LEARNING TO READ THE CONTEXTUAL TOKENS IN DIFFUSION TRANSFORMERS

Anonymous authors
Paper under double-blind review

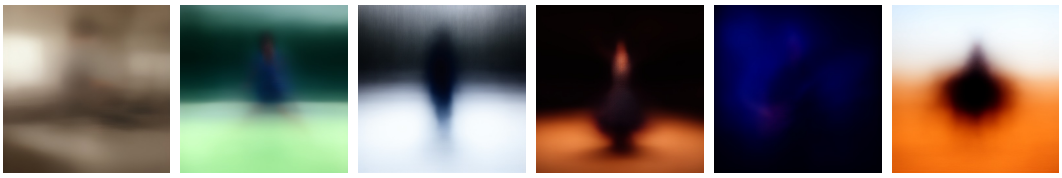

Figure 1: **Our *Contextual Reader* answers the question "What is this person doing?" directly from the contextual tokens.** We show six generations conditioned on different prompts at an early denoising stage, where the visual predictions remain ambiguous. The text below each image is decoded from the contextual tokens when the MM-DiT is provided with an empty prompt. Despite this, our reader correctly identifies the activity, demonstrating that contextual tokens encode rich semantic information about the emerging image early in generation.

ABSTRACT

Multimodal Diffusion Transformers (MM-DiTs) jointly process visual and textual representations throughout generation. These models repeatedly update the text tokens through multimodal attention, forming dynamic *contextual tokens* whose function is not well understood. In this work, we introduce a framework for *reading* this contextual space through natural-language interrogation. We train a lightweight bottleneck network that maps intermediate contextual tokens into the input space of a frozen Large Language Model (LLM), allowing the LLM to answer questions about the emerging image directly from these hidden representations. Our reader reveals that contextual tokens encode a rich, global representation of the emerging scene: generation-specific semantics, including attributes left underspecified by the prompt, are accessible surprisingly early in denoising, while increasingly fine-grained details become readable over time. Remarkably, this information remains decodable even when the MM-DiT receives an empty prompt, showing that contextual tokens accumulate substantial image-specific information from the evolving visual representation itself. We further find that generations with more readable contextual representations tend to receive higher human-preference scores. Building on these observations, we introduce Contextual Alignment, a training technique that explicitly reinforces the visual-semantic information encoded in the contextual tokens, improving generation quality and distributional coverage. Together, our results establish contextual tokens as both an interpretable view into the internal dynamics of MM-DiTs and an effective target for improving generative models.

1 INTRODUCTION

In recent years, diffusion models have fundamentally transformed visual content creation by enabling high-fidelity generation directly from text prompts (Rombach et al., 2022; Saharia et al., 2022; Ramesh et al., 2022). Building on this foundation, Multimodal Diffusion Transformers (MM-DiTs) have emerged as a dominant paradigm for state-of-the-art image and video synthesis (Esser et al., 2024; Kong et al., 2024; Chefer et al., 2026). Central to these models are multimodal attention layers, which enable bidirectional interaction between textual and visual representations throughout generation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Within this shared multimodal space, visual tokens form a spatial representation of the evolving image and are explicitly supervised by the denoising objective. Their intermediate features can therefore be directly visualized as images (Patashnik et al., 2023; Tumanyan et al., 2023), making them especially accessible to analysis and intervention. This has enabled methods that guide visual features for controlled generation (Ho & Salimans, 2022; Zhang et al., 2023; Chefer et al., 2023), manipulate them for image editing (Meng et al., 2021; Avrahami et al., 2022; Hertz et al., 2022), and align them with pretrained representations to improve generation (Yu et al., 2024; Leng et al., 2025; Jiang et al., 2025).

The text stream, however, is considerably less understood. In modern MM-DiTs, text tokens do not remain fixed conditioning vectors. Instead, they are repeatedly updated through multimodal attention, absorbing information from the emerging visual representation and evolving into what prior work termed *contextual tokens* (Dahary et al., 2026). Although this cross-modal exchange has proven effective for generation (Esser et al., 2024), the function of the resulting contextual representations remains unclear. Unlike visual tokens, contextual tokens have no explicit spatial correspondence and receive no dedicated supervision, making them inherently difficult to interpret. Consequently, it remains unknown what visual information they encode and how their dynamics ultimately steer the generation process.

In this work, we propose a framework to directly *read* and interpret the contextual space of MM-DiTs by enabling natural language interrogation of their hidden representations. Rather than inspecting raw attention maps or feature vectors, our approach enables asking plain-text questions about visual elements in the emerging image and receiving direct answers decoded straight from the contextual tokens. To achieve this, we train a lightweight bottleneck network that projects intermediate contextual tokens into the input space of a Large Language Model (LLM). Because the LLM relies exclusively on the projected contextual tokens to answer visual questions, the bottleneck learns to extract the model's internal visual semantics directly from the text-side representation. In doing so, we can read the model's evolving generative intent in plain text, revealing how it internally resolves underspecified prompt attributes over the course of generation.

Through our reading mechanism, we find that contextual tokens encode generation-specific semantics surprisingly early in denoising. Remarkably, this remains true even when no prompt is fed into the model, showing that the tokens can acquire substantial information from the evolving visual representation alone. Figure 1 demonstrates this setting: despite the MM-DiT receiving an empty prompt, the reader can identify the person's activity at an early denoising stage, before it can be readily inferred from the visual prediction. Moreover, we find that readability continues to increase over the course of denoising, with the reader recovering progressively finer details about the emerging image. Importantly, this readability can also vary across different generations from the same prompt. Generations with clearer contextual representations tend to achieve higher human-preference scores, making contextual readability a meaningful signal of generation quality.

Building on these observations, we transition from passively reading the contextual tokens to actively supervising them during model training. To this end, we introduce Contextual Alignment, a regularization objective that explicitly reinforces the visual-semantic information encoded in the contextual space. We evaluate our supervised training framework through quantitative and qualitative experiments spanning both full model training and fine-tuning. Our experiments show that contextual supervision complements existing alignment methods, improving generation quality and distributional coverage. Furthermore, while existing alignment methods provide only limited additional gains in the fine-tuning regime, contextual supervision remains highly effective when adapting an already extensively trained model to a specific dataset.

Overall, our work establishes the contextual space as a semantically rich representation that can both expose the internal dynamics of multimodal generation and provide an effective target for improving generative models beyond the standard diffusion objective.

2 Related Work

Multimodal Diffusion Transformers. Diffusion models have traditionally relied on convolutional U-Net architectures as their denoising backbone (Saharia et al., 2022; Ramesh et al., 2022; Rombach et al., 2022; Podell et al., 2023). Transformer-based denoisers later offered a more scalable

alternative (Bao et al., 2023; Peebles & Xie, 2023), leading to increasingly large transformer-based text-to-image models (Chen et al., 2024b;a). Stable Diffusion 3 (Esser et al., 2024) took this development further by introducing the Multimodal Diffusion Transformer (MM-DiT), in which textual and visual representations jointly participate in attention and continuously update one another. This bidirectional interaction has since become a prominent design in large-scale image generation and editing models, including FLUX (Labs, 2024; Labs et al., 2025), Qwen-Image (Wu et al., 2025; Zhao et al., 2026), and others (Zhuo et al., 2024; Cai et al., 2025; Xiao et al., 2025).

The same paradigm has become important in video generation, where CogVideoX (Yang et al., 2025) and HunyuanVideo (Kong et al., 2024) jointly process textual and spatiotemporal representations. Other models extend joint attention to appearance and motion (Chefer et al., 2025) or audio and video (Wang et al., 2026b; 2025; Shan et al., 2025). Together, these developments establish joint multimodal attention as a general mechanism for evolving representations across modalities.

Understanding Latent Representations in Diffusion Models A growing body of work has investigated the internal representations learned by diffusion models. Studies of visual features have shown that intermediate representations encode the semantics of individual patches and the correspondence between them, enabling a range of downstream applications (Tumanyan et al., 2023; Tang et al., 2023; Patashnik et al., 2023; Cao et al., 2023; Alaluf et al., 2024; Dahary et al., 2024; 2025; Sella et al., 2025; Luo et al., 2024). Other studies examine textual representations in text-to-image models, studying how semantic information is encoded and distributed across textual tokens (Rassin et al., 2022; Toker et al., 2024; Kaplan et al., 2026; Zhuang et al., 2024).

With the emergence of MM-DiTs, recent work has also begun to examine the evolving text-side representations formed through multimodal attention. ConceptAttention (Helbling et al., 2025) uses these representations to localize concepts in the generated image, while Contextual Repulsion (Dahary et al., 2026) demonstrates their utility as an intervention space for increasing generation diversity. Several works use causal interventions to reveal functional roles of contextual tokens under specific conditioning settings. Padding Tone (Toker et al., 2025) shows that, under textual conditioning, contextualized padding tokens can contribute additional visual information to the generated image; Vision-Language Binding (Ge et al., 2026) shows that contextual tokens can acquire semantic information from a reference image in FLUX.2 editing; Li et al. (Li et al., 2026) show that LLM-encoded template tokens in Qwen-Image can act as implicit registers for maintaining object identity. Altogether, these interventions demonstrate that contextual tokens can carry information relevant to generation. However, they do not directly reveal the semantic content of the contextual space or how it evolves over denoising. We instead decode this space into natural language, revealing a surprisingly broad range of image-specific semantics encoded within it. Remarkably, this information can be decoded in the absence of textual conditioning and even before it becomes observable in the visual prediction itself.

Representation Alignment for Generation. Recent work has shown that explicitly encouraging diffusion models to learn stronger internal representations can substantially improve both training efficiency and generation quality. REPA (Yu et al., 2024) introduced this direction by aligning intermediate visual features with representations from a pretrained vision encoder. Subsequent works extend alignment to attention features (Wang et al., 2026c), the latent space (Leng et al., 2025; Zheng et al., 2026), representations derived from the diffusion model itself (Jiang et al., 2025; Chefer et al., 2026), and text–image contrastive objectives (Lee et al., 2026). Recent analyses have also examined the role of local and global information in representation alignment. iREPA (Singh et al., 2025) shows that, for spatial visual features, alignment benefits primarily from preserving and enhancing spatial structure rather than global semantics. REG (Wu et al., 2026) instead shows that global semantic information can benefit class-conditional generation when represented separately through a high-level semantic token alongside the image latents.

We extend these insights to text-conditioned generation and find that global visual semantics are implicitly encoded in the contextual tokens. We therefore align them with a global semantic representation, complementing the local, spatial representations reinforced by existing alignment methods.

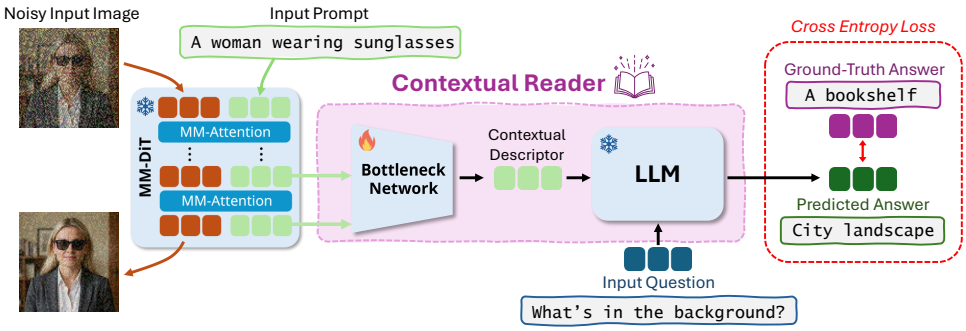

Figure 2: **Training the Contextual Reader.** At a selected denoising timestep, contextual tokens from multiple frozen MM-DiT layers are aggregated by a trainable Bottleneck Network into a compact Contextual Descriptor. A frozen LLM receives this descriptor together with a visual question and predicts an answer about the emerging image. We train only the Bottleneck Network using cross-entropy against the ground-truth answer.

3 Reading the Contextual Space

We introduce the *Contextual Reader*, a framework for decoding the internal contextual representations of Multimodal Diffusion Transformers (MM-DiTs) into natural language. We formulate this as a visual question answering task. At a selected timestep during generation, the reader receives the contextual representations produced by all MM-DiT layers together with a question about the emerging image. The question can probe the image at different levels of detail, including properties that are not explicitly specified by the prompt. The reader then generates a textual answer using the contextual representation and the question. This allows us to examine which visual semantics are encoded in the contextual space as generation progresses.

