LACON: Training Text-to-Image Model from Uncurated Data

Zhiyang Liang[1,2†], Ziyu Wan[2*], Hongyu Liu[4], Dong Chen[3], Qiu Shen[1], Hao Zhu[1], and Dongdong Chen[2]

[1] Nanjing University, Nanjing, China
[2] Microsoft Research, Redmond, USA
[3] Microsoft Research Asia, Beijing, China
[4] HKUST, Hong Kong SAR, China

Fig. 1: Generated images from LACON at a resolution of 512 × 512.

Abstract. The success of modern text-to-image generation is largely attributed to massive, high-quality datasets. Currently, these datasets are curated through a *filter-first* paradigm that aggressively discards low-quality raw data based on the assumption that it is detrimental to model performance. Is the discarded bad data truly useless, or does it hold untapped potential? In this work, we critically re-examine this question. We propose **LACON** (Labeling-and-Conditioning), a novel training framework that exploits the underlying uncurated data distribution. Instead of filtering, LACON re-purposes quality signals, such as aesthetic scores and watermark probabilities as explicit, quantitative condition labels. The generative model is then trained to learn the full spectrum of data quality, from bad to good. By learning the explicit boundary between

† Work done during internship at Microsoft Research.
* Corresponding author.

high- and low-quality content, LACON achieves superior generation quality compared to baselines trained only on filtered data using the same compute budget, proving the significant value of uncurated data. Our project page is available at https://zhiyangliang.github.io/LACON.

Keywords: Image Generation · Data Utilization · Training Strategy

1 Introduction

Large-scale generative foundation models have recently triggered a revolutionary wave of progress across diverse modalities, including image [3,4,6,17,24,35], video [5,19,28,36,45,49], and 3D asset [15,18,21,42,43,46,52]. Among these, text-to-image (T2I) synthesis has witnessed particularly dramatic advancements. State-of-the-art models, such as Qwen-Image [39], GPT-Image-1 [1], Nano Banana [2], Hunyuan-Image [7] and Seedream [33], now generate images with unprecedented visual fidelity, generation diversity, and prompt controllability. These capabilities are fundamentally reshaping entire fields, from content creation and digital art to human-computer interaction.

The success of these foundation models is inseparable from one crucial factor: massive data scale. However, raw scale alone is not the sole determinant of success; the quality and cleanliness of this data are widely considered to be equally, if not more, paramount. It is widely assumed that training on bad or uncurated data, such as images with low resolution, poor aesthetics, watermarks, or overexposure, would tends to degrade model behavior. This very belief has driven the field to adopt a filter-first paradigm, which is not a minor preprocessing step but a core, resource-intensive component in recent SOTA pipelines. For example, Stable Diffusion [31] is not trained on the full LAION-5B dataset [32], but on LAION-Aesthetics, a highly-filtered subset based on aesthetic scores. Qwen-Image [39] proposes a multi-stage filtering pipeline comprising seven sequential stages, while Hunyuan-Image [7] implemented a comprehensive three-stage process, ultimately discarding over 55% of its initial raw data.

While this filter-first paradigm has become the de facto standard, we argue that it suffers from two major drawbacks. First, it is conceptually wasteful and statistically inefficient. The vast majority of data (often over 55%) is discarded. Such bad data (e.g., images with low-aesthetic scores, watermarks, or poor text-image alignment) is not mere noise; it represents a rich, unexploited source of information that could potentially improve the underlying knowledge of foundation models. Second, perhaps most critically, filtering creates a fundamental knowledge gap in the model. By being exclusively exposed to a narrow slice of the distribution, the model never learns the full spectrum of data quality. It gains a strong prior for goodness but no explicit understanding of badness, failing to learn the crucial boundary between high-quality and low-quality content. All of these motivate us to ask the following question: Can we devise a more efficient training method that fully exploits existing data, enabling the model to learn world knowledge, rare concepts, and characteristic low-quality patterns present in so-called bad samples but underrepresented or absent in high-quality data?

In this paper, we propose LACON (Labeling-and-Conditioning), a novel training framework designed to leverage the entire uncurated dataset. Instead of discarding data based on quality signals (e.g., aesthetic scores, text-image alignment, watermark probabilities), LACON re-purposes these signals as explicit, quantitative condition labels $\mathbf{s}$. We then train a conditional text-to-image model $p(\mathbf{x}|\mathbf{y}, \mathbf{s})$ to understand not only the text condition $\mathbf{y}$, but also these fine-grained attribute labels $\mathbf{s}$. This design directly addresses the previously discussed limitations: it is data-rich, utilizing 100% of the available raw data, and knowledge-complete, as it explicitly teaches the model the full spectrum of data quality, from bad to good. Moreover, LACON provides a principled training-time counterpart to commonly used ad-hoc inference-time heuristics—such as applying classifier-free guidance with negative prompts by explicitly modeling data quality during training. Crucially, it shifts the responsibility of quality control from inference-time guesswork (e.g., tuning negative prompts) to a well-defined training-time specification. As a result, the model no longer needs to be pushed away from failure modes it never encountered; instead, it learns to navigate the quality space directly through the explicit condition $\mathbf{s}$. During inference, we can then explicitly control output quality to meet product or application requirements (e.g., aesthetic quality).

We demonstrate the effectiveness of the LACON framework using a diffusion-based text-to-image generative model through extensive experiments. Our key contributions are threefold:

- We introduce LACON (Labeling-and-Conditioning), a novel training framework that fundamentally reframes the data curation problem. Instead of the filter-first paradigm, LACON leverages 100% of the uncurated data by re-purposing quality signals as explicit, learnable conditions.
- We demonstrate that LACON achieves superior generation quality. By learning the clear boundary between high-quality and low-quality data, our model significantly outperforms baselines trained only on filtered data with the same compute budget, proving the value of bad data.
- LACON unlocks powerful and quantitative controllability. Our paradigm enables fine-grained control over generation, such as explicitly conditioning on low-quality or high-quality attribute labels as illustrated in Figure 2.

