λ-*ECLIPSE*: Multi-Concept Personalized Text-to-Image Diffusion Models by Leveraging CLIP Latent Space

Maitreya Patel*†, Sangmin Jung*, Chitta Baral, Yezhou Yang
Arizona State University

https://eclipse-t2i.github.io/Lambda-ECLIPSE/

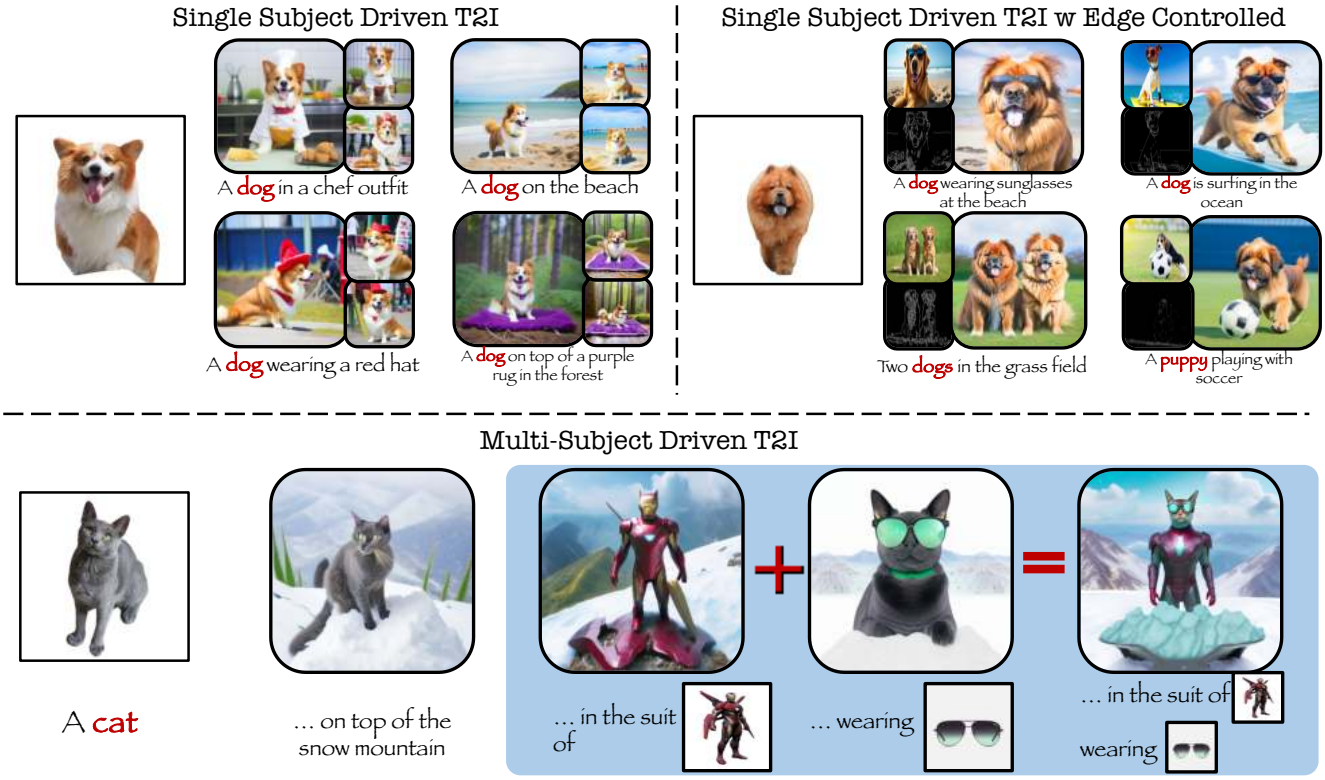

Figure 1. Our λ-*ECLIPSE* can estimate subject-specific image embeddings while maintaining the balance between concept and composition alignment. λ-*ECLIPSE* supports multi-concept personalization and provides additional controls (i.e., edge map) for image generation. In this work, we show that it is possible to perform subject-driven T2I in the latent space of a pre-trained CLIP model without depending on the diffusion UNet models; effectively reducing the heavy resource requirements.

Abstract

Despite the recent advances in personalized text-to-image (P-T2I) generative models, subject-driven T2I remains challenging. The primary bottlenecks include 1) Intensive training resource requirements, 2) Hyper-parameter sensitivity leading to inconsistent outputs, and 3) Balancing the intricacies of novel visual concept and composition alignment. We start by re-iterating the core philosophy of T2I diffusion models to address the above limitations. Predominantly, contemporary subject-driven T2I approaches hinge on Latent Diffusion Models (LDMs), which facilitate T2I mapping through cross-attention layers. While LDMs offer distinct advantages, P-T2I methods' reliance on the latent space of these diffusion models significantly escalates resource demands, leading to inconsistent results and necessitating numerous iterations for a single desired image.

* indicates equal contribution, † Corresponding Author: maitreya.patel@asu.edu.

Recently, ECLIPSE has demonstrated a more resource-efficient pathway for training UnCLIP-based T2I models, circumventing the need for diffusion text-to-image priors. Building on this, we introduce λ-ECLIPSE[1]*. Our method illustrates that effective P-T2I does not necessarily depend on the latent space of diffusion models. λ-ECLIPSE achieves single, multi-subject, and edge-guided T2I personalization with just 34M parameters and is trained on a mere 74 GPU hours using 1.6M image-text interleaved data. Through extensive experiments, we also establish that λ-ECLIPSE surpasses existing baselines in composition alignment while preserving concept alignment performance, even with significantly lower resource utilization.*

Table 1. **A quick overview of previous works on T2I personalization.** Our method is the first to offer multi-concept personalization without depending on diffusion UNet models (except for inference). We provide the extended overview table in the supplementary materials.