3.1 Reader Architecture

Figure 2 illustrates the architecture and training procedure of our Contextual Reader. The reader has two main components. First, a *Bottleneck Network* aggregates contextual tokens from multiple MM-DiT layers into a small set of tokens, which we call the *Contextual Descriptor*. Second, a frozen LLM receives this descriptor together with the visual question and generates the final answer.

Multi-Layer Contextual Extraction. At diffusion timestep t, each block l produces a sequence of contextual tokens $c^{(l)}(t) \in \mathbb{R}^{M \times d}$, where M is the text sequence length, including padding tokens, and d is the model hidden dimension. We aggregate the contextual tokens from all L transformer blocks. To retain information about where and when each representation was extracted, we augment each sequence with learnable embeddings of its layer and timestep. We then concatenate these augmented tokens along the sequence dimension to form $C(t) \in \mathbb{R}^{LM \times d}$. For full details on the feature extraction, see Appendix A.1.1.

Bottleneck Network and Language Interface. Our language interface is inspired by LLaVA (Liu et al., 2023), which connects an external representation to a language model for visual question answering through a learned projection. We adopt a similar interface for the contextual space. We use a lightweight Bottleneck Network to compress $C(t)$ into a small set of tokens. The bottleneck consists of a 2-layer Q-Former with Q learnable query tokens, where $Q \ll L \cdot M$, followed by a linear projection into the input embedding space of the frozen LLM.

The Q-Former queries attend to $C(t)$ through cross-attention, allowing them to aggregate information across layers and contextual-token positions. A linear projection maps the resulting query features to the embedding dimension of the frozen LLM, producing the *Contextual Descriptor* $p \in \mathbb{R}^{Q \times d_{\text{LLM}}}$. The descriptor is provided to the LLM together with the tokenized question. The LLM then generates the answer autoregressively. For full architectural details, see Appendix A.1.2.

3.2 Training the Contextual Reader

The Contextual Reader is trained to decode semantic properties of the corresponding final image from contextual representations at arbitrary denoising timesteps. Each training sample therefore consists of a contextual representation, a visual question, and a target answer for the corresponding generated image. We construct this supervision directly from the MM-DiT generation process.

Dataset Construction. To define a consistent set of questions that are meaningful across generations, we focus on images containing a single human subject. We curate 15 questions spanning broad scene properties, such as camera angle and background, as well as finer-grained properties, such as pose, appearance, and object interaction.

To generate the images, we use a subset of human-centric, single-subject prompts from MS-COCO (Lin et al., 2014). To encourage the reader to distinguish between different outcomes under the same textual conditioning, we generate five seeds for each prompt. Our dataset comprises 10,000 generations from 2,000 training prompts and 2,500 held-out generations from 500 prompts. For each generated image, an off-the-shelf Vision-Language Model provides answers to the 15 questions, which serve as the target supervision for the reader. For each generated image, we also store the prompt and initial noise used for generation, allowing us to retrieve its contextual representations at arbitrary denoising timesteps during training. The full list of questions and additional dataset construction details are provided in Appendix A.1.3.

Training. During reader training, we sample a denoising timestep and a generated image from our dataset, together with its associated questions and answers. Using the stored prompt and initial noise, we reconstruct the noisy latent at the sampled timestep and run the frozen MM-DiT to obtain the contextual representations $C(t)$. The reader then produces answers to the corresponding visual questions, which are supervised against the ground-truth answers using an autoregressive cross-entropy loss. We optimize only the Bottleneck Network, with all other components kept frozen. We train the Reader for 5,000 iterations with a batch size of 256; the validation loss decreases steeply during the first 2,000 iterations, with only modest improvements thereafter. Training details and validation curves are provided in Appendix A.1.4.

3.3 Semantic Readability of the Contextual Space

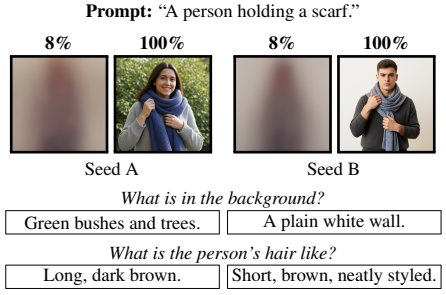

Figure 3: **Seed-specific semantic information from contextual tokens.** Two random seeds of the same prompt resolve underspecified attributes differently. The reader recovers seed-specific information from contextual tokens at the 8% timestep, even when the corresponding visual predictions remain ambiguous.

Having established a mechanism for reading the contextual tokens, we now examine what this representation reveals about the emerging generation. We study this across two pretrained MM-DiTs with different architectures: SD3.5 (Esser et al., 2024),which maintains separate image and text streams while allowing them to exchange information through multimodal attention, and FLUX.2 (Chefer et al., 2026), which additionally includes single-stream blocks that jointly process the concatenated image and text tokens.

To study how access to the prompt affects what can be decoded from the contextual space, we consider two conditioning settings. In one, the MM-DiT receives the original prompt (*conditional setting*), allowing us to study how the contextual tokens become enriched with image-specific information beyond what the prompt specifies. In the other, the MM-DiT receives an empty prompt (*unconditional setting*), allowing us to examine how much image-specific information can be recovered without access to the prompt.

Qualitative Analysis. Figure 3 shows examples from the conditional setting, where two random seeds are generated from the same prompt. Although the textual conditioning is identical, the resulting images resolve underspecified properties differently, and the reader recovers these seed-specific differences from the contextual tokens. For example, for the prompt "A person holding a scarf,"

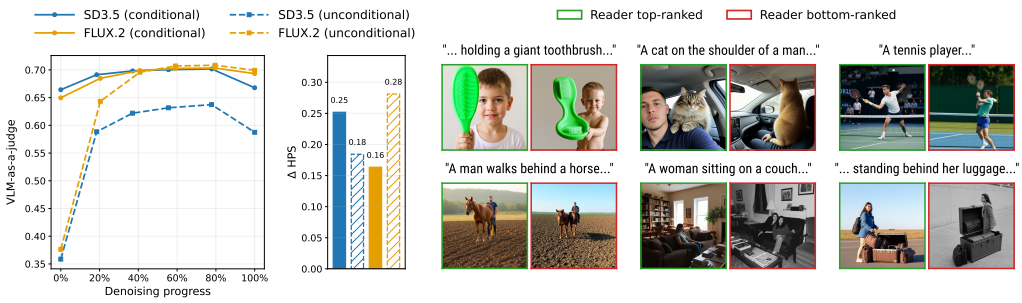

Figure 5: **Contextual readability emerges early and reflects final generation quality. Left:** VLM-as-a-judge evaluation across the denoising trajectory shows that contextual scores increase as generation progresses; with the input prompt they are already high from the start. With an empty prompt they rise rapidly as image-specific information enters the contextual space. **Middle:** Ranking seeds by reader scores separates generations with different HPS scores; ΔHPS denotes the HPS difference between the top- and bottom-ranked seeds. **Right:** Representative top- and bottom-ranked generations illustrate that higher contextual readability is associated with stronger final generations.

the reader identifies that the background is "green bushes and trees" in one seed, and "a plain white wall" in the other. These distinctions can often be decoded at an early denoising step, before they are clearly visible in the corresponding visual prediction.

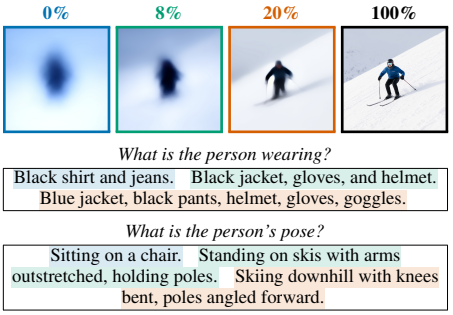

Figure 4: **Progressive semantic decoding from contextual tokens.** With an empty prompt, the reader recovers increasingly specific semantic information as denoising progresses, from coarse properties at early timesteps to finer-grained details at later timesteps. Colored highlights indicate the denoising timestep.

Figure 4 removes the prompt altogether and follows the reader over the denoising trajectory. Even without textual conditioning, highly specific semantics can be decoded early in generation: the reader recognizes that the person is skiing by 8% of the trajectory, and by 20% identifies that they are wearing goggles, before these details are clearly observable in the corresponding visual predictions. Since the model receives no prompt in this setting, these semantics can only arise from the interaction between the contextual tokens and the evolving visual representation.
Together, these examples show that the contextual tokens can expose surprisingly specific, generation-dependent semantics at early stages of denoising, often before those semantics are clearly expressed in the model's current visual prediction. Additional results are in Appendix B.1.3.

Quantifying Contextual Readability. We next quantify the behavior observed qualitatively in Figures 3 and 4. At each timestep, we use VLM-as-a-judge (Lee et al., 2024) to compare the reader's free-form answer with the ground-truth answer and assess its semantic correctness and specificity (see Appendix A.1.5 for details and Appendices B.1.1 and B.1.2 for additional quantitative analyses). The left panel of Figure 5 shows a consistent trend across both architectures. With the input prompt, the contextual representation is informative from the beginning of generation, but its readability continues to increase throughout denoising, reaching its highest value at 80%. Since some questions can already be partially answered from the prompt, this continued improvement indicates that the contextual tokens progressively capture attributes left underspecified by the textual conditioning. Readability then shows a small decrease at the final 100% timestep, as generation shifts toward refining high-frequency details rather than the semantic properties probed by our questions.

With an empty prompt, performance starts substantially lower and rises rapidly as denoising progresses. Notably, it eventually surpasses the conditional reader for FLUX.2 and approaches the conditional performance for SD3.5, showing that the contextual tokens can become highly seman-

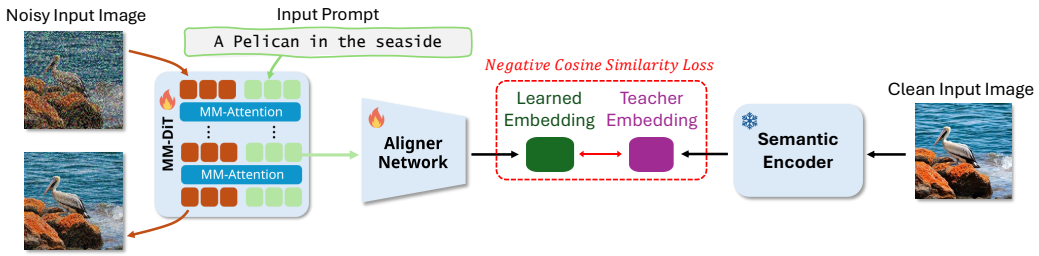

Figure 6: **Contextual Alignment.** A lightweight Aligner Network maps contextual tokens from the MM-DiT into a learned embedding, optimized via a negative cosine similarity loss to match the teacher embedding produced by a frozen semantic encoder from the clean input image.

tically informative even without direct access to the prompt. These semantics emerge despite the contextual tokens never being explicitly supervised to represent visual content. Because they arise solely under the denoising objective, this suggests that such semantic encoding is useful to the generation process itself. We therefore examine whether the clarity of this implicit representation is also related to the quality of the resulting image.

The middle and right panels of Figure 5 address this relationship. For each prompt, we first rank the five generations according to their reader score averaged across denoising timesteps. We then compare the HPS scores (Ma et al., 2025) of the top- and bottom-ranked generations. The middle panel reports ΔHPS, defined as the HPS score of the reader top-ranked generation minus that of the bottom-ranked generation. Across architectures and conditioning settings, ΔHPS is consistently positive, indicating that generations with higher contextual readability tend to receive higher HPS scores. The right panel provides qualitative examples of this relationship: for each prompt, we show the top-ranked generation in green and the bottom-ranked generation in red. The top- and bottom-ranked generations often differ in both visual fidelity and prompt alignment. For example, the bottom-ranked samples contain visible artifacts, such as a malformed toothbrush or tennis racket, and can also misinterpret the prompt, showing a person riding rather than walking behind a horse, or two women sitting on a couch rather than one. This indicates that contextual readability provides a meaningful signal for identifying higher-quality final generations. We therefore turn, in the next section, to explicitly supervising the visual-semantic information encoded in the contextual tokens.

