Subjective Camera 1.0: Bridging Human Cognition and Visual Reconstruction through Sequence-Aware Sketch-Guided Diffusion

Haoyang Chen[1,2,3]✤ Dongfang Sun[1,2]✤ Caoyuan Ma[1,2]✤ Shiqin Wang[1,2] Kewei Zhang[1,2]
Zheng Wang[1,2,3]✉ Zhixiang Wang[4]✉

[1]National Engineering Research Center for Multimedia Software, Institute of Artificial Intelligence, School of Computer Science, Wuhan University [2]Hubei Key Laboratory of Multimedia and Network Communication Engineering
[3]Zhongguancun Academy, Beijing, China. 100094 [4]CyberAgent AI Lab, Japan

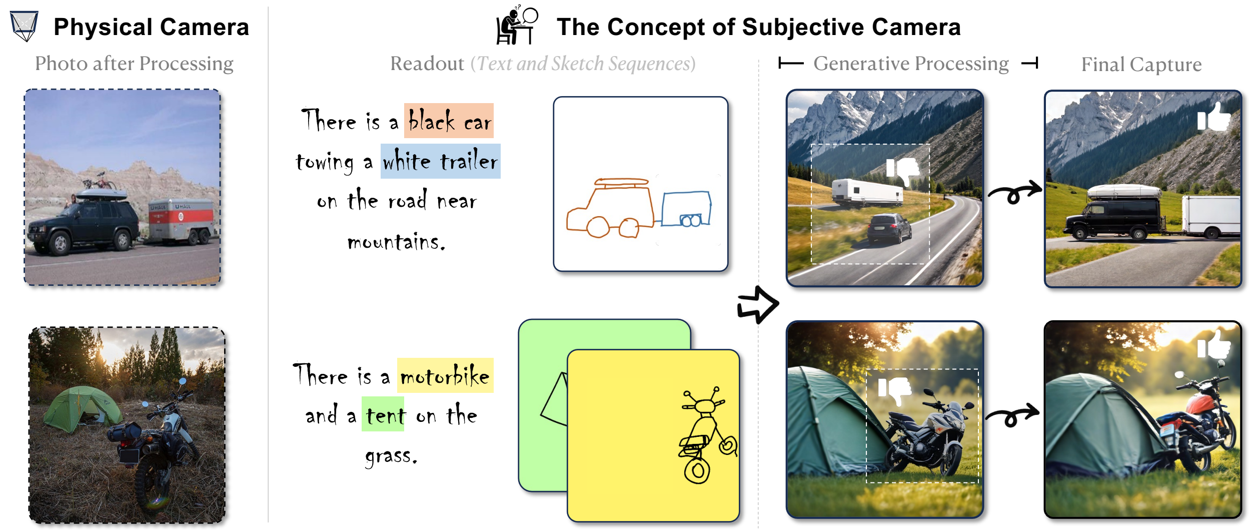

Physical Camera
Photo after Processing
The Concept of Subjective Camera
Readout (Text and Sketch Sequences)
Generative Processing
Final Capture
There is a black car towing a white trailer on the road near mountains.
There is a motorbike and a tent on the grass.

Figure 1. **What is a subjective camera?** A subjective camera transforms a person's memory into photorealistic images. We take a small yet crucial step toward this vision by leveraging generative models to sequentially decode textual and sketch-based readouts in their natural order, paving the way for reconstructing moments without physical cameras.

Abstract

We introduce the concept of a subjective camera *to reconstruct meaningful moments that physical cameras fail to capture. We propose* Subjective Camera 1.0*, a framework for reconstructing real-world scenes from readily accessible subjective readouts,* i.e.*, textual descriptions and progressively drawn rough sketches. Built on optimization-based alignment of diffusion models, our approach avoids large-scale paired training data and mitigates generalization issues. To address the challenge of integrating multiple abstract concepts in real-world scenarios, we design a Sequence-Aware Sketch-Guided Diffusion framework with three loss terms for concept-wise sequential optimization, following the natural order of subjective readouts. Experiments on two datasets demonstrate that our method achieves state-of-the-art performance in image quality as well as spatial and semantic alignment with target scenes. User studies with 40 participants further confirm that our approach is consistently preferred. Our project page is at: subjective-camera.github.io*

1. Introduction

Cameras have long been the most popular devices for preserving memorable moments. However, a physical camera

✤These authors contributed equally to this work.
✉Corresponding authors.

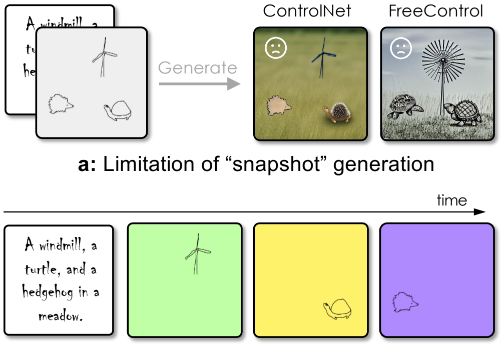
cannot always be present to capture *every* moment. This raises a question of how such unrecorded experiences can be preserved as *pixels* in the absence of a camera.

This paper proposes the concept of a *subjective camera*[1], where humans function as "imaging devices", aiming to faithfully reconstruct real-world scenes from memory. Specifically, we define a subjective camera as a system through which individuals encode sensory inputs into memory based on personal salience and emotional context, and later read out these stored representations and decode them into pixels that reconstruct scenes they experienced. We draw an analogy to physical cameras: just as a camera records, reads out, and reconstructs the visual world through optical, electronic, and computational processes, the human mind selectively captures ("records") perceptual details, which are organized into a retrievable format ("read out") and later reconstructs them as pixels.

