

Early Estimation of Language to Latent Alignment in Diffusion Models

Vasco Ramos¹, Regev Cohen², Idan Szpektor², and Joao Magalhaes¹

¹ NOVALINCS, NOVA University of Lisbon, Portugal

² Google Research

vcc.ramos@campus.fct.unl.pt

<https://novasearch.github.io/noisyclip>

Abstract. Conditional diffusion models frequently suffer from language-image misalignments. Due to the ambiguity of intermediate noise corrupted latents, assessing prompt adherence currently requires completing the entire sampling trajectory. This late-stage evaluation incurs even higher computational costs during test-time scaling strategies, such as Best-of-N (BoN) sampling, as all misaligned trajectories must finish generation before being discarded. To tackle this, we propose NoisyCLIP, a noise-aware twin-tower model that enables early language-to-latent alignment estimation. By learning a vision encoder on noise-corrupted latents, we allow the model to "see" through the ambiguity of intermediate diffusion steps. To facilitate this training, we investigate noise-data augmentation sampling strategies and introduce two new benchmark datasets: Noisy-Conceptual-Captions and Noisy-GenAI-Bench. When applied as an early-stopping criterion for BoN, NoisyCLIP at half cost matches or beats frozen CLIP at full cost. Ultimately, this transforms alignment assessment from an expensive final check into a continuous monitoring tool, drastically reducing compute costs without sacrificing semantic fidelity.

Keywords: Image Generation · Text-Image Alignment · Best-of-N

1 Introduction

The advances in image generation achieved by diffusion models [2, 15, 32, 37, 39] have opened exciting possibilities for generating realistic images from textual prompts. Reverse diffusion, coupled with cross-attention mechanisms, guides iterative removal of noise, shaping a final image that adheres to the input textual prompt. However, reverse diffusion image generation is a costly process, and there is no inherent guarantee that it will follow the optimal denoising trajectory [1], often failing to produce an image that fully adheres to the prompt.

Current methodologies for measuring alignment between textual prompts and images mainly focus on assessing alignment of the output image [14, 21, 24]. Nonetheless, the iterative nature of diffusion models raises a fundamental challenge: **accurately measuring language-to-latent alignment across different intermediate diffusion timesteps**. The core difficulty in providing real-time feedback lies in the nature of the latent space itself during the early stages

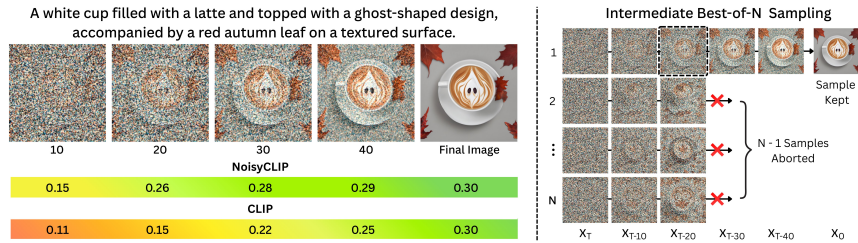


Fig. 1: Left: While standard CLIP scores require near-complete generation to stabilize, NoisyCLIP accurately predicts final image quality at early denoising steps. **Right:** We leverage this early evaluation for intermediate Best-of-N sampling. Evaluating multiple candidates early (e.g., X_{T-20}) terminates unpromising trajectories, significantly reducing compute overhead while allowing the highest-scoring candidate to reach X_0 .

of generation, where representations are dominated by Gaussian noise, making traditional vision-language encoders ineffective [36]. To address this, we propose NoisyCLIP, a noise-aware twin-tower model designed to actively measure language-to-latent alignment during the denoising process, Figure 1. By learning a vision encoder over noise-corrupted representations, we equip the model to accurately interpret intermediate diffusion stages, transforming alignment assessment from a final check into a continuous monitoring tool.

Learning such language-to-latent estimator is a non-trivial task. Although noise inversion techniques exist [8, 28], there are no obvious data resources or sampling strategies that can reliably provide the intermediate noise training sequences. This leads to our second core challenge: **identifying the training objectives and data augmentation strategies necessary to stabilize a latent-alignment estimator**. To address this critical gap, we investigate noise-data augmentation sampling strategies for training and assessing alignment of intermediate latent representations in reverse diffusion processes.

Finally, to validate the system-level utility of our approach, we address a final challenge: **establishing an alignment-based early-stopping criterion that significantly reduces Best-of-N computational costs while maintaining full-term image-text fidelity**. Evaluating solely on the final image results in wasted compute time since misaligned generations must complete the full process before being discarded. This inefficiency is amplified in test-time scaling paradigms such as Best-of-N (BoN) selection [23, 26, 45], where having multiple images incurs a substantial collective computational cost.

Our main contributions are threefold: we propose NoisyCLIP, a novel noise-aware architecture capable of measuring language-to-latent alignment during the highly ambiguous intermediate stages of diffusion; we introduce two new benchmark datasets, Noisy-Conceptual-Captions and Noisy-GenAI-Bench, containing full sequences of intermediate latents to evaluate these alignments; and we pioneer an early-stopping Best-of-N strategy that cuts computational costs by 50% while retaining 98% of post-hoc CLIP performance across multiple architectures.

2 Related Work

Text-to-image Generation. These models are trained to learn associations between linguistic and visual concepts [48]. Despite previous efforts [13, 20], reverse diffusion models, including DALL-E 2 [35], Imagen [39], and Stable Diffusion [38], have emerged as the predominant approach [2, 10, 30, 32, 38]. Although early diffusion models [15] lacked control over the generation, advances in controllability and quality have been driven by the introduction of textual prompts [19, 38] and further refined through the use of multiple text encoders [32–34].

Latent Space. Early diffusion models were constrained by the computationally demanding pixel space [15]. To overcome this inefficiency, newer approaches employ a two-stage process: an autoencoder first maps high-dimensional images into a lower-dimensional latent space, which is then used for the more efficient reverse diffusion process. This strategy significantly reduces computational demands, memory, and training time [38, 40]. Consequently, researchers have increasingly focused on understanding and manipulating this latent space, recognizing that working within it is key to substantially improving the quality of the output. [6, 29, 36, 47]. We take advantage of this latent space, using the intermediate generations to measure the alignment of the prompt and the image being generated. This allows us to understand the generation’s trajectory and predict the final alignment of the result.