4 Supervising the Contextual Space

The previous section reveals that contextual tokens encode rich, global visual-semantic information about the emerging image, and that the clarity of this information is related to the quality of the resulting generation. We therefore seek to make this implicit semantic structure an explicit training signal. Representation alignment provides a natural mechanism for doing so: a pretrained vision encoder provides a target representation, and the diffusion model is trained to align its intermediate visual features with this target (Yu et al., 2024; Chefer et al., 2026; Wang et al., 2026c). These methods primarily align spatial visual representations, and have been shown to benefit from spatially structured teacher representations rather than global semantic ones (Singh et al., 2025). Since our analysis reveals that contextual tokens encode global visual semantics, we instead use a global semantic teacher to supervise them, providing a complementary form of representation alignment. We therefore introduce **Contextual Alignment** (CoAl), which aligns contextual tokens with a global semantic representation of the corresponding clean image.

4.1 Contextual Alignment

Our training process is illustrated in Figure 6. During training, a frozen semantic encoder maps the clean image to a global teacher embedding, and a lightweight network, the *Aligner Network*, maps the contextual tokens from a single layer to a predicted embedding. The embedding is then aligned with the teacher via a negative cosine similarity loss. We incorporate this loss into the diffusion objective and optimize the MM-DiT and the Aligner Network jointly; the Aligner Network is used exclusively during training, leaving the MM-DiT's inference cost unchanged. Seeking a global, semantic, and language-aligned teacher, we use the pooled SigLIP image embedding (Zhai et al.,

2023), which SigLIP trains to align with caption embeddings through a contrastive image-text loss, making it a compact summary of the image. The Aligner Network itself is a small Q-Former, as in the Contextual Reader, but kept even more minimal: a single layer with one learned query. Its limited capacity makes closing the gap to the teacher depend on adaptation of the MM-DiT, rather than on the Aligner Network absorbing the alignment objective. Likewise, by aligning a single, early layer, we encourage the MM-DiT to encode the teacher's semantics in the contextual tokens early on, leaving the remaining layers free to refine and use them as needed. Both our choice of teacher embedding and our layer selection are confirmed empirically (Appendix B.2.1).

Null-Condition Weighting When training text-to-image diffusion models, the text prompt is occasionally dropped to support classifier-free guidance (CFG) at inference (Ho & Salimans, 2022). When the prompt is provided to the model, the alignment objective has an easy shortcut: the contextual tokens can reduce the alignment loss by leveraging information already provided by the prompt, rather than capturing image-specific semantics that the prompt leaves unspecified. On null-condition iterations, however, the tokens are initialized from an empty prompt, closing off this shortcut, meaning any alignment must instead be drawn through interaction between the contextual and image tokens. We therefore upweight the alignment loss specifically on null-condition iterations, encouraging a flow of semantic information from the image tokens into the contextual tokens. In practice, we apply this upweighting only in the later stages of full model training, or throughout fine-tuning.

4.2 Experimental Setup and Baselines

Training a generator from random initialization is the standard protocol for evaluating alignment methods, but for many practitioners the realistic use case is to adapt a pretrained generator to a dataset that reflects what they want to generate. We therefore evaluate CoAl in two complementary regimes. First, we train a dual-stream MM-DiT from random initialization on MS-COCO 2014 (Lin et al., 2014). Second, we fine-tune SD3 (Esser et al., 2024), a pretrained dual-stream MM-DiT, on the curated subset of Fine-T2I (Ma et al., 2026), a dataset specifically designed for T2I fine-tuning that pairs professional photographs with long, detailed prompts. This tests whether contextual supervision remains effective after extensive pretraining, when adapting to a new data distribution.

We compare against REPA (Yu et al., 2024) and its later variants, HASTE (Wang et al., 2026c) and SRA (Jiang et al., 2025). For full model training, we test whether CoAl complements these visual alignment methods, comparing each objective with and without CoAl. For fine-tuning, however, REPA and its variants offer no meaningful gains over the vanilla denoising objective, so we instead combine CoAl with vanilla flow matching, including REPA, HASTE, and SRA for completeness. In the main manuscript, we report FID (Heusel et al., 2017), CLIP (Radford et al., 2021), HPS (Wu et al., 2023), and DINOv2 (Oquab et al., 2023) Precision and Recall (Kynkäänniemi et al., 2019). For fine-tuning, we omit CLIP and HPS. These metrics are trained on broad, general-purpose distributions of images, captions, and preferences, so they capture what users prefer on average, which need not coincide with the characteristics of a particular fine-tuning dataset. Indeed, the ground-truth photographs score below the base model on CLIP and HPS, and therefore comparing along these axes is ill-posed. We instead evaluate how well the generated images fit and cover the target distribution, using FID, KID, and DINOv2 Precision and Recall. Our full result tables, including additional metrics and baselines, are provided in Appendix B.2.2.

4.3 Main Results

Quantiative results. Table 1a reports results for training a new model from random weights. Contextual Alignment improves all three representation-alignment baselines across all evaluated metrics, further supporting the complementary role of global semantic alignment in the contextual space. Table 1b reports results for our fine-tuning experiment. In this setting, REPA, HASTE, and SRA provide no meaningful gains over vanilla fine-tuning, with all three achieving worse FID than this baseline. Contextual Alignment continues to improve generation quality, achieving the best FID and Recall, while remaining competitive in Precision. This demonstrates the effectiveness of contextual semantic supervision for fine-tuning pretrained generators.

Qualitative results. We present qualitative comparisons with the baseline in Figure 7. These results highlight our method's ability to reduce malformed limbs (first row) and mitigate oversatu-

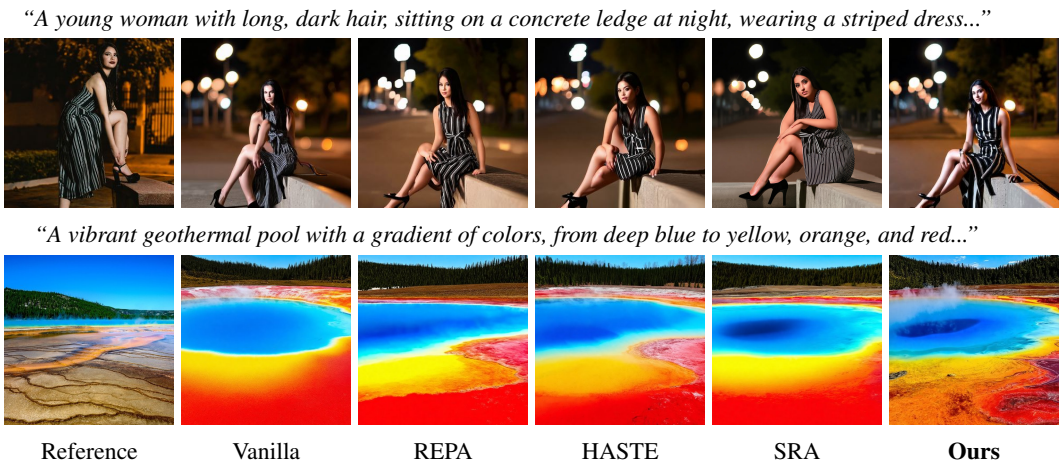

Figure 7: Qualitative comparison against visual alignment methods over SD3-Medium fine-tuning on the Fine-T2I curated dataset. The "Reference" column presents the held out images of the Fine-T2I dataset matching the prompts presented above.

Table 1: Quantitative comparisons with representation-alignment baselines.

Method	FID↓	CLIP$\times 10^2$↑	HPS$\times 10^2$↑	Prec.↑	Rec.↑
REPA	4.95	24.20	21.38	0.591	0.338
+ Ours	**4.29**	**24.40**	**21.60**	**0.628**	**0.407**
HASTE	4.87	24.21	21.42	0.609	0.361
+ Ours	**4.44**	**24.32**	**21.56**	**0.623**	**0.400**
SRA	5.89	23.49	20.37	0.515	0.261
+ Ours	**4.36**	**24.28**	**21.42**	**0.616**	**0.398**

(a) Full model training.

Method	FID↓	Prec.↑	Rec.↑
Base Model	19.30	0.993	0.939
Vanilla Flow	17.18	0.993	0.966
REPA	17.45	0.993	0.966
HASTE	17.53	**0.994**	0.963
SRA	17.48	0.991	0.964
Ours	**16.75**	0.993	**0.969**

(b) Model fine-tuning.

ration while recovering sharper, high-frequency details (second row). We present additional results in Appendix B.2.4.

5 Conclusions

We have presented a framework for reading the internal perceptual representations of generative models through natural language. More broadly, our work suggests that a generative model knows surprisingly early what it is going to generate. Semantic details of the final image are already present internally before they become perceptually apparent in the emerging image. Even more strikingly, this remains true without the text prompt: the model develops a rich semantic picture of the emerging image through the generation process itself. This observation also has a useful consequence: generations in which this internal representation is clearer tend to result in better images. This suggests that the representation is not only something we can read, but also something worth strengthening. Our Contextual Alignment does exactly this, leading to improved generations.

Our findings also highlight how much remains to be understood about the internal process of generative models. We can now read some of the semantic information that emerges during generation, but how these internal representations form, evolve, and ultimately shape the generated image remains largely unexplored. The fact that meaningful information appears so early, even before it becomes perceptually apparent, raises many new questions about the generation process itself. We believe that further studying these internal representations can deepen our understanding of generative models and reveal new opportunities for controlling and improving them.

AI USE STATEMENT

AI assistants supported implementation development, evaluation scripts, literature searches, manuscript drafting and editing, and figure preparation. The authors are responsible for verifying the implementation, references, experimental results, and final text, and retain full responsibility for the work. Model-generated images are the primary experimental data studied in this work.

ETHICS STATEMENT

Our work uses publicly available image datasets and pretrained generative and vision-language models. We do not collect or release personally identifiable information. The generated images and annotations used for training and evaluation are used solely for research purposes. As with other generative models, our method may inherit biases and undesirable behaviors from the underlying pretrained models and datasets.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the Contextual Reader and Contextual Alignment, including model architectures, training procedures, datasets, generation settings, evaluation protocols, and hyperparameters in Appendix A.

REFERENCES

Yuval Alaluf, Daniel Garibi, Or Patashnik, Hadar Averbuch-Elor, and Daniel Cohen-Or. Cross-image attention for zero-shot appearance transfer. In *ACM SIGGRAPH 2024 conference papers*, pp. 1–12, 2024.

Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18187–18197. IEEE, 2022.

Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A vit backbone for diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22669–22679. IEEE, 2023.

Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

Qi Cai, Jingwen Chen, Yang Chen, Yehao Li, Fuchen Long, Yingwei Pan, Zhaofan Qiu, Yiheng Zhang, Fengbin Gao, Peihan Xu, et al. Hidream-i1: A high-efficient image generative foundation model with sparse diffusion transformer. *arXiv preprint arXiv:2505.22705*, 2025.

Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22503–22513. IEEE, 2023.

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.

Hila Chefer, Uriel Singer, Amit Zohar, Yuval Kirstain, Adam Polyak, Yaniv Taigman, Lior Wolf, and Shelly Sheynin. Videojam: Joint appearance-motion representations for enhanced motion generation in video models. *arXiv preprint arXiv:2502.02492*, 2025.

Hila Chefer, Patrick Esser, Dominik Lorenz, Dustin Podell, Vikash Raja, Vinh Tong, Antonio Torralba, and Robin Rombach. Self-supervised flow matching for scalable multi-modal synthesis. *arXiv preprint arXiv:2603.06507*, 2026.

Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *European Conference on Computer Vision*, pp. 74–91. Springer, 2024a.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *International conference on learning representations*, volume 2024, pp. 57611–57640, 2024b.

Omer Dahary, Or Patashnik, Kfir Aberman, and Daniel Cohen-Or. Be yourself: Bounded attention for multi-subject text-to-image generation. In *European Conference on Computer Vision*, pp. 432–448. Springer, 2024.

Omer Dahary, Yehonathan Cohen, Or Patashnik, Kfir Aberman, and Daniel Cohen-Or. Be decisive: Noise-induced layouts for multi-subject generation. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–12, 2025.

