JIG: JSON-TO-IMAGE GENERATION IN PIXEL AND LATENT SPACES

Anonymous authors
Paper under double-blind review



Figure 1: Top: JIG-14M statistics. Structured annotations substantially outperform dense captions; latent outperforms pixel under both decoupled and unified architectures, though pixel nearly matches latent with an improved recipe. Bottom: three representative annotated examples.

ABSTRACT

Text conditioning is fundamental to image generation and has evolved from short alt-text to dense synthetic captions and, more recently, structured JSON, with caption lengths growing from tens to thousands of tokens. However, the large-scale JSON-annotated corpora used by recent generators remain unreleased, limiting reproducible studies of how annotation content, textual format, and generator design contribute to performance. In this paper, we present JIG, an open model family and empirical study of JSON-to-image generation in pixel and latent spaces. First, we introduce JIG-14M, a dataset of 13.7M images paired with rich structured annotations that capture global scene information, visual styles, and spatially grounded element details. Second, we investigate the influence of annotation content and textual format. Both JSON and natural-language prose realizations of the structured annotations outperform conventional dense captions, suggesting that richer annotation content is a key contributor to improved generation quality across textual formats. Third, we train and compare four JIG models spanning pixel and latent spaces using decoupled diffusion and unified Transfusion-style architectures. Under the baseline training setup, latent-space models outperform their pixel-space counterparts, while differences between architectures are comparatively small. We further analyze this gap and develop a pixel-specific training

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.

recipe that substantially narrows it in both architectures. With this improvement, pixel models approach their latent counterparts in prompt following, with the largest remaining gap in aesthetic quality. With image-generation training on JIG-14M, the resulting systems also achieve competitive performance against compact open-source generators, with strong compositional generation and prompt alignment. JIG-14M will be released to support the research community.

1 INTRODUCTION

Recent advances in text-to-image (T2I) generation have benefited from scalable diffusion transformers (Peebles & Xie, 2023; Esser et al., 2024), large-scale image–text datasets (Schuhmann et al., 2022), and improved training objectives such as flow matching (Lipman et al., 2023). As users seek increasingly complex compositions and finer-grained control, generators must follow instructions specifying multiple objects, attributes, and spatial relationships. However, models trained predominantly on short captions may receive limited explicit supervision for these requirements, a gap that increasing dataset size alone does not necessarily resolve. *Since diffusion models learn conditional image distributions, accurate and detailed captions can reduce ambiguity about the intended visual content and provide more direct supervision for following textual constraints* (Betker et al., 2023; Segalis et al., 2023). Caption supervision has therefore progressed from short alt-text to dense synthetic descriptions and, more recently, structured JSON annotations (Gutflaish et al., 2025; Ideogram AI, 2026; Reve Team, 2026; Chen et al., 2026).

Structured annotations organize scene context and spatially grounded element descriptions into explicit fields and records, specifying both what an image contains and how its elements are arranged. For example, Ideogram 4 and Reve 2 organize image content through spatially grounded records and hierarchical layouts (Ideogram AI, 2026; Reve Team, 2026). The concurrent work ContextScaling further relates caption informativeness to diffusion training loss through controlled comparisons (Chen et al., 2026). These developments motivate a closer examination of whether the benefits of structured conditioning arise from richer visual information, its textual representation format, or their combination. Disentangling these contributions requires comparisons that preserve annotation content while varying its textual format. However, limited access to large-scale structured corpora makes such comparisons difficult to reproduce.

To address this gap, we introduce JIG, an open model family and empirical study of JSON-to-image generation in pixel and latent spaces. Its data foundation, JIG-14M, comprises 13.7M images paired with structured annotations covering global scene context and spatially grounded element details, with a JSON structure design that closely follows (Ideogram AI, 2026; Reve Team, 2026). Deterministic converters express the same annotations as JSON or natural-language prose while preserving field values and element associations, enabling content-preserving comparisons of textual formats. We further hold the structured JSON annotations fixed and construct four generators by crossing two image representations—VAE latents and raw pixels—with decoupled diffusion and unified Transfusion-style architectures. All four models are trained using image-generation supervision only, providing a common setting for studying generator design under structured conditioning.

We first study a fundamental question: *What makes structured captions more effective than conventional dense captions?* Models trained on finalized annotations outperform the dense-caption baseline in both prompt following and aesthetic quality, with comparable results for JSON and content-preserving natural-language prose. These findings suggest that richer annotation content, explicitly describing scene elements, attributes, and spatial relationships, provides informative supervision and is a key contributor to improved generation across textual formats.

We next study how generator design affects learning under structured conditioning. *Under our baseline training setup, latent-space models outperform their pixel-space counterparts, while differences between the two architectures are comparatively small.* To narrow this gap, we analyze the different learning demands of the two representations and develop a pixel-specific training recipe by integrating prior advances in noise scheduling, pixel-level refinement, and perceptual supervision. The gains in both architectures highlight the importance of representation-specific training when assessing the relative strengths of pixel and latent generators. The resulting JIG systems also perform competitively against the latest open-source small-scale generators. Using only JIG-14M for training, our

model can outperform most external systems on GenEval2 and achieve strong PRISM alignment, demonstrating the practical value of this data foundation and model family.

To summarize, our contributions are threefold:

- We introduce JIG-14M, a dataset of 13.7M images with spatially grounded structured annotations, establishing a common data foundation for studying structured conditioning.
- We investigate the respective roles of annotation content and JSON syntax through content-preserving comparisons, showing that the benefits of structured annotations over conventional dense captions extend beyond JSON-formatted inputs.
- We compare four JIG generators across image representations and architectural configurations, characterize their performance differences, and develop a pixel-specific training recipe that substantially narrows the pixel–latent gap.

We will release JIG-14M, its annotation schema and format converters, and all four JIG models to support reproducible research on structured conditioning for image generation.

2 Related Work

Caption Supervision. Detailed synthetic captions at scale have shown improvements on image–text alignment (Betker et al., 2023; Segalis et al., 2023; Chen et al., 2025; 2024; Singla et al., 2024; Yu et al., 2024). Consistent prose templates in Re-LAION-Caption (Merchant et al., 2025) and region-grounded descriptions in ROVI, SA-PanCap, and GroundCap (Peng et al., 2025; Lin et al., 2025; Oliveira et al., 2025) give these annotations additional structure. FIBO, Ideogram 4, and Cosmos 3 adopt JSON representations (Gutflaish et al., 2025; Ideogram AI, 2026; NVIDIA, 2026); Reve 2 uses hierarchical layouts (Reve Team, 2026). ContextScaling relates caption informativeness to diffusion training loss and constructs natural-language controls from shared intermediate annotation evidence (Chen et al., 2026). To examine the role of textual format, we compare JSON and deterministic prose realizations of the same finalized records.

Generator Architectures. Conventional image generators condition image denoisers on features from separate pretrained text encoders (Chen et al., 2023; Esser et al., 2024; Wu et al., 2025a; Z-Image Team et al., 2025). Transfusion processes text and images in a shared transformer (Zhou et al., 2024), while BAGEL and Cosmos 3 use specialized transformer branches coupled through joint attention (Deng et al., 2025; NVIDIA, 2026). We compare decoupled diffusion and unified Transfusion-style architectures in both pixel and latent spaces.

Pixel-Space Training. Latent diffusion denoises autoencoder representations (Rombach et al., 2022), whereas pixel-space models operate directly in RGB space (Hoogeboom et al., 2023). Recent multimodal systems also adopt pixel-space generation (Cai et al., 2026; Liu et al., 2026; Diao et al., 2026). Work on clean-image prediction and noise schedules (Li & He, 2025), pixel-level refinement (Yu et al., 2026), and perceptual supervision (Ma et al., 2026) identifies effective designs for direct RGB prediction. Jiang et al. (2026) compare large-scale pixel- and latent-space pretraining and develop a latent-to-pixel adaptation recipe. We study direct pixel-space training on JIG-14M under both architectures, evaluating a common pixel-specific recipe against latent generators.

3 JIG-14M: An Open JSON-to-Image Generation Dataset

JIG-14M comprises 13.7M images paired with spatially grounded structured annotations. Each annotation combines global scene context with element-level descriptions, locations, and visual attributes, providing a common information source for our caption and generator studies. In the following sections, we describe the annotation schema and its textual realizations, the image-grounded construction pipeline, and the dataset curation and quality assessment.

3.1 Annotation Schema and Textual Realizations

We adopt the publicly documented JSON caption design of Ideogram 4 (Ideogram AI, 2026) as the structured-annotation schema throughout our ablation studies. This schema organizes visual information into image-level context and spatially grounded element records, as illustrated in Figure 1.

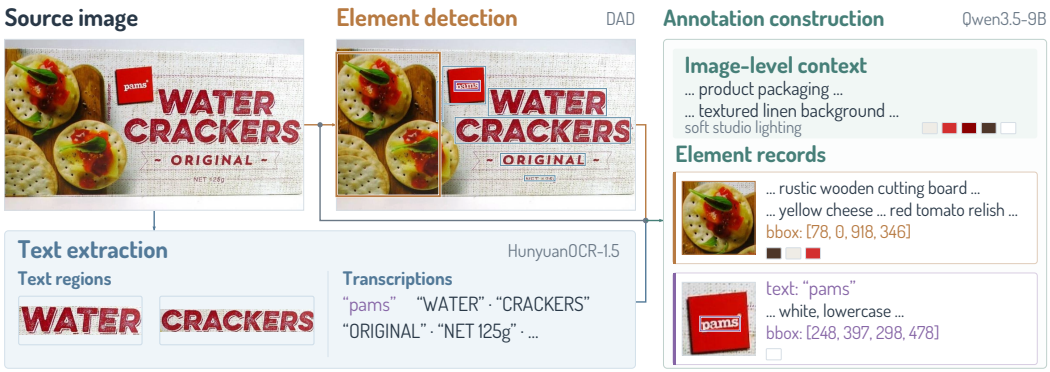


Figure 2: **Image-grounded annotation pipeline.** HunyuanOCR-1.5 extracts text and DAD detects element boxes. Qwen3.5-9B integrates both outputs with the source image to construct structured JSON annotations. Matching colors link selected image regions and output records.

Specifically, three top-level fields describe the overall scene (*high_level_description*), visual style (*style_description*), and composition (*compositional_deconstruction*). The composition block separates the background from individual element records. Element records associate descriptions of objects or text regions with bounding boxes and local color palettes where applicable. For text elements, *text* stores the literal characters, while *desc* describes their typography and visual context. For example, the "Travel Farther" record in Figure 1 links the phrase to its bounding box and describes its bold black sans-serif lettering. Section A.1 details the schema.

We further distinguish the underlying annotation content from its textual realization. The same finalized records can be serialized as JSON or deterministically rendered as natural-language prose, preserving field values like coordinates, colors, and their associations with individual elements.

In the system-level comparison (Section 6.3), we additionally evaluate Reve 2-style annotations (Reve Team, 2026), which use element identifiers in the global description and parent references to encode a hierarchy. Section A.1 describes the additional fields, and Figure 7 compares the two annotation datasets.

3.2 Image-Grounded Annotation Pipeline

Our Ideogram 4-style annotation pipeline consists of three stages: text extraction, element detection, and image-grounded JSON construction (Figure 2). The first two stages process the source image independently; the final stage integrates their outputs with the image to construct structured annotations.

Text extraction. We use HunyuanOCR-1.5 (Li et al., 2026) to extract visible text. The transcriptions populate explicit text fields, separating literal content from visual descriptions.