Alignment Metrics. Alignment between the generated image and the text input is a key measure to assess the quality of the generative model. Dual encoders such as CLIP [14, 33] and SigLIP [51] provided an early solution by directly measuring the cosine similarity between the image and text representations. More recently, approaches that leverage language models [34] have achieved state-of-the-art results in measuring image-text similarity, demonstrating a high degree of agreement with human judgments [24]. Moreover, utilizing Vision Language Models for evaluation improves generation performance by incorporating models feedback into the process [53]. However, this approach requires a verifier model that may be larger than the generation model, potentially slowing the generation, and has so far only been evaluated on autoregressive models [46], excluding more common diffusion models [38]. To mitigate this, recent work [3] employs a CLIP model to directly receive latent information from the latest diffusion step, bypassing the need to process the full image through the decoder. Additionally, another work [28] presents similar approaches to use latent information to determine the optimal classifier free guidance during the reverse diffusion process.

While current metrics successfully assess final image similarity, there’s a critical need for efficient, in-process assessment during reverse diffusion. Existing latent alignment efforts either require full pipeline processing [3] or simply treat alignment as a means to an end [28]. These studies lack a dedicated analysis, open-source models, and standardized comparison data. Our work fills this gap by proposing a novel method and open dataset that leverage intermediate generation data to assess semantic alignment throughout the diffusion latent space.

3 Aligning Language with Noisy Latents

The core challenge of our work is to measure the alignment between the source text and the ambiguous and noisy latent representations generated during reverse diffusion. To address this, we propose a solution based on twin-tower similarity models, such as CLIP, to measure language-to-latent alignment as a similarity problem. This section starts by formalizing the language-to-latent alignment problem (Section 3.1). We then explain how to mitigate the limited availability of noise-corrupted data (Section 3.2), required to train NoisyCLIP as described in Section 3.3. Following this, we propose a new benchmark to evaluate language-to-latent assessment methods (Section 3.4).

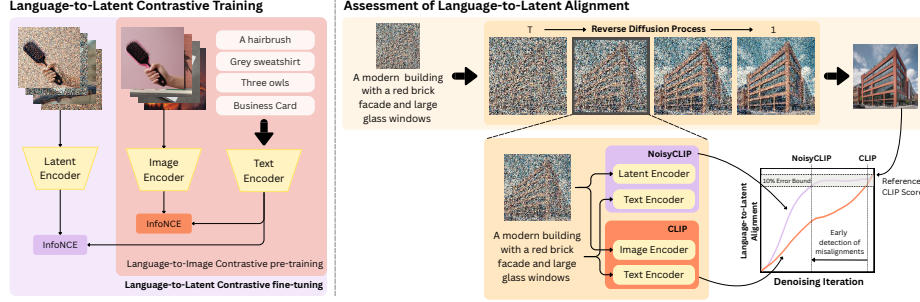


Fig. 2: A prompt is used to generate an image, and at an intermediate step, the generated image and the original prompt are encoded and their similarity is measured. As demonstrated, our method generates a similarity score that is closely aligned with the score assessed for the final produced image, while at intermediate stages of generation, allowing for the early identification of misalignments.

3.1 Language-to-Latent Alignment Problem

In image generation, earlier techniques such as Classifier-Free Guidance [16] enhance the output quality of diffusion models, but these approaches inherently sacrifice in-process feedback. This limitation leads to the adoption of post-hoc language-image alignment methods, which evaluate the final image after generation, introducing computational inefficiencies and hindering iterative control over the generation process [24, 26]. Despite Latent Diffusion Models [38] grounding generation in textual prompts, there remains no mechanism to leverage this information for real-time feedback during generation. In this section, we explore how to enable direct language-to-latent alignment evaluation during generation, eliminating post-hoc assessment and offering a more scalable and responsive approach to image synthesis. We begin by recalling the objective function of the Latent Diffusion Model:

$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x), y, \epsilon \sim \mathcal{N}(0,1), t} \left[\|\epsilon - \epsilon_{\theta}(z_t, t, \tau_{\theta}(y))\|_2^2 \right] \quad (1)$$

In this formulation, x is the original image, and $z = \mathcal{E}(x)$ is its latent representation generated by the encoder $\mathcal{E}$. z_t represents the noisy latent state after t diffusion steps, where t ranges from 1 to T . ϵ is the Gaussian noise used to corrupt the latent state, and ϵ_θ is the noise prediction model trained to estimate the noise ϵ added to z_t . The reverse diffusion steps are conditioned by the input text y , encoded by the text encoder $\tau_\theta(y)$. The text embedding $\tau_\theta(y)$ is then used in cross-attention mechanisms to interact with noisy latent states z_t , guiding the denoising process towards a text-aligned image. We maintain this standard notation throughout the paper.

To evaluate similarity in the image space, we calculate the cosine similarity between the image and text prompt [33]. Given the final generated image, x , and the text prompt, y , we employ the text encoder, τ_θ , and an image encoder, ν_θ , to compute the similarity score,

$$S_{\text{final}}(x, y) = \tau_\theta(y) \cdot \nu_\theta(x). \quad (2)$$

To generalize this operator to intermediate diffusion steps, we extend $S_{\text{final}}(x, y)$ which operates on pixel-space images (x), to support latent representations z_t that cannot be directly inputted into ν_θ . This extension is defined as

$$S_{\text{latent}}(z_t, y) = \tau_\theta(y) \cdot \nu_\theta(\Phi(z_t)), \quad (3)$$

where z_t is the latent representation at step t , and Φ is a transformation (inverse of $\mathcal{E}$) that converts z_t into an RGB-like space consumable by the image encoder ν_θ , as illustrated in Figure 2. The text prompt y and the text encoder parameters remain the same as in S_{final} .

The definition of S_{latent} allows for the assessment of how well image representations align with text prompts during the reverse diffusion process. This approach relies on having access to noise-corrupted data. In the following section, we will explore the scarcity of publicly available data of this kind and explore our approach to generate this kind of data.