2 Related Work

2.1 Text-to-Image Model

Text-to-image generation has advanced rapidly in recent years, producing visually compelling and semantically aligned outputs. Current approaches largely fall into two dominant paradigms: autoregressive models and diffusion-based models. Autoregressive methods [9, 40, 48, 51] represent images as sequences of discrete tokens and leverage large-scale Transformer architectures to model the joint text–image distribution. This formulation enables strong compositionality and scalability but often struggles with fine-grained spatial coherence due

Fig. 2: Comparison of generation from the baseline trained on full raw data and LACON under different attribute conditions. From left to right: (1) model trained on the full raw dataset without quality conditioning; (2) LACON conditioned on a low aesthetic score s_{aes}; (3) LACON conditioned on a high aesthetic score s_{aes}; (4) LACON jointly conditioned on high aesthetic score s_{aes} and high clarity score s_{cla}.

to the sequential nature of generation. In contrast, diffusion-based approaches [8, 14, 26, 29, 31, 33, 37, 39] synthesize images by progressively denoising Gaussian noise under text conditioning. Variants such as latent diffusion [31] and DiT-style architectures [37, 39] have set new benchmarks for photorealism and high-resolution fidelity. These models excel at capturing rich visual details and global consistency, though they typically require iterative sampling, which can be computationally expensive. Despite their impressive generative capabilities, both paradigms—especially state-of-the-art systems like [20,22,39] depend heavily on massive, high-quality image–text datasets for training.

2.2 Data Utilization

Existing state-of-the-art text-to-image models overwhelmingly adopt a filter-first paradigm, where massive raw datasets are aggressively pruned through multi-stage pipelines before training [7, 12, 39]. For example, Qwen-Image [39] applies sequential thresholds to eliminate images with low brightness, poor clarity, or artifacts such as watermarks and mosaics. Seedream [12] goes further by training defect detectors and masking defective regions during gradient computation, while Hunyuan-Image [7] employs aesthetic scoring models to enforce unified quality thresholds across all content types. These strategies share a common goal: distill a small, high-quality subset from web-scale corpora to maximize visual fidelity. However, this paradigm is inherently wasteful and statistically biased. By discarding over half of the raw data, models lose exposure to rare concepts and characteristic low-quality patterns, creating a blind spot in their understanding of the full quality spectrum. In contrast, LACON explores a new strategy: rather than filtering out bad data, we repurpose quality signals as explicit conditioning labels, enabling the model to learn both high-quality and low-quality distributions. This approach transforms data curation from a destructive process into a constructive one, leveraging almost 100% of available data while preserving controllability.

2.3 Controllable Generation

Controllability [50] has long been a central theme in text-to-image synthesis, with prior work introducing adapter modules or conditional branches to steer generation beyond raw text prompts [31]. For instance, IP-Adapter [47] injects image prompts via decoupled cross-attention, Elite [38] encodes user-specific visual concepts into textual embeddings, and VMix [41] introduces aesthetics adapters for disentangled style control. Recent works have begun to turn data-side signals into training conditions rather than using them only for data selection. Coherence-Aware Diffusion [10] makes a strong case for conditioning on annotation coherence instead of filtering out low-coherence samples, while MIRO [11] extends this idea to multi-reward-conditioned T2I pretraining. Concept-aware batch sampling [13] also exploits data-side concepts to organize vision-language pretraining batches. These works show the benefit of reusing data-side information, but they focus on annotation reliability, reward-conditioned preference modeling, or batch construction rather than explicitly modeling the quality spectrum of uncurated T2I data. LACON is related in spirit, but focuses more on repurposing signals typically used for data filtering and curation, such as aesthetics, watermark probability, clarity, entropy, and luminance, as conditioning labels. This turns filtering-oriented scores into controllable quality dimensions and allows the model to learn from diverse raw samples instead of treating quality only as a criterion for sample removal. At inference time, LACON exposes these dimensions as controllable inputs, enabling fine-grained and interpretable control over attributes such as watermark level and aesthetic score.

3 Method

This section details our proposed method, LACON (LAbeling-and-CONditioning), for training text-to-image models on uncurated data. We begin in Section 3.1 by providing a brief overview of the prerequisite generative model framework used in our experiments. In Section 3.2, we introduce the core components of LACON, detailing our data attribute labeling process, the mechanism for label-conditioning, and the associated training strategy. Finally, we describe the flexible test-time inference strategies enabled by our approach in Section 3.3.

3.1 Preliminaries

Text-to-image (T2I) generative models aim to learn the complex conditional probability distribution $p(\mathbf{x}|\mathbf{y})$, where $\mathbf{x}$ is the image and $\mathbf{y}$ is corresponding text condition. After training, the model allows the sampling of high-quality and diversified visual contents from this distribution given prompts. In the past, most successful T2I models [7, 39] have been trained using filtered high-quality data. In this paper, we propose LACON, a model-agnostic framework that can be generalized to various generative methods [16, 30, 34, 44], to fully leverage the potential of uncurated dataset.

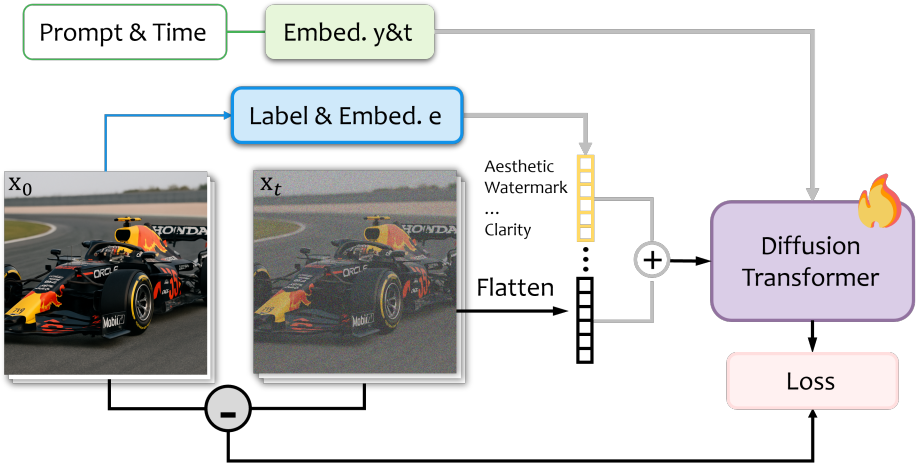

Fig. 3: Overview of LACON. High-level training pipeline that repurposes quality signals as explicit attribute conditioning during training.