It is generally challenging for humans *without* professional training to directly reconstruct a dense, pixel-wise image. Instead, the most practical approach is to *read out* the mental scene through textual descriptions and/or freehand sketches and leverage computational tools to decode these readouts. However, text alone often fails to fully convey a mental image, particularly in terms of object layout and fine details. Freehand sketches provide complementary information beyond text but remain sparse and often contain noise and uncertainty [11, 15]. With the advancement of generative models [20, 23], this long-standing vision is becoming increasingly feasible by decoding such subjective readouts into photorealistic imagery.

Although not intended to reconstruct real-world scenes, several works have explored generating images from combined text and sketch inputs [4, 25, 30, 35–37]. These approaches either adapt model weights using *large-scale* paired data [19, 25, 28, 30, 33, 36, 37] or modify the latent variables through *per-scene* optimization [4, 28, 35]. Training-based methods directly learn the mapping from sketches to pixels. However, they require extensive paired data and struggle with user-specific biases, often generalizing poorly to abstract sketches beyond the training distribution. In contrast, training-free methods avoid the heavy data requirements, high computational costs, and generalization challenges of training-based approaches.

However, optimization-based methods often fail to reconstruct *multiple* concepts from *idiosyncratic* sketches, as would be required for recreating real-world scenes. This often leads to missing concepts or misaligned spatial relationships (see Figure 2a). The main obstacle lies in simultaneously interpreting multi-concept sketches, which inherently combine *varying* levels of abstraction and randomness across different concepts. For instance, freehand sketches may only roughly indicate object boundaries, omit fine details such as texture and geometry, or distort proportions in non-standard ways. While generative priors can fill sparse regions and mitigate randomness during the transformation from sketches to pixels, attempting to integrate all concepts at once often reduces the problem to a least-squares solution [1, 4], ultimately blurring their distinct characteristics.


A windmill, a turtle, and a hedgehog in a meadow.
Generate
ControlNet
FreeControl
time
A windmill, a turtle, and a hedgehog in a meadow.

a: Limitation of "snapshot" generation

b: Inspiration from human readout order

Figure 2. **Motivation for concept-wise sequential optimization.** (a) Attempting to "snap" all concepts into a complete image at once using generative diffusion priors often leads to suboptimal results, such as distortions and misaligned semantics or layouts. (b) In contrast, humans reconstruct mental scenes step by step. This natural readout order inspires our design[2], making it better suited to leverage current generative diffusion priors.

Our key idea is simple — to apply optimizations to each concept *individually* and do so *sequentially*. This design considers not only the limitations of the "snapshot" generation process but also human cognition. Humans tend to describe their mental images step-by-step [13], typically starting with a textual description of abstract concepts or the overall scene structure, and then progressively refining individual elements with increasing detail (see Figure 2b).

Building upon this idea, we introduce a sequence-aware, sketch-guided diffusion framework that sequentially optimizes the latent noise of a pre-trained text-to-image (T2I) diffusion model [23] to align individual subjective readouts. The order of these optimizations mirrors the natural process of reading out subjective impressions: it begins with a text-reward optimization (Section 3.2) to align the generated image with the initial textual description, followed by concept-wise sequential optimizations (Section 3.5) to incorporate progressively provided sketches that may encode shapes, poses, and spatial arrangements. The sequential optimization also incorporates a loss to ensure that previously optimized concepts are not significantly altered.

Two additional components are proposed to improve

[1] While the term "subjective camera" appears in cinematic contexts, our usage here differs significantly in meaning.

[2] This design philosophy is reminiscent of the progressive decoding used in rolling-shutter image sensors, which arises from the need to accommodate the limited processing capacity of the circuits.

the concept-wise sequential optimization. First, each input sketch is individually encoded into the latent space to provide spatial guidance. To bridge the gap between rough sketch shapes and their corresponding real-world forms, we introduce an optimization-based inversion strategy for refining this spatial guidance (Section 3.3). Second, because input sketches often lack appearance details, directly optimizing with their guidance over multiple iterations may lead to unnatural textures and degraded visual fidelity. We therefore leverage the latent representation obtained from text-reward optimization to define an appearance loss (Section 3.4), thereby preserving visual style quality.

By processing sketches individually and sequentially, our approach naturally aligns with the human cognition process, enabling users to incrementally construct mental images while ensuring each addition respects and complements previous elements. This sequence-aware generation strategy provides *fine-grained* control over complex multi-concept scenes while preserving the user's subjective intent encoded in the drawing order. Evaluations on two datasets demonstrate that our method achieves state-of-the-art performance in recreating a real-world scene from human memory, demonstrating the potential of subjective cameras. Notably, our framework is entirely *training-free*, eliminating the need for large-scale paired datasets and avoiding generalization.

To sum up, our contributions are twofold:

- We propose the concept of a subjective camera, which functions the human as an "imaging device" and leverages computational tools to transform cognitive impressions into photographs.
- We propose a sequence-aware diffusion-guided generation framework that enables faithful reconstruction of complex, multi-concept scenes. This framework aligns with human cognition and surpasses "snapshot" generation methods, thereby establishing a new paradigm for cognition-driven image generation.

2. Related Work

Aligning Image Synthesis Models Despite their success, T2I generative models [20, 22, 23] often fail to reproduce the fine-grained semantics and compositional details described in complex prompts. Reward-based alignment has emerged as a promising direction [6, 7, 12, 14, 26, 31, 32], using human preference models [12, 29, 32] to guide generation. The human preference models were trained on paired human preference data. Early works [12, 32] fine-tune diffusion models with the reward models, whereas later methods [6, 14] sidestep costly fine-tuning by directly optimizing the latent noise [6, 31]. However, these techniques largely improve global prompt adherence while neglecting *spatial* control. Our model not only aligns the T2I generative models with the text prompt, but also ensures the spatial information to respect to user-provided sketches.