Omer Dahary, Benaya Koren, Daniel Garibi, and Daniel Cohen-Or. On-the-fly repulsion in the contextual space for rich diversity in diffusion transformers. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–11, 2026.

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.

Chris Ge, Rohit Gandikota, Antonio Torralba, and Tamar Rott Shaham. Vision-language binding in in-context image generation. In *Mechanistic Interpretability Workshop at ICML 2026*, 2026.

Alec Helbling, Tuna Han Salih Meral, Ben Hoover, Pinar Yanardag, and Duen Horng Chau. Conceptattention: Diffusion transformers learn highly interpretable features. *arXiv preprint arXiv:2502.04320*, 2025.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

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.

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

Dengyang Jiang, Mengmeng Wang, Liuzhuozheng Li, Lei Zhang, Haoyu Wang, Wei Wei, Guang Dai, Yanning Zhang, and Jingdong Wang. No other representation component is needed: Diffusion transformers can provide representation guidance by themselves. *arXiv preprint arXiv:2505.02831*, 2025.

Guy Kaplan, Michael Toker, Yuval Reif, Yonatan Belinkov, and Roy Schwartz. Follow the flow: On information flow across textual tokens in text-to-image models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 34139–34157, 2026.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems*, 32, 2019.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Jaa-Yeon Lee, Byunghee Cha, Jeongsol Kim, and Jong Chul Ye. Aligning text to image in diffusion models is easier than you think. *Advances in Neural Information Processing Systems*, 38:157106–157136, 2026.

Seongyun Lee, Seungone Kim, Sue Park, Geewook Kim, and Minjoon Seo. Prometheus-vision: Vision-language model as a judge for fine-grained evaluation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 11286–11315, 2024.

Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. Repa-e: Unlocking vae for end-to-end tuning with latent diffusion transformers. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18262–18272. IEEE, 2025.

Maohua Li, Qirui Li, Yanke Zhou, Yiduo Li, Zhaosheng Chi, Chao Xu, Cuifeng Shen, Yixuan Xu, Hanlin Tang, Kan Liu, et al. Text template tokens are implicit semantic registers in diffusion transformers. *arXiv preprint arXiv:2607.19139*, 2026.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

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. In *European Conference on Computer Vision*, pp. 366–384. Springer, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *ArXiv*, abs/2304.08485, 2023.

Grace Luo, Trevor Darrell, Oliver Wang, Dan B Goldman, and Aleksander Holynski. Readout guidance: Learning control from diffusion features. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8217–8227. IEEE, 2024.

Xu Ma, Yitian Zhang, Qihua Dong, and Yun Fu. Fine-t2i: An open, large-scale, and diverse dataset for high-quality t2i fine-tuning. *arXiv preprint arXiv:2602.09439*, 2026.

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15086–15095. IEEE, 2025.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Or Patashnik, Daniel Garibi, Idan Azuri, Hadar Averbuch-Elor, and Daniel Cohen-Or. Localizing object-level shape variations with text-to-image diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22994–23004. IEEE, 2023.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

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.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

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*, pp. 8748–8763. PmLR, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Royi Rassin, Shauli Ravfogel, and Yoav Goldberg. Dalle-2 is seeing double: Flaws in word-to-concept mapping in text2image models. In *Proceedings of the Fifth BlackboxNLP workshop on analyzing and interpreting neural networks for NLP*, pp. 335–345, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

Etai Sella, Yanir Kleiman, and Hadar Averbuch-Elor. Instancegen: Image generation with instance-level instructions. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–10, 2025.

Sizhe Shan, Qiulin Li, Yutao Cui, Miles Yang, Yuehai Wang, Qun Yang, Jin Zhou, and Zhao Zhong. Hunyuanvideo-foley: Multimodal diffusion with representation alignment for high-fidelity foley audio generation. *arXiv preprint arXiv:2508.16930*, 2025.

Jaskirat Singh, Xingjian Leng, Zongze Wu, Liang Zheng, Richard Zhang, Eli Shechtman, and Saining Xie. What matters for representation alignment: Global information or spatial structure? In *The Fourteenth International Conference on Learning Representations*, 2025.

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.

Michael Toker, Hadas Orgad, Mor Ventura, Dana Arad, and Yonatan Belinkov. Diffusion lens: Interpreting text encoders in text-to-image pipelines. *arXiv preprint arXiv:2403.05846*, 2024.

Michael Toker, Ido Galil, Hadas Orgad, Rinon Gal, Yoad Tewel, Gal Chechik, and Yonatan Belinkov. Padding tone: A mechanistic analysis of padding tokens in t2i models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7618–7632, 2025.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1921–1930. IEEE, 2023.

Jun Wang, Xijuan Zeng, Chunyu Qiang, Ruilong Chen, Shiyao Wang, Le Wang, Wangjing Zhou, Pengfei Cai, Jiahui Zhao, Nan Li, et al. Kling-foley: Multimodal diffusion transformer for high-quality video-to-audio generation. *arXiv preprint arXiv:2506.19774*, 2025.

Kaibo Wang, Jianda Mao, Tong Wu, and Yang Xiang. Towards a golden classifier-free guidance path via foresight fixed point iterations. *Advances in Neural Information Processing Systems*, 38: 124994–125030, 2026a.

Le Wang, Jun Wang, Chunyu Qiang, Feng Deng, Chen Zhang, and Kun Gai. Audiogen-omni: A unified multimodal diffusion transformer for video-synchronized audio, speech, and song generation. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 16047–16051. IEEE, 2026b.

Ziqiao Wang, Wangbo Zhao, Yuhao Zhou, Zekai Li, Zhiyuan Liang, Mingjia Shi, Xuanlei Zhao, Pengfei Zhou, Kaipeng Zhang, Zhangyang Wang, et al. Repa works until it doesn't: Early-stopped, holistic alignment supercharges diffusion training. *Advances in Neural Information Processing Systems*, 38:136854–136887, 2026c.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Ge Wu, Shen Zhang, Ruijing Shi, Shanghua Gao, Zhenyuan Chen, Lei Wang, Zhaowei Chen, Hongcheng Gao, Yao Tang, Ming-Ming Cheng, et al. Representation entanglement for generation: Training diffusion transformers is much easier than you think. *Advances in Neural Information Processing Systems*, 38:7714–7743, 2026.

Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Better aligning text-to-image models with human preference. *arXiv preprint arXiv:2303.14420*, 1(3), 2023.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13294–13304. IEEE, 2025.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025.

Shai Yehezkel, Omer Dahary, Andrey Voynov, and Daniel Cohen-Or. Navigating with annealing guidance scale in diffusion space. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–11, 2025.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3813–3824. IEEE, 2023.

Bing Zhao, Chenfei Wu, Deqing Li, Hao Meng, Jiahao Li, Jie Zhang, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kuan Cao, et al. Qwen-image-2.0 technical report. *arXiv preprint arXiv:2605.10730*, 2026.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *International Conference on Learning Representations*, volume 2026, pp. 35791–35820, 2026.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Chenyi Zhuang, Ying Hu, and Pan Gao. Magnet: We never know how text-to-image diffusion models work, until we learn how vision-language models function. *Advances in Neural Information Processing Systems*, 37:57115–57149, 2024.

Le Zhuo, Ruoyi Du, Han Xiao, Yangguang Li, Dongyang Liu, Rongjie Huang, Wenze Liu, Xiangyang Zhu, Fu-Yun Wang, Zhanyu Ma, et al. Lumina-next: Making lumina-t2x stronger and faster with next-dit. *Advances in Neural Information Processing Systems*, 37:131278–131315, 2024.

APPENDIX

A IMPLEMENTATION DETAILS

A.1 CONTEXTUAL READER

A.1.1 CONTEXTUAL FEATURE EXTRACTION AND ENCODING

We extract contextual representations from the attention operator of the frozen MM-DiT at the selected diffusion timestep. Specifically, we extract the output of the attention operator before the subsequent block-specific attention gate, residual addition, normalization, or MLP update. We retain only the contextual/text positions and exclude image-token positions.

For SD3.5, we capture the text-side output of joint attention from all transformer blocks. For FLUX.2, we capture the text-side output from all dual-stream blocks and the text positions from all single-stream blocks.

For the conditional reader, we use the contextual representations produced by the positive-prompt branch. For the unconditional reader, we use the model's empty-prompt setting. No image tokens are supplied to the reader.

Because the reader aggregates contextual representations from multiple MM-DiT layers, we add a learned layer embedding to the contextual tokens from each layer before concatenating them. We also provide the denoising timestep to the reader. The continuous noise level σ is encoded using a 256-dimensional sinusoidal embedding followed by an MLP with dimensions $256 \to 256 \to D$ and SiLU activation, where D is the hidden dimension of the corresponding diffusion backbone. The resulting timestep embedding is added to every contextual token. We do not average or otherwise mix the representations across layers before the Q-Former; all layer-token pairs remain independently accessible.

A.1.2 BOTTLENECK NETWORK AND LANGUAGE-MODEL INTERFACE

The Bottleneck Network is a two-block, eight-head Q-Former with 16 learned queries. Its hidden dimension matches that of the corresponding diffusion backbone. Each Q-Former block uses a cross-first ordering. The learned queries first attend to all contextual tokens through pre-normalized cross-attention, followed by a residual connection. This is followed by pre-normalized self-attention among the queries and a residual connection, and then a pre-normalized feed-forward network with dimensions $D \to 2048 \to D$, GELU activation, and a residual connection. A final layer normalization is applied after the second block. No dropout is used.

The Q-Former outputs are mapped individually into the input embedding space of the frozen Qwen2.5-7B-Instruct (Qwen et al., 2025) language model using a learned affine projection. Each projected token is L2-normalized and rescaled to the mean norm of the LLM's vocabulary embeddings, excluding the padding token.

The resulting Contextual Descriptor is provided to the LLM as a sequence of 16 continuous prefix embeddings, followed by the textual question and its answer-format instruction. The original generation prompt is not provided to the LLM.

During teacher-forced training, the target consists of the ground-truth answer and the tokenizer EOS token. Prefix and question positions are masked, and the cross-entropy loss is computed only over answer tokens.

1. What is the person's apparent gender presentation?
2. What age group does the person appear to belong to?
3. What is the person wearing?
4. What is the person's hair like?
5. Where is the person located in the image?
6. What is the person's detailed pose?
7. Which direction is the person facing?
8. What is the person doing?
9. What is in the background?
10. What is the camera angle?
11. What object is the person holding or using?
12. Which parts of the person are visible in the image?
13. How large is the person relative to the full image?
14. What is the overall color scheme of the image?
15. What objects are visible in the image?

Figure 8: **Visual questions used by the Contextual Reader.** We manually curate 15 questions spanning diverse semantic aspects of human-centric images.

A.1.3 Dataset Construction

We construct the dataset for training and evaluating the Contextual Reader from the 2014 MS-COCO training split. We first restrict the source images to those containing exactly one non-crowd person instance according to the COCO instance annotations. We then filter the associated captions based on their textual content: each caption is classified by the text-only model Qwen2.5-7B-Instruct (Qwen et al., 2025) as describing either zero, one, multiple, or an ambiguous number of real people. We retain captions classified as describing exactly one real person without ambiguity. This procedure produces 62,069 eligible captions.

From these captions, we select 2,500 prompts from distinct COCO source images using deterministic stratified sampling. We first select 500 prompts for the held-out set and then select 2,000 prompts from the remaining images, ensuring no source-image overlap between the two sets. Each prompt is rendered with five randomly assigned seeds, such that no seed is shared between prompts. This results in 10,000 training generations and 2,500 held-out generations per diffusion backbone.

We generate images at 1024×1024 resolution. SD3.5 uses 28 FlowMatch Euler denoising steps and guidance scale 7.0, while FLUX.2 uses 50 FlowMatch Euler steps and guidance scale 4.0. The same prompt-to-seed mapping is used for both backbones.

We manually curate 15 visual questions designed to span diverse semantic aspects of human-centric images, including appearance, pose, activity, spatial configuration, and scene properties. The complete set of questions is shown in Figure 8.