Table 1: Configuration of cluster-centroid anchors for each attribute embedding. The number and placement of anchors (p_i) are selected based on the value range of the corresponding quality signals.

Metric (s_{metric})	**Meaning**	**Value Range**	**Clipping**	**Tokens** (N)	**Centroid Anchors** (p_i^{metric})
Aesthetic (s_{aes})	Aesthetic quality	0–10	None	10	$\{0.5, 1.5, \ldots, 9.5\}$
Watermark (s_{wat})	Watermark probability	0–1	None	10	$\{0.05, 0.15, \ldots, 0.95\}$
Clarity (s_{cla})	Perceptual sharpness	0–1M+	> 3000	10	$\{150, 450, \ldots, 2850\}$
Entropy (s_{ent})	Information density	0–8	None	8	$\{0.5, 1.5, \ldots, 7.5\}$
Luminance (s_{luma})	Visual brightness	0–1	None	10	$\{0.05, 0.15, \ldots, 0.95\}$

We demonstrate its efficacy by integrating LACON into a diffusion model. Following standard practice, we adopt the flow matching objective [23, 25] with linear interpolation path $\mathbf{x}_t = (1-t)\mathbf{x}+t\boldsymbol{\epsilon}$, where $\mathbf{x} \sim p(\mathbf{x})$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and train the target generative model to predict the velocity $v_\theta(\mathbf{x}_t, t, \mathbf{y}) = \boldsymbol{\epsilon} - \mathbf{x}$. To ensure the training efficiency, we use Sana [44], a highly efficient DiT-based [27] T2I framework, as our model backbone.

3.2 LACON

As established in the previous sections, the filter-first paradigm suffers from data waste and creates a knowledge gap inside the model. Instead of viewing quality signals (e.g., aesthetic scores, watermarks) as a pre-training filter, LACON simply re-purposes them as explicit, learnable conditions. Technically, this approach reframes the T2I generation task. We shift from the standard objective $v_\theta(\mathbf{x}_t, t, \mathbf{y})$ trained on a filtered dataset, to a multi-conditional objective $v_\theta(\mathbf{x}_t, t, \mathbf{s}, \mathbf{y})$ trained on the entire uncurated dataset. Here, $\mathbf{s}$ is a vector representing the quantitative attribute labels for each sample.

Data Attribute Labeling. Without loss of generality, we consider five condition signals, i.e.,`aesthetic`, `watermark`, `clarity`, `entropy`, and `luminance`, which are commonly used in image-filtering pipelines and capture a diverse range of semantic and low-level image properties. Our training corpus consists of 110M images collected from publicly available sources. To implement LACON, we build an automated labeling pipeline that annotates each image with a five-dimensional conditioning vector $\mathbf{s} = [s_{\text{aes}}, s_{\text{wat}}, s_{\text{cla}}, s_{\text{ent}}, s_{\text{luma}}]$. More details and label configurations could be found in Table 1.

Condition Embedding. To leverage the conditional vector $\mathbf{s}$, we must convert them into learnable embeddings that can be ingested by the generative model. We propose a flexible embedding strategy that converts each continuous scalar score into a soft vector representation. We process each of the $K = 5$ attributes independently. For the k-th attribute (e.g., $s_{\texttt{aes}}$), we define two components: 1) Fixed anchors: A set of N fixed, non-learnable scalar anchor points, $\mathbf{p}^{(k)} = \{p_1^{(k)}, \ldots, p_N^{(k)}\}$, where N is chosen based on the value range and granularity of attribute. These anchors partition the continuous space of the attribute, linearly spaced across the expected score range of the specific attribute. 2) Learnable centroids: For each anchor $p_i^{(k)}$, we attach a learnable embedding vector, or centroid token $\mathbf{c}_i^{(k)} \in \mathbb{R}^d$, where the embedding dimension d matches the channel size of the noisy latent $\mathbf{x}_t$.

Given k-th attribute s_k, we compute its affnity scores $u_i^{(k)}$ over different anchor points $p_i^{(k)}$ using a Gaussian Radial Basis Function (RBF) kernel:

$$u_i^{(k)} = \exp\left(-\frac{(s_k - p_i^{(k)})^2}{2\sigma_k^2}\right)$$

Here, σ_k is an attribute-specific parameter determined by half the distance between adjacent centroid anchors of the k-th attribute, enabling independent smoothness control for each attribute. We then normalize the affinity scores to produce the final weights $w_i^{(k)}$:

$$w_i^{(k)} = \frac{u_i^{(k)}}{\sum_{j=1}^{N} u_j^{(k)}}$$

The final embedding $\mathbf{e}_k$ for the k-th attribute is the weighted sum of all N learnable centroid tokens:

$$\mathbf{e}_k = \sum_{i=1}^{N} w_i^{(k)} \mathbf{c}_i^{(k)}$$

Training. The LACON framework accepts two distinct conditional inputs: the text prompt $\mathbf{y}$ and our new attribute vector $\mathbf{s}$. For prompt $\mathbf{y}$ we feed its text embedding to the model using cross attention to guide the semantics of the generation. For our new attribute embedding, we adopt a more straightforward strategy, i.e. context conditioning, which simply concatenates the noisy latents $\mathbf{x}_t$ with condition embedding $\mathbf{e}$ to form a new token sequence as the DiT input

so that the model could perceive the complete context from the very first layer. An high-level illustration of our training framework is provided in Figure 3. We found that this simple concatenation approach works well and effectively conditions the entire denoising process on the desired quality attributes.