Method	Multi Concepts	Finetuning Free	Diffusion Free	Total opt. params	GPU Hours
Textual Inversion [12]	✗	✗	✗	768	1
DreamBooth [38]	✗	✗	✗	0.9B	0.5
ELITE [48]	✗	✓	✗	77M	-
BLIP-Diffusion [21]	✗	✓	✗	1.5B	2304
IP-Adapter [50]	✗	✓	✗	22M	-
Custom Diffusion [20]	✓	✗	✗	57M	0.2
Subject-Diffusion [27]	✓	✓	✗	252M	-
Kosmos-G [29]	✓	✓	✗	1.9B	12300
λ-ECLIPSE (ours)	✓	✓	✓	34M	74

1. Introduction

The field of text-to-image (T2I) diffusion models has recently witnessed remarkable advancements, achieving greater photorealism and enhanced adherence to textual prompts. This has catalyzed the emergence of diverse applications, notably in controllable T2I models and personalized T2I (P-T2I) models. P-T2I, in particular, encompasses the intricate task of learning and reproducing novel visual concepts or subjects in varied contexts, beyond mere replication of reference images. This personalization task is challenging, necessitating a nuanced balance between concept and compositional alignment. The complexity escalates further when multi-subject personalization is desired.

Early works in this field, such as Textual Inversion [12] and DreamBooth [38], employed per-concept optimization strategies involving fine-tuning certain parameters within T2I diffusion models. However, these methods struggle with generalization challenges and are time-intensive. Contemporary research is pivoting towards fast personalization techniques that involve training hypernetworks and integrating new layers or parameters within pre-trained diffusion UNet models. As summarized in Table 1, despite their relative parameter efficiency, these methods are still resource-intensive (especially for multi-concept personalization), largely due to their reliance on the latent spaces of diffusion models. Additionally, controlling the inherent randomness in diffusion models remains a significant challenge, often resulting in inconsistent outputs and necessitating multiple attempts to generate a single accurate image.

Upon further investigation, we find that most subject-driven T2I approaches build upon variants of the Latent Diffusion Model (LDM) [37], specifically Stable Diffusion models. LDMs employ cross-attention layers to condition diffusion models with text embeddings, necessitating a mapping of target subject images to latent spaces compatible with the diffusion models, either as text embeddings or as supplementary model parameters. This process involves backpropagation through the entire diffusion model, often comprising over a billion parameters, contributing to the inefficiency of existing P-T2I methods.

In contrast, the UnCLIP [35] T2I family of LDMs offers an intriguing alternative, wherein text embeddings are mapped to vision embeddings through a diffusion transformer prior network before passing the conditioning to the diffusion UNet image generator. This mechanism, as illustrated in Figure 2, enables the generator to accurately reproduce original images from their visual embeddings alone. Consequently, our research pivots towards leveraging the UnCLIP T2I stack for P-T2I strategies. However, like their predecessors, these T2I diffusion priors are also parameter-heavy, with around 1 billion parameters. Until recently, *ECLIPSE* [31] showed that it is possible to compress this 1B parameter of the T2I prior model down to 33M by removing the diffusion modeling.

ECLIPSE posits that text-to-image mapping can be optimized through contrastive pre-training. It inputs text embeddings and estimates corresponding image embeddings, ensuring strong alignment with textual features. Building on these insights, to enhance this framework and deepen the comprehension of novel visual subjects, we introduce a novel pre-training strategy and propose the *λ-ECLIPSE* T2I prior model for the UnCLIP framework. Specifically, we propose a subject-driven instruction tuning task involving the image-text interleaved data. This involves creating 1.6 million high-quality image-text pairs, where text tokens linked to subjects are substituted with image embeddings. Subsequently, *λ-ECLIPSE* is trained to estimate image embeddings that not only harmonize with text semantics but also encapsulate subject representations.

This model facilitates multi-subject-driven image generation using UnCLIP models without imposing excessive re-

[1]The designation *λ-ECLIPSE* is inspired by its conceptual alignment with the λ-calculus. In this context, the *λ-ECLIPSE* model functions similarly to a functional abstraction within λ-calculus, where it effectively binds variables. These variables, in our case, represent novel visual concepts that are integrated through composition prompts.

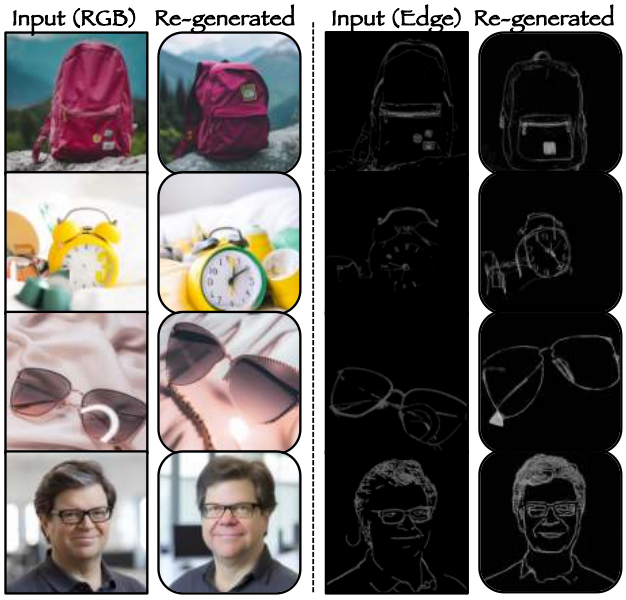

Figure 2. **CLIP(vision) features capture the semantics and fine-grained visual details.** UnCLIP models (e.g., Kandinsky v2.2) can effectively leverage this latent space to reconstruct the image from its embedding.

source demands. λ-*ECLIPSE* bypasses the reliance on diffusion latent space and operates within a pre-trained CLIP latent space [33]. We leverage Kandinsky v2.2 diffusion UNet model's [36] capacity to interpret image structures from CLIP embeddings (refer to Figure 2). Importantly, CLIP image embeddings even accurately capture the details of the canny edge map [6]. Therefore, we extend λ-*ECLIPSE* to incorporate canny edge maps as auxiliary guides for subject-driven T2I. Our extensive experimental validation demonstrates that λ-*ECLIPSE* achieves unparalleled performance in single and multi-subject composition alignment while preserving the integrity of reference subjects, competitively rivaling Multimodal Large Language Models (MLLM) based methods that require intense computing resources. At last, we find that λ-*ECLIPSE* inherits the continuous latent space from the CLIP; for the first time allowing the interpolations between multiple concepts for P-T2I models. Figure 1 illustrates the qualitative capabilities of λ-*ECLIPSE*.