Controllable Image Synthesis T2I models inherently rely on highly *compressed* textual tokens, which limits their controllability and often prevents them from meeting user expectations [20, 22, 23]. To overcome these limitations, a growing body of research has explored controllable generation by incorporating fine-grained conditioning signals, such as reference images [2], edge maps or contours or skeletons [33], semantic layouts [1, 3, 38], and lighting specifications [34]. One line of work fine-tunes pretrained diffusion models by adding trainable modules [19, 33, 37], while others directly retrain the diffusion backbone by minimizing reconstruction objectives [2]. In parallel, training-free techniques [5, 7, 9, 26] have been developed to enhance controllability without modifying model weights. These include attention injection or latent optimization strategies, such as Prompt-to-Prompt editing [9] and diffusion self-guidance [5]. Despite these advances, existing methods often fail in complex scenes involving multiple guidance given by freehand sketches, where interactions between elements lead to interference and degraded synthesis quality. Building on these insights, our method incorporates attention-based guidance extraction but *extends* it to better handle multi-concept freehand sketches.

Sketch-to-Image Synthesis It is the task most relevant to our concept of the subjective camera. It aims to generate images from freehand sketches along with textual prompts. Several recent efforts have attempted to adapt T2I diffusion models for sketch-to-image synthesis [4, 17, 25, 28, 35, 36]. Training-based approaches [25, 36] fine-tune the T2I models with paired data. However, this training-based method incurs substantial computational cost and suffers from poor generalization. In contrast, training-free approaches [4, 17, 35] avoid fine-tuning by guiding the diffusion denoising process with sketches. However, these methods still *jointly* create multiple concepts, causing interference between their differing levels of randomness and abstraction. More critically, both training-based and training-free approaches fail to address the *subjective biases* inherent in user-provided sketches and are particularly vulnerable to highly abstract inputs, often leading to appearance distortions and semantic inconsistencies in the generated outputs. Furthermore, unlike creative generation methods, our focus is on reconstructing real-world scenes.

3. Sequence-Aware Sketch-Guided Diffusion

3.1. Overview

Our goal is to render photorealistic images from subjective readouts, consisting of a textual description p and a sequence $S = \{s_i\}_{i=1}^{N}$ of N freehand sketches, where each sketch represents a distinct concept. As illustrated

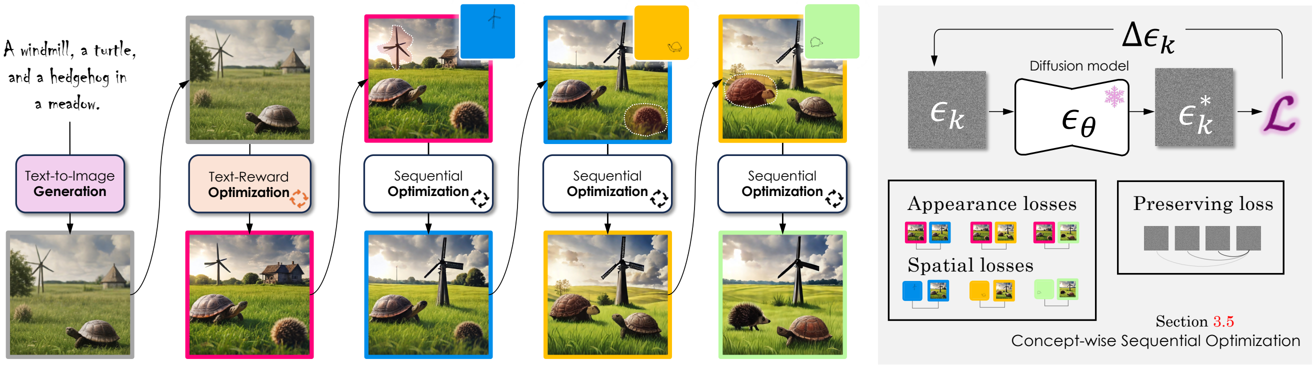


Figure 3. **Sequence-aware sketch-guided diffusion.** (*Left*) User-provided textual description undergoes text-to-image generation and text-reward optimization to obtain the initial latent. Then, the following generation progresses under three constraints — spatial layout conforming to sketch topology while appearance details adhere to the initial image from text-reward optimization, with consistency maintained through ordered latent propagation and three loss terms, ensuring coherent integration of emerging and established scene elements. (*Right*) The concept-wise sequential optimization.

in Figure 3, we design a Sequence-Aware Sketch-Guided Diffusion framework to accomplish this challenging task. Our method processes the subjective readouts individually and sequentially, following the order in which users describe them. It employs a sequence-aware optimization strategy to align a pre-trained generative diffusion model, $\mathcal{G}_{:} = \epsilon_\theta(\epsilon_\theta(\ldots\epsilon_\theta(\epsilon, p)))$, and ϵ_θ denotes the pre-trained U-Net, with these readouts, naturally mirroring the human cognitive process of step-by-step scene construction. The process begins with a text-reward optimization, which establishes an initial latent representation and appearance prior for subsequent concept-wise optimizations. The appearance prior is extracted via style encoding, while sketch encoding, implemented using diffusion inversion, provides spatial guidance for individual sketches. In addition, a preserving loss is introduced to maintain previously optimized concepts throughout the sequential process. These components collectively enable our framework to decode subjective readouts into coherent, photorealistic scenes.

3.2. Text-Reward Optimization

We randomly sample a latent noise ϵ and feed it, together with the textual description p, into a pre-trained T2I diffusion model, which renders an image through the full reverse diffusion process, $\mathcal{G}_\theta$. To ensure that the generated image contains the concepts in the given textual information, without concept omission or significant misalignment, we formulate a text-reward optimization procedure as existing practice [6, 14] that iteratively refines the latent noise to maximize a text-image alignment objective:

$$\epsilon^{0\star} = \arg\max_{\epsilon} \mathcal{C}_F\left(\mathcal{G}_\theta(\epsilon, p), p\right), \tag{1}$$

where $\mathcal{C}_F$ represents a differentiable reward model that measures alignment between the generated image and the text prompt p. The optimization process iteratively refines ϵ through gradient ascent:

$$\epsilon_{t+1} = \epsilon_t + \eta \nabla_\epsilon \mathcal{C}_F(\mathcal{G}_\theta(\epsilon_t, p)), \tag{2}$$

where η is the learning rate.