3.3 Inference Strategy

LACON's explicit attribute conditioning $\mathbf{s}$ could unlock quantitative control over the generated output to improve the generated visual quality. In this section, we introduce how to do inference with LACON. We still use standard classifier-free guidance (CFG) for the text condition $\mathbf{y}$, but we will introduce two ways: LACON-S and LACON-A to leverage the learned quality attributes.

- *LACON-S* (standard guidance): The standard and most efficient inference method. It performs CFG on the text condition $\mathbf{y}$ while holding the vector of quality signals $\mathbf{s}$ constant at a single target value (e.g., high aesthetics, no watermark).
- *LACON-A* (aggressive multi-condition guidance): a more advanced and computationally intensive scheme that applies CFG to both the text condition $\mathbf{y}$ and each quality attribute s_k independently, which allows for dynamic, per-attribute guidance scaling at inference time. At each denoising step t, we compute $K + 2$ parallel predictions and combine all these guidance vectors by:

$$\begin{aligned}\mathbf{v}_a^t = \mathbf{v}_{\text{base}} &+ \omega_c \cdot (\mathbf{v}_{\text{text}} - \mathbf{v}_{\text{base}}) \\ &+ \omega_{\text{aes}} \cdot (\mathbf{v}_{\text{aes}} - \mathbf{v}_{\text{text}}) \\ &+ \omega_{\text{wat}} \cdot (\mathbf{v}_{\text{wat}} - \mathbf{v}_{\text{text}}) \\ &+ \dots \\ &+ \omega_{\text{luma}} \cdot (\mathbf{v}_{\text{luma}} - \mathbf{v}_{\text{text}})\end{aligned}$$

  where $\mathbf{v}_{\text{base}} = v_\theta(\mathbf{x}_t, t, \varnothing, \mathbf{s}_{\text{base}})$ $\mathbf{v}_{\text{text}} = v_\theta(\mathbf{x}_t, t, \mathbf{y}, \mathbf{s}_{\text{base}})$. We compute one prediction $\mathbf{v}_{\mathbf{s}}$ for each of our $K = 5$ attributes, where only that single attribute is set to its high-quality target, e.g., $\mathbf{v}_{\text{aes}} = v_\theta(\mathbf{x}_t, t, \mathbf{y}, \mathbf{s}_{\text{high-aes}})$. ω_c is the text-guidance scale, while $\omega_{\text{aes}}, \dots, \omega_{\text{luma}}$ are individual guidance scales for each quality attribute. This aggressive method provides maximum controllability, allowing a user to dial up or dial down each specific quality metric at inference time to achieve the preferred results.

4 Experiment

4.1 Experiments Setting

Implementation Details. Our main experiments are conducted on Sana-0.6B and Sana-1.6B, both pre-trained on 512×512 images. Sana-1.6B is pre-trained for 150K steps with a global batch size of 2048 on 32 NVIDIA A100 GPUs, whereas Sana-0.6B is pre-trained for 100K steps with the same global batch size

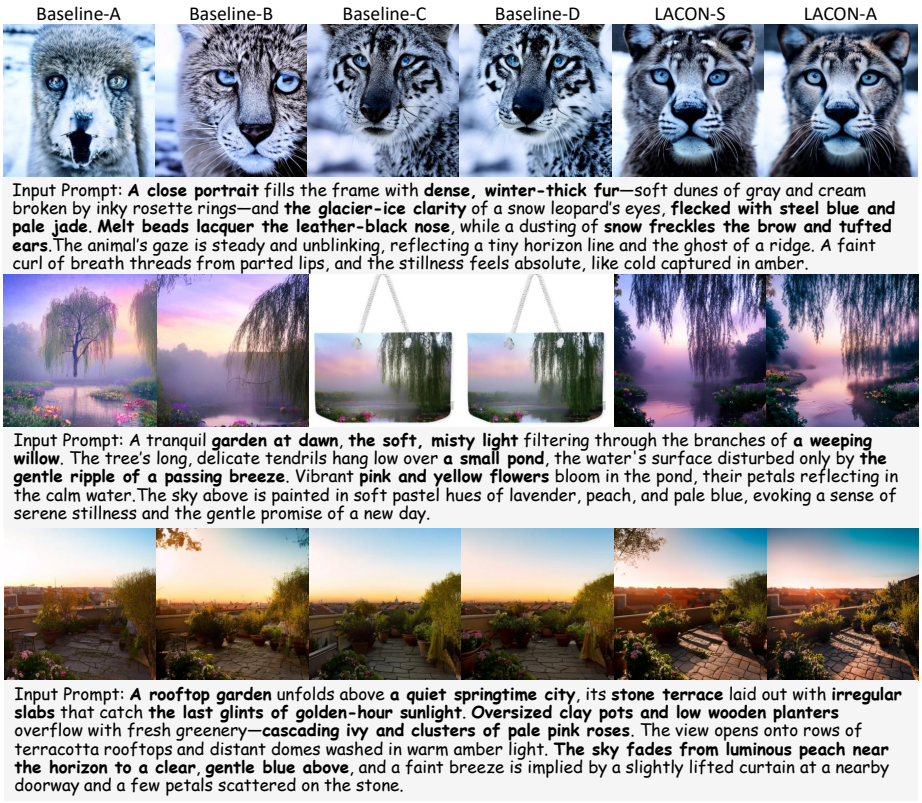

Fig. 4: Visualization comparison of LACON-S and LACON-A against Baselines, which demonstrate LACON can still achieve superior visual generation quality even when training on the full set images without filtering.

on 16 NVIDIA A100 GPUs. For high-resolution setting, we further fine-tune the corresponding pre-trained models for an additional 20K steps. We report GenEval, DPG, and FID as evaluation metrics. FID is computed on MJHQ-30K, while GenEval and DPG use 533 and 1,065 test prompts, respectively.

Dataset. Our training dataset contains 110M images collected from a subset of LAION-2B-EN and our organization's internal collection. Approximately 100M images are sampled from the LAION-2B-EN subset, while the remaining images come from internal data sources. To mitigate inconsistencies caused by heterogeneous raw captions, all image captions are generated offline using GPT-4o, providing consistent and standardized textual supervision.