In summary, our contributions are: (1) we introduce a straightforward and yet very effective diffusion-free multi-subject personalization strategy by estimating subject-aligned image embeddings in CLIP(vision) latent space, (2) despite significantly reducing resource requirements, our method achieves state-of-the-art performance, and (3) we demonstrate the potential of utilizing diffusion image generators as standalone pre-trained entities, enabling controlled outputs without necessitating explicit fine-tuning.

2. Related Works

Text-to-Image Generative Models. Pioneering efforts in image generation, notably DALL-E [34] and CogView [11], leveraged autoregressive models to achieve significant results. Recent advancements predominantly feature diffusion models, acclaimed for their high image fidelity and diversity in text-to-image (T2I) generation. A notable example is Stable Diffusion, which builds upon the Latent Diffusion Model (LDM) [37] and excels in semantic and conceptual understanding by transitioning training to latent space. Imagen [40], Pixart-α [8], and DALL-E 3 [5] propose using a large T5 language model to improve language understanding. DALL-E 2 [35] along with its UnCLIP variation models such as Kandinsky [36] and Karlo, uses a diffusion prior and diffusion UNet modules to generate images using the pre-trained CLIP [33] model.

Personalized T2I Methods. Approaches like Textual Inversion [12], DreamBooth [38], and Custom Diffusion [20] focus on training specific parameters to encapsulate visual concepts. LoRA [16] and Perfusion [45] target efficient fine-tuning adjustments, particularly rank 1 modifications. However, these methods are constrained by their requirement for concept-specific tuning. ELITE [48] was the first approach addressing fast personalization for single-subject T2I. BLIP-Diffusion [21] adapts the BLIP2 encoder [22], training approximately 1.5B parameters to enable zero-shot, subject-driven image generation. IP-Adapter [50] introduces a decoupled cross-attention mechanism, negating the need to train the foundational UNet model by permitting fine-tuning of a reduced parameter set (22M).

Mix-of-Show [14] and Zip-LoRA [42] train individual concepts and then combine them to generate multiple subjects. Break-A-Scene [4] shows multi-concept capability but requires single images containing diverse objects. Subject Diffusion [27] creates a high-quality dataset and presents the precision control for fast multi-subject image generation. Kosmos-G [29], akin to Subject-Diffusion, employs a Multimodal Large Language Model (MLLM) for text-image embedding alignment, though it necessitates extensive parameter optimization (1.9B). These multi-subject P-T2I methods are not only demanding in terms of parameters but also depend on a massive number of frozen parameters of the diffusion UNet model, increasing training computational loads. In contrast, our model, λ-*ECLIPSE*, forgoes test-time fine-tuning and training-time reliance on the diffusion UNet model for single and multi-concept, control-guided P-T2I, positioning it as a resource-efficient solution.

At last, methods like GLIGEN [23], ControlNet [51], and UniControl [32] incorporate additional controls (i.e., edge map, depth, segmentations) into the diffusion model to generate the desired images. BLIP-Diffusion, IP-Adapter,

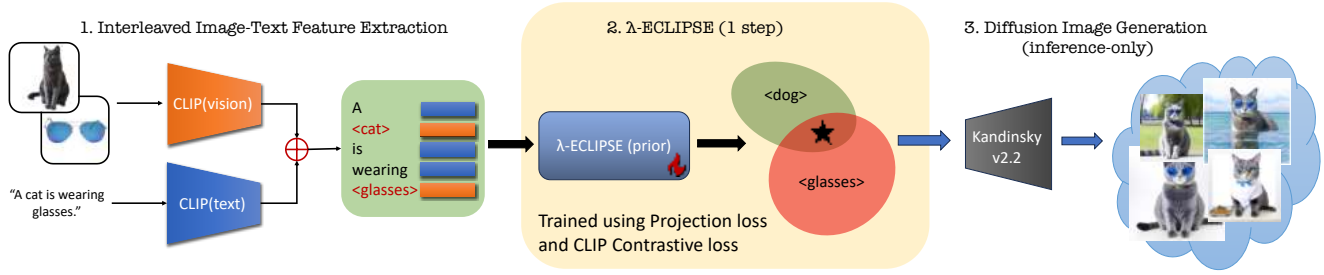

Figure 3. This figure illustrates the three stages of the λ-*ECLIPSE* pipeline. 1) Create the image-text interleaved features by utilizing the pre-trained CLIP encoders. 2) Pre-train the λ-*ECLIPSE* (34M parameters) using Eq. 1, which ensures the mapping to the desired latent space given the image-text interleaved data. During inference, we just need a single-forward pass to estimate the corresponding image embedding. 3) During inference, the Kandinsky v2.2 diffusion UNet model takes the output from the λ-*ECLIPSE* (prior) and generates the corresponding images.

and Kosmos-G can leverage such pre-trained controls. However, in many scenarios, these controls are too strong which makes generated images lose subject-specific details. While λ-*ECLIPSE* can be plugged with these pre-trained methods, we explore the possibility of learning canny edge as the additional guidance for P-T2I. We show that λ-*ECLIPSE* learns to balance the edge map, subjects, and text composition.

3. Method

In this section, we introduce λ-*ECLIPSE*, our approach to multi-subject text-to-image personalization. This method extends the capabilities of the existing *ECLIPSE* framework through a novel image-text interleaved pretraining strategy, notably omitting the need for explicit diffusion modeling. Our approach mainly capitalizes on the efficient utilization of the CLIP latent space. Figure 3 outlines the end-to-end framework.