3.3. Sketch Encoding

Before proceeding to the sketch-guided optimization, we first encode the sketch to extract high-level semantic guidance, rather than relying on low-level line details. Given a list of input sketches $\{s_i\}_{i=1}^N$, each sketch s_i is encoded into a latent representation $z_i \in \mathbb{R}^{h\times w\times d}$ using the variational autoencoder (VAE) encoder E_ϕ of the diffusion model: $z_i = E_\phi(s_i)$. Then, we can leverage arbitrary diffusion inversion algorithms, such as DDIM inversion [18] or SDEdit [16], to map the sketch into latent noise. To address the significant domain gap between abstract planar sketches and realistic images, we introduce an optimization procedure that iteratively refines the latent representation z_i by aligning it with the rich prior knowledge embedded in the diffusion model.

At each optimization step k, we perturb the latent sketch representation with a Gaussian noise $\epsilon_k \sim \mathcal{N}(0, \mathbf{I})$ to obtain a noisy latent representation

$$z_t^{(k)} = \alpha_{t_k} z_i^{(k)} + \sigma_{t_k}\epsilon_k, \tag{3}$$

where α_{t_k} and σ_{t_k} are time-dependent scaling factors following the noise schedule of the diffusion process. The noisy latent $z_t^{(k)}$ is then fed into the pre-trained U-Net ϵ_θ, which predicts the noise component $\hat{\epsilon}_\theta(z_t^{(k)}, t_k)$. The noise prediction error $\Delta\epsilon_k$ is computed as:

$$\Delta\epsilon_k = \hat{\epsilon}_\theta(z_t^{(k)}, t_k) - \epsilon_k. \tag{4}$$

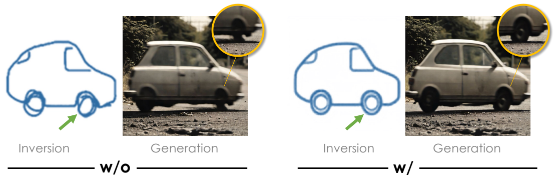


Figure 4. **The effectiveness of sketch optimization.** Without optimization (*left*), the generated image fails to accurately respect the wheel placement and proportions indicated in the sketch. With our optimization (*right*), the generated result better aligns with the sketch structure, particularly in the wheel positions (green arrows) and overall shape, demonstrating improved spatial coherence.

This noise gap serves as a gradient signal to iteratively refine the latent sketch features. Specifically, we update the latent representation $z_i^{(k)}$ as

$$z_i^{(k+1)} = z_i^{(k)} - \eta_k \left(\Delta\epsilon_k + \lambda(z_i^{(k)} - z_i^{(0)}) \right), \tag{5}$$

where λ is a regularization term that anchors the optimization to the initial latent $z_i^{(0)}$, preventing semantic drift.

Through this iterative optimization in Eq. (3–5), the latent $\dot{z}_i$is pushed to align with the rich prior knowledge embedded in the diffusion model.[3] This alignment effectively infuses the sketch with realistic details, textures, and structural coherence, thereby establishing a robust mapping from human hand-drawn sketches to physically grounded visual representations (see Figure 4).

However, we observed that the optimization method is less effective when the sketch is drawn relatively small. In such cases, the generated result often fails to align with the sketch and instead produces a shape-independent object in a central position. To address this, we suggest scaling the original sketch using linear interpolation or, alternatively, discarding the optimization for that step.

3.4. Appearance Encoding

While the sketch provides spatial information, additional appearance cues are required to ensure visual quality during optimization. We first define a operator $\mathcal{A}(\cdot)$ to obtain the appearance descriptor:

$$\mathcal{A}(x) = \left\{ \frac{\sum_{u,v} \text{sigmoid}\left([x]_{uv}\right) [f_x]_{uv}}{\sum_{u,v} \text{sigmoid}\left([x]_{uv}\right)} \mid f_x \right\}, \tag{6}$$

which is a collection of weighted spatial means of diffusion features f_x and f_x are from different layer and time steps. Then, we treat the descriptor extracted from the optimized latent $z^{0\star}$, derived from $\epsilon^{0\star}$ in Eq. (1), as our appearance prior. We will compare the appearance descriptor of the current latent and the appearance prior during optimization.

3.5. Concept-wise Sequential Optimization

Next, we leverage this order information of subjective readouts to progressively construct the scene while maintaining coherence between newly introduced and previously established elements.

Spatial Attention-guided Control We leverage the spatial attention mechanisms within the diffusion model's U-Net architecture. For each sketch concept s_i, we extract cross-attention maps that highlight regions in the latent space corresponding to specific semantic concepts:

$$\mathcal{M}(x) = \text{softmax}\left(\frac{Q(x)K(x)^T}{\sqrt{d}}\right) > \tau, \tag{7}$$

where $Q(x)$ and $K(x)$ are query and key matrices, d is the dimensionality of the key vectors, and τ is a threshold parameter. The attention maps $\mathcal{M}(\dot{z}_i)$ extracted from $\dot{z}_i$, the latent of the sketch s_i (Section 3.3), serve as spatial indicators to guide the concept's layout. The control is given through the spatial loss, defined as:

$$\mathcal{L}_{\text{spatial}}(z_i, \dot{z}_i) = \frac{\sum_{u,v} [\mathcal{M}(\dot{z}_i)]_{uv} \, \|[\dot{z}_i]_{uv} - [z_i]_{uv}\|_2^2}{\sum_{u,v} [\mathcal{M}(\dot{z}_i)]_{uv}}, \tag{8}$$

where u, v are spatial positions indices.

Appearance Control We define the appearance loss as:

$$\mathcal{L}_{\text{app}}(z_i, z^{0\star}) = \left\|\mathcal{A}(z_i) - \mathcal{A}(z^{0\star})\right\|_2^2, \tag{9}$$

where $\mathcal{A}$ is the operator of exacting appearance descriptors.