Element detection. We apply Detect Anything in Graphic Design (DAD) (Zhu et al., 2026) to detect image elements and obtain their bounding boxes. These detections provide spatial anchors for associating element-level descriptions with image regions.

Image-grounded JSON construction. We provide Qwen3.5-9B with the source image, bounding boxes, and OCR transcriptions to construct the final annotation. The model combines global scene and style descriptions with element records, associating detected regions with descriptions, visual attributes, and recognized text where applicable. The construction rules in Section A.2.1 govern box selection and OCR integration. The Reve-style JSON annotation adopts a different strategy.

3.3 Dataset Curation and Annotation Quality

We construct JIG-14M from the short- and long-caption partitions of BLIP3o (Chen et al., 2025), pairing the retained images with structured annotations. We exclude AI-generated images originating from JourneyDB and apply the image-quality filtering, deduplication, and image-validity checks

described in Section A.3. Element bounding boxes are adjusted to remain aligned with the images after resizing and cropping. The final corpus contains 13.7 million image–annotation pairs.

Annotation quality. Following the assessment protocol in Section A.5, we sample 10,000 images proportionally from the two source partitions and assess their finalized annotations using GPT-5.6-Luna. The model judges 96.0% of element-description facts as correct and 98.6% of bounding boxes as matching their intended instances. Table 1 also reports attribute accuracy, object coverage, and text exact match.

Table 1: Annotation quality statistics.

Metric	Rate (%)
Description accuracy	96.0
Box–instance match	98.6
Attribute accuracy	97.5
Object coverage	61.1
Text exact match	78.3

4 Experimental Setup

We use JIG-14M to study the effects of annotation content, textual format, image representation, and generator architecture, followed by system-level comparisons with compact open-source generators. All ablation studies use the Ideogram 4-style annotation schema, while system-level experiments evaluate both Ideogram 4-style and Reve 2-style annotations.

Model configurations. We construct four JIG baselines by crossing two image representations—VAE latents and raw pixels—with decoupled diffusion (Peebles & Xie, 2023) and unified Transfusion-style architectures (Zhou et al., 2024). Figure 3 illustrates these architectures, which share the core configuration of Qwen3-VL-2B but differ in text conditioning and initialization.

JIG-Diffusion pairs a frozen Qwen3-VL-2B text encoder with a randomly initialized 1.4B single-stream DiT (Peebles & Xie, 2023). Hidden states from selected encoder layers are summed, projected, and concatenated with noisy-image tokens. The DiT applies bidirectional attention across the combined sequence. Timestep embeddings modulate the attention and feed-forward sublayers through adaptive normalization (AdaLN), using a modulation network shared across blocks.

JIG-Transfusion uses a shared Transformer initialized from Qwen3-VL-2B to process text and denoise images. Caption tokens, a dedicated timestep token, and noisy-image tokens enter this backbone together. Caption tokens attend causally, while image tokens attend to the preceding context and bidirectionally within the image. The timestep token provides noise-level conditioning through attention without introducing AdaLN into the pretrained backbone.

Image representations. JIG-Latent encodes images with the frozen FLUX.2 VAE and applies 4×4 patchify to the noisy latent grid. JIG-Pixel adds noise directly to RGB images and applies 32×32 patchify. Both representations yield 256 image tokens at 512×512 resolution.

Training configurations. We train all JIG models on JIG-14M for 150k steps using the configurations in Sections B.2 and B.3, with caption sequences capped at 2,048 tokens. Training uses image–caption pairs, with a flow-matching loss applied only to image outputs and no language-modeling loss or text-only training data. Latent models predict velocity directly. Pixel models predict clean images and convert both predictions and targets to velocity before computing the MSE loss.

Inference-time conditioning. Following Ideogram 4's publicly released magic-prompt design (Ideogram AI, 2026), we use Gemini-3.5-Flash-Lite to convert each benchmark's original text prompt into structured JSON records. The system prompt instructs the model to separate scene-level context from individual object and text elements, specify their layout where appropriate, and preserve the exact strings requested for rendering. Within each annotation schema, we reuse the same precomputed records across corresponding JIG comparisons and deterministically render prose inputs from those records.

Evaluation protocols. The evaluation protocol in Section C uses GenEval2 (Kamath et al., 2025) to measure compositional prompt following and PRISM-Bench (Fang et al., 2025) to assess prompt alignment and aesthetic quality. PRISM total averages alignment and aesthetic scores. JIG comparisons use a 100-step Euler sampler, classifier-free guidance scale 4.0, and matched sampling seeds. Comparisons with compact open-source generators are conducted at the system level, with each system using its own input-processing and inference pipeline.

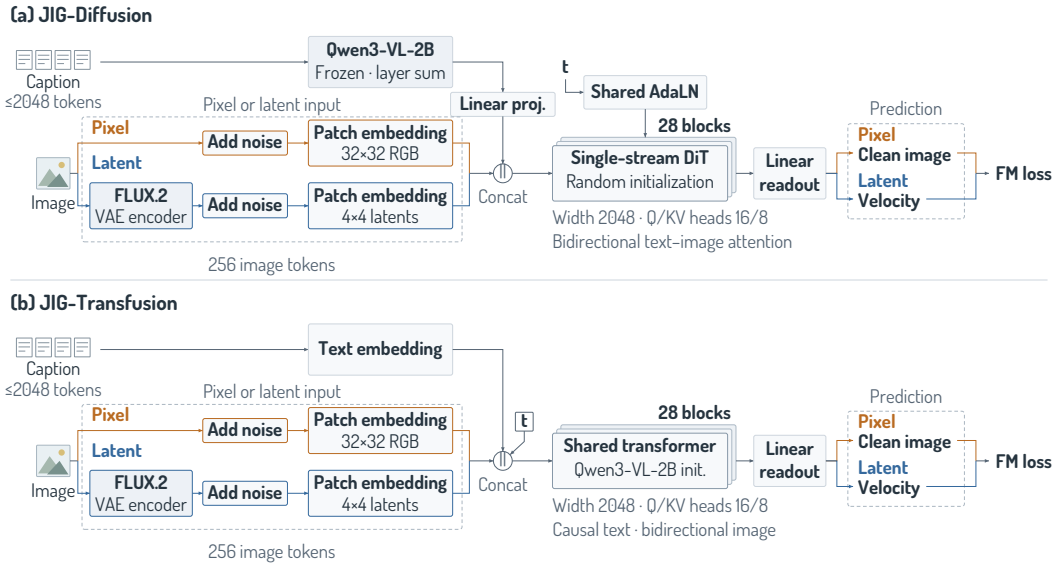


Figure 3: **JIG architectures.** (a) A frozen text encoder conditions a separate DiT. (b) A shared Transformer processes text and image tokens. Dashed boxes indicate pixel- and latent- alternatives.

5 What Makes Structured Captions Effective?

Structured annotations associate element descriptions with their locations and visual attributes. We first test whether this supervision improves generation over dense captions, then ask whether the gains persist across different textual realizations. All comparisons use JIG-Latent-Diffusion trained on the same image corpus with common training settings.

Table 2: **Caption supervision under latent diffusion.** Dense rows share weights. Record desc. are rendered from complete evaluation records.

Conditioning		Score ↑	
Training	Eval. input	GenEval2	PRISM
Dense	Dense desc.	36.26	5.79
Dense	Record desc.	44.74	5.94
JSON	JSON	62.98	**6.29**
Prose	Prose	**63.23**	**6.29**

Table 3: **Cumulative pixel adaptations.** Components are added in order under both architectures.

Setting	Diffusion		Transfusion	
	GenEval2	PRISM	GenEval2	PRISM
Pixel	48.35	5.32	52.14	5.26
+ Schedule	57.19	5.38	56.38	5.37
+ PiT	58.37	5.54	59.61	5.61
+ Perceptual	62.01	5.80	64.27	5.86
Latent	**62.98**	**6.29**	**65.11**	**6.43**

5.1 Structured Annotations Improve Image Generation

We compare dense-caption supervision with finalized annotations expressed in JSON format that explicitly associate individual elements with their locations and local attributes. Compared with the dense-caption model, the JSON model raises both GenEval2 and PRISM score substantially, and the improvements exist in both prompt following and aesthetic quality (Table 13).

To examine the contribution of richer inference inputs, we keep the dense-caption model's weights and sampling fixed and provide *record-based descriptions*: free-form paragraphs rendered from the complete evaluation records used by the JSON and prose models. In this setting, GenEval2 score rises to 44.74, with positioning and attribute scores approaching those of the structured-annotation models(Section D.1). This reveals that richer descriptions at inference thus recover part of the benefit, but do not reproduce the performance of models trained on the finalized annotations.

Reconstruction from image-derived captions provides a complementary test of recovering image-specific content. Across 256 held-out references, mean DINOv2 cosine similarity (Oquab et al., 2023) increases from 0.7203 with dense captions to 0.7654 with JSON annotations. The generated layouts in Figure 4 illustrate the benefit: JSON better preserves the magazine's text hierarchy and the arrangement of bowls in the overhead food scene.

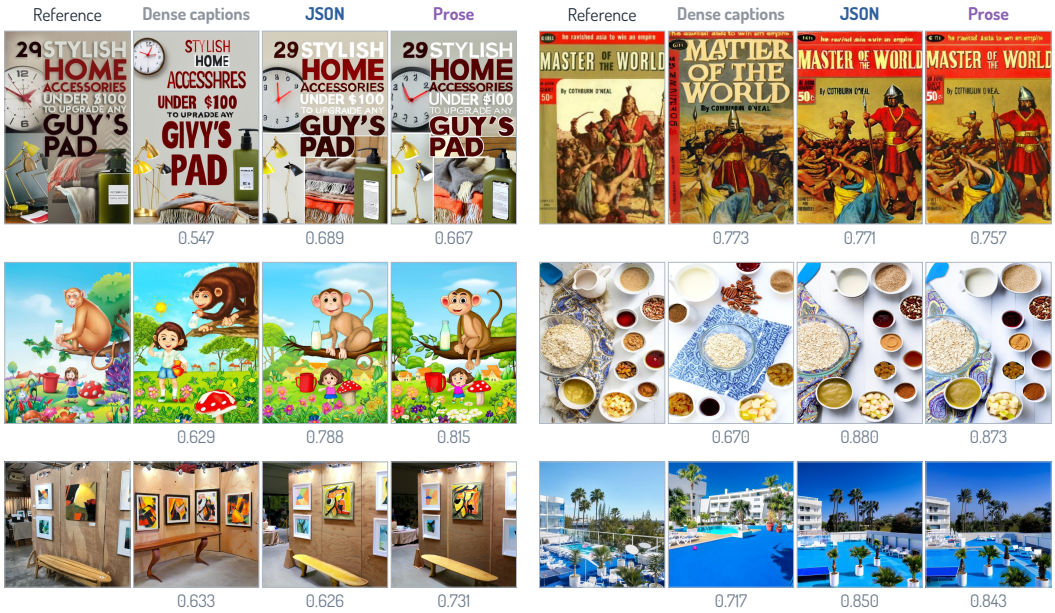


Figure 4: **Reconstruction of held-out references.** Generations use captions derived from the corresponding reference image. Numbers report DINOv2 cosine similarity to that reference.

5.2 Annotation Benefits Persist Across Textual Realizations

We then hold annotation content fixed and change its textual realization. A deterministic converter expresses the same records as prose, preserving field values and their element associations in both training and evaluation (Section B.1).