The primary objective of λ-*ECLIPSE* is to facilitate both single and multi-subject T2I generation processes, accommodating controlled conditional guidance such as edge maps. Initially, we detail the problem formulation and elaborate on the extended design of the *ECLIPSE* T2I prior module. Subsequently, we delve into the image-text interleaved training methodology. This fine-tuning process enables the *ECLIPSE* model to harness semantic correlations between CLIP image and text latent spaces while preserving subject-specific visual features. Upon attaining the target embedding, the diffusion UNet model, derived from Kandinsky v2.2, is employed to reconstruct the image.

3.1. Text-to-Image Prior Mapping

In the UnCLIP T2I models, the objective of the text-to-image prior model (f_θ) is to establish a proficient text-to-image embedding mapping. This model is designed to adeptly map textual representations to their corresponding visual embeddings, denoted as ($f_\theta : z_y \rightarrow z_x$), where $z_{x/y}$ represent the embeddings for images and text, respectively. The visual embedding predictions ($\hat{z_x} = f_\theta(z_y)$) are then effectively utilized by the diffusion image generators (h_ϕ), which are inherently conditioned on these vision embeddings ($h_\phi : z_x \rightarrow x$).

As illustrated in Figure 2, these diffusion image generators possess a remarkable capability for precise image reproduction, since z_x encompasses comprehensive image information. Our goal is to accurately estimate the image embedding $\hat{z}_x$, incorporating the subject representations, thereby eliminating reliance on h_ϕ during training. Existing LDM-based P-T2I methods are limited by the LDM's singular module approach ($h_\phi : z_y \rightarrow x$). Consequently, mastering the latent space of h_ϕ becomes essential for effective P-T2I for the baseline methodologies.

We propose a mapping function, f'_θ, which processes text representations (z_y) alongside subject (x_i) specific representations (z_{x_i}), to derive an image embedding that encapsulates both text prompts and subject visuals ($\hat{z}_x$). The challenge lies in harmonizing z_{x_i} and z_y within $f'_\theta : (z_y, z_{x_i}) \rightarrow \hat{z}_x$, ensuring alignment while preventing overemphasis on either aspect, as this could compromise composition alignment. To address this, we employ the contrastive pretraining strategy after [31]:

$$\mathcal{L}_{prior} = \mathop{\mathbb{E}}_{\substack{\epsilon \sim \mathcal{N}(0,I) \\ z_y, \hat{z}_x}} \Big[||z_x - f_\theta(\epsilon, z_y)||_2^2 \Big] - \frac{\lambda}{N} \sum_{i=0}^{N} \log \frac{\exp(\langle \hat{z}_x^i, z_y^i \rangle / \tau)}{\sum_{j \in [N]} \exp(\langle \hat{z}_x^i, z_y^j \rangle / \tau)}. \tag{1}$$

Here, λ serves as the hyperparameter. The first loss term measures the mean-squared error between the estimated and actual image embeddings, primarily ensuring concept alignment. However, our preliminary studies reveal that exclusive reliance on this term diminishes composition alignment, echoing observations made by [31]. Therefore, we

stick with the contrastive loss component to bolster compositional generalization, with λ balancing concept and composition alignment.

Additional Control-based T2I Prior Mapping. Acknowledging the limitations in existing methods, which necessitate learning the diffusion latent space even for additional control inputs, we endeavor to achieve a more nuanced balance between subject, text, and supplementary conditions. Consequently, we have augmented λ-*ECLIPSE* to accommodate an additional modality, a canny edge map, providing more refined control over subject-driven image generation. This entails modifying the prior model to accept additional conditions ($f'_\theta : (z_y, z_{x_i}, z_c) \rightarrow \hat{z}_x$, where z_c symbolizes the additional modality embedding). During training, we drop z_c for 1% to improve the unconditional generations.

Incorporating canny edge maps during training enhances stability and broadens the generalization capabilities of λ-*ECLIPSE*, yielding benefits even in the absence of these controls during inference. Our findings indicate that λ-*ECLIPSE* successfully learns a unified mapping function, effectively estimating target image representations from a synergy of text, subjects, and edge map.

3.2. Image-text Interleaved Training

Our approach targets developing a versatile prior model capable of processing diverse inputs to estimate visual outputs. Drawing from earlier methodologies, a straightforward solution involves concatenating different inputs, like combining text ("a dog wearing sunglasses") with respective concept-specific images. However, preliminary experiments indicated that this method does not effectively capture the intricate relationships between target text tokens (e.g., "dog") and the corresponding concept images.

To address this, we adopt the interleaved pre-training strategy used in Kosmos-G, but with a notable modification to enhance resource efficiency. We incorporate pretrained CLIP text and vision encoders for extracting modality-specific embeddings—separating text-only from subject-specific images. The key refinement in our process is the substitution of subject token embeddings with corresponding vision embeddings instead of introducing additional tokens to handle the image embeddings via resampler [2]. This alteration allows us to bypass the need to train the prior or MLLMs, significantly improving the model's proficiency in handling interleaved data.

For the generation of high-quality training datasets, we carefully selected 2 million high-quality images from the LAION dataset [41], each with a resolution of 1024x1024. Utilizing BLIP2, we generate captions for these images and employ SAM [19] for extracting noun or subject-specific segmentation masks. Given the CLIP model's requirement for 224x224 resolution images, we avoid resizing the masks within their original resolutions. Instead, we opt for cropping the area of interest using DINO [24], followed by resizing the masked object while preserving its aspect ratio. This technique is crucial in retaining maximum visual information for each subject during the training phase. We provide more details about the filters used in the supplementary material.

In summary, the *ECLIPSE* prior, trained with our image-text interleaved data and supplementary conditions, presents an efficient pathway for P-T2I. The culmination of this process yields the λ-*ECLIPSE* model, a refined version characterized by enhanced effectiveness in handling and interpreting multimodal data.

4. Experiments

In this section, we first introduce the experimental setup, evaluations, and baselines. Later, we delve into the qualitative and quantitative results.