Baselines. We study how training on subsets constructed using different filtering thresholds affects performance on Sana-1.6B. The thresholds and resulting retained ratios are summarized in Table 2, and the results on GenEval, DPG, and FID are reported in Table 3. The results reveal a clear trade-off between data quantity and data quality: training on an overly small subset leads to sub-

(a) Aesthetic & Luminance. (b) Clarity & Entropy.

Fig. 5: (a) **Comparison of images jointly conditioned on aesthetic score s_{aes} and HSV-luma score s_{luma}**. From left to right, the columns correspond to $s_{\text{aes}} = 3, 5, 7$. From top to bottom, the rows correspond to $s_{\text{luma}} = 0.3, 0.4, 0.5$. (b) **Comparison of images jointly conditioned on clarity lapvar score s_{cla} and entropy score s_{ent}**. From left to right, the columns correspond to $s_{\text{cla}} = 200, 1000, 2500$. From top to bottom, the rows correspond to $s_{\text{ent}} = 5, 6, 7$.

optimal performance due to insufficient coverage and diversity. As the retained ratio increases, the training set tends to contain a larger fraction of low-quality samples. Without explicit quality-aware supervision, these samples can degrade model performance.

Based on this sweep, we define four baselines. Baseline-A is trained from scratch on a heavily filtered subset ($\approx 5\%$ of the raw data) obtained with strict quality thresholds; this choice corresponds to the most aggressive filtering setting in our ratio sweep and achieves the worst performance in Table 3. Baseline-B is trained on a more loosely filtered subset ($\approx$ 65% of the raw data) under relaxed thresholds; this choice matches the best-performing retention ratio in Table 3 and serves as the strongest filtered-data reference. Baseline-C is trained on the full set of 110M unfiltered images. Baseline-D uses the same training data as Baseline-C, but applies negative-prompt guidance at inference time, which is conceptually related to our LACON strategy of using low-quality signals to steer generation toward higher-quality outputs. Specifically, we use the following negative prompt: *low aesthetic score, ugly, bad composition, watermark, logo.*

4.2 Main Experiments

Performance Comparison. We compare LACON with the baselines described in Section 4.1 and report results on GenEval, DPG, and FID for Sana-0.6B and Sana-1.6B in Table 4. Consistent with the filtering-ratio study, Baseline-

Table 2: The filtering thresholds for different filtering ratios.

Ratio	$s_{\texttt{aes-min}}$	$s_{\texttt{wat-min}}$	$s_{\texttt{cla-min}}$	$s_{\texttt{ent-min}}$	$s_{\texttt{luma-min}}$	$s_{\texttt{luma-max}}$
$\approx 5\%$	5.0	0.3	800.0	6.0	0.1	0.9
$\approx 30\%$	4.0	0.5	600.0	4.0	0.1	0.9
$\approx 50\%$	3.5	0.6	500.0	3.0	0.1	0.9
$\approx 65\%$	3.0	0.7	400.0	2.0	0.1	0.9
$\approx 80\%$	3.0	0.8	200.0	2.0	0.1	0.9

Table 3: Quantitative comparisons on Sana-1.6B are conducted between models trained on subsets obtained under different filtering thresholds, evaluated on GenEval, DPG and FID. The results reveal a clear trade-off between data quantity and data quality in terms of model performance.

Ratio	**GenEval** $\uparrow$	**DPG** $\uparrow$	**FID** $\downarrow$
$\approx 5\%$	58.3	73.4	14.0
$\approx 30\%$	62.1	73.6	13.7
$\approx 50\%$	64.5	74.5	13.4
$\approx 65\%$	**68.0**	**76.1**	**13.1**
$\approx 80\%$	67.5	75.8	13.5
100%	67.0	75.4	13.5

B (filtered data) consistently outperforms Baseline-C/D (full unfiltered data), whereas the heavily filtered Baseline-A underperforms, reflecting the expected quality–quantity tension. In contrast, LACON achieves the best performance across GenEval, DPG, and FID for both 512×512 and 1024×1024 resolutions, while being trained on the full raw dataset with explicit quality conditioning. This suggests that samples typically discarded as low quality can still be valuable when their quality is modeled and leveraged rather than removed. LACON-A consistently outperforms LACON-S across most metrics and resolutions, as its stronger guidance yields larger improvements in generation quality and often produces higher-quality visuals.

Visualization Comparison. The visualization comparisons in Figure 4 further demonstrate that LACON produces higher-quality images while enabling explicit attribute control at inference time. In contrast, training on a heavily filtered subset can create a pronounced knowledge gap, as evidenced by Baseline-A in Figure 4. More cases are available in Section 4.4. Conversely, training directly on the raw dataset may introduce characteristic artifacts at generation time. Beyond inherited watermarks, extremely low-entropy samples can lead to a distinctive failure mode that manifests as blank boundary regions, as observed for Baseline-C/D in Figure 4. LACON alleviates both issues while fully leveraging the raw dataset by learning to disentangle high- and low-quality patterns across multiple quality metrics. At inference time, conditioning on the corresponding quality tokens enables the model to suppress artifacts associated with

Table 4: Quantitative comparisons on Sana-0.6B and Sana-1.6B, evaluated using GenEval, DPG, and FID, demonstrate that training on the full unfiltered dataset with explicit conditioning enables LACON to consistently outperform both filter-first baselines and unconditioned baselines.

512×512 resolution						
Model	**Sana-0.6B**			**Sana-1.6B**		
Metric	**GenEval ↑**	**DPG ↑**	**FID ↓**	**GenEval ↑**	**DPG ↑**	**FID ↓**
Baseline-A	55.0	68.1	14.9	58.3	73.4	14.0
Baseline-B	62.8	70.4	13.4	68.0	76.1	13.1
Baseline-C	60.8	69.7	15.4	67.0	75.4	13.5
Baseline-D	61.0	69.2	15.3	67.1	75.5	13.5
LACON-S (Ours)	64.4	**71.9**	11.9	70.9	77.3	12.0
LACON-A (Ours)	**65.6**	71.8	**11.8**	**71.6**	**78.1**	**11.2**
1024×1024 resolution						
Model	**Sana-1.6B**					
Metric	**GenEval ↑**		**DPG ↑**		**FID ↓**	
Baseline-B	68.3		76.3		12.9	
LACON-S (Ours)	70.9		77.6		11.4	
LACON-A (Ours)	**71.5**		**78.8**		**11.3**	

low-entropy samples (e.g., blank boundary regions), resulting in cleaner and more visually coherent outputs.