JSON and prose achieve nearly identical GenEval2 scores and the same PRISM total (Table 2). Their reconstruction similarities are also close: 0.7601 for prose versus 0.7654 for JSON. The dynamics curves in Figure 10 show that this observation persists throughout the training process.

The benefits of the finalized annotation content therefore persist across JSON and natural-language realizations, supporting its use beyond a single serialization format.

6 JSON-to-Image Generation in Pixel and Latent Spaces

Rich annotations specify the desired image, but realizing that specification also depends on generator design and training. We first compare four JIG baselines along two axes: image representation (latent versus pixel) and architecture (decoupled diffusion versus unified Transfusion-style), holding the structured annotations and their JSON realization fixed.

6.1 Comparing Pixel and Latent Generators

Under the baseline recipe, JIG-Latent outperforms JIG-Pixel in both architectures (Table 3). The gap is roughly 13–15 GenEval2 points and around one PRISM point, whereas differences between the two architectures within either representation are comparatively small. This separation persists across the training checkpoints in Figure 11.

The two representations impose different learning demands. Latent models denoise compact autoencoder representations and rely on a pretrained decoder to recover pixel-level appearance (Rombach et al., 2022). Pixel models must learn both scene structure and fine visual detail directly in RGB space (Jiang et al., 2026). The visual comparisons in Figure 12 also show differences in texture and rendering style even when both representations capture the main objects and broad arrangements.

6.2 Adapting Training to Pixel Space

We next test how far adapting specific design to RGB prediction can narrow this gap. The three adaptations are added cumulatively as shown inTable 3, with implementation details in Section B.4.

Pixel-adapted noise schedule. Following JiT (Li & He, 2025), we double the noise standard deviation to 2 at 512^2 resolution, change the logit-normal parameters from $(0, 1)$ to $(-0.8, 0.8)$, and remove the baseline time shift of 2. Together, these settings alter the signal-to-noise ratios encountered during training. GenEval2 improves in both architectures while aesthetic scores change little (Table 16), showing that adapting the corruption process recovers part of the compositional gap before modifying the output head or supervision.

Pixel-level refinement. We replace the linear readout for each 32×32 RGB patch with a Pixel Transformer (PiT) head (Yu et al., 2026). It refines per-pixel features from the noisy image using backbone and timestep conditioning, enabling updates within each patch while retaining the same backbone token count. The sustained PRISM gains over the schedule-only model support finer-grained output processing alongside noise adaptation (Figure 11).

Perceptual supervision. We supplement the RGB velocity objective with LPIPS and DINOv2 losses (Ma et al., 2026). These compare predicted clean images with their targets through frozen visual networks, providing appearance and structure supervision in learned feature spaces. Adding them after noise adaptation and PiT refinement can further reduce performance gaps.

The full recipe brings pixel models within one GenEval2 point of their latent counterparts and roughly halves the PRISM gap in both architectures. Much of the baseline separation can therefore be reduced by adapting generation to RGB space, with compositional performance approaching the latent reference more closely than aesthetic quality.

6.3 System-Level Comparison

We compare JIG-Pixel with the full diffusion recipe and JIG-Latent-Diffusion from Table 3 against eight open-source generators with approximately 0.9B–3B backbones (Table 4). We additionally evaluate JIG-Latent† with Reve 2-style annotations as a system-level extension; all controlled ablations use Ideogram 4-style annotations. The comparison covers complete generation systems, including model-specific prompt enhancement where provided and structured prompt conversion for JIG. All systems generate one image for each of the same 800 GenEval2 and 700 PRISM prompts and follow the benchmark scoring protocols. Input processing and sampling settings appear in Section C. Trained on the 13.74M pairs in JIG-14M, JIG-Pixel and JIG-Latent outperform all eight external systems on GenEval2 and seven of the eight on PRISM alignment.

JIG-Latent† further improves system-level performance, reaching 73.82 on GenEval2 and 7.42 on PRISM total, gains of 10.84 and 1.13 points over the default latent system. This configuration differs from the default setup on both the training and inference sides. First, its training annotations combine hierarchical region records with detailed local descriptions. Second, at inference time, Reve 2's official API converts the original benchmark prompts into structured JSON conditioning. The observed gains therefore characterize the joint training-and-inference configuration; they do not isolate the contributions of annotation content, organization, or inference-time conversion. We report this configuration as a system-level extension, separate from the controlled textual-format and pixel–latent comparisons. These results suggest substantial headroom for improving structured-conditioning pipelines and motivate future iterations of JIG-14M toward richer Reve 2-style annotations.

Figure 5 presents qualitative comparisons across generation systems. The JIG models demonstrate stronger prompt adherence in these examples, integrating city buildings into the bird's wings and

combining the requested lettering with the specified garment colors. JIG-Latent† renders more distinct architectural details in the wings and finer surface textures in the living room, particularly in the sofa upholstery and the coffee table's wood grain. Additional comparisons appear in Section E.3.

Table 4: **System-level comparison with compact open-source generators.** Best and second-best scores in each metric are bold and underlined, respectively. †: Reve 2-style annotations.

Method	Backbone	Text encoder	Res.	Training data	GenEval2 ↑	PRISM ↑		
						Total	Alignment	Aesthetic
Pixel baselines								
MiniT2I-L/16	0.912B	0.34B	512^2	12M	26.02	4.90	5.57	4.22
PixelDiT	1.3B	2.61B	1024^2	26M	24.27	6.23	6.48	5.99
PixelGen-XXL/16	1.1B	1.72B	512^2	36M	18.33	5.71	5.76	5.66
PixNerd-XXL/16	1.2B	1.72B	512^2	45M	23.02	5.27	5.70	4.84
Latent baselines								
SANA-1.5	1.6B	2.61B	1024^2	50M	24.51	6.41	6.86	5.97
Infinity-2B	2.0B	1.22B	1024^2	—	23.55	5.98	6.17	5.80
i1-3B	3.0B	2.61B	512^2	163M	38.39	**7.60**	**8.06**	**7.13**
OpenUni-L	1.6B	1.54B	512^2	23M	19.95	6.20	6.27	6.13
Ours								
JIG-Pixel	1.4B	2B	512^2	14M	62.01	5.80	6.95	4.65
JIG-Latent	1.4B	2B	512^2	14M	<u>62.98</u>	6.29	7.17	5.42
JIG-Latent†	1.4B	2B	512^2	14M	**73.82**	<u>7.42</u>	<u>7.96</u>	<u>6.87</u>



Figure 5: **System-level visual comparison on PRISM.** Each row uses the same original prompt, processed by each system's inference pipeline. †: Reve 2-style diffusion.

7 Conclusion

We presented JIG, an open model family and empirical study of JSON-to-image generation, supported by JIG-14M with 13.7M spatially grounded image–annotation pairs. Our experiments show that structured annotations improve generation in both JSON and natural-language prose, highlighting the importance of annotation content beyond JSON syntax. Across four generators, baseline performance differs more between pixel and latent representations than between decoupled diffusion and unified Transfusion-style architectures. Within our controlled setting, pixel-specific training substantially narrows the representation gap in both architectures, bringing prompt following close to latent baselines while leaving a larger gap in aesthetic quality. The resulting systems perform competitively against compact open-source generators. Together, these findings suggest that JSON-to-image generation could become the foundation for next-generation image generation systems. We will release JIG-14M and all four JIG models to support reproducible research.

AI USE STATEMENT

Large Language Models assisted the research process in the following ways. For research ideation and execution, we used LLMs as a discussion partner to stress-test the experimental design and the interpretation of results, including feedback on which comparisons and controls were needed. For retrieval and discovery, we used LLMs with web search to identify candidate related work; every cited paper was read and verified by the authors, and the characterization of prior work is our own. In addition, we used LLMs to refine language, improve readability, and check grammar across the manuscript, including sentence rephrasing and adjustments to the flow of the text.

The authors take full responsibility for the content of the manuscript, including any text generated or polished by the LLM. We have ensured that the LLM-generated text adheres to ethical guidelines and does not contribute to plagiarism or scientific misconduct.

REFERENCES

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. Improving image generation with better captions. Technical report, OpenAI, 2023. URL https://cdn.openai.com/papers/dall-e-3.pdf.

Qi Cai, Jingwen Chen, Chengmin Gao, Zijian Gong, Yehao Li, Yingwei Pan, Yi Peng, Zhaofan Qiu, Kai Yu, Yiheng Zhang, Hao Ai, Siying Bai, Yang Chen, Zhihui Chen, Fengbin Gao, Ying Guo, Dong Li, Zhen Shen, Leilei Shi, Jing Wang, Siyu Wang, Yimeng Wang, Rui Zheng, Ting Yao, and Tao Mei. HiDream-O1-Image: A natively unified image generative foundation model with pixel-level unified transformer. *arXiv preprint arXiv:2605.11061*, 2026. URL https://arxiv.org/abs/2605.11061.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, Le Xue, Caiming Xiong, and Ran Xu. BLIP3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025. URL https://arxiv.org/abs/2505.09568.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. PixArt-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023. URL https://arxiv.org/abs/2310.00426.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. ShareGPT4V: Improving large multi-modal models with better captions. In *Computer Vision – ECCV 2024*, pp. 370–387, 2024. doi: 10.1007/978-3-031-72643-9_22. URL https://link.springer.com/chapter/10.1007/978-3-031-72643-9_22.

Zilong Chen, Chaorui Deng, Kunchang Li, Hongyi Yuan, and Haoqi Fan. Scaling properties of text conditioning in visual generation. *arXiv preprint arXiv:2607.29679*, 2026. URL https://arxiv.org/abs/2607.29679.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025. URL https://arxiv.org/abs/2505.14683.

Haiwen Diao, Jiahao Wang, Chenjing Ding, Hanming Deng, Jiangnan Chen, Ruixi Zhang, Ruohui Wang, Wenwen Tong, Xiangyu Fan, Yubo Wang, Yue Zhu, Yuwei Niu, Zhengqi Bai, Zhiqian Lin, Zhitao Yang, Zhongang Cai, Bo Yang, Chen Feng, Chengguang Lv, Guangjia Liu, Guanlin Wang, Hanyu Zhang, Haojia Yu, Hongcan Xiao, Hongli Wang, Huan Wu, Huaping Zhong, Jian Fang, Jianan Fan, Jiaqi Li, Jiefan Lu, Jing Zuo, Jingcheng Ni, Junxiang Xu, Linjun Dai, Mutian Xu, Peishen Yan, Penghao Wu, Ruijie Mao, Ruisi Wang, Shihao Bai, Shuang Yang, Shuya Yang, Shuyan Zheng, Silei Wu, Siying Li, Tao Chu, Tianbo Zhong, Tongxi Zhou, Weichao Luo, Weichen Fan, Wenhao Jia, Wenjie Gao, Xiangli Kong, Yan Li, Yang Yong, Zimo Wen, Zixuan Qian, Wenxiu Sun, Ruihao Gong, Quan Wang, Lewei Lu, Lei Yang, Ziwei Liu, and Dahua Lin.

SenseNova-U1.5: Towards native unified visual intelligence. *arXiv preprint arXiv:2609.11929*, 2026. URL https://arxiv.org/abs/2609.11929.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12606–12633. PMLR, 2024. URL https://proceedings.mlr.press/v235/esser24a.html.