Training and inference details. We initialize our model, λ-*ECLIPSE*, equipped with 34M parameters. We train our model on an image-text interleaved dataset of 2M instances, partitioned into 1.6M for training and 0.4M for validation. The model is specifically tuned for the Kandinsky v2.2 diffusion image decoder. Therefore, we use pre-trained OpenCLIP-ViT-G/14[2] as the text and vision encoders, ensuring alignment with Kandinsky v2.2 image embeddings. Training is executed on 2 x A100 GPUs, leveraging a per-GPU batch size of 512 and a peak learning rate of 0.00005, across approximately 100,000 iterations, summing up to 74 GPU hours. During inference, the model employs 50 DDIM steps and 7.5 classifier-free guidance for the Kandinsky v2.2 diffusion image generator. Adhering to baseline methodologies, we perform the P-T2I following the baseline papers' protocols. For λ-*ECLIPSE*, target subject pixel regions in reference images are segmented before embedding extraction via the CLIP(vision) encoder. We drop the canny edge map during inference unless specified explicitly.

Evaluation setup. We primarily utilize Dreambench (encompassing 30 unique concepts with 25 prompts per concept) for qualitative and quantitative evaluations using DINO and CLIP-based metrics [38]. Due to their limitations, we extend our evaluations on the ConceptBed [30] benchmark (covering 80 diverse imagenet concepts and a total of 33K composite prompts), where we report performance on concept replication, concept, and composition alignment using the Concept Confidence Deviation ($CCD(\downarrow)$) metric.

[2] https://huggingface.co/laion/CLIP-ViT-g-14-laion2B-s12B-b42K

Figure 4. Qualitative examples of λ-*ECLIPSE*.

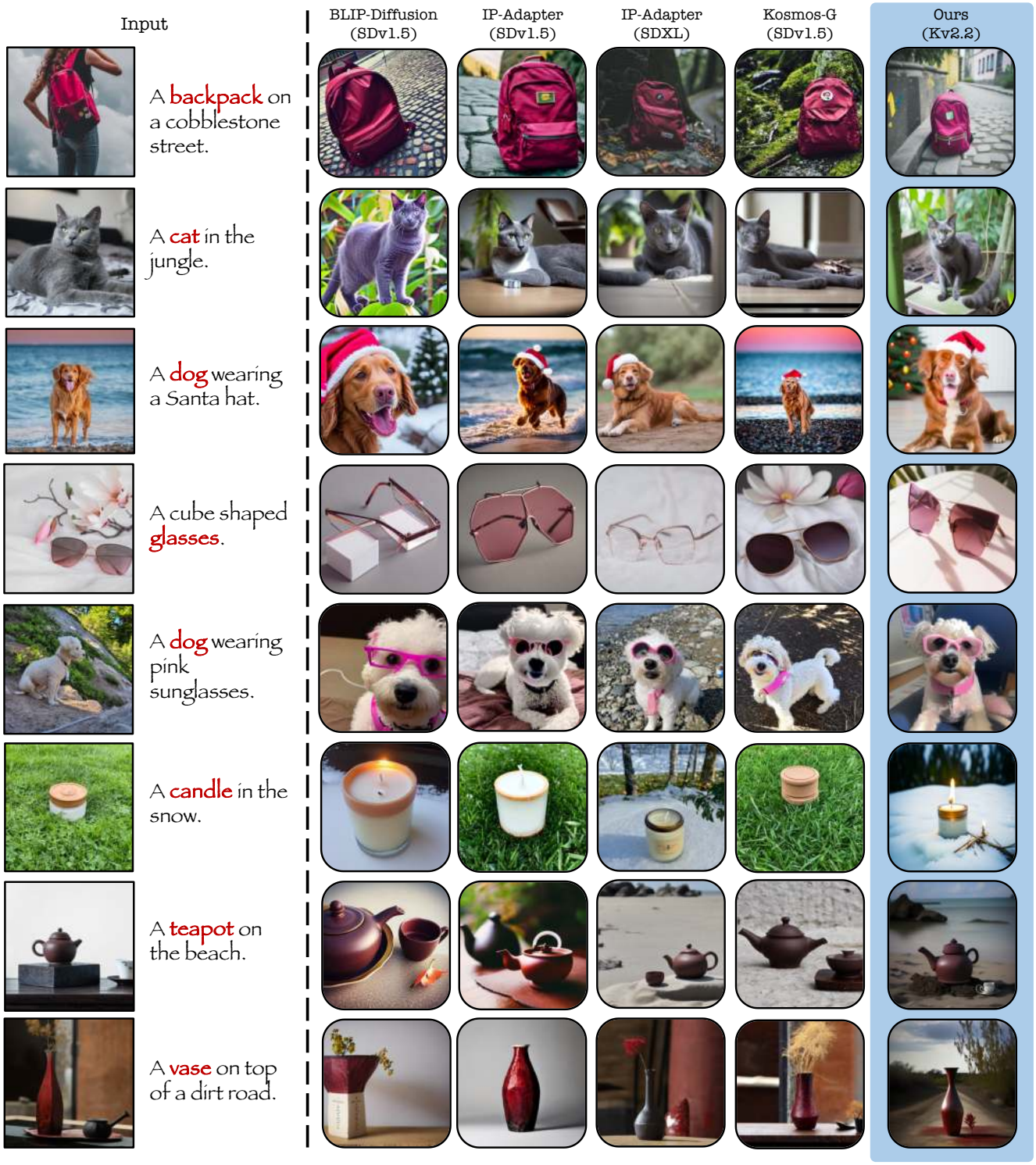

Figure 5. **This figure illustrates the qualitative comparison of λ-*ECLIPSE w.r.t.* the state-of-the-art approaches for single-subject T2I generations.** We choose concepts and prompts from the Dreambench dataset. For all methods, concepts, and prompts, we generate the four images and select the best one that represents the concept and composition more accurately.

Table 2. **Quantitative comparisons of different methodologies on Dreambench.** The **Bold** and underline represent the metric-specific first and second-ranked methods, respectively. * represents that we reproduce the open-source methodologies. Other results are borrowed from the Subject-Diffusion [27].