4.3 Autoregressive Architecture

We further evaluate LACON against the strongest baseline (Baseline-B) with Qwen3-0.6B and Qwen3-1.7B using LlamaGen's VQ-VAE, verifying that our approach generalizes to autoregressive architectures. Qwen3-1.7B is pre-trained for 120K steps with a global batch size of 2048 on 32 NVIDIA A100 GPUs, while Qwen3-0.6B is pre-trained for 100K steps with a global batch size of 1024 on 16 NVIDIA A100 GPUs. We report results on GenEval, DPG, and FID for both models in Table 5. Across GenEval, DPG, and FID, LACON consistently outperforms Baseline-B on both Qwen3-0.6B and Qwen3-1.7B, with LACON-A further surpassing LACON-S at both scales. These results suggest that LACON's gains transfer reliably from diffusion to autoregressive generation.

4.4 Knowledge Gap

We evaluate the world-knowledge capabilities of the strongest baseline (Baseline-B) and LACON. Figure 6 compares images generated for long-tail concepts by Baseline-B and LACON. Although Baseline-B retains approximately 65% of the full dataset after filtering, it exhibits a noticeable knowledge gap relative to LACON. We attribute this gap to the fact that filtering can discard concept-relevant

Table 5: Quantitative comparisons on Qwen3-0.6B and Qwen3-1.7B, evaluated using GenEval, DPG, and FID, demonstrate that training on the full unfiltered dataset with explicit conditioning enables LACON to consistently outperform the strongest baseline.

Model	**Qwen3-0.6B**			**Qwen3-1.7B**		
Metric	**GenEval** ↑	**DPG** ↑	**FID** ↓	**GenEval** ↑	**DPG** ↑	**FID** ↓
Baseline-B	58.9	73.1	12.4	66.4	78.2	12.1
LACON-S (Ours)	**61.3**	74.9	11.4	69.7	79.4	11.0
LACON-A (Ours)	**61.3**	**75.4**	**11.2**	**70.3**	**80.1**	**10.9**

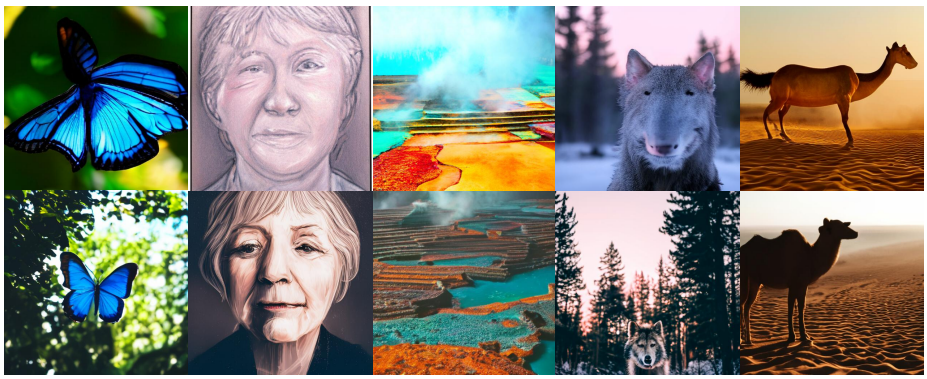

Fig. 6: The evidence of knowledge gap between Baseline-B (the first row) and LACON (the second row). From left to right, the concepts are about morpho butterfly, pastel portrait of elderly woman, mineral terraces, wolf and camel.

Table 6: Comparison of different token-injection strategies: linear interpolation, discrete binning, Fourier-feature embedding, and Gaussian-weighted Cluster Centroids (GCC), evaluated on GenEval, DPG, and FID.

Method	**GenEval** ↑	**DPG** ↑	**FID** ↓
Linear Interpolation	69.5	76.1	12.7
Discrete Binning	69.8	77.2	12.4
Fourier Feature	70.4	77.0	12.5
GCC (Ours)	**71.6**	**78.1**	**11.2**

training samples, thereby reducing coverage of rare or long-tail concepts. These observations suggest a limitation of the filter-first paradigm: samples deemed low-quality can still provide non-trivial, largely untapped supervision that helps models acquire broader world knowledge. By incorporating explicit quality signals, LACON can leverage such long-tail supervision while better separating high- and low-quality content, ultimately improving generation performance.

Table 7: Ablation of individual quality signals in LACON-S, where each signal is individually removed during inference.

Strategy	GenEval ↑	DPG ↑	FID ↓
LACON-S (w/o aes)	69.4	76.6	12.6
LACON-S (w/o wat)	70.2	77.2	12.1
LACON-S (w/o cla)	69.7	76.8	12.5
LACON-S (w/o ent)	69.8	77.0	12.2
LACON-S (w/o luma)	70.4	**77.3**	12.1
LACON-S	**70.9**	**77.3**	**12.0**

4.5 Ablation Studies

Token-injection Strategies. We compare four token-injection strategies: linear interpolation, discrete binning, Fourier-feature embedding, and Gaussian-weighted Cluster Centroids (GCC). For linear interpolation, we initialize two learnable tokens per metric at the minimum and maximum of its score range, and obtain each sample's token by interpolating between the two endpoints according to its score. For discrete binning, we use the same number of tokens per metric as GCC, but quantize scores into bins and assign each sample the token corresponding to its bin. For Fourier-feature embedding, we encode the quality score with Fourier features and map the resulting representation to the token-embedding space via an MLP. GCC is described in detail in Section 3.2.