Rongyao Fang, Aldrich Yu, Chengqi Duan, Linjiang Huang, Shuai Bai, Yuxuan Cai, Kun Wang, Si Liu, Xihui Liu, and Hongsheng Li. FLUX-Reason-6M & PRISM-Bench: A million-scale text-to-image reasoning dataset and comprehensive benchmark. *arXiv preprint arXiv:2509.09680*, 2025. URL https://arxiv.org/abs/2509.09680.

Eyal Gutflaish, Eliran Kachlon, Hezi Zisman, Tal Hacham, Nimrod Sarid, Alexander Visheratin, Saar Huberman, Gal Davidi, Guy Bukchin, Kfir Goldberg, and Ron Mokady. Generating an image from 1,000 words: Enhancing text-to-image with structured captions. *arXiv preprint arXiv:2511.06876*, 2025. URL https://arxiv.org/abs/2511.06876.

Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. *arXiv preprint arXiv:2412.04431*, 2024. URL https://arxiv.org/abs/2412.04431.

Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. Simple diffusion: End-to-end diffusion for high resolution images. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 13213–13232. PMLR, 2023. URL https://proceedings.mlr.press/v202/hoogeboom23a.html.

Ideogram AI. Ideogram 4.0 technical details: Open model at the forefront of design. Official technical blog, June 2026. URL https://ideogram.ai/blog/ideogram-4.0/. Accessed September 19, 2026.

Dengyang Jiang, Ruoyi Du, Zhennan Chen, Dongyang Liu, Zanyi Wang, Mingzhe Zheng, Xiangpeng Yang, Huanqia Cai, Aiming Hao, Yuming Jiang, Peng Gao, Harry Yang, and Steven Hoi. An empirical study of training pixel-space text-to-image diffusion models. *arXiv preprint arXiv:2608.16887*, 2026. URL https://arxiv.org/abs/2608.16887.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL https://kellerjordan.github.io/posts/muon/.

Amita Kamath, Kai-Wei Chang, Ranjay Krishna, Luke Zettlemoyer, Yushi Hu, and Marjan Ghazvininejad. GenEval 2: Addressing benchmark drift in text-to-image evaluation. *arXiv preprint arXiv:2512.16853*, 2025. URL https://arxiv.org/abs/2512.16853.

Gengluo Li, Xingyu Wan, Shangpin Peng, Weinong Wang, Hao Feng, Yongkun Du, Binghong Wu, Zheng Ruan, Zhiqiong Lu, Liang Wu, Pengyuan Lyu, Huawen Shen, Zibin Lin, Shijing Hu, Jieneng Yang, Hongbing Wen, Guanghua Yu, Hong Liu, Bochao Wang, Can Ma, Han Hu, Chengquan Zhang, and Yu Zhou. HunyuanOCR-1.5: Making lightweight OCR VLMs faster and better. *arXiv preprint arXiv:2607.04884*, 2026. URL https://arxiv.org/abs/2607.04884.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. *arXiv preprint arXiv:2511.13720*, 2025. URL https://arxiv.org/abs/2511.13720.

Kun-Yu Lin, Hongjun Wang, Weining Ren, and Kai Han. Panoptic captioning: An equivalence bridge for image and text. In *Advances in Neural Information Processing Systems*, volume 38, pp. 78553–78584, 2025. doi: 10.52202/085713-2631. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/7116cda41d75d580bae15d9e484a8466-Abstract-Conference.html.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Zhiheng Liu, Weiming Ren, Xiaoke Huang, Shoufa Chen, Tianhong Li, Mengzhao Chen, Yatai Ji, Sen He, Jonas Schult, Belinda Zeng, Tao Xiang, Wenhu Chen, Ping Luo, Luke Zettlemoyer, and Yuren Cong. Tuna-2: Pixel embeddings beat vision encoders for multimodal understanding and generation. *arXiv preprint arXiv:2604.24763*, 2026. URL https://arxiv.org/abs/2604.24763.

Zehong Ma, Ruihan Xu, and Shiliang Zhang. PixelGen: Improving pixel diffusion with perceptual supervision. *arXiv preprint arXiv:2602.02493*, 2026. URL https://arxiv.org/abs/2602.02493.

Nicholas Merchant, Haitz Sáez de Ocáriz Borde, Andrei Cristian Popescu, and Carlos Garcia Jurado Suarez. Structured captions improve prompt adherence in text-to-image models (Re-LAION-Caption 19m). *arXiv preprint arXiv:2507.05300*, 2025. URL https://arxiv.org/abs/2507.05300.

NVIDIA. Cosmos 3: Omnimodal world models for physical AI. *arXiv preprint arXiv:2606.02800*, 2026. URL https://research.nvidia.com/labs/cosmos-lab/cosmos3/technical-report.pdf.

Daniel A. P. Oliveira, Lourenço Teodoro, and David Martins de Matos. GroundCap: A visually grounded image captioning dataset. *arXiv preprint arXiv:2502.13898*, 2025. URL https://arxiv.org/abs/2502.13898.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. URL https://arxiv.org/abs/2304.07193.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Peebles_Scalable_Diffusion_Models_with_Transformers_ICCV_2023_paper.html.

Cihang Peng, Qiming Hou, Zhong Ren, and Kun Zhou. ROVI: A VLM-LLM re-captioned dataset for open-vocabulary instance-grounded text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20204–20214, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Peng_ROVI_A_VLM-LLM_Re-Captioned_Dataset_for_Open-Vocabulary_Instance-Grounded_Text-to-Image_Generation_ICCV_2025_paper.html.

Reve Team. The layout bet. Reve research blog, June 2026. URL https://blog.reve.com/posts/the-layout-bet/. Accessed September 19, 2026.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.html.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 25278–25294, 2022. doi:

10.52202/068431-1833. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/a1859debfb3b59d094f3504d5ebb6c25-Abstract-Datasets_and_Benchmarks.html.

Eyal Segalis, Dani Valevski, Danny Lumen, Yossi Matias, and Yaniv Leviathan. A picture is worth a thousand words: Principled recaptioning improves image generation. *arXiv preprint arXiv:2310.16656*, 2023. URL https://arxiv.org/abs/2310.16656.

Vasu Singla, Kaiyu Yue, Sukriti Paul, Reza Shirkavand, Mayuka Jayawardhana, Alireza Ganjdanesh, Heng Huang, Abhinav Bhatele, Gowthami Somepalli, and Tom Goldstein. From pixels to prose: A large dataset of dense image captions. *arXiv preprint arXiv:2406.10328*, 2024. URL https://arxiv.org/abs/2406.10328.

Shuai Wang, Ziteng Gao, Chenhui Zhu, Weilin Huang, and Limin Wang. PixNerd: Pixel neural field diffusion. *arXiv preprint arXiv:2507.23268*, 2025. URL https://arxiv.org/abs/2507.23268.

Xianbang Wang, Hanhong Zhao, Yiyang Lu, Kangyang Zhou, Linrui Ma, and Kaiming He. MiniT2I: A minimalist baseline for text-to-image generation, 2026. URL https://peppaking8.github.io/#/post/minit2i.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-Image technical report. *arXiv preprint arXiv:2508.02324*, 2025a. URL https://arxiv.org/abs/2508.02324.

Size Wu, Zhonghua Wu, Zerui Gong, Qingyi Tao, Sheng Jin, Qinyue Li, Wei Li, and Chen Change Loy. OpenUni: A simple baseline for unified multimodal understanding and generation. *arXiv preprint arXiv:2505.23661*, 2025b. URL https://arxiv.org/abs/2505.23661.

Enze Xie, Junsong Chen, Yuyang Zhao, Jincheng Yu, Ligeng Zhu, Chengyue Wu, Yujun Lin, Zhekai Zhang, Muyang Li, Junyu Chen, Han Cai, Bingchen Liu, Daquan Zhou, and Song Han. SANA 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer. *arXiv preprint arXiv:2501.18427*, 2025. URL https://arxiv.org/abs/2501.18427.

Qiying Yu, Quan Sun, Xiaosong Zhang, Yufeng Cui, Fan Zhang, Yue Cao, Xinlong Wang, and Jingjing Liu. CapsFusion: Rethinking image-text data at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14022–14032, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yu_CapsFusion_Rethinking_Image-Text_Data_at_Scale_CVPR_2024_paper.html.

Yongsheng Yu, Wei Xiong, Weili Nie, Yichen Sheng, Shiqiu Liu, and Jiebo Luo. PixelDiT: Pixel diffusion transformers for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14273–14282, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Yu_PixelDiT_Pixel_Diffusion_Transformers_for_Image_Generation_CVPR_2026_paper.html.

Z-Image Team, Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Aiming Hao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Yuming Jiang, Xin Jin, Liangchen Li, Zhen Li, Zhong-Yu Li, David Liu, Dongyang Liu, Qilong Wu, Feng Yu, Zechao Zhan, Chi Zhang, Shifeng Zhang, Ruikai Zhou, and Shilin Zhou. Z-Image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025. URL https://arxiv.org/abs/2511.22699.

Boya Zeng, Tianze Luo, Shu Pu, Jucheng Shen, Taiming Lu, Gabriel Sarch, and Zhuang Liu. i1: A simple and fully open recipe for strong text-to-image models. *arXiv preprint arXiv:2606.11289*, 2026. URL https://arxiv.org/abs/2606.11289.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_The_Unreasonable_Effectiveness_CVPR_2018_paper.html.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024. URL https://arxiv.org/abs/2408.11039.

Jiangning Zhu, Bowen Li, Shenyu Qiao, Yima Gu, Zhao Zhang, Yuhui Yuan, and Shixia Liu. Detect anything in graphic design: Element-level rewards for autoregressive detection. In *Advances in Neural Information Processing Systems*, 2026. URL https://arxiv.org/abs/2609.07072.

A JIG-14M Annotations and Data

A.1 Annotation Schema

The default Ideogram 4-style annotation contains a global description, style fields, and a composition block for the background and individual elements (Table 5). Each element record associates a description with its spatial extent, local color palette, and visible text where applicable.

Table 5: **Default Ideogram 4-style annotation schema.** Global and style fields describe the image as a whole. Composition fields organize the background and element records, which associate descriptions with boxes, colors, and visible text.

Scope	Field	Description
Global	`high_level_description`	Overall scene and main subjects.
Style	`aesthetics`	Overall visual character and atmosphere.
	`lighting`	Lighting conditions.
	`medium`	Visual medium, e.g., photograph or illustration.
	`photo / art_style`	Photographic properties or artistic style; mutually exclusive.
	`color_palette`	Dominant colors as hexadecimal RGB strings.
Composition	`background`	Background scene and environment.
	`elements`	List of element records.
Element	`type`	`obj` for objects or `text` for visible text.
	`bbox`	Four normalized coordinates in $[y_1, x_1, y_2, x_2]$ order, when applicable.
	`desc`	Element identity, appearance, attributes, and state.
	`text`	Literal visible characters; present for text elements.
	`color_palette`	Element colors as hexadecimal RGB strings, when applicable.

Bounding boxes use a top-left origin and integer coordinates normalized to $[0, 1000]$. For text elements, `text` preserves the visible characters without translation or correction, while `desc` describes their appearance and context, such as color and typography.

Reve 2-style annotations. The system-level extension (Section 6.3) uses a hierarchical variant with the same three top-level fields and shared description, box, palette, and literal-text fields. Element identifiers link the global description to individual records; parent references encode the element hierarchy (Table 6). Records span multiple granularities, including parts and scene regions.