3.2 Sampling Noise-Corrupted Data

There is a notable lack of open data involving noisy latent images, as prior works often rely on on-the-fly noise inversion [8, 44, 49]. This hinders reproducibility and consistent benchmarking. To mitigate this, our training dataset $\mathcal{D}$ is structured as a collection of samples where each sample within $\mathcal{D}$ is a tuple of the form $(y, x, \{z_t\}_{t=1}^T)$, where y represents the image generation prompt, x is the corresponding generated image and $\{z_t\}_{t=1}^T$ represents the sequence of latent representations sampled from the image’s reverse diffusion process.

We optimize our encoder using a rewritten version of CC12M [9, 25]. These refined captions provide dense, detailed prompts that better align with the requirements of image generation tasks, based on the diverse visual concepts present in the original CC12M [5]. Using SDXL [32], we generate images from these prompts and extract latent representations for training (as detailed in Section 4.5).

Although this process successfully captures the complete latents necessary for optimization, we cannot simply sample latents at random throughout the sequence. Early steps are dominated by Gaussian noise, whereas late steps lack the ambiguity required to train an early-stopping monitor. Determining the optimal noise distribution therefore becomes a highly non-trivial challenge, motivating our systematic evaluation in Section 5.2.

3.3 Language-to-Latent Contrastive Training

To improve noise robustness in pre-trained vision encoders, we introduce Noisy-CLIP, a twin-tower encoder capable of measuring similarity earlier in the reverse diffusion process.

We select a fixed latent interval (20–29) for contrastive training to enable stable language-to-latent alignment across diverse representations using only a subset of the data, as validated in Section 5.2. Using the generated noisy latents z_t , we exclusively fine-tune the image encoder (see Figure 2). During the contrastive learning, each noisy latent representation z_t forms a positive pair with its corresponding text prompt y , while being contrasted against in-batch negatives $\mathcal{N}$ (all image-text pairs in the batch except the positive pair).

By applying the contrastive loss (InfoNCE) [27] we maximize similarity between positive pairs (z_t, y) while suppressing similarity with negatives $\mathcal{N}$. Freezing the text encoder maintains its general-purpose functionality while optimizing language-to-latent alignment through image encoder tuning. The crucial advantage of this approach is that the frozen text encoder can be shared with the diffusion model, which ensures consistency between textual representations and the diffusion model’s conditioning space. This grounding forces the latent encoder to align better with the diffusion representations, effectively mimicking the cross-attention interaction during denoising while providing an early projection of final image alignment [19].

3.4 Benchmark Proposal

Lastly, to overcome the lack of standardized data to evaluate representations in inner diffusion processes, we propose a benchmark built on top of image captioning [5] and image generation [23] datasets. We derive two subsets, Noisy-Conceptual-Captions and Noisy-GenAI-Bench to enable open, systematic community evaluation across three critical dimensions.

Diverse visual representations. To evaluate inner diffusion representations, we generate multiple images for each prompt p , allowing us to quantify how alignment metrics choose the output that most closely matches the prompt across diverse visual representations.

Factual consistency. We employ a classification task where a single image is evaluated against k prompts, one correct and $k - 1$ intentionally corrupted to

systematically violate factual consistency (as detailed in the Appendix). This approach improves the robustness of the evaluation compared to random prompt sampling by introducing targeted factual violations rather than random information.

Noisy latent information. To support the assessment of the previously described tasks in the reverse diffusion process, we store the noisy latent information at each denoising step during generation.

4 Experimental Setup

In this section, we provide a detailed description of our evaluation settings, including tasks, datasets, baseline models, metrics, and implementation details.

4.1 Evaluation Tasks

To evaluate the proposed method, we conducted the following experiments:

- **Language-to-Latent Alignment.** We start by examining the language-to-latent alignment throughout the denoising process of N competing random seeds. We measure the alignment with NoisyCLIP and CLIP, and compare the difference between the best and the worst images (VQAScore is used to rank images by their alignment with the prompt). This allows us to observe how early NoisyCLIP can detect misalignments.
- **Latent-to-Text Retrieval:** As an additional experiment, we assess the model’s ability to assign the highest score to the most factually accurate prompt. This experiment aims to analyze NoisyCLIP in the presence of captions that include non-factual distractors. This evaluation is crucial as it tests the model’s capacity to distinguish the correct caption from highly similar, yet non-factual, errors. Ultimately, success on this task suggests that the model possesses the semantic understanding required to prevent non-factual image selections during generation.
- **Early-Stopping Best-of-N:** Our main experiment follows recent work [24, 26, 45] that demonstrated a significant improvement in image quality by generating multiple images and selecting the most aligned one. In our setting, instead of completing all denoising processes, we leverage NoisyCLIP early misalignment detection capabilities, and evaluate how early it can stop poorly aligned denoising processes and keep the best language-to-latent aligned one.

4.2 Datasets

For the experiments explained above, we use two datasets. Specifically, the test sets are:

- **Noisy-Conceptual-Captions:** This subset from our benchmark comprises a set of 1000 captions sampled to enable objective assessment of factual caption selection. This dataset comprises the original prompts, four generated images per original prompt as well as four non-factual prompts (the creation of these prompts is detailed in the Appendix).
- **Noisy-GenAI-Bench:** This dataset includes a total of 1600 prompts, consisting of 730 basic prompts and 870 more advanced ones, going from simple compositions such as objects (e.g. "a dog") to more complex compositions such as counting or negations. To evaluate the performance on the BoN task, we expand this dataset by generating 10 images along with the corresponding latent information for each prompt.

4.3 Baselines

We evaluate NoisyCLIP against several established baselines. Because our approach builds upon CLIP [33], we utilize pre-trained CLIP and SigLIP [51] as primary points of comparison, employing their base versions for ablations and large versions for main results. We also compare against LatentCLIP [3] (versions 8 and 4-plus) to benchmark against latent-optimized representations. All primary generation experiments are conducted using Stable Diffusion XL [32], with generalization tests performed on two DiT-based [30] architectures: Stable Diffusion 3.5 [10] and FLUX.1 [dev] [22].

4.4 Evaluation Metrics

In the Latent-to-Text Retrieval task, we presented each image with a single correct prompt and multiple non-factual prompts, measuring performance using Recall@1 (R@1) [11, 31, 52, 54]. This metric quantifies the model’s accuracy by measuring the proportion of images in which the correct prompt is ranked first among all candidates.