Sequential Concept Integration To maintain coherence across the progressive addition of concepts, we implement a sequential integration process that updates the latent noise representation while preserving previously incorporated elements. For each sketch s_i in the sequence, we compute:[4]

$$z_i = z_{i-1} + \Delta z_i, \tag{10}$$

where z_{i-1} is the optimized latent from the previous concept, and Δz_i represents the incremental update required to incorporate the current sketch concept s_i. This update is derived from our loss function:

$$\mathcal{L}_i = \alpha_i \mathcal{L}_{\text{new}}(z_i, \dot{z}_i, z^{0\star}) + \beta_i \mathcal{L}_{\text{preserve}}(z_i, \dot{z}_{0,\ldots,i-1}), \tag{11}$$

where α_i and β_i are balancing coefficients. $\mathcal{L}_{\text{new}}$ ensures proper integration of the current z_i and is defined as:

$$\mathcal{L}_{\text{new}}(z_i, \dot{z}_i, z^{0\star}) = \mathcal{L}_{\text{spatial}}(z_i, \dot{z}_i) \, + \mathcal{L}_{\text{app}}(z_i, z^{0\star}), \tag{12}$$

while $\mathcal{L}_{\text{preserve}}$ maintains the integrity of previously optimized concepts, given by:

$$\mathcal{L}_{\text{preserve}}(z_i, \dot{z}_{0,\ldots,i-1}) = \sum_{j=1}^{i-1} \gamma^{i-j} \mathcal{L}_{\text{spatial}}(z_i, \dot{z}_j), \tag{13}$$

[3] To avoid ambiguity, we use $\dot{z}_i$ to denote the optimized latent for the i-th sketch concept, distinguishing it from latent z_i would be used in the concept-wise sequential optimization.

[4] In fact, an inner-loop optimization is omitted here for simplicity.

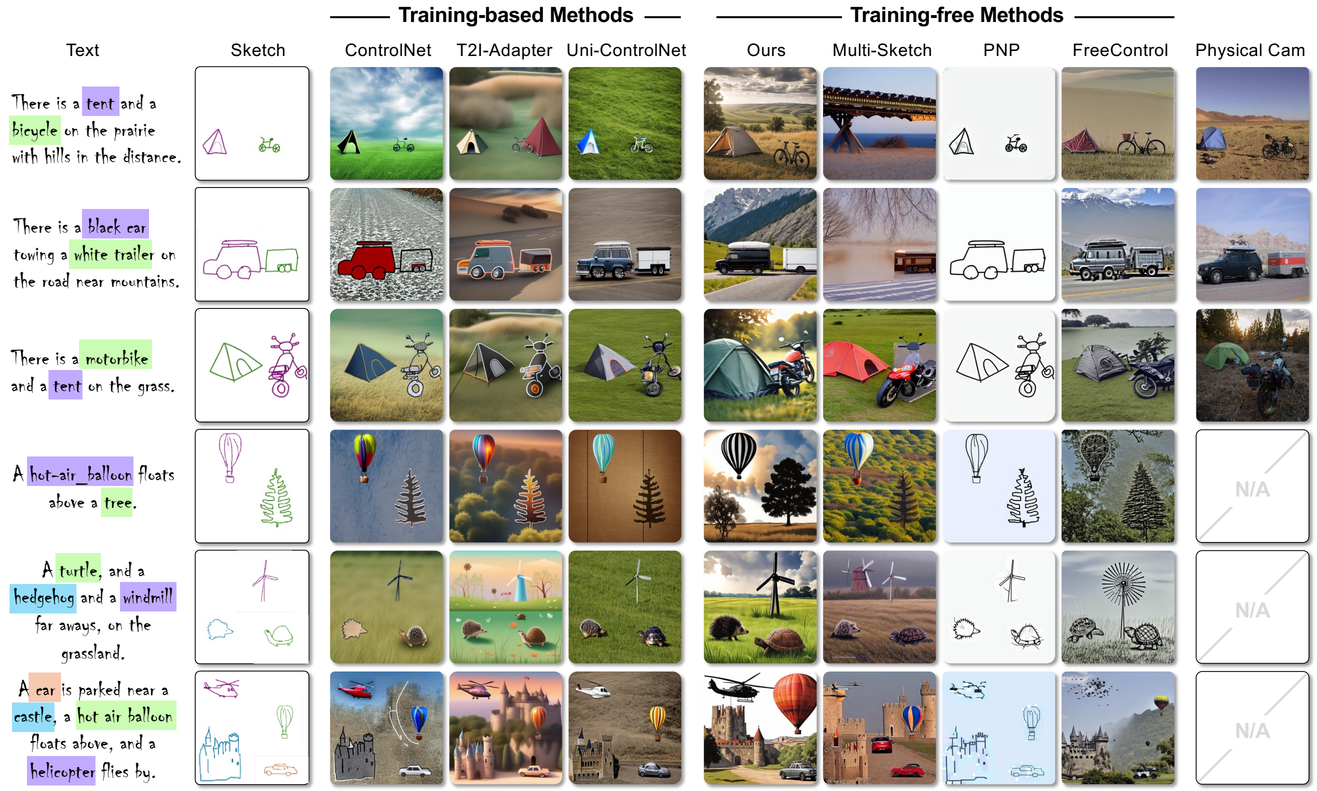

Training-based Methods
Training-free Methods
Text
Sketch
ControlNet
T2I-Adapter
Uni-ControlNet
Ours
Multi-Sketch
PNP
FreeControl
Physical Cam
There is a tent and a bicycle on the prairie with hills in the distance.
There is a black car towing a white trailer on the road near mountains.
There is a motorbike and a tent on the grass.
A hot-air_balloon floats above a tree.
A turtle, and a hedgehog and a windmill far aways, on the grassland.
A car is parked near a castle, a hot air balloon floats above, and a helicopter flies by.
N/A
N/A
N/A

Figure 5. **Qualitative comparison** of existing sketch-to-image generation methods, and our method on FMC and CMC datasets. Existing training-based methods tend to overfit to sketch lines, often failing to reconstruct scenes realistically, while training-free methods frequently misinterpret or overlook key concepts. In comparison, our approach more faithfully represents real-world physical scenes. The reading and processing order of the sketches is indicated by colors as follows: 🟣 🟢 🔵 🟠.

where j means preserving established concepts, and γ^{i-j} is a decay factor that weights the importance of earlier concepts based on their step gap from the current step.