For each generated image, we obtain an answer to each question using InternVL3-8B (Zhu et al., 2025). The teacher receives only the generated image and the visual question, without access to the original generation prompt, and is instructed to answer using visible image evidence. Answers are generated independently using deterministic decoding. We automatically validate the generated responses and retry invalid answers up to two times, replacing answers that remain invalid with "unknown."

A.1.4 Reader Training

We train a separate reader for each diffusion backbone and conditioning setting. The diffusion backbone, text encoders, VAE, and LLM remain frozen throughout training. Only the Q-Former and the projection layer are optimized.

At each training iteration, we sample a continuous noise level $\sigma \sim U[0, 1]$, allowing the reader to observe contextual representations throughout the denoising trajectory. For each generated image, we store its final latent x_0 and corresponding initial noise ϵ, and construct the noisy latent as $z_t = (1 - \sigma)x_0 + \sigma\epsilon$. We then run the frozen diffusion model on this latent and extract the contextual

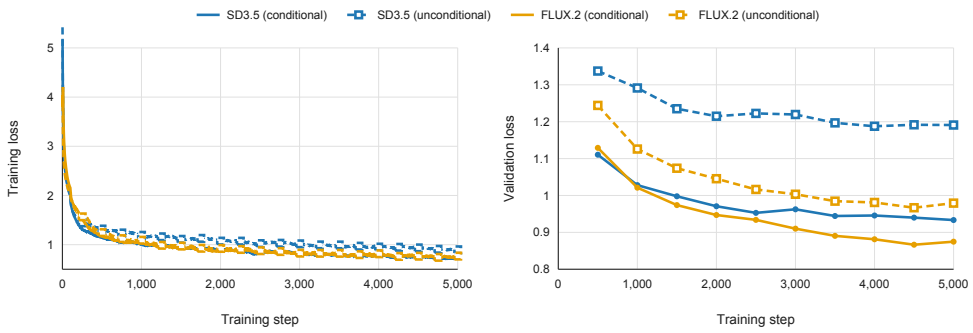

Figure 9: **Reader training and validation loss.** Training and validation loss over 5,000 optimization steps for the conditional and unconditional readers across both diffusion backbones. The validation set consists of 5% of the training data and is separate from the held-out data used for reader evaluation.

representations at the sampled noise level. Thus, contextual representations are extracted on the fly rather than stored for every timestep.

We train for 5,000 optimizer updates using AdamW with learning rate 10^{-4}, weight decay 10^{-4}, and gradient-norm clipping at 1.0. No learning-rate scheduler or warmup is used. Training uses bfloat16 mixed precision with an effective global batch size of 256. We use 8 NVIDIA A100 GPUs for training. Training takes approximately 4 hours for the SD3.5 reader and 8 hours for the FLUX.2 reader.

The resulting contextual representations are passed through the Bottleneck Network and LLM, and the reader is optimized using the autoregressive cross-entropy objective.

We additionally monitor training and validation loss during reader training, as shown in Figure 9. The validation loss decreases steeply until approximately 2,000 training steps, with further training yielding only modest additional reductions. The validation loss is computed on 5% held-out prompts from the training data, separate from the data used for reader evaluation.

A.1.5 VLM-AS-A-JUDGE DETAILS

We evaluate the reader's predictions using a VLM-as-a-judge. We use Prometheus-Vision 13B (Lee et al., 2024) to compare each reader prediction with the corresponding reference answer produced by InternVL3-8B. The judge receives the visual question, the reader's prediction, the reference answer, and the final generated image. The generation prompt is not provided to the judge. The exact evaluation prompt and scoring rubric are shown in Figure 10.

The judge assigns an integer score from 1 to 5 based on semantic correctness, completeness, and consistency with the image. A score of 5 indicates a semantically equivalent, fully correct, and image-consistent answer, while lower scores reflect increasing degrees of incorrectness, omission, ambiguity, or contradiction. Empty predictions receive a score of 1, and all other predictions are evaluated by Prometheus-Vision. The resulting scores are normalized to $[0, 1]$.

For each denoising timestep, we compute the reader's readability as the mean normalized score across all validation images and questions.

A.2 CONTEXTUAL ALIGNMENT

A.2.1 FEATURE EXTRACTION AND ALIGNER NETWORK ARCHITECTURE

Across architectures, contextual features are extracted in the same manner as in the Contextual Reader (Appendix A.1.1): tokens from the contextual indices are pooled immediately after the shared attention block. Unlike the Reader however, in CoAl we extract only the contextual features from a single MM-DiT layer l ($l = 4$ by default) and no positional embedding is added to it.

The Aligner network is made up of a Q-Former followed by a small MLP. The Q-Former is 24 head single-layer, single-query Q-Former which attends to the contextual features via pre-normalized cross-attention between the learnable query and the features. The MLP is made up of two linear layers and SiLU activation such that the dimension of the vector produced by the Q-Former is $D \to 1024 \to d_y$. With D being the hidden dimension of the contextual tokens (768 in the MM-DiT used in the full model training setting, 1536 in the SD3-Medium model used in the fine-tune setting). As explained in the main paper, the Aligner network, along with the MM-DiT, are optimized via a negative cosine similarity loss between the Aligner Network output and the teacher embedding.

A.2.2 Null-Condition Weighting

Null-Condition Weighting involves upweighting the CoAl loss for training samples in which the text condition has been dropped. In practice, this mechanism is activated after a set amount of training iterations i_{start} has passed. After i_{start} the CoAl loss is multiplied by a factor $w_{null} > 1.0$ for samples with dropped text conditioning and multiplied by $w_{text} < 1.0$ for text guided samples. These factors are applied on top of the overall CoAl loss weight w_{CoAl}. For full model training we use $w_{CoAl} = 1.3$, $i_{start} = 125{,}000$, $w_{null} = 2.0$, and $w_{text} = 1.0$; for fine-tuning we use $w_{CoAl} = 0.75$, $i_{start} = 0$, $w_{null} = 6.0$, and $w_{text} = 0.25$.

A.2.3 Teacher Embeddings

Our default teacher encoder is SigLIP, specifically the `google/siglip-base-patch16-256` model, producing 768-dimensional embeddings. In our ablation study (Appendix B.2.1), we additionally consider CLIP (`openai/clip-vit-large-patch14`), DINOv2 (`dinov2_vitb14`, the same encoder standard REPA itself uses as its own teacher), and Qwen3-VL (`Qwen/Qwen3-VL-Embedding-2B`).

A.2.4 Training Configurations

Full Model Training In the full model training setting the MM-DiT architecture we train is based on the Stable Diffusion 3 model (Esser et al., 2024), but is trained entirely from scratch (no pre-trained weights) at a reduced scale: 24 transformer layers, matching SD3's own depth, with a hidden dimension of 768 and 24 attention heads (32-dimensional per head), compared to SD3-Medium's 1536-dimensional hidden size at the same head count (64-dimensional per head). Training data is MS-COCO 2014 (Lin et al., 2014), downloaded from this source, using the official `train2014`/`val2014` split (82,783/40,504 images, each with ∼5 captions), at 256×256 resolution. We use the AdamW optimizer with a constant learning rate of 1×10^{-4}, no weight decay, and $\epsilon = 1 \times 10^{-8}$, with a global batch size of 256. We maintain an exponential moving average (EMA) of the model weights with decay 0.9999, initialized as an exact copy of the model at step 0, and report all results using the EMA weights. We train for 150,000 iterations using `fp16` mixed precision. Using 8 NVIDIA B200 GPUs training takes approximately 12.5 hours.

Fine-Tuning In the fine-tuning setting we fine-tune the full Stable Diffusion 3 Medium model (Esser et al., 2024) (all weights trainable, no LoRA or partial freezing), with 24 transformer layers, a hidden dimension of 1536, and 24 attention heads (64-dimensional per head). Training data is the curated subset of the Fine-T2I dataset (Ma et al., 2026), available at huggingface.co/datasets/ma-xu/fine-t2i, using 162,979 training images and a held-out set of 5,000 validation images, at 1024×1024 resolution. We use the AdamW optimizer with a constant learning rate of 1×10^{-5}, weight decay 1×10^{-4}, and $\epsilon = 1 \times 10^{-8}$, with a global batch size of 32. As in the full model training setting, we maintain an EMA of the weights with decay 0.9999 and report results using the EMA weights. Text conditioning is dropped with probability 0.15 during training (for classifier-free guidance). We train for 30,000 iterations using `bf16` mixed precision. Using 8 NVIDIA B200 GPUs training takes approximately 5.6 hours.

A.2.5 Evaluation

We report Fréchet Inception Distance (FID) and Kernel Inception Distance (KID) using the `pytorch-fid` package's InceptionV3 network with 2048-dimensional pool3 features; the standard configuration for both metrics. KID uses a degree-3 polynomial kernel averaged over 100 ran-

Evaluation Prompt

Instruction to evaluate:
`<question>`

Response to evaluate:
`<reader prediction>`

Reference Answer (Score 5):
`<ground-truth VLM answer>`

Output format:
`Feedback: (feedback) [RESULT] (1--5)`

Scoring Rubric

1. The answer is empty, unrelated, or contradicts the reference answer or image on the requested attribute.
2. The answer contains some relevant content, but an essential fact is wrong or missing.
3. The answer is partially correct but has a material ambiguity, omission, or unsupported addition.
4. All core facts are correct and image-supported; only a nonessential detail is omitted or added.
5. The answer is semantically equivalent to the reference answer, fully correct, and image-consistent.

Figure 10: **VLM-as-a-judge prompt and scoring rubric.** Prometheus-Vision receives the visual question, the reader's prediction, the reference answer, and the generated image, and assigns a score from 1 to 5 according to the rubric.

dom subsets of size 1000. CLIP similarity between generated images and their prompts is computed with `openai/clip-vit-large-patch14`. DINO precision and recall follow the Improved Precision and Recall formulation, using `facebook/dinov2-base` features with $k = 5$ nearest neighbors. VQAScore is computed with the `t2v_metrics` package using `clip-flant5-xxl`. For Human Preference Score we use HPS v2.1, via the `hpsv2` package (`xswu/HPSv2` checkpoint, the latest official release, downloaded through `huggingface_hub`).

In the full model training setting, FID/KID/CLIP similarity/DINO precision-recall and HPS are computed over $40{,}192$ generated samples against the full MS-COCO 2014 validation set, while VQAScore is computed over a fixed random subset of $4{,}096$ samples. In the fine-tuning setting, all metrics are computed over the full $5{,}000$-image held-out validation set.

B Additional Results and Experiments

B.1 Contextual Reader

B.1.1 Comparing Semantic Readability Across Representations

We compare the semantic readability of the contextual space with that of the latent and the prompt. We follow the evaluation protocol described in Appendix A.1.5 and consider two complementary questions: whether contextual tokens provide semantic information beyond what is recoverable from the latent at early denoising stages, and whether they encode information beyond what can be inferred from the textual conditioning alone.

Semantics Emerge Early in Contextual Tokens. We first compare the contextual space with two alternative sources of information. First, we train a Latent Reader using the same LLM and reader architecture as the Contextual Reader, but provide it with the output latent representation at the corresponding denoising timestep instead of contextual tokens. Second, we use InternVL3-8B to answer the visual question from the predicted clean image at each timestep; we refer to this baseline as x_0+VLM. We note that the ground-truth answers used for evaluation are also generated by InternVL3-8B from the final images, making x_0+VLM more closely aligned with the evaluation target.