As illustrated in Table 6, GCC consistently outperforms the other strategies on GenEval, DPG, and FID. Notably, discrete binning surpasses linear interpolation, indicating that representing each metric with only two learnable tokens and a purely linear parameterization is overly restrictive, limiting the model's ability to exploit its capacity. In contrast, GCC retains the same token budget as discrete binning but models the score distribution more expressively by softly assigning each sample to multiple anchors. This smooth, overlap-aware encoding captures local structure in the score space and yields more stable, sample-efficient learning from quality-diverse data. Compared with Fourier-feature embedding, GCC is also more robust: Fourier features impose a fixed oscillatory encoding that can be sensitive to small score variations and frequency choices, whereas GCC learns data-adaptive anchors with smooth Gaussian weighting, leading to more stable gradients and better-calibrated conditioning.

Joint Controllability. Our LACON framework supports joint conditioning on multiple metric scores, allowing each attribute to vary smoothly from low to high and enabling fine-grained control along different quality dimensions. Figure 5(a) illustrates joint control over aesthetic and luminance, while Figure 5(b) illustrates joint control over clarity and entropy. For each attribute pair, we vary both attributes across three discrete levels (low/medium/high), resulting in a 3×3 grid with nine combinations. The generated images change smoothly and coherently along both axes, indicating that LACON provides compositional and fine-grained control beyond standard text-only conditioning.

The entropy score reflects image information density. Setting it to a low value at inference time encourages the model to produce low-information outputs, often with blank boundary regions, as shown in the first row of Figure 5(b). This behavior is consistent with the blank-boundary patterns observed in the second row of Figure 4. Our LACON can alleviate these artifacts by conditioning on the corresponding quality tokens with a high entropy score at generation time; similarly, conditioning on a high watermark score suppresses watermark-related patterns. Thus, even though LACON is trained on raw data that includes low-entropy or watermarked images, it can still generate clean, high-quality outputs at inference time through explicit quality control.

Quality Signal Ablation Table 7 shows an inference-time ablation of individual quality signals using LACON-S. The quantitative results reveal that all five attributes are mutually beneficial. Notably, aesthetic quality and perceptual sharpness emerge as the most influential signals in the conditioning space.

Uncurated Data vs. Curated Data Scaling. Scaling high-quality data ultimately depends on scaling raw data, as curated high-quality samples must be drawn from a sufficiently large and diverse data pool. LACON provides a paradigm for efficiently exploiting the broader raw-data pool by modeling quality variations with conditioning labels, rather than relying solely on strict filtering. Table 3 first shows that proper filtering improves unconditioned training over directly using raw data, while Table 4 further demonstrates that LACON consistently outperforms both the strongest filter-first baseline and unconditioned baselines. In this way, LACON turns massive raw data into quality-aware and controllable supervision for text-to-image training.

5 Conclusion

In this paper, we revisit the prevailing filter-first paradigm in text-to-image generation and propose LACON, a novel training framework that leverages the full uncurated data distribution by converting quality signals (e.g., aesthetic scores) into learnable conditioning tokens. This design enables the model to learn the complete spectrum of data quality across multiple dimensions, from bad to good, and to internalize an explicit boundary between them. Across GenEval, DPG and FID, LACON consistently outperforms baselines trained either on filtered high-quality subsets or on full raw data without quality conditioning. These results suggest that data typically discarded as low-quality can be highly valuable when modeled and controlled rather than removed, underscoring the broader potential of uncurated data for building stronger and more controllable generative models.

Acknowledgements

This research is supported by the NSFC (No. U25B2046), the Fundamental Research Funds for the Central Universities (No. 481014380011), the “111 Center” (No. B26023), and the SKL-NST at Nanjing University (No. ZZKT2026A20).

References

1. Gpt-image-1. https://openai.com/index/image-generation-api/
2. Nano banana. https://aistudio.google.com/models/gemini-2-5-flash-image
3. Baldridge, J., Bauer, J., Bhutani, M., Brichtova, N., Bunner, A., Castrejon, L., Chan, K., Chen, Y., Dieleman, S., Du, Y., et al.: Imagen 3. arXiv preprint arXiv:2408.07009 (2024)
4. Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., et al.: Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. arXiv e-prints pp. arXiv–2506 (2025)
5. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
6. Cai, Q., Chen, J., Chen, Y., Li, Y., Long, F., Pan, Y., Qiu, Z., Zhang, Y., Gao, F., Xu, P., et al.: Hidream-i1: A high-efficient image generative foundation model with sparse diffusion transformer. arXiv preprint arXiv:2505.22705 (2025)
7. Cao, S., Chen, H., Chen, P., Cheng, Y., Cui, Y., Deng, X., Dong, Y., Gong, K., Gu, T., Gu, X., et al.: Hunyuanimage 3.0 technical report. arXiv preprint arXiv:2509.23951 (2025)
8. Chang, H., Zhang, H., Barber, J., Maschinot, A., Lezama, J., Jiang, L., Yang, M.H., Murphy, K., Freeman, W.T., Rubinstein, M., et al.: Muse: Text-to-image generation via masked generative transformers. arXiv preprint arXiv:2301.00704 (2023)
9. Dong, R., Han, C., Peng, Y., Qi, Z., Ge, Z., Yang, J., Zhao, L., Sun, J., Zhou, H., Wei, H., et al.: Dreamllm: Synergistic multimodal comprehension and creation. arXiv preprint arXiv:2309.11499 (2023)
10. Dufour, N., Besnier, V., Kalogeiton, V., Picard, D.: Don't drop your samples! coherence-aware training benefits conditional diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6264–6273 (2024)
11. Dufour, N., Degeorge, L., Ghosh, A., Kalogeiton, V., Picard, D.: Miro: Multi-reward conditioned pretraining improves t2i quality and efficiency. arXiv preprint arXiv:2510.25897 (2025)
12. Gao, Y., Gong, L., Guo, Q., Hou, X., Lai, Z., Li, F., Li, L., Lian, X., Liao, C., Liu, L., et al.: Seedream 3.0 technical report. arXiv preprint arXiv:2504.11346 (2025)
13. Ghosh, A., Udandarao, V., Nguyen, T., Farina, M., Cherti, M., Jitsev, J., Oh, S., Ricci, E., Schmidt, L., Bethge, M.: Concept-aware batch sampling improves language-image pretraining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3056–3068 (2026)
14. Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., Yuan, L., Guo, B.: Vector quantized diffusion model for text-to-image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10696–10706 (2022)
15. He, X., Zou, Z.X., Chen, C.H., Guo, Y.C., Liang, D., Yuan, C., Ouyang, W., Cao, Y.P., Li, Y.: Sparseflex: High-resolution and arbitrary-topology 3d shape modeling. arXiv preprint arXiv:2503.21732 (2025)
16. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)