A.2 Annotation Pipelines

A.2.1 Ideogram 4-style annotations

Qwen3.5-9B constructs structured annotations from the source image, DAD bounding boxes, and HunyuanOCR-1.5 transcriptions. The construction prompt specifies these integration rules.

Table 6: **Reve 2-style schema extensions.** Additional fields and conventions relative to Table 5.

Field	Convention
`type`	`coarse`, `medium`, and `fine` indicate granularity; `face`, `hand`, and `text` identify specialized elements.
`label`	Element identifier, also referenced in `high_level_description`.
`parent`	Parent element's label, or `ROOT` for a top-level record.
`entity`	Optional name for a named entity.

Element selection and geometry. The model is instructed to identify coherent visual or text elements and select the best matching candidate boxes. Matching coordinates are copied as complete arrays in the normalized $[y_1, x_1, y_2, x_2]$ convention; boxes that would be misleading are omitted. Scene-wide surfaces and distant scenery belong in the background field, while repeated decorative motifs can be grouped into element records. For text-rich designs, the prompt prioritizes useful text-layer boxes without imposing a fixed element-count cap.

Text integration. OCR fragments may be merged when they belong to the same word, line, or continuous text layer, preserving visual reading order. Separate labels and containers remain distinct. For merged text, the model is instructed to use the matching layer box when it tightly covers the text, or the OCR box or union of OCR boxes otherwise. The prompt requires visually supported transcriptions and permits direct image reading when OCR is missing. Illegible or contradicted candidates are omitted. Literal characters populate `text`; `desc` records typography and appearance.

A.2.2 Reve 2-style annotations

Gemini-3.5-Flash-Lite generates hierarchical JSON annotations directly from source images. The prompt prioritizes tight boxes and recursively separates groups, distinguishable members, and salient parts, including faces and hands. Parent references link these regions, and the global description references their labels to connect scene context with localized elements.

Within each box, descriptions specify geometry, pose, material, color, lighting, and occlusion. Legible text is transcribed verbatim, preserving script, case, punctuation, and line breaks.

A.2.3 Dense captions

Gemini-3.5-Flash-Lite generates dense training captions directly from source images. The prompt requests descriptions of subjects, counts, attributes, spatial relationships, composition, lighting, and style, with detail scaled to image complexity. Legible text is quoted verbatim in context, preserving its original script, case, and punctuation; illegible characters must not be guessed.

A.3 Dataset Curation

We construct JIG-14M from 31.98M source records in BLIP3o's short- and long-caption partitions (Chen et al., 2025). Table 7 reports curation stages and retained counts by source partition.

To reduce exposure to portraits with pixelated faces and AI-generated imagery, we exclude the upstream subset labeled as "people" from the long-caption partition. Within this partition, we also use source identifiers to remove AI-generated images originating from JourneyDB.

We adjust bounding-box coordinates for image resizing and cropping. Records with missing or empty annotations after adjustment are discarded. Image filtering uses quality metadata on clarity, texture, compression artifacts, exposure, and color. We also remove duplicate records using existing exclusion lists and discard missing or unreadable images. Retained images must have dimensions divisible by the training patch size. These stages yield 13.74M image–annotation pairs.

A.4 Dataset Statistics and Comparison

Content distribution. We estimate content statistics from 50,000 Ideogram 4-style records sampled at random in proportion to the two source partitions. We extract 24 topics from the global scene

Table 7: **Dataset curation and final size of JIG-14M.** Records retained at each stage by source partition; the final row gives the training image–annotation pairs.

Stage	Short-caption	Long-caption	Total
Packaged source partitions	4,783,500	27,195,389	31,978,889
After excluding "people" (long)	4,783,500	16,876,572	21,660,072
After JourneyDB exclusion	4,783,500	13,966,587	18,750,087
Records with adjusted annotations	4,076,224	12,329,154	16,405,378
Final training pairs	**2,989,350**	**10,749,817**	**13,739,167**

descriptions using non-negative matrix factorization of TF–IDF features. Characteristic terms and representative descriptions guide topic naming and grouping into the five broader categories. Topic shares are calculated from dominant-topic assignments over the full sample.

The resulting distribution spans built spaces, nature, objects, people-related subjects, and designed content (Figure 6). Built spaces form the largest group (38.9%), while designed content (18.4%) includes graphics, text, cartoons, and printed materials.



Figure 6: **Visual content coverage of JIG-14M.** Five content groups form the inner ring, with 24 topics and their estimated shares on the outer ring. Each topic links to a dataset example.

Field coverage and annotation statistics. Figure 7 compares the Ideogram sample with 50,000 independently and uniformly sampled records from the 13.58M-record Reve 2-style BLIP3o corpus. We use inverse-inclusion weights for Ideogram and uniform weights for Reve. Statistics use stored annotations before training-time transformations; samples are not paired by image. Element counts include hierarchy nodes, parts, and scene regions, and box-area distributions use valid coordinates.

We tokenize compact JSON objects with Qwen3-VL-2B-Instruct, excluding storage wrappers, chat formatting, and special tokens. Annotated text length sums characters in the `text` fields of text elements, including spaces and punctuation; panel (d) includes only images with annotated text.

Ideogram and Reve annotations average 5.48 and 13.08 elements per image and 479 and 1,640 JSON tokens, respectively. Text annotations appear in 37.5% and 43.4% of images; median non-text box areas are 8.35% and 4.03% of image area. In the default Ideogram annotations, 99.1% of element records include boxes and 88.4% include local color palettes.

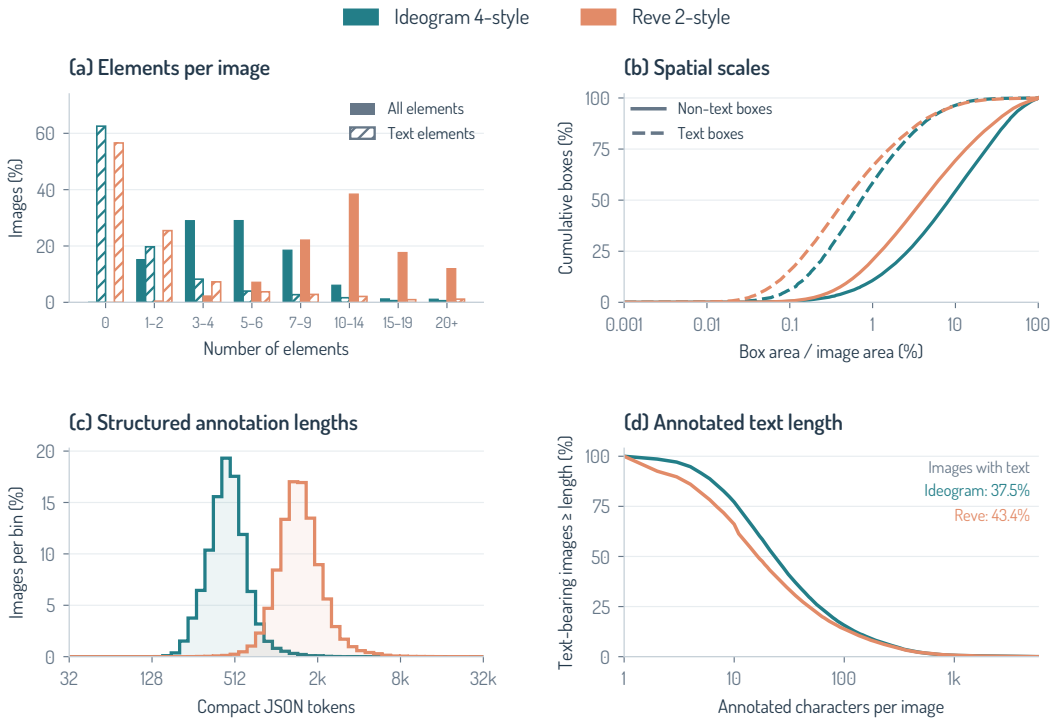


Figure 7: **Ideogram 4-style and Reve 2-style annotation statistics.** (a) All and text elements per image. (b) Cumulative box-area distributions, normalized by image area. (c) Compact JSON token lengths. (d) Percentage of text-bearing images with at least the indicated number of annotated characters. Colors identify formats throughout; panels (b–d) use logarithmic horizontal axes.

Comparison with public caption datasets. Table 8 situates JIG-14M among public caption datasets in scale and annotation format. Alongside scene descriptions, JIG-14M makes element locations and literal text directly accessible through explicit fields.

Table 8: **Annotation coverage of public caption datasets.** The comparison reports image count, caption format, and the presence of explicit element, bounding-box, and literal-text fields. Checkmarks refer to fields provided by the annotation format.

Dataset	Images	Format	Element records	Boxes	Text field
PixelProse (Singla et al., 2024)	16.9M	Dense prose	–	–	–
BLIP3o (long-caption) (Chen et al., 2025)	27.0M	Dense prose	–	–	–
Re-LAION-Caption (Merchant et al., 2025)	19.0M	Templated prose	–	–	–
ROVI (Peng et al., 2025)	1.01M	Grounded prose	✓	✓	–
SA-PanCap (Lin et al., 2025)	9.6K	Panoptic	✓	✓	–
GroundCap (Oliveira et al., 2025)	52.0K	Grounded caption	✓	✓	–
JIG-14M	**13.7M**	Structured JSON	✓	✓	✓

We will release the structured JSON annotations linked to their source images, together with the annotation schema and deterministic format converters.

A.5 Annotation Quality

We assess the default Ideogram 4-style annotations on a source-proportional random sample of 10,000 images using GPT-5.6-Luna. Visual checks use full images and region crops, yielding 9,483 valid image assessments; 517 images with invalid checker outputs are excluded. Object inventories are generated from images alone; separate model calls transcribe text regions without candidate strings or descriptions. Text checks complete for 9,999 images.

Description and attribute accuracy measure the fraction of atomic claims visually supported by the image; palette colors are checked for coarse compatibility. Box–instance match checks whether each box corresponds to its intended instance, separately from box extent. Object coverage matches salient single objects from image-only inventories to element records, excluding groups, parts, background, and text; global-only mentions do not count. Text coverage requires content and spatial correspondence. Redundancy flags records describing the same entity at the same granularity without additional information. Text exact match compares complete-region transcriptions for 10,778 of 14,283 records with fully readable references and valid annotation strings (75.5%). Unicode NFC and whitespace normalization preserve case, punctuation, and spelling.

Table 9 reports source-weighted rates and 95% percentile intervals from 2,000 within-source bootstrap resamples of images. Applicable denominators include unverifiable judgments; non-applicable checks are excluded. Intervals quantify sampling variability with the VLM judgments fixed.

Table 9: **Annotation assessment details.** Counts report unweighted positives over applicable units; rates and 95% intervals are source-weighted.

Measure	Positive / Applicable	Rate (%) [95% CI]
Description accuracy	118,005/122,985	96.0 [95.8, 96.1]
Box–instance match	49,724/50,406	98.6 [98.4, 98.8]
Attribute accuracy	62,433/64,054	97.5 [97.3, 97.6]
Object coverage	26,486/43,369	61.1 [60.4, 61.7]
Text exact match	8,443/10,778	78.3 [77.0, 79.7]
Global/style/background fact support	167,315/184,058	90.9 [90.8, 91.0]
Reasonable box extent	46,424/50,406	92.1 [91.8, 92.4]
Image-only text inventory coverage	9,690/13,538	71.6 [70.2, 73.0]
Records flagged as redundant	24/50,777	0.05 [0.02, 0.08]

B Model and Training Details

B.1 Caption Variant Construction

Training in Section 5 uses dense captions or finalized structured annotations expressed as JSON or natural-language prose. Figure 8 illustrates these three forms for the same training image.