Method	Base Model	DINO (↑)	CLIP-I (↑)	CLIP-T (↑)
Textual Inversion	SDv1.5	0.569	0.780	0.255
DreamBooth	SDv1.5	**0.668**	**0.803**	**0.305**
Custom Diffusion	SDv1.5	0.643	0.790	**0.305**
Re Imagen	Imagen	0.600	0.740	0.270
ELITE	SDv1.4	0.621	0.771	0.293
Subject-Diffusion	SDv1.5	**0.711**	0.787	0.293
BLIP-Diffusion*	SDv1.5	0.603	0.793	0.291
IP-Adapter*	SDv1.5	0.629	**0.827**	0.264
IP-Adapter*	SDXL	0.613	0.810	0.292
Kosmos-G*	SDv1.5	0.618	0.822	0.250
λ-ECLIPSE*	Kv2.2	0.604	0.777	**0.307**

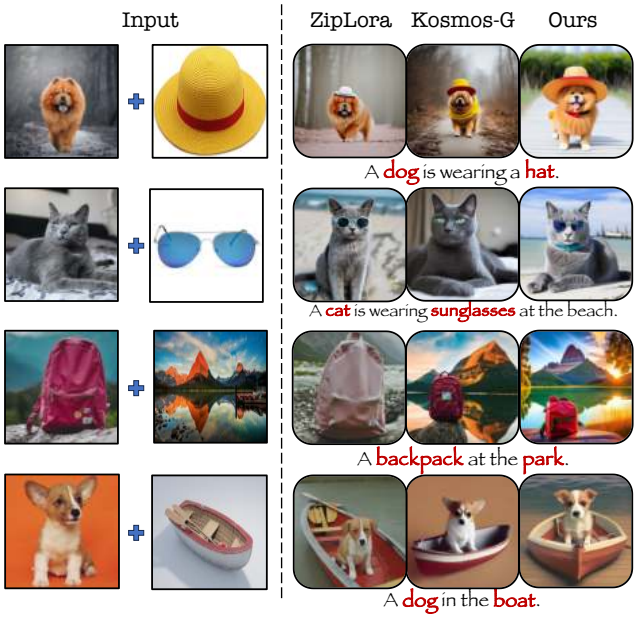

Figure 6. **Multi-subject qualitative examples** illustrating the comparison between *λ-ECLIPSE* and other SOTA baselines.

4.1. Results & Analysis

Quantitative comparison. The quantitative assessments are detailed in Table 2 and Table 3, focusing on the single-concept T2I task. For Dreambench, we generate and evaluate four images per prompt, reporting average performance on three metrics (DINO, CLIP-I, and CLIP-T). In the case of ConceptBed, we process each of the 33K prompts to generate a single image. The results, as depicted in these tables, highlight *λ-ECLIPSE*'s superior performance in composition tasks and its competitive edge over even fine-tuning-based methodologies, while maintaining robust concept alignment. *λ-ECLIPSE* exhibits a notable proficiency

Table 3. **Quantitative comparisons of different methodologies on ConceptBed.** We present results on CCD (↓) across three evaluation categories. The **Bold** and underline represent the metric-specific first and second-ranked methods, respectively. * represents that we reproduce the open-source methodologies. Other results are borrowed from the ConceptBed [30].

Method	Base Model	Concept Replication	Concept Alignment	Composition Alignment
Textual Inversion	SDv1.4	**0.0662**	**0.1163**	0.1436
Dreambooth	SDv1.4	0.0880	0.3551	0.0360
Custom Diffusion	SDv1.4	0.2309	0.4882	**0.0204**
ELITE*	SDv1.4	0.3195	0.4666	0.1832
BLIP-DIffusion*	SDv1.5	0.3510	**0.3245**	0.1589
IP-Adapter*	SDXL	0.3665	0.3571	0.0641
λ-ECLIPSE*	Kv2.2	**0.2853**	0.3619	**-0.0200**

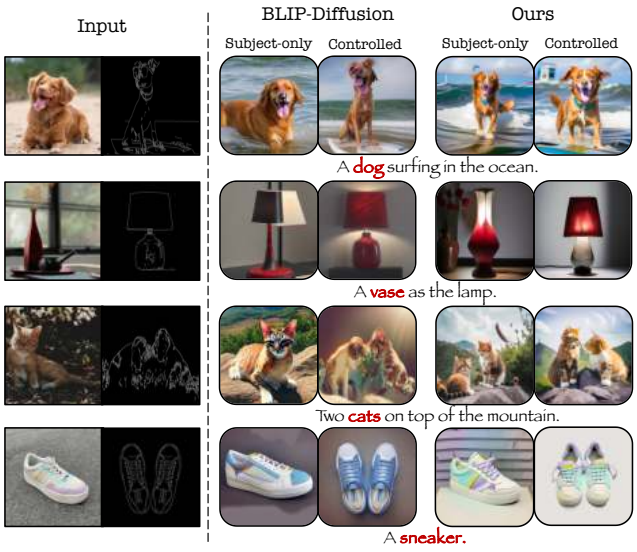

Figure 7. **Qualitative examples for edge-guided P-T2I** showcasing the comparison between BLIP-Diffusion *vs. λ-ECLIPSE*.

in concept replication, albeit with a marginal trade-off in concept alignment for enhanced composition fidelity. Comparatively, IP-Adapter, ELITE, and BLIP-Diffusion prioritize concept alignment, often at the expense of composition alignment. Noteworthy is *λ-ECLIPSE*'s efficiency, achieved with significantly fewer resources.

