. Itigen: Inclusive text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3969–3980, 2023. 3

[51] Gong Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1755–1764, 2024. 3

[52] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *Advances in Neural Information Processing Systems*, 36, 2024. 3