Dense captions. The dense baseline uses the image-derived `full_description` captions from Section A.2.3, paired with the same training images.

JSON realization. The JSON realization serializes the finalized structured annotations with fields and element records in a predefined order.

Prose realization. The prose converter uses fixed sentence templates for nonempty fields, describing the global scene and style, then the background and individual elements. It preserves descriptions, normalized numerical coordinates, literal text, and hexadecimal color values, together with their element associations. Conversion is deterministic and uses only the structured record.

B.2 Model Configuration

Table 10 lists the core Transformer settings shared by the baseline JIG models. In JIG-Diffusion, the frozen text encoder sums hidden states from layers 1, 4, 7, 10, 13, 16, 19, 22, 25, and 28 into a 2,048-dimensional feature per token. The summed features are projected into the denoiser and concatenated with noisy-image tokens. Timestep embeddings modulate the attention and feed-forward sublayers through AdaLN, using a modulation network shared across Transformer blocks.

JIG-Transfusion retains the pretrained language backbone and token embeddings, omitting the original vision tower. A continuous timestep embedding replaces the embedding of a dedicated token preceding the image. Caption tokens provide conditioning and carry no language-modeling loss.

Training image



• RYANAIR

Dense caption

The image captures a Ryanair Boeing 737-800 passenger aircraft in the final stages of its landing approach. The plane is positioned in the midground, captured from a side profile view as it moves from left to right across the frame. The fuselage is predominantly white, featuring the prominent, bold blue text "**RYANAIR**" displayed along the upper portion of the aircraft's side. The underside of the plane is painted in a deep blue, which continues toward the tail section. The aircraft's landing gear is deployed and visible, hanging beneath the wings and center fuselage. A single, large, yellow-tipped winglet is visible at the end of the wing on the right side of the frame, angled upward. [...] The photograph appears to be taken with a telephoto lens, compressing the distance between the plane and the background scenery.

JSON

{ **"high_level_description"**: "A Ryanair Boeing 737-800 passenger aircraft is captured in a side profile view during its landing approach, flying low over an airfield landscape under a cloudy sky.",
"style_description": { **"lighting"**: "soft, diffused daylight under an overcast sky", **"medium"**: "photograph" } }

{ **"type"**: "text", **"bbox"**: [458, 68, 518, 500], **"text"**: "**RYANAIR**", **"desc"**: "large bold blue sans-serif lettering displayed along the upper side of the white fuselage", **"color_palette"**: ["#1E3A8A"] }

Prose

A Ryanair Boeing 737-800 passenger aircraft is captured in a side profile view during its landing approach, flying low over an airfield landscape under a cloudy sky. The image style is described by the following details. [...] soft, diffused daylight under an overcast sky. [...] The medium is photograph. [...]

The second element is text, and occupies normalized vertical coordinates 458 through 518 and horizontal coordinates 68 through 500. Its exact text is **RYANAIR**. large bold blue sans-serif lettering displayed along the upper side of the white fuselage. Its color palette contains #1E3A8A.

Selected fields are shown; [...] marks omitted text.

Figure 8: **Caption realizations for one training image.** JSON and prose express selected image-level fields followed by one complete text element. Rows and colors link corresponding content. Caption passages and field values are reproduced verbatim, with omissions marked by [...].

Table 10: **Common baseline backbone configuration.** The two architectures share these core Transformer settings. Pixel-specific architectural changes are described in Section B.4.

Setting	Value
Transformer blocks	28
Hidden dimension	2,048
FFN intermediate dimension	6,144
Attention heads / key–value heads	16 / 8
Head dimension	128
Feed-forward activation	SwiGLU
Normalization	RMSNorm with QK normalization
Positional encoding	Multimodal RoPE

Image tokenization. For a 512×512 image, the FLUX.2 VAE encoder produces a 64×64 latent grid with 32 channels, corresponding to $8\times$ spatial downsampling. We apply 4×4 patchify to the noisy latent grid and linearly embed each patch, yielding a 16×16 grid of 256 image tokens. Pixel models instead apply 32×32 patchify directly to noisy RGB images, yielding the same token count. Linear readouts predict latent velocities or clean RGB patches in the baseline models.

B.3 Training Objectives and Optimization

For a clean image or latent representation x_0, Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$, and timestep t, the baseline noisy input follows the linear flow-matching path (Lipman et al., 2023), $x_t = t\,x_0 + (1-t)\epsilon$.

Conditioned on caption c, latent models predict velocity directly and minimize

$$\mathcal{L}_{\text{latent}} = \mathbb{E}\big[\|v_\theta(x_t, t, c) - (x_0 - \epsilon)\|_2^2\big]. \tag{1}$$

Pixel models predict the clean image $\hat{x}_{0,\theta}(x_t, t, c)$. We convert the prediction and target to velocity using the same denominator floor, $d(t) = \max(1 - t, t_\epsilon)$ with $t_\epsilon = 0.05$:

$$\hat{v}_\theta = \frac{\hat{x}_{0,\theta}(x_t, t, c) - x_t}{d(t)}, \qquad v^\star = \frac{x_0 - x_t}{d(t)}. \tag{2}$$

Thus, pixel loss is a squared velocity error, or a timestep-weighted clean-image error:

$$\mathcal{L}_{\text{pixel}} = \mathbb{E}\big[\|\hat{v}_\theta - v^\star\|_2^2\big] = \mathbb{E}\left[\frac{\|\hat{x}_{0,\theta}(x_t, t, c) - x_0\|_2^2}{d(t)^2}\right]. \tag{3}$$

For baseline timestep sampling, we draw $u = \text{sigmoid}(z)$ with $z \sim \mathcal{N}(0, 1)$ and apply the shift $t = u/[u + 2(1 - u)]$.

Optimization. Table 11 summarizes the baseline training settings. Muon (Jordan et al., 2024) updates Transformer-block weight matrices and selected timestep and modulation projections, while AdamW updates the remaining trainable parameters. We use a constant base learning rate of 2×10^{-4} and maintain an exponential moving average of the denoiser. Training runs on 16 H200 GPUs with variable-length sequence packing and a global batch size of up to 640 image–caption pairs. Captions are truncated to 2,048 tokens, and attention is isolated between packed samples.

Table 11: **Baseline training configuration.** Pixel-specific adaptations build on these settings.

Setting	Value
Optimizer	Muon + AdamW
Base learning rate	2×10^{-4}
Learning-rate schedule	Constant
Muon momentum	0.95
AdamW (β_1, β_2)	$(0.9, 0.95)$
Weight decay	0
Global-norm gradient clipping	1.0
EMA decay	0.9999
Timestep distribution	Logit-normal $(0, 1)$
Timestep shift	2.0
Noise standard deviation	1.0
Velocity denominator floor t_ϵ	0.05
Caption dropout	0.1
Global batch size	640

B.4 Pixel-Space Training Adaptations

Starting from the baseline settings in Section B.3, we add the following adaptations cumulatively in both architectures, following the order in Table 3.

Pixel-adapted noise schedule. At 512^2 resolution, we adopt JiT's noise scaling and logit-normal timestep distribution (Li & He, 2025):

$$\epsilon \sim \mathcal{N}(0, 4I), \qquad z \sim \mathcal{N}(-0.8, 0.8^2), \qquad t = \text{sigmoid}(z), \qquad x_t = t\, x_0 + (1 - t)\epsilon. \tag{4}$$

Relative to Section B.3, this doubles the noise standard deviation, changes the logit-normal parameters, and removes the time shift. The clean-image prediction and velocity conversion in Equations (2) and (3) remain unchanged. At inference, the sampler starts from the same scaled Gaussian distribution and uses an unshifted time grid.

Pixel-level refinement. The Pixel Transformer (PiT) head from PixelDiT (Yu et al., 2026) replaces the linear output layer. For each backbone image token h_i, we form the condition $s_i = \text{SiLU}(h_i + e(t))$, where $e(t)$ is the timestep embedding. A learned splitter adds four distinct offsets to s_i,

yielding one condition for each 16×16 subregion of the original 32×32 patch. Each noisy RGB pixel is embedded into 16-dimensional features with two-dimensional positional embeddings.

Each of the two PiT blocks produces per-pixel AdaLN scale, shift, and gating parameters from the regional conditions. Attention groups 16×16 pixel features and projects each group to a 1,536-dimensional token. Global 12-head attention processes the resulting 1,024 tokens at 512^2 resolution, after which a learned projection restores the pixel features. A per-pixel MLP with expansion ratio 4 completes the block, and the final linear readout predicts clean RGB values. With PiT, JIG-Diffusion uses 25 backbone blocks to keep denoiser parameters near the baseline count; JIG-Transfusion retains all 28 pretrained blocks. Both use 256 backbone image tokens.

Perceptual supervision. Following the image-space perceptual supervision in PixelGen (Ma et al., 2026), we compare the predicted clean image $\hat{x}_{0,\theta}$ with the training image x_0 using LPIPS (Zhang et al., 2018) and DINOv2 (Oquab et al., 2023). The combined objective is

$$\mathcal{L} = \mathcal{L}_{\text{pixel}} + 0.1\,\mathcal{L}_{\text{LPIPS}} + 0.01\,\mathcal{L}_{\text{DINO}}. \tag{5}$$

The perceptual terms are averaged over samples with $t \geq 0.3$ and are zero when no sample meets this condition. LPIPS uses a frozen VGG-16 network on the native-resolution RGB images. The DINO term averages cosine distance between corresponding patch features from the final block of a frozen DINOv2 ViT-B/14. Its inputs are mapped to $[0, 1]$, ImageNet-normalized, and bicubically resized by a factor of $14/32$ in each spatial dimension (512×512 becomes 224×224). Gradients pass through the prediction branches of these networks to the generator; target features are detached.

C Evaluation Protocols

Benchmark inputs. We evaluate on 800 GenEval2 prompts and 700 English PRISM prompts, with input construction illustrated in Figure 9.

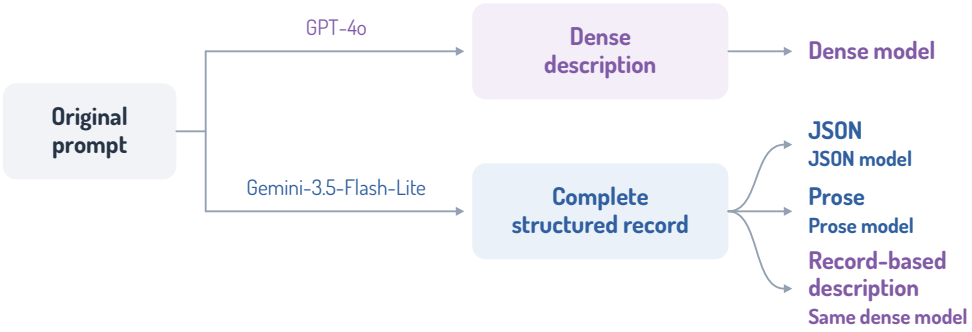
Gemini-3.5-Flash-Lite converts each benchmark's original prompt directly into structured JSON, retaining the requested content and organizing it into scene-level fields and element records. Individual object instances receive separate records with descriptions and bounding boxes that express their spatial relationships. Requested visible strings are copied verbatim into text elements. Within each schema, JIG experiments reuse the same precomputed records.