Qualitative comparisons. In Figure 5, we present a range of single subject-specific images generated by various methodologies including BLIP-Diffusion, IP-Adapter, Kosmos-G, and *λ-ECLIPSE*. *λ-ECLIPSE* demonstrates exemplary proficiency in composition while ensuring concept alignment. In contrast, the baselines often overemphasize reference images (notably in rows 2 and 4) or exhibit concept dilution (as seen in rows 6 and 8), leading to higher concept alignment but compromised composition. Addi-

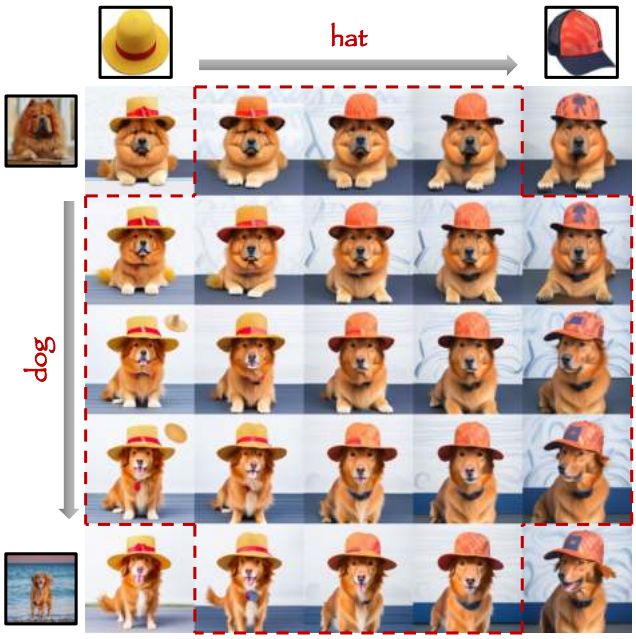

Figure 8. Interpolation between four concepts. Here, we estimate the image embedding using λ-*ECLIPSE* corresponding to each corner and then interpolate from top to bottom and left to right. At last, we use the Kandinsky v2.2 diffusion UNet model to generate the images with fixed random seeds from these sets of image embeddings.

tionally, Figure 6 exhibits λ-*ECLIPSE*'s multi-concept generation prowess, in comparison to ZipLoRA (fine-tuning-based approach) and Kosmos-G (Multimodal LLM-based approach), underscoring its capability to rival compute-intensive methodologies.

Figure 2 previously demonstrated Kandinsky v2.2's ability to partially reconstruct canny edge maps while preserving semantic integrity. Building on this, Figure 7 offers qualitative insights, comparing BLIP-Diffusion and λ-*ECLIPSE* in edge-guided subject-driven T2I scenarios. Here, BLIP-Diffusion adheres strictly to the imposed conditions, often at the cost of concept retention (rows 1, 3, and 4). Conversely, λ-*ECLIPSE* not only complies with these conditions but also adeptly balances the need for high concept and composition alignment.

Ablations. We extend our study to evaluate the individual contributions of different components in λ-*ECLIPSE*. Initially, the model's performance with solely the projection loss (referenced in Eq.1) is assessed. Subsequent experiments involve training λ-*ECLIPSE* variants with varying hyperparameters for the contrastive loss, specifically λ values of 0.2 and 0.5. A comparative analysis of these baselines is conducted against the fully equipped λ-*ECLIPSE* model, which incorporates $\mathcal{L}_{prior}$ (Eq.1) with

Table 4. **Ablation studies w.r.t.** to the key components of λ-*ECLIPSE* design. We report the concept and composition alignment for single-subject T2I using CCD ($\downarrow$) on the ConceptBed benchmark.

Model	Concept Alignment	Composition Alignment
Projection loss	0.394	0.008
w/ contrastive loss (lambda=0.5)	0.435	**-0.043**
w/ contrastive loss (lambda=0.2)	0.402	-0.026
w/ edge conditions (lambda=0.2)	**0.362**	-0.020

$\lambda = 0.2$ and utilizes canny edge maps during training. As Table 4 illustrates, relying solely on projection loss results in high concept alignment but compromises compositions. The contrastive loss variant with $\lambda = 0.5$ enhances composition alignment at the expense of concept alignment, whereas $\lambda = 0.2$ achieves a more balanced performance. Crucially, the integration of canny edge maps during training optimally balances both alignments and, specifically, improves the concept alignment. Here, the negative values indicate that the CCD oracle model performs similarly to real images having the same set of compositions.

Multi-subject interpolation. A key attribute of the CLIP latent space is the ability to perform smooth interpolation between two sets of embeddings. We conducted experiments to demonstrate λ-*ECLIPSE*'s ability to learn and replicate this smooth latent space transition. We selected two distinct dog breeds (<dog1>, <dog2>) and two types of hats (<hat1>, <hat2>) as the concepts. λ-*ECLIPSE* was then used to estimate image embeddings for all four possible combinations, each corresponding to the input phrase "a <dog> wearing a <hat>." Figure 8 showcases a gradual and seamless transition in the synthesized images from the top left to the bottom right. Unlike current diffusion models, which often exhibit sensitivity to input variations necessitating numerous iterations for desired outcomes, λ-*ECLIPSE* inherits CLIP's smooth latent space. This inheritance not only facilitates progressive changes in the conceptual domain but also extends the model's utility by enabling personalized **multi-subject interpolations**, a feature that is not inherently available in previous P-T2I methods.

5. Conclusion

In this paper, we have introduced a novel diffusion-free methodology for personalized text-to-image (T2I) applications, utilizing the latent space of the pre-trained CLIP model with high efficiency. Our λ-*ECLIPSE* model, trained on an image-text interleaved dataset of superior quality, marks a significant advancement in the field. This model

outperforms existing methods by its capability to execute single-concept, multi-concept, and edge-guided controlled P-T2I tasks using a singular model framework, while simultaneously minimizing resource utilization. Notably, λ-*ECLIPSE* sets a new benchmark in achieving state-of-the-art results in composition alignment, coupled with exceptional maintenance of subject fidelity. Furthermore, our research illuminates the potential of λ-*ECLIPSE* in exploring and leveraging the smooth latent space. This capability unlocks new avenues for interpolating between multiple concepts and their amalgamation, thereby generating entirely novel concepts. Our findings underscore a promising pathway where pre-trained diffusion image generators can be effectively controlled without necessitating extra supervision.