For the BoN task, we assess the model’s ability to select the most relevant image for a given prompt. VQA-based evaluation methods have shown improved reliability and alignment with human assessments [7, 17, 18, 23, 50]. Similarly to previous research, we adopt VQAScore [24] as our primary metric to measure how well encoders can select the most similar image over others for a certain prompt [2, 4, 12, 41, 43].

4.5 Implementation Details

Training. We fine-tune solely the visual encoder of the dual-encoder on a dataset of 50000 latents, which are uniformly sampled from the generated images. Training is done with a learning rate of $5.7e-5$, using a cosine annealing scheduler with a warm-up ratio of 0.1 and a weight decay of 0.1. All computations are performed on an NVIDIA A100 40G GPU over 10 epochs with a batch size of 128. **Latent RGB Representation.** Although SDXL [32] latent space utilizes

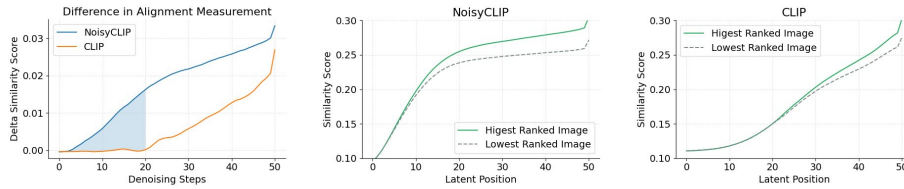


Fig. 3: Analysis of alignment measurement throughout the reverse diffusion process shows that our model distinctly separates aligned from misaligned images as soon as latent 20.

a four-channel structure, direct visualization requires transforming these latents into RGB space. This transformation is done by decoding latent representations via a linear transformation to RGB [42]. For our generalization experiments on DiTs, the highly compressed nature of their latent spaces requires using the model’s VAE for RGB decoding.

5 Results and Discussion

In this section, we evaluate how NoisyCLIP addresses the challenges established in our introduction. We begin by assessing NoisyCLIP’s Language-to-Latent Alignment to validate its capacity for monitoring intermediate denoising steps. To understand the data strategies required for this monitoring, we then analyze the Impact of Training Latent Ranges on Inference. Following this, we demonstrate practical computational reductions by employing our method for Early-Stopping in Best-of-N Sampling. Finally, we present a Generalization Analysis across varying backbone architectures (further ablations in the Appendix).

5.1 Language-to-Latent Alignment

Language-to-Latent Alignment Estimators. To evaluate alignment effectiveness, we sample prompts from the Noisy-Conceptual-Captions dataset, generating four images per prompt. Using VQAScore to establish a reference ranking, we report the average similarity score assigned to each rank by both NoisyCLIP and CLIP.

As illustrated in Figure 3, our model effectively discriminates the top image from sub-optimal ones as early as latent 20, as the backbone can only do so by latent 40. We can see that our method increases the delta between the highest and lowest scoring images surpassing 2% in the middle of the generation. Although our approach focuses on the noisy images, the results do not show any deterioration in the image dimensions. Furthermore, our approach achieves even greater differentiation by the end of the process compared to the backbone trained on images alone.

In contrast, CLIP exhibited persistent difficulty maintaining a separation between the highest and lowest scoring images. Although it reaches a 2.5% delta

in the image dimension, the gap remains under 1% for most of the generation (up to the mid-30s) and only emerges after latent 20. These results demonstrate that our approach reliably measures similarity in intermediate latents while achieving final scores comparable to frozen CLIP. Moreover, our model produces scores that increase almost linearly from the start of generation, whereas CLIP only shows this behavior from the midpoint onward.

Latent-to-Text Retrieval. This experiment evaluates models ability to select the correct caption for an image from five candidates (one correct, four non-factual). As shown in Figure 4, NoisyCLIP discriminates the correct caption from the non-factual ones remarkably early, by latent 10. In contrast, CLIP requires an additional 20 denoising iterations to achieve a similar separation (colored dots, Figure 4).

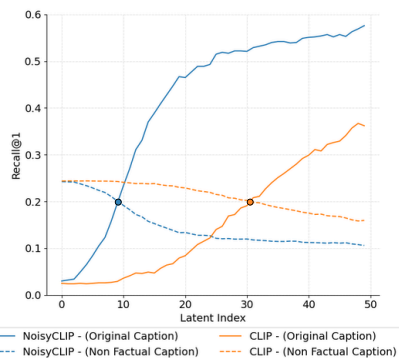


Fig. 4: Recall@1 for Caption Selection.

While both models increasingly favor the correct caption as denoising progresses, NoisyCLIP demonstrates significantly faster and higher convergence. Within the first 25 iterations, NoisyCLIP’s recall@1 for the correct caption surges from near zero to over 0.5, whereas the baseline remains below 0.1. Furthermore, NoisyCLIP’s recall stabilizes during the second half of generation, whereas the baseline never exceeds 0.4 (further analysis of non-factual caption selection is provided in the Appendix).

5.2 Impact of Training Latent Ranges on Inference

In this experiment we examine how different training latent ranges impact NoisyCLIP performance. Results are reported on the Noisy-Concept-Captions dataset. As shown in Figure 5, the alignment performance correlate with the inference latent range: lower latent ranges result in less accurate alignment, underscoring the challenge of assessing language-to-latent alignment under the presence of noise.

The baseline CLIP model (top row), performs poorly in latents ranges 1-49 achieving a top performance of 0.696, despite a VQAScore of 0.747 on the final image. This disparity highlights the limitation of lacking noisy data training. By comparison, NoisyCLIP inference performance generally correlates

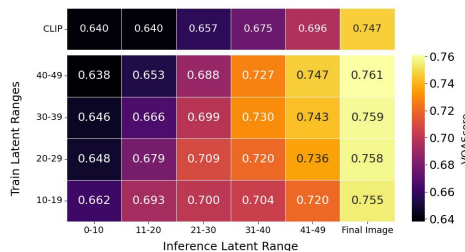


Fig. 5: Performance across inference ranges (columns) when trained with different latent ranges (rows).

with the level of noise present during training. Low-noise training (ranges 40-49) consistently yields the best overall results in both the final latent range and the final image. However, the model trained on initial latent ranges (10-19) performs best early on (up to 4 points gain) but declines sharply in the final two ranges. Middle-range training (20-29) strikes the best balance, achieving stable performance: it is comparable to the best models (40-49) in later inference ranges while performing better in initial ranges. These results indicate that moderate training noise (20-29) optimizes balanced performance across inference stages.