Figure 11 shows the resulting readability scores for both diffusion backbones. The conditional Contextual Reader consistently outperforms the Latent Reader throughout denoising, while the unconditional Contextual Reader outperforms the Latent Reader from 10% onward despite having no access to the textual conditioning. In comparison, x_0+VLM achieves the highest readability from

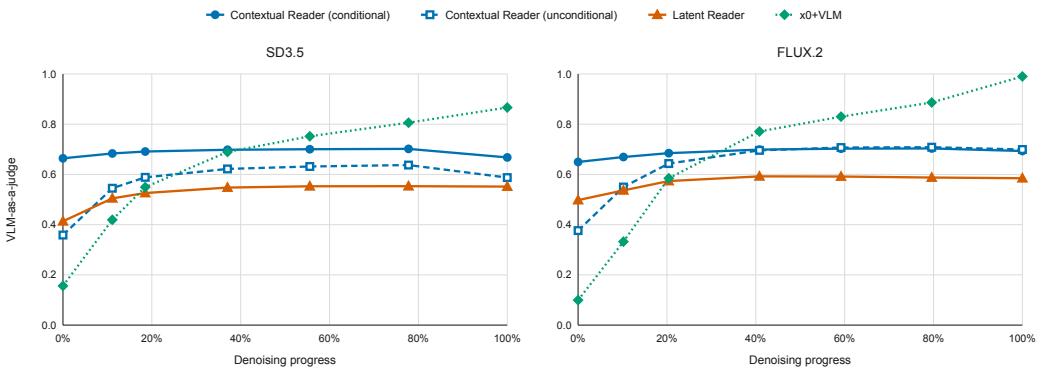

Figure 11: **Comparing semantic readability with the predicted image.** We compare the Contextual Reader with a Latent Reader operating on the output latent representation and an x_0+VLM baseline that answers questions directly from the predicted clean image. Results are shown for both conditional and unconditional readers across two diffusion backbones. The contextual representations provide higher readability during the early stages of denoising, while x_0+VLM achieves the highest readability from approximately 40% onward.

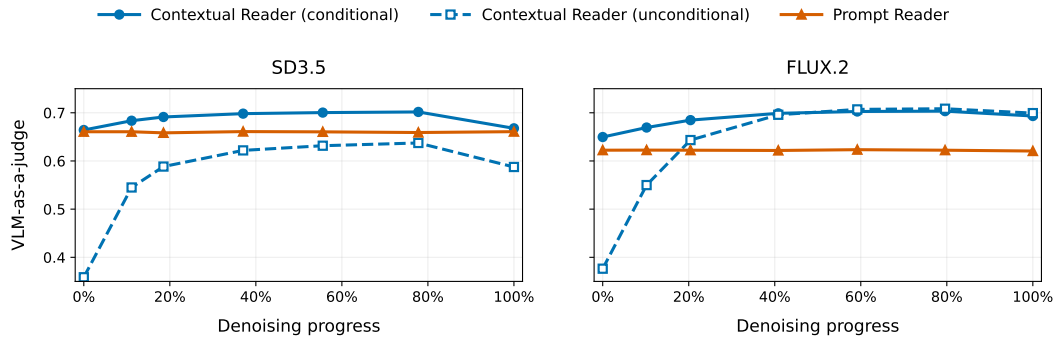

Figure 12: **Comparing semantic readability with the prompt.** We compare the Contextual Reader with a Prompt Reader using the same architecture but receiving only the input prompt. Results are shown for both conditional and unconditional readers across two diffusion backbones. The conditional Contextual Reader achieves higher readability than the Prompt Reader throughout denoising, while the unconditional Contextual Reader surpasses it for FLUX.2 and approaches it for SD3.5 as denoising progresses.

approximately 40% onward. Remarkably, before this point, both the conditional and unconditional Contextual Readers achieve higher readability despite having no direct access to the image. These results show that contextual tokens provide a particularly informative representation of semantic information during the early stages of denoising.

Contextual Tokens Encode Generation-Specific Semantics. We next examine how much of the semantic information decoded by the Contextual Reader can be attributed to the information available from the textual conditioning and the LLM's prior alone. To this end, we train a Prompt Reader using the same LLM and reader architecture as the Contextual Reader, but provide it only with the input prompt rather than contextual tokens. The Prompt Reader is trained with the same visual questions and target answers as the Contextual Reader.

Figure 12 compares the resulting readability scores for both diffusion backbones and conditioning settings. The Prompt Reader achieves constant readability throughout denoising, as expected since its input prompt remains unchanged. In contrast, the conditional Contextual Reader achieves higher readability than the Prompt Reader throughout denoising, with the largest gap between 20% and 80%. The unconditional Contextual Reader, which receives no textual conditioning, surpasses the

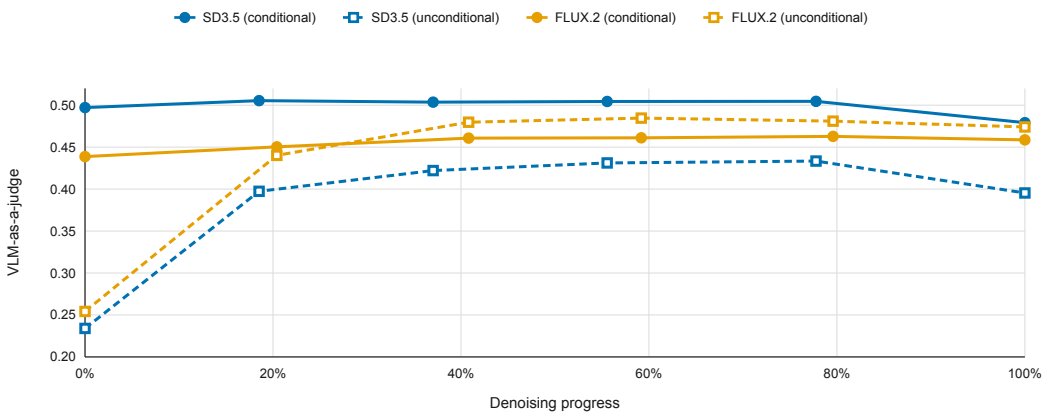

Figure 13: **Generalization to unseen questions.** For each question, we train a separate reader on the remaining 14 questions and evaluate it on the held-out question. We report the average semantic readability across all 15 held-out evaluations. Despite the reduced supervision, the reader maintains meaningful readability across denoising timesteps for both conditioning settings and diffusion backbones.

Prompt Reader as denoising progresses for FLUX.2 and approaches it for SD3.5. These results show that contextual tokens encode semantic information beyond what can be inferred from the textual conditioning and the LLM's prior.

B.1.2 Generalization to Unseen Questions

To evaluate whether the reader generalizes beyond the questions used during training, we perform a held-out-question evaluation. For each of the 15 questions, we train a separate reader on the remaining 14 questions and evaluate it exclusively on the held-out question. We repeat this procedure for all questions and report the average readability across the 15 held-out evaluations.

Figure 13 shows the resulting readability scores for both diffusion backbones and conditioning settings. Compared to the evaluation on questions seen during training, readability decreases, as expected, but remains meaningful. The conditional readers reach scores of approximately 0.50, corresponding to responses that are typically judged as partially correct, with some omissions. We observe the same trend for the unconditional readers as in our earlier experiments: their readability approaches that of the conditional readers as denoising progresses. SD3.5 reaches approximately 0.40 (generally relevant but with some omissions), while FLUX.2 eventually surpasses the conditional reader later in denoising. These results show that the reader retains the ability to decode semantic information even for questions that were never observed during training.

B.1.3 Additional Qualitative Results

Final images corresponding to Figure 1. Figure 14 shows the final images corresponding to the generations presented in Figure 1. While the early visual predictions are ambiguous, the corresponding denoising trajectories eventually resolve into the activities identified by the Contextual Reader.



cooking | playing tennis | skateboarding | dancing | playing guitar | riding a horse

Figure 14: **Final images corresponding to the early visual predictions in Figure 1.**

Table 2: Ablation study conducted in both the full model training setting ($150k$ steps from random initialization on MS-COCO) and the fine-tuning setting ($30k$ steps on the Fine-T2I curated subset, starting from SD3). In full model training, Ours refers to REPA + CoAl; in fine-tuning, it refers to vanilla flow matching + CoAl. Bold/underline mark the best/second-best value per column within each setting.

	Full model training					Fine-tuning		
Variant	FID↓	CLIP$\times 10^2$↑	HPS$\times 10^2$↑	Prec.↑	Rec.↑	FID↓	Prec.↑	Rec.↑
Semantic Teacher: CLIP	4.71	24.16	21.30	0.599	0.375	16.87	<u>0.994</u>	<u>0.970</u>
Semantic Teacher: DINOv2	4.67	23.96	21.03	0.604	<u>0.392</u>	16.92	0.993	0.964
Semantic Teacher: Qwen3-VL-2B	5.03	24.09	21.23	0.588	0.352	<u>16.81</u>	0.991	0.969
Contextual Layer: L3	4.54	24.20	21.30	0.607	0.388	17.04	0.992	0.968
Contextual Layer: L5	4.78	24.18	21.23	0.597	0.364	16.97	<u>0.994</u>	0.970
Contextual Layer: L12	5.30	23.99	21.05	0.575	0.318	17.27	0.991	0.968
Contextual Layer: L20	5.19	24.04	21.12	0.566	0.315	17.70	0.986	0.960
w/o Null-Condition Weighting	<u>4.44</u>	<u>24.31</u>	<u>21.45</u>	<u>0.617</u>	0.389	17.60	**0.994**	**0.970**
Ours	**4.29**	**24.40**	**21.60**	**0.628**	**0.407**	**16.75**	0.993	0.970

Qualitative analysis. We provide additional qualitative results from our Contextual Reader in Figures 15 and 17 for FLUX.2, and in Figures 16 and 18 for SD3.5. For the conditional setting, we repeat the experiment in which we generate two images from the same prompt, where underspecified attributes are realized differently. As shown in Figures 15 and 16, the reader can recover these seed-specific differences from the contextual tokens at early denoising timesteps. For the unconditional setting, we again examine how the reader's predictions evolve as denoising progresses. Figures 17 and 18 show that the reader first recovers coarse semantic properties and progressively provides more fine-grained answers at later timesteps. Altogether, these results demonstrate that contextual tokens can capture image-specific information through their interaction with the visual tokens.

B.2 Contextual Alignment

B.2.1 Ablation Studies

To validate our design choices, we ablate the semantic teacher, the contextual layer at which alignment is applied, and the null-condition weighting scheme, in both the full model training and fine-tuning settings (Table 2). Across both settings, CoAl is fairly robust to the choice of teacher, though SigLIP consistently edges out CLIP, DINOv2, and Qwen3-VL-2B. Aligning earlier contextual layers also consistently outperforms aligning later ones, likely because doing so gives more of the network the opportunity to make use of the well-aligned representation as it propagates forward. Finally, null-condition weighting proves especially important in the fine-tuning setting: since a pre-trained model already aligns reasonably well with the semantic teacher whenever the prompt is present, upweighting the loss contributes most on the text-dropped iterations, where that shortcut disappears.

B.2.2 Extended Quantitative Comparisons

In addition to the results in the main paper, we provide an extended quantitative comparison for both the full model training and fine-tuning settings (Tables 3 and 4), reporting VQAScore (Lin et al., 2024) and KID (Bińkowski et al., 2018) alongside the metrics used in the main paper. KID is shown to be more informative than FID on smaller evaluation sets, which is relevant here since the Fine-T2I validation set contains only 5k images. We also add REG (Wu et al., 2026) as an additional baseline in both settings. Similar to CoAl, it supervises a global semantic representation of the image rather than local, patch-level structure, which made it a natural point of comparison. However, it is designed for class-conditional architectures with no text branch, and it does not seem to translate well to MM-DiT, which already carries this kind of information through a fully functional contextual branch. Its FID and KID both come out worse than plain diffusion training. We therefore do not include it in the main manuscript or test it in combination with CoAl. For fine-tuning, we also report Reference and Base Model rows, where Reference denotes the real Fine-T2I validation images and