17. Huang, M., et al.: Dialoggen: multi-modal interactive dialogue system for multi-turn text-to-image generation. multimodal dialogue systems workshop, lncs, vol. 13999 (2024)
18. Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D.: Scaling laws for neural language models. arXiv preprint arXiv:2001.08361 (2020)
19. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
20. Li, D., Kamko, A., Akhgari, E., Sabet, A., Xu, L., Doshi, S.: Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation. arXiv preprint arXiv:2402.17245 (2024)
21. Li, Y., Zou, Z.X., Liu, Z., Wang, D., Liang, Y., Yu, Z., Liu, X., Guo, Y.C., Liang, D., Ouyang, W., et al.: Triposg: High-fidelity 3d shape synthesis using large-scale rectified flow models. arXiv preprint arXiv:2502.06608 (2025)
22. Li, Z., Zhang, J., Lin, Q., Xiong, J., Long, Y., Deng, X., Zhang, Y., Liu, X., Huang, M., Xiao, Z., et al.: Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. arXiv preprint arXiv:2405.08748 (2024)
23. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
24. Liu, B., Akhgari, E., Visheratin, A., Kamko, A., Xu, L., Shrirao, S., Lambert, C., Souza, J., Doshi, S., Li, D.: Playground v3: Improving text-to-image alignment with deep-fusion large language models. arXiv preprint arXiv:2409.10695 (2024)
25. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022)
26. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
27. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
28. Peng, X., Zheng, Z., Shen, C., Young, T., Guo, X., Wang, B., Xu, H., Liu, H., Jiang, M., Li, W., et al.: Open-sora 2.0: Training a commercial-level video generation model in $200 k. arXiv preprint arXiv:2503.09642 (2025)
29. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
30. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International conference on machine learning. pp. 8821–8831. Pmlr (2021)
31. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
32. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in neural information processing systems **35**, 25278–25294 (2022)
33. Seedream, T., Chen, Y., Gao, Y., Gong, L., Guo, M., Guo, Q., Guo, Z., Hou, X., Huang, W., Huang, Y., et al.: Seedream 4.0: Toward next-generation multimodal image generation. arXiv preprint arXiv:2509.20427 (2025)
34. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)

35. Team, K.: Kolors: Effective training of diffusion model for photorealistic text-to-image synthesis. arXiv preprint (2024)
36. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
37. Wei, T., Chen, D., Zhou, Y., Pan, X.: Enhancing mmdit-based text-to-image models for similar subject generation. arXiv preprint arXiv:2411.18301 (2024)
38. Wei, Y., Zhang, Y., Ji, Z., Bai, J., Zhang, L., Zuo, W.: Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15943–15953 (2023)
39. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
40. Wu, C., Chen, X., Wu, Z., Ma, Y., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C., et al.: Janus: Decoupling visual encoding for unified multimodal understanding and generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12966–12977 (2025)
41. Wu, S., Ding, F., Huang, M., Liu, W., He, Q.: Vmix: Improving text-to-image diffusion model with cross-attention mixing control. arXiv preprint arXiv:2412.20800 (2024)
42. Wu, S., Lin, Y., Zhang, F., Zeng, Y., Yang, Y., Bao, Y., Qian, J., Zhu, S., Cao, X., Torr, P., et al.: Direct3d-s2: Gigascale 3d generation made easy with spatial sparse attention. arXiv preprint arXiv:2505.17412 (2025)
43. Xiang, J., Lv, Z., Xu, S., Deng, Y., Wang, R., Zhang, B., Chen, D., Tong, X., Yang, J.: Structured 3d latents for scalable and versatile 3d generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21469–21480 (2025)
44. Xie, E., Chen, J., Chen, J., Cai, H., Tang, H., Lin, Y., Zhang, Z., Li, M., Zhu, L., Lu, Y., et al.: Sana: Efficient high-resolution image synthesis with linear diffusion transformers. arXiv preprint arXiv:2410.10629 (2024)
45. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)
46. Ye, C., Wu, Y., Lu, Z., Chang, J., Guo, X., Zhou, J., Zhao, H., Han, X.: Hi3dgen: High-fidelity 3d geometry generation from images via normal bridging. arXiv preprint arXiv:2503.22236 **3**, 2 (2025)
47. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)
48. Yu, J., Xu, Y., Koh, J.Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B.K., et al.: Scaling autoregressive models for content-rich text-to-image generation. arXiv preprint arXiv:2206.10789 **2**(3), 5 (2022)
49. Zhang, S., Wang, J., Zhang, Y., Zhao, K., Yuan, H., Qin, Z., Wang, X., Zhao, D., Zhou, J.: I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. arXiv preprint arXiv:2311.04145 (2023)
50. Zhao, S., Chen, D., Chen, Y.C., Bao, J., Hao, S., Yuan, L., Wong, K.Y.K.: Unicontrolnet: All-in-one control to text-to-image diffusion models. Advances in Neural Information Processing Systems **36**, 11127–11150 (2023)
51. Zhou, C., Yu, L., Babu, A., Tirumala, K., Yasunaga, M., Shamis, L., Kahn, J., Ma, X., Zettlemoyer, L., Levy, O.: Transfusion: Predict the next token and diffuse images with one multi-modal model. arXiv preprint arXiv:2408.11039 (2024)

52. Zou, Z., Cheng, W., Cao, Y.P., Huang, S.S., Shan, Y., Zhang, S.H.: Sparse3d: Distilling multiview-consistent diffusion for object reconstruction from sparse views. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 7900–7908 (2024)