5.3 Early-Stopping in Best-of-N Sampling

To further evaluate our fine-tuning approach with noisy latent information, we analyze NoisyCLIP performance as a ranker on a BoN denoising task. We evaluate model’s ability to select a candidate image by denoising multiple seeds with the same prompt, and scoring the latents according to its alignment with the prompt in the GenAI-Bench basic prompts. This enables the early stopping of denoising processes that are estimated to be sub-optimal, reducing the aggregated number of denoising iterations. A detailed analysis of the performance on the basic and advanced prompts is provided in the Appendix.

Defining Denoising Iterations as a Metric for Cost. We define cost by the number of reverse diffusion steps required to produce an image; in our baseline [32], this is 50 steps per image. If BoN is applied at a certain stage, only one image proceeds with the generation. For example, if two images begin generation but only one is selected to continue at latent stage 25, the total cost becomes $(25 \times 2) + (25 \times 1) = 75$, representing a 25% saving compared to a full "best-of-2" approach ($2 \times 50 = 100$).

Performance Gains Versus Costs Reduction. In Figure 6 we analyze the trade-off between prompt alignment and computational cost within a specific best-of-6 selection scenario. The standard baseline, which uses CLIP to select the final image after all 6 generations are complete, achieves a VQAScore of 0.85 at a cost of 300 ($= 6 \times 50$). At this same 300-step cost, the best result achieved by LatentCLIP matches the performance of our method (0.86).

By reducing the computational budget by 50% (to 150 denoising steps), our method incurs a minimal 2% drop in VQAScore compared to 300-step CLIP and LatentCLIP-8 baselines, and less than a 3% drop compared to LatentCLIP-4-plus. Furthermore, when evaluated at this reduced 150-step budget, our approach

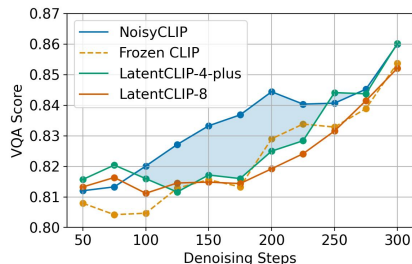


Fig. 6: Alignment (VQAScore) vs Best-of-N denoising cost. Cost is normalized for a scenario where $N = 6$.

outperforms all baselines by nearly 2%. This advantage persists at a 200-step budget (a one-third reduction), where our method continues to achieve the highest latent selection score. Overall, our approach consistently outperforms the baselines across the reverse diffusion process, demonstrating a superior balance between alignment and efficiency.

Selecting images in the latent space. In this analysis, we compare the performance of our method on Best-of-N denoising task, with $N = \{1, \dots, 10\}$. When applying a dual encoder to rank the N generations in latent 20 (Figure 7, top), the performance of all baselines mirror the mean alignment of the entire candidate set. This suggests that applying current approaches at this stage yields negligible benefit. In contrast, our method improves alignment by up to 2%, demonstrating the effectiveness of our noise-aware approach in handling early-stage latent representations.

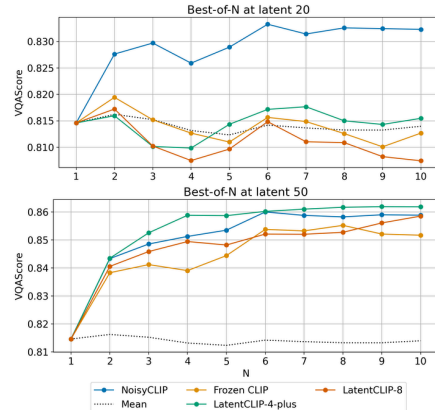


Fig. 7: Best-of-N results for latent 20 and final image. Alignment is assessed using VQAScore.

Figure 7 (bottom) presents results for the compute-intensive final image selection, where all images are fully generated. While CLIP baseline benefits from the absence of noise in the final images, our method consistently achieves superior prompt alignment across all values of N . Even compared to LatentCLIP (trained on clean z_0 latents), our model consistently outperforms their version with a patch size of 8 and remains highly competitive with the 4-plus variant.

In summary, NoisyCLIP enables effective mid-generation BoN selection, while baselines fail in high-noise settings by mirroring random sampling. Though competitive with LatentCLIP in final stages, our novelty lies in unlocking computational efficiency via early-step ranking, where current CLIP methods fail.

Qualitative Results. Figure 8 demonstrates that our method achieves high-fidelity image selection in a best-of-6 setting using only 150 diffusion steps, consistently matching or surpassing CLIP and LatentCLIP at an equivalent or higher computational cost. For simple queries (e.g., the third prompt), all approaches successfully select text-aligned images, confirming that standard generative models adequately handle low-complexity prompts.

However, our method excels on complex compositions. For *"a tall cactus in a terracotta pot next to a short succulent in a ceramic bowl"*, only NoisyCLIP maintains all attributes (plant heights, container materials) using half the computation. Similarly, NoisyCLIP resolves subtle relationships, identifying a cat