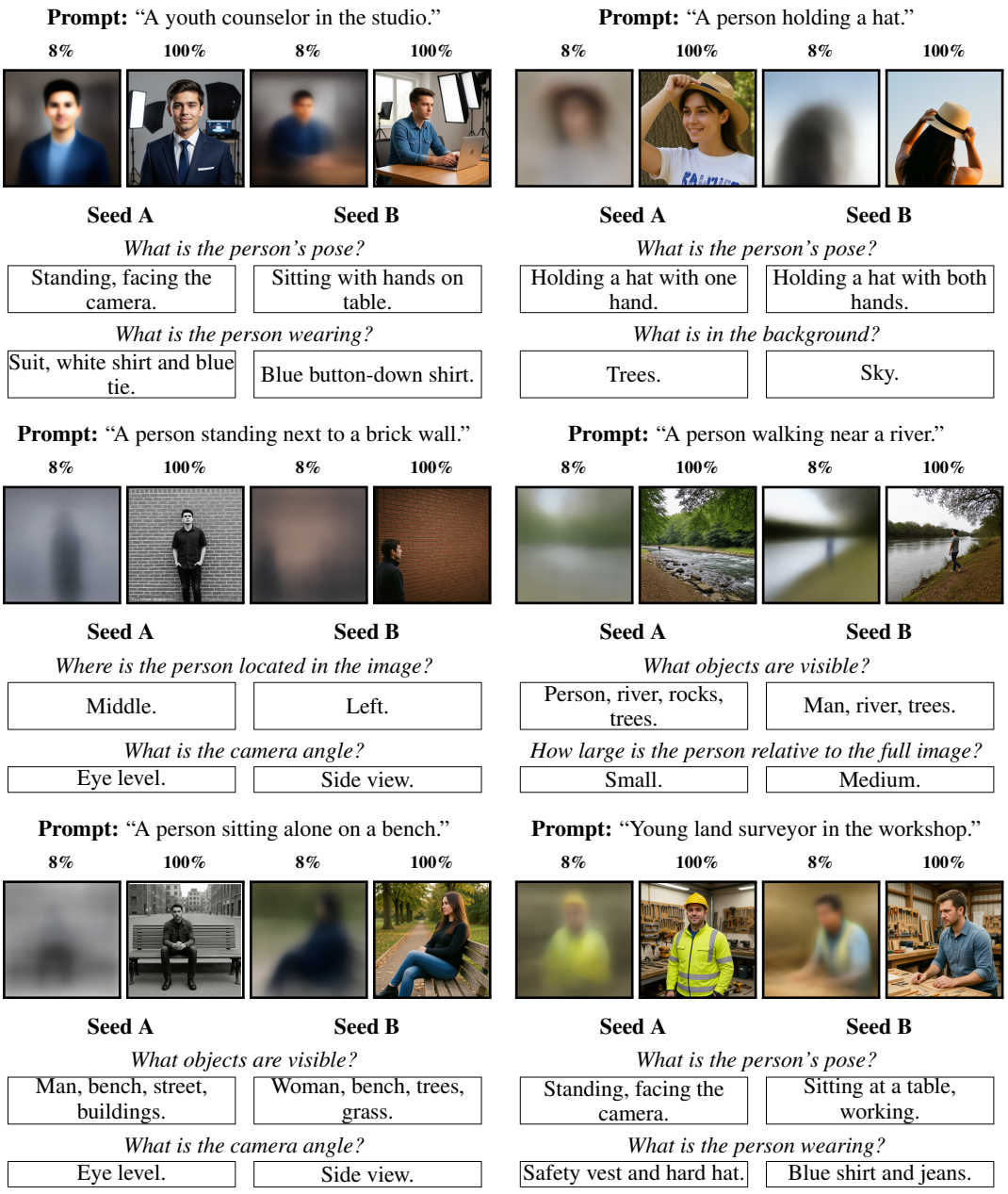

Figure 15: **Additional plain-text predictions from our conditional Contextual Reader on Flux.2.** For each prompt, we compare two random seeds that resolve underspecified attributes differently. We show the predicted clean images at 8% of the denoising trajectory for both seeds, followed by their final images. The reader predictions are obtained from the contextual tokens at the same 8% timestep.

Base Model is the pretrained SD3-Medium checkpoint before fine-tuning. For every configuration combined with CoAl, we report both how it compares to its own non-CoAl counterpart, colored green where it improves and red where it regresses, with the percent change alongside it, and how it ranks against every other configuration in the table, with bold/underline marking the best/second-best per column.

As already noted in the main paper for CLIP and HPS, Reference in fact scores below Base Model on all three preference and alignment style metrics in this extended comparison, including VQAScore. This reinforces that these axes should be read with caution in the fine-tuning setting: a real

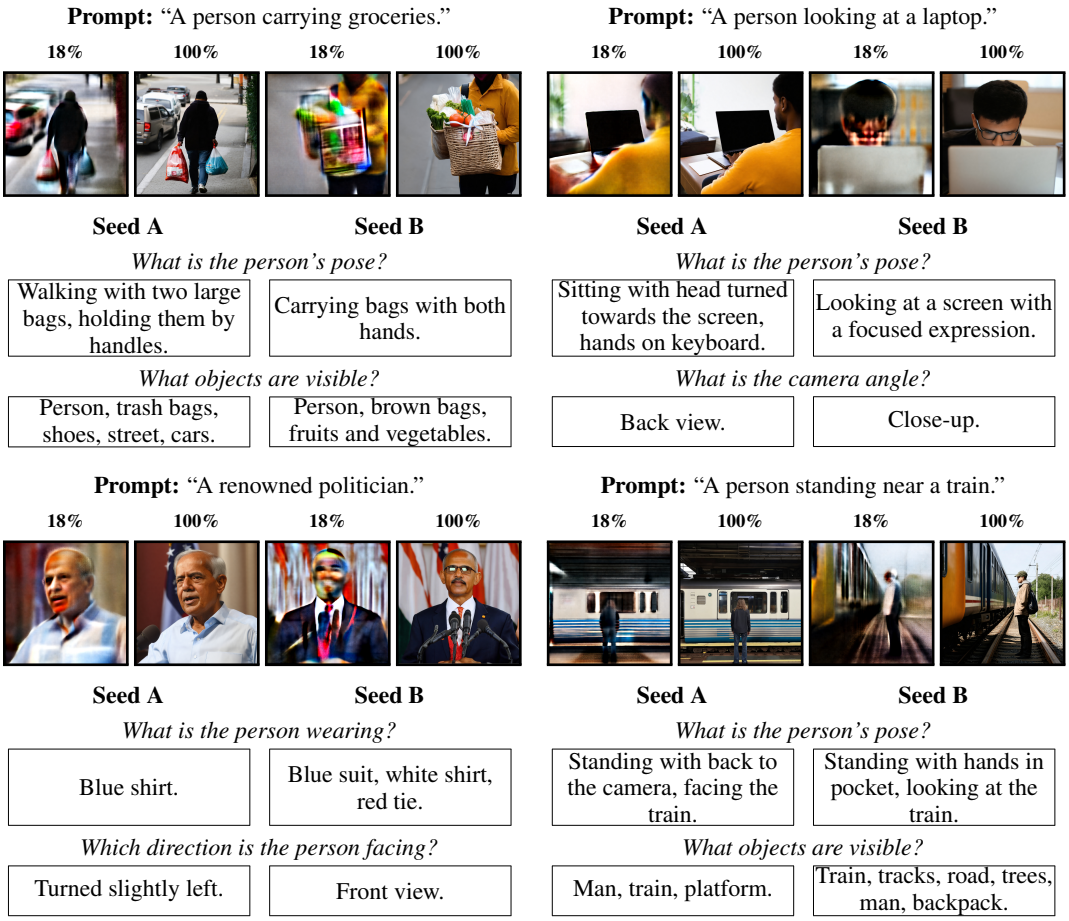

Figure 16: **Exemplary plain-text predictions from our conditional Contextual Reader on SD3.5.** For each prompt, we compare two random seeds that resolve underspecified attributes differently. We show the predicted clean images at 18% of the denoising trajectory for both seeds, followed by their final images. The reader predictions are obtained from the contextual tokens at the same 18% timestep.

Table 3: Extended quantitative comparison against baselines in the full model training setting ($150k$ steps from random initialization on MS-COCO 2014). Baseline rows: **bold**/underline mark best/second-best among the baselines. CoAl rows: value colored green/red if it improves/regresses on that metric relative to its own non-CoAl counterpart, with the percent change in parentheses.

Method	FID↓	KID×10^3↓	CLIP×10^2↑	VQA↑	HPS×10^2↑	Prec.↑	Rec.↑
Vanilla	6.52	1.53	23.27	0.630	20.10	0.496	0.206
REG	7.33	2.10	23.79	0.613	20.81	0.486	0.268
SRA	5.89	1.23	23.49	0.637	20.37	0.515	0.261
HASTE	4.87	0.93	24.21	0.691	21.42	0.609	0.361
REPA	4.95	0.94	24.20	0.691	21.38	0.591	0.338
Vanilla + Ours	6.23 (-4.4%)	1.52 (-0.9%)	23.30 (+0.1%)	0.629 (-0.2%)	20.17 (+0.3%)	0.515 (+3.7%)	0.274 (+33.2%)
SRA + Ours	4.36 (-25.9%)	**0.76 (-38.7%)**	24.28 (+3.4%)	0.688 (+8.1%)	21.42 (+5.1%)	0.616 (+19.5%)	0.398 (+52.5%)
HASTE + Ours	4.44 (-8.8%)	0.79 (-15.1%)	24.32 (+0.4%)	**0.696 (+0.8%)**	21.56 (+0.7%)	0.623 (+2.3%)	0.400 (+10.6%)
REPA + Ours	**4.29 (-13.4%)**	0.76 (-19.4%)	**24.40 (+0.8%)**	0.694 (+0.4%)	**21.60 (+1.0%)**	**0.628 (+6.2%)**	**0.407 (+20.5%)**

photograph scoring below a model's output does not mean the model produces more faithful images, only that these metrics reward image and caption conventions particular to their own training distributions.

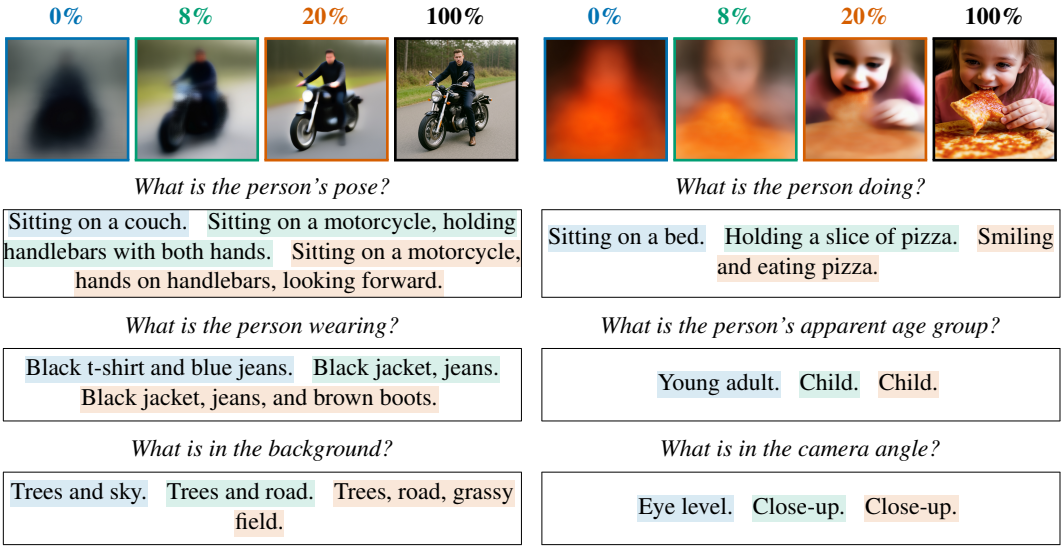

Figure 17: **Additional plain-text predictions from the unconditional Contextual Reader on Flux.2.** We show predicted clean images at 0%, 8%, and 20% of the denoising trajectory, together with the final generation. Colored highlights associate each intermediate timestep with the corresponding reader answer below. Despite receiving no text prompt, the reader recovers increasingly specific semantic information about the eventual image as denoising progresses.