For both benchmarks, each complete record yields three input forms. JSON serializes the record directly, while prose applies the deterministic templates in Section B.1, preserving field values and their element associations. These forms condition the JSON- and prose-trained models, respectively.

For the dense-caption model, Gemini-3.5-Flash-Lite renders the complete record as a detailed, free-form *record-based description*. The record is the sole content source, and the rewriter is instructed to retain scene context, style, and all element information, including attributes, counts, relationships, and exact visible text. Bounding boxes become descriptions of position, relative size, and spatial extent, while palettes become color words. A separate verification pass checks information coverage and unsupported additions. Separately, GPT-4o expands the original prompts into dense descriptions for the dense-caption baseline, following Kamath et al. (2025). The dense-caption model uses the same weights and sampling settings with these expanded descriptions and the record-based inputs.

Scoring. GenEval2 uses the original question–answer annotations and Qwen3-VL-8B for Soft-TIFA evaluation. We report the geometric-mean score and per-skill results on a 0–100 scale. For PRISM, GPT-4.1 scores generations against the original prompts using the benchmark's category-specific alignment criteria and shared aesthetic criteria (Fang et al., 2025). We report alignment and aesthetic scores on a ten-point scale and their arithmetic mean as the total.

External baselines. Table 12 summarizes the evaluated weights and sampling settings; JIG uses the sampler in Section 4. Infinity uses its GPT-4o rewriter and the returned guidance scale. i1 uses its Qwen3-30B-A3B rewriter, with released rewrites for PRISM. The remaining external systems use GPT-4o-expanded GenEval2 descriptions and original English PRISM prompts. PixelDiT and SANA retain their text-encoder instruction prefixes. All systems use the scoring procedure above.



Example prompt

The logo features a green leaf symbol above the bold text "**RAINFOREST PARTNERSHIP™**" centered against a white background.

Dense description

[...] **Below the leaf**, the bold text "**RAINFOREST PARTNERSHIP™**" is centered in a strong, sans-serif font. The text is in a **deep, forest green** color, complementing the leaf symbol above. [...]

Record-based description

[...] **Directly below this leaf symbol, in the lower-middle section**, is the bold, centered uppercase brand name typography reading "**RAINFOREST PARTNERSHIP™**" in a **dark green** color. [...]

JSON

{
"type": "text",
"text": "**RAINFOREST PARTNERSHIP™**",
"bbox": **[500, 220, 620, 780]**,
"desc": "Bold, centered uppercase brand-name typography beneath the leaf symbol.",
"color_palette": ["**#1F6F43**"]
}

Prose

[...] The second element is text, and occupies normalized vertical coordinates **500 through 620** and horizontal coordinates **220 through 780**. Its exact text is **RAINFOREST PARTNERSHIP™**. Bold, centered uppercase brand-name typography beneath the leaf symbol. Its color palette contains **#1F6F43**.

The same brand-name element is shown in all four forms. [...] marks omitted text.

Figure 9: **Benchmark evaluation inputs.** Original prompts independently yield dense descriptions and structured records. Four input forms show the same brand-name element; [...] marks omissions.

Table 12: **Inference settings for external baselines.** Resolutions and model sizes appear in Table 4.

Method	Weights	Sampler	Steps	CFG
MiniT2I-L/16 (Wang et al., 2026)	Pretrained, 250k	Euler	100	6.0
PixelDiT (Yu et al., 2026)	1300M	DPM++	50	3.5
PixelGen-XXL/16 (Ma et al., 2026)	EMA	AdamLM	25	4.0
PixNerd-XXL/16 (Wang et al., 2025)	EMA	AdamLM	25	4.0
SANA-1.5 (Xie et al., 2025)	1.6B	DPM++	20	4.5
Infinity-2B (Han et al., 2024)	2B reg.	Autoregressive	—	1.0 or 3.0
i1-3B (Zeng et al., 2026)	512-resolution	Euler	250	12.0
OpenUni-L (Wu et al., 2025b)	BLIP3o-60k	DPM++	20	4.5

DPM++ uses second-order updates with flow-matching time grids and shift 4 for PixelDiT or 3 for SANA and OpenUni. AdamLM uses second-order updates and shift 3. i1 uses time shift 0.3 and guidance rescaling 1.0. Infinity uses temperature 0.5 and prompt-dependent guidance.

D Additional Caption Analyses

D.1 Benchmark Breakdowns

Table 13 expands the aggregate comparison in Table 2. JSON and prose improve both prompt alignment and aesthetic quality over the dense-caption model, including when that model receives descriptions of the complete evaluation records.

Table 13: **Annotation content and textual realization under latent diffusion.** The two dense-caption rows share model weights and differ only in evaluation input. JSON and prose express the same finalized annotations. Best scores are bold.

Training caption	Evaluation input	GenEval2 ↑	PRISM ↑		
			Total	Alignment	Aesthetic
Dense captions	Dense descriptions	36.26	5.79	6.52	5.07
Dense captions	Record-based descriptions	44.74	5.94	6.59	5.29
Structured annotations (JSON)	Records (JSON)	62.98	**6.29**	**7.17**	**5.42**
Structured annotations (Prose)	Records (Prose)	**63.23**	**6.29**	**7.17**	**5.42**

GenEval2 skill profile. Richer inference descriptions recover much of the dense-caption model's positioning gap and bring its attribute score close to the structured-annotation models (Table 14). Counting remains the clearest distinction: JSON and prose score 78.95 and 78.76, compared with 61.51 for the dense model with record-based inputs. The dense model already leads on object presence with its original inputs and on verbs with record-based inputs. Structured annotations thus yield a different skill balance, with counting gains in both JSON and prose.

Table 14: **GenEval2 skill scores for caption variants under latent diffusion.** Scores use a 0–100 scale; higher is better. Best scores are bold.

Training caption	Evaluation input	Object	Attribute	Count	Position	Verb
Dense captions	Dense descriptions	**93.78**	90.58	56.14	69.31	32.87
Dense captions	Record-based descriptions	92.21	91.26	61.51	81.57	**38.96**
Structured annotations (JSON)	Records (JSON)	90.05	91.24	**78.95**	81.56	35.78
Structured annotations (Prose)	Records (Prose)	90.12	**91.43**	78.76	**83.23**	34.58

PRISM breakdown. The effect of richer inference inputs also varies across PRISM categories (Table 15). Record-based descriptions improve the dense model's Imagination score most strongly, while Composition and Text change little. The largest remaining advantages of the structured-annotation models are in Text and Entity, showing that richer descriptions alone do not reproduce their gains on these categories. JSON and prose again exhibit similar profiles.

Table 15: **PRISM category scores for caption variants.** Scores average prompt alignment and aesthetic quality on a 0–10 scale; higher is better. *Long text* concerns complex, extended prompts, whereas *Text* concerns visible-text rendering. Best scores are bold.

Training caption	Evaluation input	Affection	Composition	Entity	Imagination	Long text	Style	Text
Dense captions	Dense descriptions	6.46	7.33	5.13	6.41	5.55	6.81	2.85
Dense captions	Record-based descriptions	**6.59**	7.37	4.87	7.25	5.71	6.90	2.89
Structured annotations (JSON)	Records (JSON)	6.37	7.69	5.50	**7.58**	**6.14**	6.93	**3.83**
Structured annotations (Prose)	Records (Prose)	6.49	**7.71**	**5.51**	7.52	6.01	**7.01**	3.81

D.2 Training Dynamics

Figure 10 compares caption variants from 60k to 150k training steps under latent diffusion. JSON and prose follow similar trajectories and outperform dense captions on both benchmarks throughout. The endpoint comparison in Table 2 thus reflects this consistent trend.



Figure 10: **Training dynamics of caption variants under latent diffusion.** The dense baseline uses prompt-derived descriptions at evaluation; JSON and prose express the same complete records.

E Additional Generator Analyses

E.1 Benchmark Breakdowns

Table 16 reports alignment and aesthetic scores for all configurations in Table 3. Noise adaptation primarily improves GenEval2, while PiT refinement and perceptual supervision improve both PRISM components. The full recipe closes more of the latent gap in alignment than in aesthetics.

Table 16: **Cumulative training adaptations for pixel generation.** Cells report JIG-Diffusion / JIG-Transfusion scores, with the best per architecture in bold. JIG-Latent provides the reference.

Setting	GenEval2 ↑	PRISM ↑		
		Total	Alignment	Aesthetic
JIG-Pixel	48.35 / 52.14	5.32 / 5.26	6.45 / 6.28	4.20 / 4.25
+ Pixel-adapted schedule	57.19 / 56.38	5.38 / 5.37	6.57 / 6.44	4.20 / 4.29
+ Pixel-level refinement	58.37 / 59.61	5.54 / 5.61	6.71 / 6.68	4.37 / 4.53
+ Perceptual supervision	62.01 / 64.27	5.80 / 5.86	6.95 / 6.95	4.65 / 4.77
JIG-Latent	**62.98 / 65.11**	**6.29 / 6.43**	**7.17 / 7.19**	**5.42 / 5.68**

The full recipe approaches latent generation in overall GenEval2 score with a distinct skill profile (Table 17). It improves all five skills in both architectures, with the largest gains over the pixel baselines in verbs and object presence. Pixel models then exceed their latent counterparts in positioning and verbs and come close in counting, while attributes retain a larger gap.

Table 17: **GenEval2 skill scores for pixel and latent generators.** Baseline and full-recipe results are reported for both architectures on a 0–100 scale; higher is better. Best scores are bold.

Setting	Object	Attribute	Count	Position	Verb
JIG-Latent-Diffusion	90.05	91.24	78.95	81.56	35.78
JIG-Pixel-Diffusion	79.13	84.29	69.17	74.81	24.59
JIG-Pixel-Diffusion (full recipe)	89.34	89.08	78.61	83.21	**39.22**
JIG-Latent-Transfusion	**92.76**	**91.83**	**80.16**	82.61	30.10
JIG-Pixel-Transfusion	81.46	83.13	72.46	77.20	25.16
JIG-Pixel-Transfusion (full recipe)	91.25	87.51	79.95	**83.32**	39.00

PRISM likewise shows uneven recovery across prompt categories (Table 18). The full recipe brings pixel models close to their latent counterparts on Text, with pixel diffusion scoring higher than latent diffusion. Imagination and Composition retain the largest gaps in both architectures.

Table 18: **PRISM category scores for pixel and latent generators.** Each score averages prompt alignment and aesthetic quality on a 0–10 scale. Best scores are bold.

Setting	Affection	Composition	Entity	Imagination	Long text	Style	Text
JIG-Latent-Diffusion	6.37	7.69	5.50	7.58	**6.14**	6.93	3.83
JIG-Pixel-Diffusion	5.86	6.46	4.50	6.21	5.07	6.00	3.17
JIG-Pixel-Diffusion (full recipe)	6.17	6.88	5.14	6.54	5.42	6.31	**4.14**
JIG-Latent-Transfusion	**6.74**	**7.84**	**5.59**	**7.62**	5.79	**7.38**	4.08
JIG-Pixel-Transfusion	5.67	6.51	4.50	6.04	4.89	6.17	3.08
JIG-Pixel-Transfusion (full recipe)	6.28	7.00	5.19	6.46	5.50	6.58	4.01

E.2 Training Dynamics and Visual Comparisons

Training dynamics. Figure 11 shows sustained gains from pixel adaptations over 60k to 150k steps. The schedule raises GenEval2 in both architectures, and PiT adds consistent PRISM gains over the schedule-only model. Perceptual supervision further improves both benchmarks.