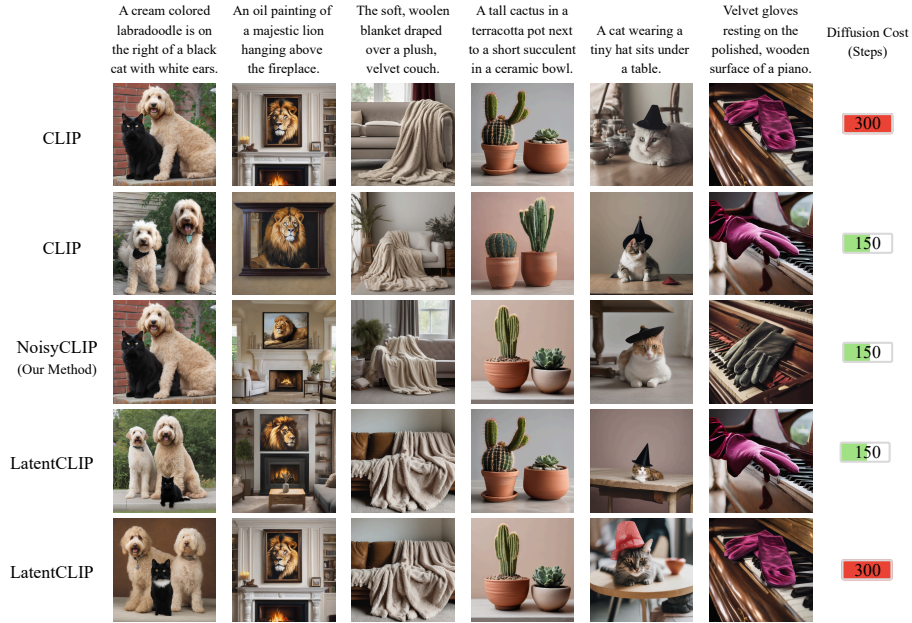


Fig. 8: Images generated using a best-of-6 strategy. We compare NoisyCLIP with the CLIP and LatentCLIP-4-plus under two computation budgets: 150 and 300 diffusion steps (generating all 6 images). Our approach surpasses the baseline at the 150-step limit and achieves comparable or superior alignment to the full 300-step baselines.

under a table (fifth prompt) that baselines miss at full computation. Furthermore, it distinguishes resting gloves on a piano (sixth prompt) at half the budget, a nuance where baselines fail completely. By capturing these details early in the latent space, our results validate that alignment can be measured without the overhead of full image generation (see Appendix for failure cases).

5.4 Generalization Analysis

Diffusion Architectures. To evaluate the robustness of NoisyCLIP, we validate its generalization capabilities on two DiT architectures in a zero-shot setting. By applying NoisyCLIP to VAE-decoded latent representations, we observe consistent improvements over the frozen CLIP baseline across these distinct models. Specifically, NoisyCLIP achieves a +2% alignment gain on Stable Diffusion

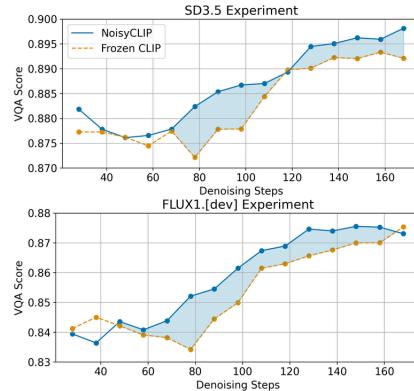


Fig. 9: Alignment versus computational cost for zero-shot generalization on DiTs.

3.5 [10] and a +1.8% gain on FLUX.1 [22], alongside a 57% reduction in denoising steps (Fig.9).

These findings reinforce two core strengths of our method. First, they demonstrate that although NoisyCLIP is trained exclusively on data generated by SDXL, our noise-aware contrastive training effectively transfers to different generative architectures. Second, the observed alignment gains mirror those achieved on SDXL (Sec. 5.3), demonstrating that our significant reductions in computational cost persist even when applied to newer models configured with fewer denoising iterations (e.g., 28 steps).

NoisyCLIP Backbones. Table 1 shows our approach substantially improves intermediate alignment. In factual consistency (R@1), NoisyCLIP boosts CLIP performance by $3.5\times$. On the Best-of- N task, it achieves a 0.709 VQAScore (+5% over baseline), reaching 95% of final-image performance midway through generation.

Evaluating backbone transferability, NoisySigLIP matches NoisyCLIP on our primary Best-of- N visual ranking task (0.718 vs. 0.709). Although NoisySigLIP trails in R@1, this discrepancy is an expected architectural consequence. R@1 tests fine-grained text perturbations, and SigLIP’s sigmoid loss inherently lacks the discriminative sharpness of CLIP’s softmax for subtle textual nuances [55]. Thus, our method transfers robustly across backbones for visual selection, with text-discrimination performance constrained by the base encoder.

Table 1: Performance comparison in latent and image spaces. We report Factual Consistency (R@1) and Best-of-N (VQAScore).

Model	Fact. Consist.	BoN
<i>Denoised Latent (it. 21-30)</i>		
NoisyCLIP	0.506	<u>0.709</u>
CLIP L14	0.145	0.657
NoisySigLIP	<u>0.194</u>	0.718
SigLIP L16	0.147	0.652
<i>Final Image (it. 50)</i>		
CLIP L14	–	0.747
SigLIP L16	–	0.764

6 Conclusions

In this work, we introduced NoisyCLIP, a framework for evaluating language-to-latent alignment during the reverse diffusion process by quantifying similarity across intermediate latent representations. Empirical evaluations demonstrate that NoisyCLIP provides substantial advantages in Best-of- N generation strategies by enabling early-stage discrimination of higher-quality samples. This capability leads to more efficient use of computational resources, reducing the number of required diffusion iterations by up to 50% while preserving the fidelity of the final outputs relative to a full cost solution.

Collectively, these findings establish the feasibility of continuous, in-process alignment evaluation within latent diffusion models. By leveraging iterative latent information, NoisyCLIP transforms alignment assessment from a purely post-hoc procedure into an integral component of the generation process advancing the development of adaptive, controllable, and computationally efficient multimodal generative systems.

Acknowledgments. We thank anonymous reviewers for their valuable comments and suggestions. This work was partially supported by the FCT Ph.D. scholarship grant Ref. 2025.02103.BD, by the FCT project NOVA LINC Ref. (UID/04516/2025) and by a Google Research Gift.