4. Experiments and Results

4.1. Experiment Setup

Datasets To rigorously evaluate our method, we have undertaken extensive efforts to develop two specialized datasets: the Composed Multi-Concept (CMC) dataset and the Freehand Multi-Concept (FMC) dataset. Both datasets are designed to comprehensively assess the capabilities of our framework in handling multi-concept scene understanding and generation. **CMC** dataset comprises 142 scene sketches, each containing 2-4 distinct concepts sampled from the 150 semantic classes in the Sketchy dataset [24]. Each is accompanied by annotated textual prompts with the help of multimodal large language models. The CMC dataset does not provide real-world images captured by physical cameras for reference. **FMC** dataset is constructed from freehand sketches drawn by volunteers, based on real-world images sourced from publicly available outdoor scene datasets. It comprises 42 hand-drawn sketches, each containing 2-3 distinct concepts. Additionally, each sketch is paired with a corresponding real-world image as a reference for validation and evaluation purposes.

Evaluation Metrics To conduct a comprehensive quantitative evaluation of the quality of generated images, four widely-used evaluation metrics were employed, including Fréchet Inception Distance (FID) [10], CLIP-T Score (CLIP-T) [21], Human Preference Score (HPS) [29], and CLIP-I distance (CLIP-I) [21]. **FID** measures the feature distribution similarity between generated and real images. Lower FID values indicate better distribution matching (↓). **CLIP-T** quantifies the image-text alignment degree within the CLIP-T embedding space. A higher CLIP-T score reflects greater semantic alignment between the generated image and text prompt (↑). **HPS** measures the subjective quality of the generated images based on human evaluations. Higher HPS values suggest that the generated images are more favorable or preferred by human evaluators (↑). **CLIP-I** measures the alignment between two images. A higher value suggests that the two images are more consistent in terms of their semantic content (↑).

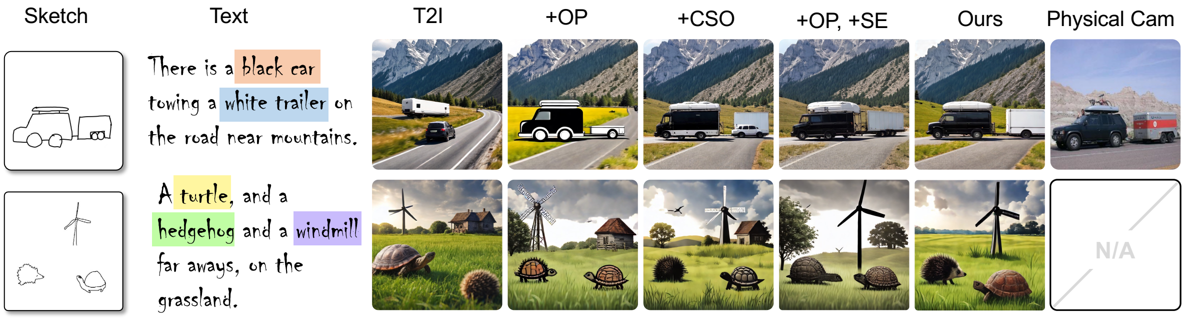

Sketch
Text
T2I
+OP
+CSO
+OP, +SE
Ours
Physical Cam
There is a black car towing a white trailer on the road near mountains.
A turtle, and a hedgehog and a windmill far aways, on the grassland.
N/A

Figure 6. **Ablation study on FMC and CMC datasets.** Column headers are defined as follows: T2I denotes the baseline text-to-image diffusion model; OP (Overall Perception) indicates the single-step guidance approach utilizing complete sketch as a holistic conditioning signal; CSO refers to our proposed Concept-wise Sequential Optimization (Section 3.5); SE represents the Sketch Encoding module (Section 3.3) that enhances structural plausibility by injecting diffusion priors into imperfect sketches. The qualitative comparison demonstrates that CSO achieves finer layout control than OP. While SE effectively bridges the domain gap between freehand sketches and physical reality through latent space regularization, ultimately elevating generation quality.

Metric / Method	CMC			FMC			
	FID ↓	CLIP-T ↑	HPS ↑	FID ↓	CLIP-T ↑	HPS ↑	CLIP-I ↑
Training-based methods							
ControlNet [33]	15.28	61.71	0.794	14.45	66.80	0.802	0.622
T2I-Adapter [19]	18.13	61.61	0.745	15.37	61.61	0.745	0.689
Uni-ControlNet [37]	19.60	62.58	0.761	18.53	62.58	0.761	0.712
Training-free methods							
Multi-Sketch [4]	13.05	64.11	0.740	11.59	72.02	0.775	0.690
PNP [27]	17.01	63.27	0.840	16.26	71.24	0.852	0.593
FreeControl [17]	14.52	64.74	0.757	11.17	66.67	0.759	0.775
Ours	**11.56**	**66.64**	**0.850**	**10.18**	**74.25**	**0.862**	**0.797**

Table 1. **Quantitative results on CMC and FMC datasets.** Our method consistently surpasses all training-free methods in distribution matching, image-text alignment, subjective quality, and appearance details as measured by FID, CLIP-T, HPS, and CLIP-I. Our method achieves superior image-text alignment compared to training-based methods (‘↓’ indicates lower values are better, while ‘↑’ is opposite).