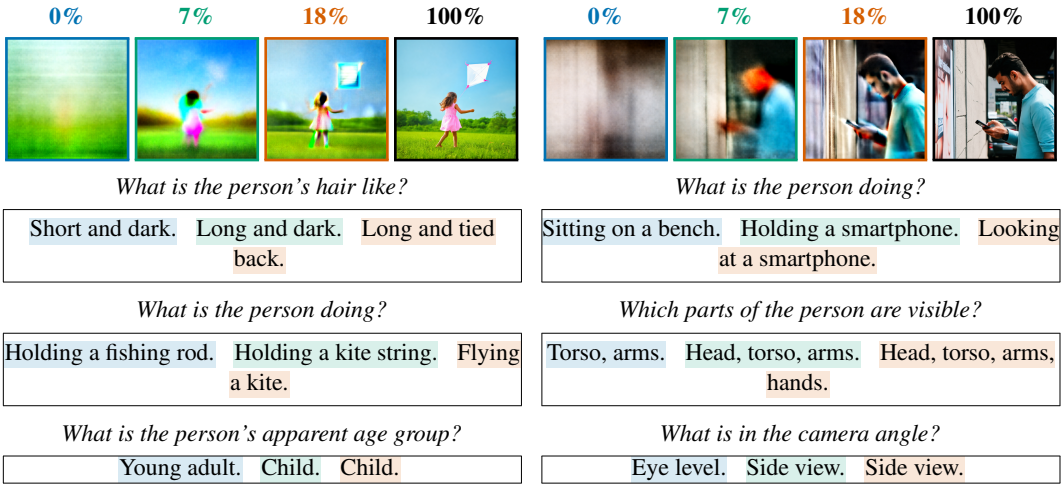

Figure 18: **Exemplary plain-text predictions from the unconditional Contextual Reader on SD3.5.** We show predicted clean images at 0%, 7%, and 18% of the denoising trajectory, together with the final generation. Colored highlights associate each intermediate timestep with the corresponding reader answer below. Despite receiving no text prompt, the reader recovers increasingly specific semantic information about the eventual image as denoising progresses.

In the full model training setting, the orthogonal comparison confirms the pattern from the main paper: every representation-alignment baseline improves on most metrics when combined with CoAl. Looking across the table as a whole rather than baseline by baseline, REPA + CoAl is not just the best combination relative to plain REPA, but the single best configuration overall, achieving the best value in every column we report. For fine-tuning, we additionally report REPA, HASTE, and SRA combined with CoAl for completeness. All three combinations improve KID substantially over their respective baselines (5.6 to 20.7%), and where they do decrease Precision or Recall, the drop is marginal: REPA + CoAl is the only one of the three to regress on either metric, and only by 0.0% and 0.1%, respectively.

Table 4: Extended quantitative comparison against baselines on the Fine-T2I curated benchmark (val5k, 30k fine-tuning steps starting from SD3-Medium). Reference refers to the real Fine-T2I validation set images, and Base Model is the pretrained SD3-Medium checkpoint before fine-tuning. **Bold**/underline mark the best/second-best value per column across all rows. CoAl rows are additionally colored green/red if they improve/regress on that metric relative to their own non-CoAl counterpart, with the percent change in parentheses.

Method	FID↓	KID×10^3↓	CLIP×10^2↑	VQA↑	HPS×10^2↑	Prec.↑	Rec.↑
Reference	–	–	26.84	0.925	25.53	–	–
Base Model	19.30	3.05	28.70	0.940	**29.04**	0.993	0.939
Vanilla	17.18	2.23	28.87	**0.941**	27.78	0.993	0.966
REG	21.13	3.68	28.43	0.923	24.18	0.976	0.961
SRA	17.48	2.45	28.87	0.940	27.71	0.991	0.964
HASTE	17.53	2.34	28.80	0.938	27.71	0.994	0.963
REPA	17.45	2.13	28.83	0.939	27.77	0.993	0.966
Vanilla + Ours	**16.75** (-2.5%)	**1.82** (-18.4%)	**28.92** (+0.2%)	0.938 (-0.3%)	27.92 (+0.5%)	0.993 (-0.0%)	**0.969** (+0.3%)
SRA + Ours	17.10 (-2.2%)	1.94 (-20.7%)	28.81 (-0.2%)	0.937 (-0.2%)	27.69 (-0.1%)	0.992 (+0.1%)	0.964 (0.0%)
HASTE + Ours	17.08 (-2.6%)	2.09 (-10.5%)	28.81 (+0.0%)	0.937 (-0.1%)	27.84 (+0.5%)	**0.994** (+0.0%)	0.966 (+0.3%)
REPA + Ours	17.44 (-0.1%)	2.01 (-5.6%)	28.87 (+0.1%)	0.937 (-0.2%)	27.80 (+0.1%)	0.993 (-0.0%)	0.965 (-0.1%)

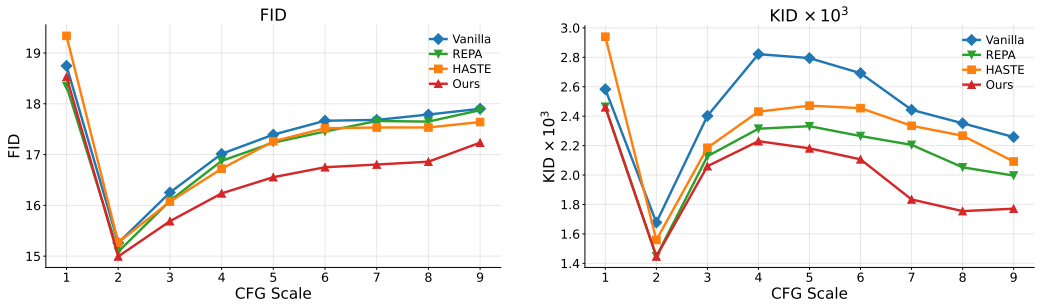

Figure 19: Effect of classifier-free guidance (CFG) scale on FID and KID on the Fine-T2I curated benchmark (val5k), comparing Vanilla, REPA, HASTE, and CoAl.

B.2.3 Performance Across CFG Scales

One might argue that our null-condition weighting scheme simply shifts the effective classifier-free guidance scale reached during training, and that evaluating at a different CFG value could reveal different behavior. To test this, we sweep CFG scale from 1 to 9 for Vanilla, REPA, HASTE, and Vanilla + CoAl on the Fine-T2I curated benchmark (Figure 19). CoAl is consistently the strongest of the four across the entire sweep. CFG$=2$ gives the best FID and KID for every method, ourselves included. We nonetheless use CFG$=7$ for all evaluations and generations elsewhere in the paper, since it offers a better balance between prompt adherence and distributional fidelity. Recent work ties image quality under classifier-free guidance to how closely the conditional and unconditional predictions agree with each other over the course of sampling: as the distance between them grows, the guidance direction takes large, sharp turns that can push generation off the data manifold (Wang et al., 2026a; Yehezkel et al., 2025). CoAl handles large CFG values noticeably better than the other methods, which could suggest that its conditional and unconditional predictions agree with each other more closely. This is indeed what our null-condition weighting is designed to encourage: by encouraging the contextual branch to draw on information from the image tokens when the prompt is omitted, we in turn encourage similarity between the null and conditioned predictions, assuming the image is aligned with the prompt.

"A man stands in waist-deep water, holding a fishing rod, wearing a green jacket and black waders..."



*"A **couple** walking along a serene beach, dressed in white, under a vibrant blue sky..."*



"A close-up of a hand holding an old-fashioned Yashica 35mm camera..."



"A minimalist photograph of the word 'SALE' spelled out in white tiles against a coral pink background..."



"A neon sign in blue text reading 'SWEET DREAMS ARE MADE OF THIS' hangs on a wall..."



"A close-up of a hand wiping the words 'Happy birthday' off a black chalkboard..."



Ground Truth | Vanilla | REPA | HASTE | SRA | **Ours**

Figure 20: Additional qualitative comparison against visual alignment methods over SD3-Medium fine-tuning on the Fine-T2I curated dataset.

B.2.4 Additional Qualitative Results

We provide two additional qualitative figures in the supplementary material, Figure 20 and Figure 21. The first follows the same format as the qualitative figure in the main paper, comparing Reference, the held-out Fine-T2I validation images matching each prompt, against Vanilla, REPA, HASTE,

SRA, and CoAl, on six additional examples. The second instead pairs each baseline directly against itself combined with CoAl, for Vanilla, REPA, and HASTE, with five examples per pair.

Figure 20 includes further examples of CoAl fixing limb artifacts and adding high-level detail, most clearly in the fisherman (row 1) and camera (row 3) examples. It also highlights CoAl's alignment with the text prompt: in the beach couple example (row 2), CoAl is the only method that renders two people as the prompt describes, and the last three rows, SALE (row 4), the neon sign (row 5), and the birthday chalkboard (row 6), all require the model to render specific words legibly, where CoAl again fares noticeably better than the baselines. Figure 21 shows further examples of these same effects, high-level detail, artifact prevention, and text alignment, this time paired baseline by baseline against Vanilla, REPA, and HASTE.

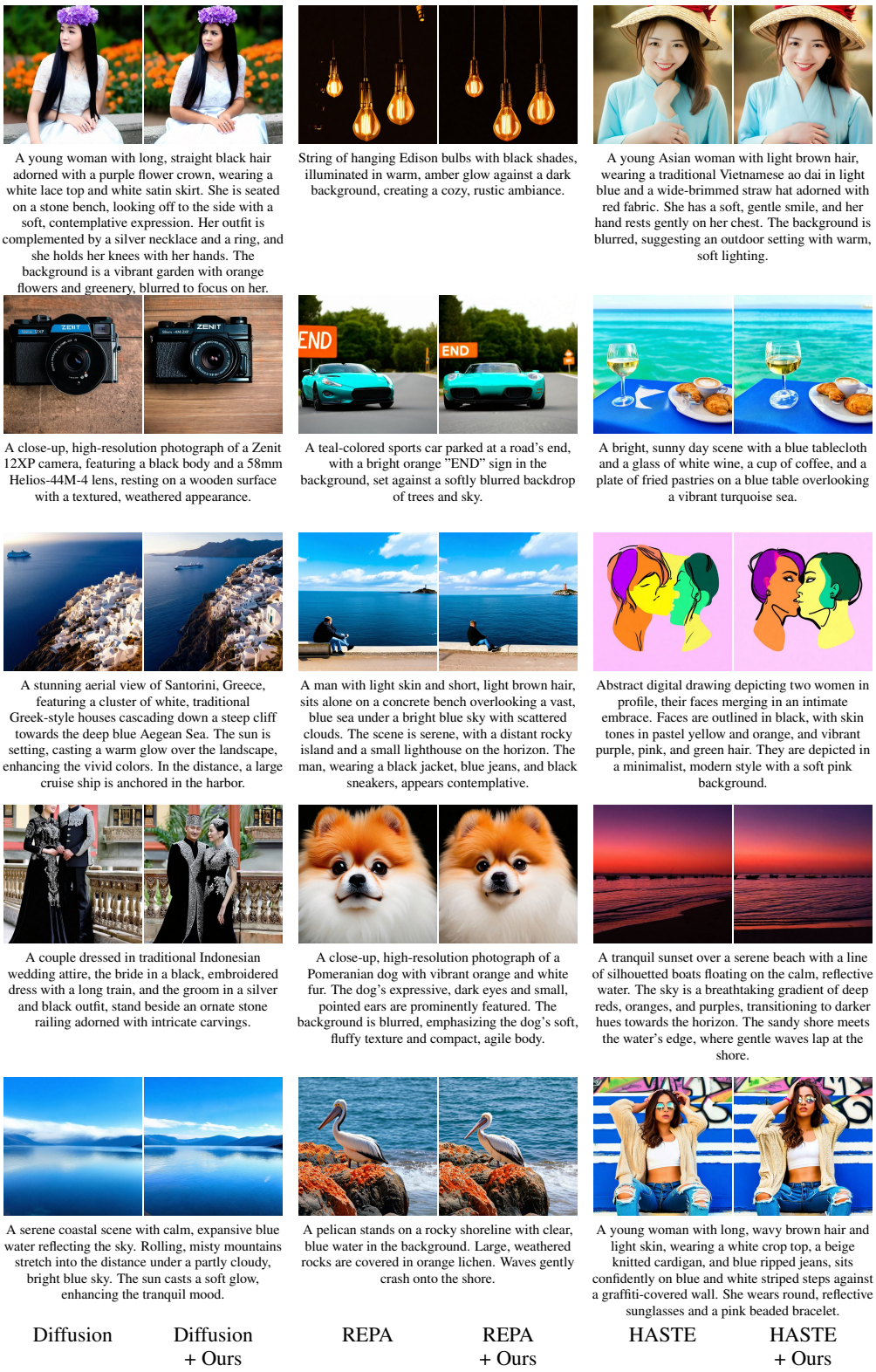

Figure 21: Qualitative comparisons on Fine-T2I val5k prompts: each pair shows a baseline (left) against the same baseline with our method added (right). Each pair uses its own set of example prompts.