References

1. Agarwal, A., Karanam, S., Joseph, K.J., Saxena, A., Goswami, K., Srinivasan, B.V.: A-STAR: test-time attention segregation and retention for text-to-image synthesis. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023. pp. 2283–2293. IEEE (2023). <https://doi.org/10.1109/ICCV51070.2023.00217>
2. Baldridge, J., Bauer, J., Bhutani, M., Brichtova, N., Bunner, A., Chan, K., Chen, Y., Dieleman, S., Du, Y., Eaton-Rosen, Z., Fei, H., de Freitas, N., Gao, Y., Gladchenko, E., Colmenarejo, S.G., Guo, M., Haig, A., Hawkins, W., Hu, H., Huang, H., Igwe, T.P., Kaplanis, C., Khodadadeh, S., Kim, Y., Konyushkova, K., Langner, K., Lau, E., Luo, S., Mokrá, S., Nandwani, H., Onoe, Y., van den Oord, A., Parekh, Z., Pont-Tuset, J., Qi, H., Qian, R., Ramachandran, D., Rane, P., Rashwan, A., Razavi, A., Riachi, R., Srinivasan, H., Srinivasan, S., Strudel, R., Uria, B., Wang, O., Wang, S., Waters, A., Wolff, C., Wright, A., Xiao, Z., Xiong, H., Xu, K., van Zee, M., Zhang, J., Zhang, K., Zhou, W., Zolna, K., Aboubakar, O., Akbulut, C., Akerlund, O., Albuquerque, I., Anderson, N., Andreetto, M., Aroyo, L., Bariach, B., Barker, D., Ben, S., Berman, D., Biles, C., Blok, I., Botadra, P., Brennan, J., Brown, K., Buckley, J., Bunel, R., Bursztein, E., Butterfield, C., Caine, B., Carpenter, V., Casagrande, N., Chang, M., Chang, S., Chaudhuri, S., Chen, T., Choi, J., Churbanau, D., Clement, N., Cohen, M., Cole, F., Dektiarev, M., Du, V., Dutta, P., Eccles, T., Elue, N., Feden, A., Fruchter, S., Garcia, F., Garg, R.: Imagen 3. CoRR **abs/2408.07009** (2024). <https://doi.org/10.48550/ARXIV.2408.07009>
3. Becker, J., Wendler, C., Baylies, P., West, R., Wressnegger, C.: Controlling latent diffusion using latent CLIP. CoRR **abs/2503.08455** (2025). <https://doi.org/10.48550/ARXIV.2503.08455>
4. Cascante-Bonilla, P., Hou, Y., Cao, Y.T., III, H.D., Rudinger, R.: Natural language inference improves compositionality in vision-language models. In: The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025. OpenReview.net (2025)
5. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021. pp. 3558–3568. Computer Vision Foundation / IEEE (2021). <https://doi.org/10.1109/CVPR46437.2021.00356>
6. Chefer, H., Lang, O., Geva, M., Polosukhin, V., Shocher, A., Irani, M., Mosseri, I., Wolf, L.: The hidden language of diffusion models. In: The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net (2024)
7. Cho, J., Zala, A., Bansal, M.: Visual programming for step-by-step text-to-image generation and evaluation. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023 (2023)

8. Dong, W., Xue, S., Duan, X., Han, S.: Prompt tuning inversion for text-driven image editing using diffusion models. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023. pp. 7396–7406. IEEE (2023). <https://doi.org/10.1109/ICCV51070.2023.00683>
9. Emporium, C.: conceptual-captions-cc12m-llavanext. <https://huggingface.co/datasets/CaptionEmporium/conceptual-captions-cc12m-llavanext> (2024)
10. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., Rombach, R.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024. OpenReview.net (2024)
11. Ganon, K., Alper, M., Mikulinsky, R., Averbuch-Elor, H.: WAFFLE: multimodal floorplan understanding in the wild. In: IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2025, Tucson, AZ, USA, February 26 - March 6, 2025. pp. 1488–1497. IEEE (2025). <https://doi.org/10.1109/WACV61041.2025.00152>
12. Girrbach, L., Alaniz, S., Smith, G., Akata, Z.: A large scale analysis of gender biases in text-to-image generative models. CoRR **abs/2503.23398** (2025). <https://doi.org/10.48550/ARXIV.2503.23398>
13. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K. (eds.) *Advances in Neural Information Processing Systems*. vol. 27. Curran Associates, Inc. (2014)
14. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. In: Moens, M., Huang, X., Specia, L., Yih, S.W. (eds.) *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*. pp. 7514–7528. Association for Computational Linguistics (2021). <https://doi.org/10.18653/V1/2021.EMNLP-MAIN.595>
15. Ho, J., Jain, A., Abbeel, P.: Denoising Diffusion Probabilistic Models. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) *Advances in Neural Information Processing Systems*. vol. 33, pp. 6840–6851. Curran Associates, Inc. (2020)
16. Ho, J., Salimans, T.: Classifier-free diffusion guidance. In: *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications* (2021)
17. 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: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023. pp. 20349–20360. IEEE (2023). <https://doi.org/10.1109/ICCV51070.2023.01866>
18. Huang, K., Sun, K., Xie, E., Li, Z., Liu, X.: T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023* (2023)
19. Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., Carreira, J.: Perceiver: General perception with iterative attention. In: Meila, M., Zhang, T. (eds.) *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event. Proceedings of Machine Learning Research*, vol. 139, pp. 4651–4664. PMLR (2021)

20. Kingma, D.P., Welling, M.: Auto-encoding variational bayes (2022), <https://arxiv.org/abs/1312.6114>
21. Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., Levy, O.: Pick-a-pic: An open dataset of user preferences for text-to-image generation. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023* (2023)
22. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., et al.: Flux. 1 kontekst: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742* (2025)
23. Li, B., Lin, Z., Pathak, D., Li, J., Fei, Y., Wu, K., Ling, T., Xia, X., Zhang, P., Neubig, G., Ramanan, D.: Genai-bench: Evaluating and improving compositional text-to-visual generation. *CoRR* **abs/2406.13743** (2024). <https://doi.org/10.48550/ARXIV.2406.13743>
24. 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: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part IX. Lecture Notes in Computer Science, vol. 15067*, pp. 366–384. Springer (2024). https://doi.org/10.1007/978-3-031-72673-6_20
25. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llava-next: Improved reasoning, ocr, and world knowledge (January 2024)
26. Ma, N., Tong, S., Jia, H., Hu, H., Su, Y., Zhang, M., Yang, X., Li, Y., Jaakkola, T.S., Jia, X., Xie, S.: Scaling inference time compute for diffusion models. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*. pp. 2523–2534. Computer Vision Foundation / IEEE (2025). <https://doi.org/10.1109/CVPR52734.2025.00241>
27. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. *CoRR* **abs/1807.03748** (2018)
28. Papalampidi, P., Wiles, O., Ktena, I., Shtedritski, A., Bugliarello, E., Kajic, I., Albuquerque, I., Nematzadeh, A.: Dynamic classifier-free diffusion guidance via online feedback. *CoRR* **abs/2509.16131** (2025). <https://doi.org/10.48550/ARXIV.2509.16131>
29. Park, Y., Kwon, M., Choi, J., Jo, J., Uh, Y.: Understanding the latent space of diffusion models through the lens of riemannian geometry. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023* (2023)
30. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*. pp. 4172–4182. IEEE (2023). <https://doi.org/10.1109/ICCV51070.2023.00387>
31. Pei, G., Chen, T., Wang, Y., Cai, X., Shu, X., Zhou, T., Yao, Y.: Seeing what matters: Empowering CLIP with patch generation-to-selection. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*. pp. 24862–24872. Computer Vision Foundation / IEEE (2025). <https://doi.org/10.1109/CVPR52734.2025.02315>

32. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: SDXL: improving latent diffusion models for high-resolution image synthesis. In: The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net (2024)
33. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (2021)
34. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **21**, 140:1–140:67 (2020)
35. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)
36. Ramos, V., Bitton, Y., Yarom, M., Szpektor, I., Magalhaes, J.: Contrastive sequential-diffusion learning: Non-linear and multi-scene instructional video synthesis. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 4645–4654 (2025). <https://doi.org/10.1109/WACV61041.2025.00456>
37. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-Resolution Image Synthesis with Latent Diffusion Models. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10674–10685. IEEE, New Orleans, LA, USA (Jun 2022). <https://doi.org/10.1109/CVPR52688.2022.01042>
38. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. pp. 10674–10685. IEEE (2022). <https://doi.org/10.1109/CVPR52688.2022.01042>
39. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, S.K.S., Lopes, R.G., Ayan, B.K., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic text-to-image diffusion models with deep language understanding. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022 (2022)
40. Sauer, A., Boesel, F., Dockhorn, T., Blattmann, A., Esser, P., Rombach, R.: Fast high-resolution image synthesis with latent adversarial diffusion distillation. In: Igarashi, T., Shamir, A., Zhang, H.R. (eds.) SIGGRAPH Asia 2024 Conference Papers, SA 2024, Tokyo, Japan, December 3-6, 2024. pp. 106:1–106:11. ACM (2024). <https://doi.org/10.1145/3680528.3687625>
41. Sun, H., Xia, B., Chang, Y., Wang, X.: Generalizing alignment paradigm of text-to-image generation with preferences through f-divergence minimization. In: Walsh, T., Shah, J., Kolter, Z. (eds.) AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA. pp. 27644–27652. AAAI Press (2025). <https://doi.org/10.1609/AAAI.V39I26.34978>

42. Vass, T.A.: Explaining the SDXL latent space. <https://huggingface.co/blog/TimothyAlexisVass/explaining-the-sdxl-latent-space> (May 2024), accessed on 01.03.2025
43. Wang, A.Z., Ge, S., Karras, T., Liu, M., Balaji, Y.: A comprehensive study of decoder-only llms for text-to-image generation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025. pp. 28575–28585. Computer Vision Foundation / IEEE (2025). <https://doi.org/10.1109/CVPR52734.2025.02661>
44. Wang, Z., Bao, J., Zhou, W., Wang, W., Hu, H., Chen, H., Li, H.: DIRE for diffusion-generated image detection. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023. pp. 22388–22398. IEEE (2023). <https://doi.org/10.1109/ICCV51070.2023.02051>
45. Xie, E., Chen, J., Zhao, Y., YU, J., Zhu, L., Lin, Y., Zhang, Z., Li, M., Chen, J., Cai, H., Liu, B., Zhou, D., Han, S.: SANA 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer. In: Forty-second International Conference on Machine Learning (2025)
46. Xie, J., Mao, W., Bai, Z., Zhang, D.J., Wang, W., Lin, K.Q., Gu, Y., Chen, Z., Yang, Z., Shou, M.Z.: Show-o: One single transformer to unify multimodal understanding and generation. In: The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025. OpenReview.net (2025)
47. Xu, K., Zhang, L., Shi, J.: Good seed makes a good crop: Discovering secret seeds in text-to-image diffusion models. In: IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2025, Tucson, AZ, USA, February 26 - March 6, 2025. pp. 3024–3034. IEEE (2025). <https://doi.org/10.1109/WACV61041.2025.00299>
48. Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Zhang, W., Cui, B., Yang, M.H.: Diffusion models: A comprehensive survey of methods and applications. *ACM computing surveys* **56**(4), 1–39 (2023)
49. Yang, Z., Ding, G., Wang, W., Chen, H., Zhuang, B., Shen, C.: Object-aware inversion and reassembly for image editing. In: The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net (2024)
50. Yarom, M., Bitton, Y., Changpinyo, S., Aharoni, R., Herzig, J., Lang, O., Ofek, E., Szpektor, I.: What you see is what you read? improving text-image alignment evaluation. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023* (2023)
51. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 11975–11986 (2023)
52. Zhang, B., Zhang, P., Dong, X., Zang, Y., Wang, J.: Long-clip: Unlocking the long-text capability of CLIP. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LI*. *Lecture Notes in Computer Science*, vol. 15109, pp. 310–325. Springer (2024). https://doi.org/10.1007/978-3-031-72983-6_18
53. Zhang, R., Tong, C., Zhao, Z., Guo, Z., Zhang, H., Zhang, M., Liu, J., Gao, P., Li, H.: Let’s verify and reinforce image generation step by step. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025. pp. 28662–28672. Computer Vision Foundation / IEEE (2025). <https://doi.org/10.1109/CVPR52734.2025.02669>

54. Zhang, Y., Wang, J., Wu, H., Xu, W.: Distinctive image captioning via CLIP guided group optimization. In: Karlinsky, L., Michaeli, T., Nishino, K. (eds.) *Computer Vision - ECCV 2022 Workshops - Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part IV*. Lecture Notes in Computer Science, vol. 13804, pp. 223–238. Springer (2022). https://doi.org/10.1007/978-3-031-25069-9_15
55. Zohra, F., Zhao, C., Itani, H., Ghanem, B.: b-clip: Text-conditioned contrastive learning for multi-granular vision-language alignment. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 680–689 (June 2026)