Implementation Details In the implementation of our proposed method, we set the learning rate η in Eq. (2) and the threshold parameter τ in Eq. (7) is 5.0 and 0.3, respectively. Meanwhile, hyperparameters α_i and β_i in Eq. (11) are set as 0.8 and 0.2, respectively. γ in Eq. (13) is set as 0.9. Our method is performed on one NVIDIA GeForce RTX 4090, and input sketch images are 512×512 in resolution.

4.2. Comparisons

We conducted a comprehensive quantitative and qualitative evaluation of the proposed method and competing methods on the CMC and FMC datasets. Our competing methods include the training-based approaches ControlNet [33], T2I-Adapter [19], and Uni-ControlNet [37], as well as the training-free approaches Multi-Sketch [4], PNP [27], and FreeControl [17]. The quantitative results are presented in Table 1 and qualitative results are shown in Figure 5.

We observe that training-based methods [19, 33, 37] are highly susceptible to subjective biases in sketches and often fail to adhere to text prompts, producing distorted and unrealistic images with lower CLIP-T scores. Training-free methods [4, 8, 17] suffer from concept omissions, conspicuous artifacts, and inconsistencies both among concepts and between concepts and their backgrounds, resulting in poor image realism. These issues stem from mutual interference during the “snapshot” generation of multiple concepts. In contrast, our proposed method faithfully reconstructs real-world scenes and achieves the best performance across all evaluation metrics on both the CMC and FMC datasets, demonstrating its effectiveness.

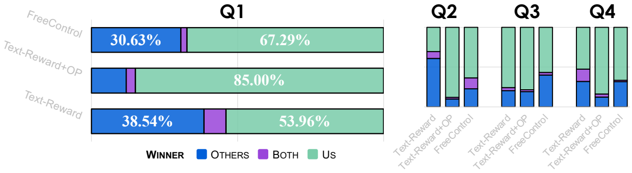


Figure 7. **User preference experiment.** We asked participants to answer four questions based on the reference images, subjective readouts, and generated images, which included one produced by our method and one by a competing method. Results show that users significantly preferred our outputs over those of competing methods in most cases.

4.3. Ablation Study

To validate the effectiveness of each module in our method, we conducted ablation experiments on the FMC and CMC datasets, with the visualization results shown in Figure 6.

It is observed that traditional T2I diffusion models (T2I) generate images that are close to text prompts. While incorporating sketches as guidance further constrains the consistency of concept positioning, deviations in appearance and spatial alignment persist (T2I, OP). The sequential concept understanding facilitated by Concept-wise Sequential

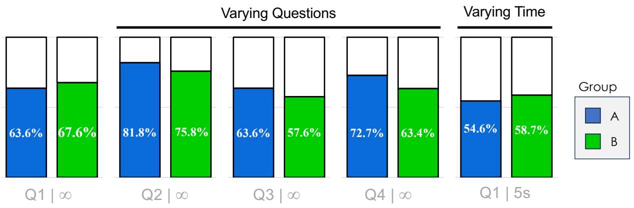


Figure 8. **Control experiments on user preferences.** We controlled three factors to examine their influence: whether the rater drew the photo, the availability timing of the reference image (∞ *vs*. 5s), and the types of questions asked. In general, our method is consistently better than the competing method, FreeControl.

Optimization (CSO) in Section 3.5 enhances the alignment of concept positioning and appearance with input sketches (CSO). Simultaneously, the Sketch Encoding (SE) module in Section 3.3 aligns content latent variables with real-world objects, mitigating geometric misalignment (T2I, OP, SE). When integrating all modules, our method faithfully reconstructs real-world scenes and achieves optimal results (Ours), demonstrating the effectiveness of each module and underscoring the indispensable nature of their design.

4.4. User Preference Experiments

We invited 12 participants (group A) to provide subjective readouts and an additional 28 participants (group B) to evaluate the reconstructed images. Participants in group A were first asked to select a deeply impressive scene from their photo gallery and then provide a textual description along with sketches within two minutes. Our method and competing approaches, including two baselines (text-reward optimization and text-reward optimization with sketch-guided optimization based on all concepts) as well as an optimal method, FreeControl [17], were used to generate images from these readouts. Participants in both groups were then shown the subjective readouts, the reference image, and the generated images for the same scene. They were asked to answer four single-choice questions in a single-blind manner: Q1: Which image is closer to the original image (both style and content)? Q2: Which image has higher visual quality (*e.g*., more realistic, more natural)? Q3: Which image better matches the structure of the sketch? Q4: Which image better matches the textual description?

Figure 7 presents the statistics of the collected results. Participants judged our reconstructions as being closer to the corresponding real-world scenes compared to competing methods. Similar conclusions were observed for the other three questions. However, there was an exception in the comparison between our method and text-reward optimization regarding image quality (Q2). We observed that text-reward optimization without additional control tends to hallucinate details that enhance perceived realism, leading some users to prefer it for Q2. This also explains why our advantage over text-reward optimization in Q1 was not statistically significant and why some participants expressed hesitation in this case.

We further conducted controlled experiments to examine the effects of three factors: whether the rater provided the subjective readouts, the availability timing of the reference image, and the types of questions asked. Results in Figure 8 show that our method consistently outperformed FreeControl [17], regardless of these variations. Notably: 1) Participants who provided subjective readouts showed slightly lower preference for our method in terms of similarity to the reference image compared to those who did not, while showing a higher preference in terms of alignment with the readouts and perceived image quality. We suppose this is mainly because of the subjectivity in reading out. 2) Reducing the time participants could view the reference image decreased the preference for our method; when allowed to compare freely, participants preferred our results more strongly.

5. Conclusions

This paper presented Subjective Camera, a paradigm for reconstructing real-world scenes from humans' mental impressions. Our framework relies on the most accessible readouts of mental imagery, *i.e*., textual descriptions and sketches, and employs a training-free, sequence-aware optimization process. This approach eliminates the need for costly fine-tuning of text-to-image diffusion models on large-scale data while mitigating challenges such as dependence on sketch quality and geometric misalignment. Extensive experiments on both synthetic and real-world sketch datasets demonstrate that Subjective Camera achieves state-of-the-art performance, outperforming both training-free and training-based sketch-to-image generation methods across all evaluation metrics. Furthermore, user preference studies confirm that participants consistently favor our method.