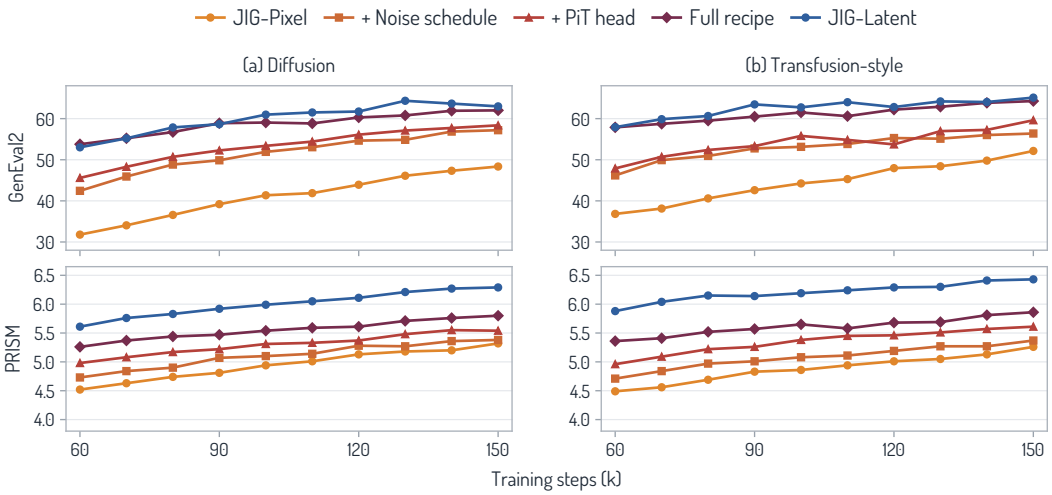


Figure 11: **Training dynamics of the pixel-specific recipe.** Intermediate configurations cumulatively add the adapted noise schedule and PiT head; the full recipe additionally uses perceptual supervision. JIG-Latent provides the reference in each architecture.

Pixel and latent outputs. Figure 12 compares the four JIG variants under the same structured prompts. Differences in appearance remain visible when the outputs contain the same main subjects: latent models render more varied textures across the toast, toppings, and sauce, and softer washes in the watercolor café. Pixel models use stronger outlines in the latter example.

Visual effects of the adaptations. Figure 13 illustrates how the cumulative adaptations affect different visual details under diffusion. In the mountain scene, the adapted schedule makes the winding wall more distinct. In the sweatshirt example, the configurations with PiT refinement and perceptual supervision render the graphic and lettering more clearly.

E.3 Additional System-Level Comparisons

Figure 14 extends the visual comparison to eight additional PRISM prompts, covering object counts, typography, recognizable subjects, and varied rendering styles.

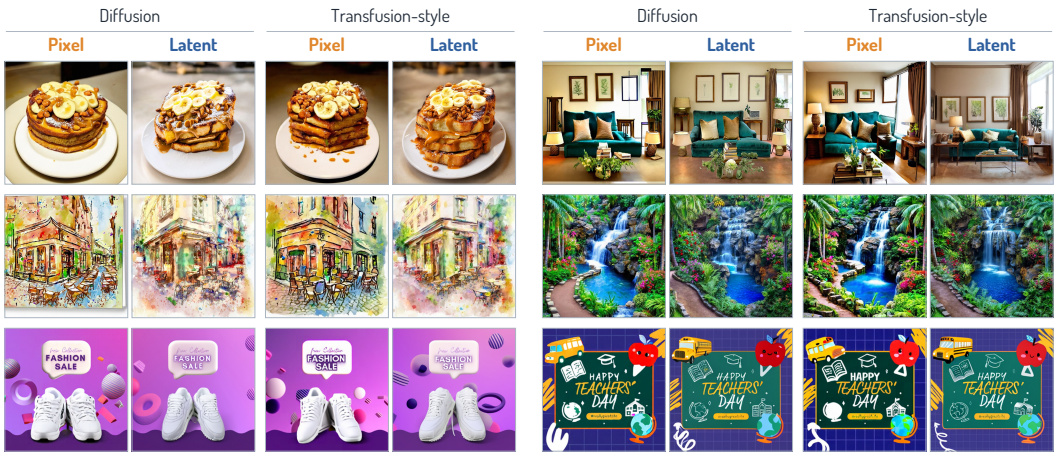


Figure 12: **Visual comparison of pixel and latent generators.** Four JIG models share each structured prompt. Six cases cover photography, watercolor, and graphic design.

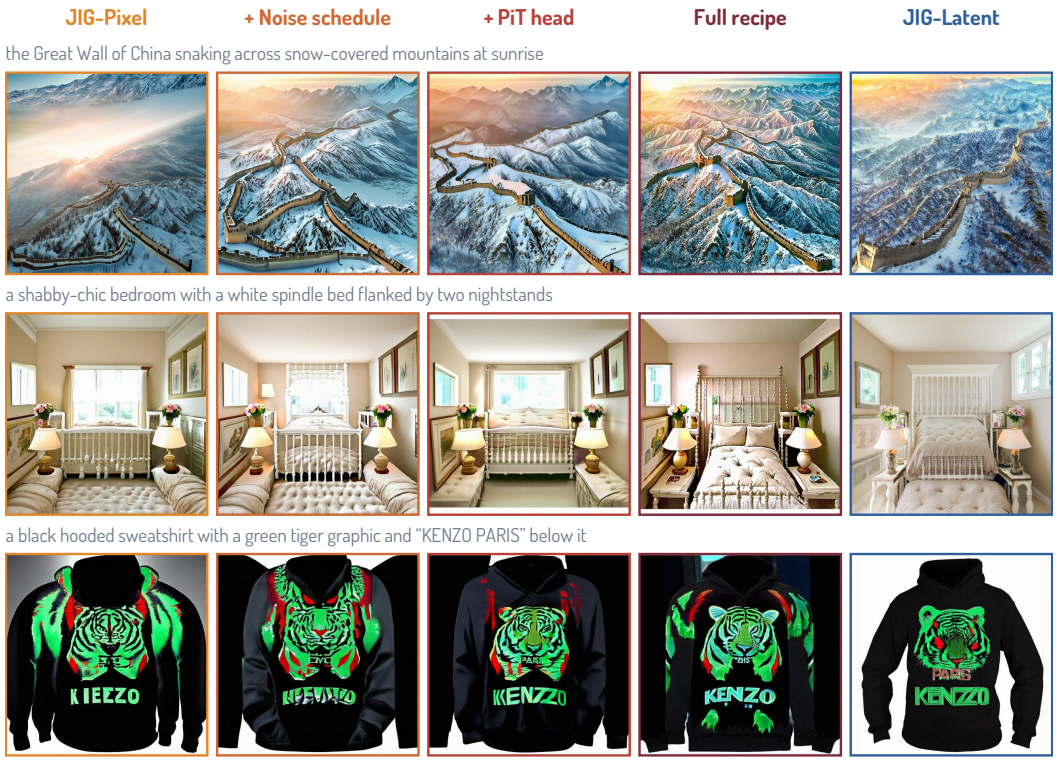


Figure 13: **Pixel-space training adaptations under diffusion.** Starting from JIG-Pixel-Diffusion, columns add the pixel-adapted schedule, pixel-level refinement, and perceptual supervision cumulatively, with JIG-Latent-Diffusion on the right. Each row uses the same prompt and sampling seed across columns; the displayed text summarizes the prompt.

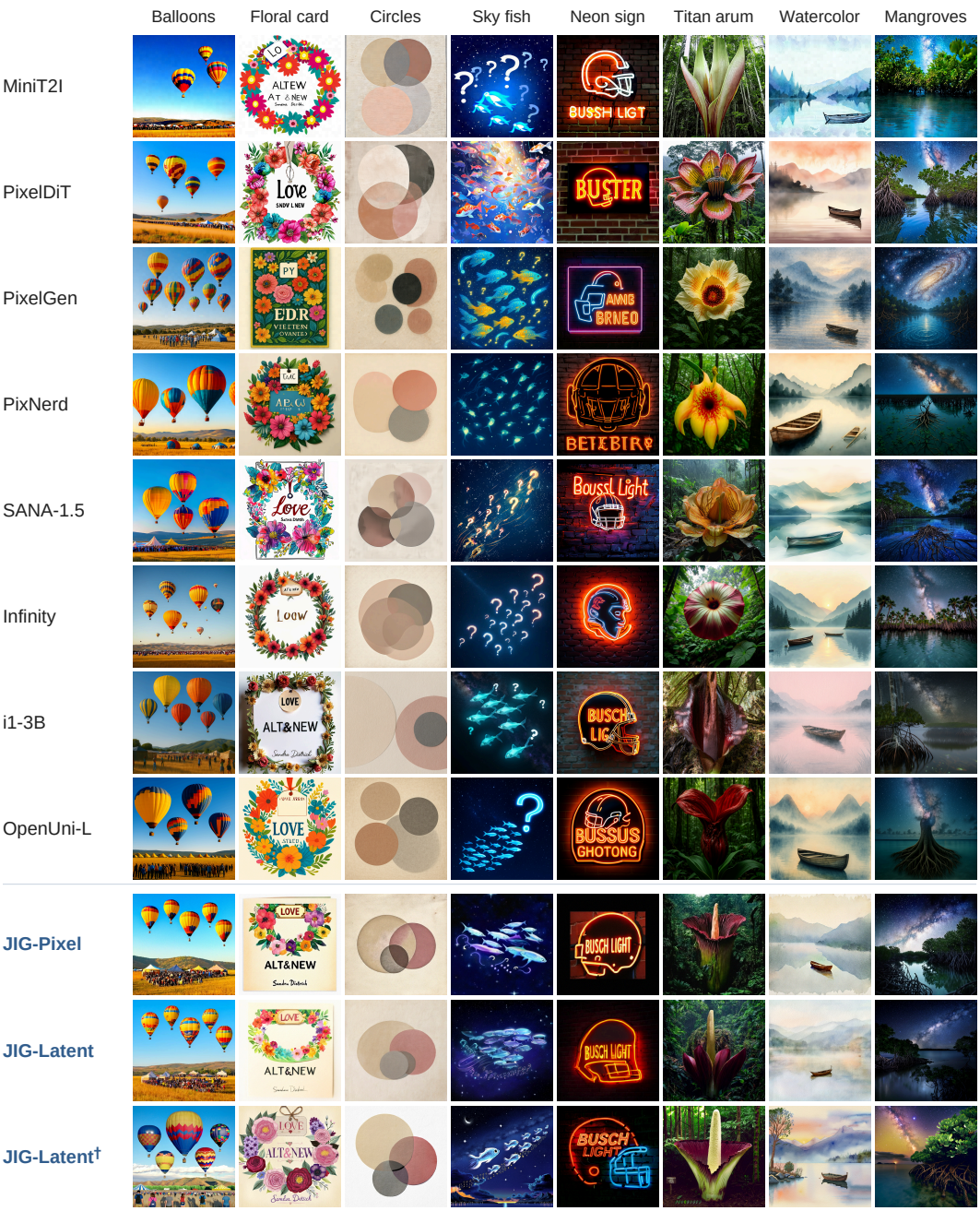


Figure 14: **Additional system-level comparisons on PRISM.** Each row shows one method; each column uses the same original prompt across methods. †: Reve 2-style annotations.