In closing, we address a natural question: "With the rapid advancement of generative AI, will physical cameras ever become unnecessary" For now, the answer is *no*. First, the reconstructed images still fall short of perfectly matching actual scenes. This limitation largely stems from the fact that current T2I models are trained on highly diverse data, creating gaps between the prior and the specific scene to be reconstructed. One promising direction to mitigate this is to personalize the prior models using user-specific galleries. Second, we observe that variations in the order and quality of sketches have a significant impact on reconstruction results. To address this, we plan to enhance our framework with computational tools that guide users in providing more precise and suitable subjective readouts. For these reasons, we designate this system as Subjective Camera 1.0, leaving these improvements for future iterations.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China (62225113, 624B2109, 623B2079) and the Zhongguancun Academy Project (20240308). We thank Po-Hsun Huang for providing the questionnaire system for our user study, Kaijin Zheng for assistance with data processing, and all participants who took part in the studies.

References

[1] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. *Proceedings of Machine Learning Research*, 202:1737–1752, 2023. 2, 3

[2] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402, 2023. 3

[3] Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5343–5353, 2024. 3

[4] Zhenwei Cheng, Lei Wu, Changshuo Wang, and Xiangxu Meng. Scene sketch-to-image synthesis based on multi-object control. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3775–3779. IEEE, 2024. 2, 3, 7

[5] Dave Epstein, Allan Jabri, Ben Poole, Alexei Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. *Advances in Neural Information Processing Systems*, 36:16222–16239, 2023. 3

[6] Luca Eyring, Shyamgopal Karthik, Karsten Roth, Alexey Dosovitskiy, and Zeynep Akata. Reno: Enhancing one-step text-to-image models through reward-based noise optimization. *Advances in Neural Information Processing Systems*, 37:125487–125519, 2025. 3, 4

[7] Xiefan Guo, Jinlin Liu, Miaomiao Cui, Jiankai Li, Hongyu Yang, and Di Huang. Initno: Boosting text-to-image diffusion models via initial noise optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9380–9389, 2024. 3

[8] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*. 7

[9] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control.(2022). *URL https://arxiv. org/abs/2208.01626*, 3, 2022. 3

[10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[11] Zechao Hu, Zhengwei Yang, Hao Li, Yixiong Zou, Fengbin Zhu, and Zheng Wang. Unified category and style generalization for instance-level sketch retrieval. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 864–873. Association for Computing Machinery, 2025. 2

[12] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36: 36652–36663, 2023. 3

[13] Stephen M Kosslyn. Aspects of a cognitive neuroscience of mental imagery. *Science*, 240(4859):1621–1626, 1988. 2

[14] Seung Hyun Lee, Yinxiao Li, Junjie Ke, Innfarn Yoo, Han Zhang, Jiahui Yu, Qifei Wang, Fei Deng, Glenn Entis, Junfeng He, et al. Parrot: Pareto-optimal multi-reward reinforcement learning framework for text-to-image generation. In *European Conference on Computer Vision*, pages 462–478. Springer, 2024. 3, 4

[15] Kejun Lin, Zhixiang Wang, Zheng Wang, Yinqiang Zheng, and Shin'ichi Satoh. Beyond domain gap: Exploiting subjectivity in sketch-based person retrieval. In *Proceedings of the 31st ACM international conference on multimedia*, pages 2078–2089, 2023. 2

[16] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 4

[17] Sicheng Mo, Fangzhou Mu, Kuan Heng Lin, Yanli Liu, Bochen Guan, Yin Li, and Bolei Zhou. Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7465–7475, 2024. 3, 7, 8

[18] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6038–6047, 2023. 4

[19] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4296–4304, 2024. 2, 3, 7

[20] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 2, 3

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 6

[22] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. 3

[23] 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*, pages 10684–10695, 2022. 2, 3

[24] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG)*, 35(4):1–12, 2016. 6

[25] Vishnu Sarukkai, Lu Yuan, Mia Tang, Maneesh Agrawala, and Kayvon Fatahalian. Block and detail: Scaffolding sketch-to-image generation. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, pages 1–13, 2024. 2, 3

[26] Aravindan Sundaram, Ujjayan Pal, Abhimanyu Chauhan, Aishwarya Agarwal, and Srikrishna Karanam. Cocono: Attention contrast-and-complete for initial noise optimization in text-to-image synthesis. *arXiv preprint arXiv:2411.16783*, 2024. 3

[27] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023. 7

[28] Andrey Voynov, Kfir Aberman, and Daniel Cohen-Or. Sketch-guided text-to-image diffusion models. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 2, 3

[29] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2096–2105, 2023. 3, 6

[30] Zhenbei Wu, Haoge Deng, Qiang Wang, Di Kong, Jie Yang, and Yonggang Qi. Sketchscene: Scene sketch to image generation with diffusion models. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 2087–2092. IEEE, 2023. 2

[31] Xin Xie and Dong Gong. Dymo: Training-free diffusion model alignment with dynamic multi-objective scheduling. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13220–13230, 2025. 3

[32] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023. 3

[33] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2, 3, 7

[34] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Scaling in-the-wild training for diffusion-based illumination harmonization and editing by imposing consistent light transport. In *The Thirteenth International Conference on Learning Representations*, 2025. 3

[35] Tianyu Zhang and Haoran Xie. Sketch-guided text-to-image generation with spatial control. In *2024 2nd International Conference on Computer Graphics and Image Processing (CGIP)*, pages 153–159. IEEE, 2024. 2, 3

[36] Tianyu Zhang, Xiaoxuan Xie, Xusheng Du, and Haoran Xie. Sketch-guided scene image generation. *arXiv preprint arXiv:2407.06469*, 2024. 2, 3

[37] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36:11127–11150, 2023. 2, 3, 7

[38] Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22490–22499, 2023. 3