Acknowledgments

This work was supported by NSF RI grants #1750082, #2132724, and CPS grant #2038666. The views and opinions of the authors expressed herein do not necessarily state or reflect those of the funding agencies and employers.

References

[1] Yuval Alaluf, Elad Richardson, Gal Metzer, and Daniel Cohen-Or. A neural space-time representation for text-to-image personalization. *ACM Transactions on Graphics (TOG)*, 42(6):1–10, 2023. 14

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 5

[3] Moab Arar, Rinon Gal, Yuval Atzmon, Gal Chechik, Daniel Cohen-Or, Ariel Shamir, and Amit H. Bermano. Domain-agnostic tuning-encoder for fast personalization of text-to-image models. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–10, 2023. 14

[4] Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. *arXiv preprint arXiv:2305.16311*, 2023. 3, 14

[5] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2:3, 2023. 3

[6] John F. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8:679–698, 1986. 3

[7] Hong Chen, Yipeng Zhang, Xin Wang, Xuguang Duan, Yuwei Zhou, and Wenwu Zhu. Disenbooth: Disentangled parameter-efficient tuning for subject-driven text-to-image generation. *arXiv preprint arXiv:2305.03374*, 2023. 14

[8] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023. 3

[9] Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. Re-imagen: Retrieval-augmented text-to-image generator. In *The Eleventh International Conference on Learning Representations*, 2022. 14

[10] Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Rui, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. *arXiv preprint arXiv:2304.00186*, 2023. 14

[11] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34:19822–19835, 2021. 3

[12] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2022. 2, 3, 14

[13] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Encoder-based domain tuning for fast personalization of text-to-image models. *ACM Transactions on Graphics (TOG)*, 42(4):1–13, 2023. 14

[14] Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, et al. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. *arXiv preprint arXiv:2305.18292*, 2023. 3, 14

[15] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. *arXiv preprint arXiv:2303.11305*, 2023. 14

[16] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021. 3

[17] Xuhui Jia, Yang Zhao, Kelvin CK Chan, Yandong Li, Han Zhang, Boqing Gong, Tingbo Hou, Huisheng Wang, and Yu-Chuan Su. Taming encoder for zero fine-tuning image customization with text-to-image diffusion models. *arXiv preprint arXiv:2304.02642*, 2023. 14

[18] Changhoon Kim, Kyle Min, Maitreya Patel, Sheng Cheng, and Yezhou Yang. Wouaf: Weight modulation for user attribution and fingerprinting in text-to-image diffusion models. *arXiv preprint arXiv:2306.04744*, 2023. 16

[19] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 5, 13

[20] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023. 2, 3, 14

[21] Dongxu Li, Junnan Li, and Steven CH Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *arXiv preprint arXiv:2305.14720*, 2023. 2, 3, 14

[22] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 3

[23] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023. 3

[24] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 5, 13

[25] Zhiheng Liu, Ruili Feng, Kai Zhu, Yifei Zhang, Kecheng Zheng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones: Concept neurons in diffusion models for customized generation. *arXiv preprint arXiv:2303.05125*, 2023. 14

[26] Zhiheng Liu, Yifei Zhang, Yujun Shen, Kecheng Zheng, Kai Zhu, Ruili Feng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Customizable image synthesis with multiple subjects. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 14

[27] Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. *arXiv preprint arXiv:2307.11410*, 2023. 2, 3, 8, 13, 14

[28] Yiyang Ma, Huan Yang, Wenjing Wang, Jianlong Fu, and Jiaying Liu. Unified multi-modal latent diffusion for joint subject and text conditional image generation. *arXiv preprint arXiv:2303.09319*, 2023. 14

[29] Xichen Pan, Li Dong, Shaohan Huang, Zhiliang Peng, Wenhu Chen, and Furu Wei. Kosmos-g: Generating images in context with multimodal large language models. *arXiv preprint arXiv:2310.02992*, 2023. 2, 3, 14

[30] Maitreya Patel, Tejas Gokhale, Chitta Baral, and Yezhou Yang. Conceptbed: Evaluating concept learning abilities of text-to-image diffusion models. *arXiv preprint arXiv:2306.04695*, 2023. 5, 8

[31] Maitreya Patel, Changhoon Kim, Sheng Cheng, Chitta Baral, and Yezhou Yang. Eclipse: A resource-efficient text-to-image prior for image generations. *arXiv preprint arXiv:2312.04655*, 2023. 2, 4

[32] Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, et al. Unicontrol: A unified diffusion model for controllable visual generation in the wild. *arXiv preprint arXiv:2305.11147*, 2023. 3

[33] 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. 3

[34] 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

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

[36] Anton Razzhigaev, Arseniy Shakhmatov, Anastasia Maltseva, Vladimir Arkhipkin, Igor Pavlov, Ilya Ryabov, Angelina Kuts, Alexander Panchenko, Andrey Kuznetsov, and Denis Dimitrov. Kandinsky: an improved text-to-image synthesis with image prior and latent diffusion. *arXiv preprint arXiv:2310.03502*, 2023. 3

[37] 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

[38] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 2, 3, 5, 14

[39] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. *arXiv preprint arXiv:2307.06949*, 2023. 14

[40] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 3

[41] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, 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. 5

[42] Viraj Shah, Nataniel Ruiz, Forrester Cole, Erika Lu, Svetlana Lazebnik, Yuanzhen Li, and Varun Jampani. Ziplora: Any subject in any style by effectively merging loras. *arXiv preprint arXiv:2311.13600*, 2023. 3, 14

[43] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. *arXiv preprint arXiv:2304.03411*, 2023. 14

[44] James Seale Smith, Yen-Chang Hsu, Lingyu Zhang, Ting Hua, Zsolt Kira, Yilin Shen, and Hongxia Jin. Continual diffusion: Continual customization of text-to-image diffusion with c-lora. *arXiv preprint arXiv:2304.06027*, 2023. 14

[45] Yoad Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. Key-locked rank one editing for text-to-image personalization. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 3, 14

[46] 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. 14

[47] Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. $p+$: Extended textual conditioning in text-to-image generation. *arXiv preprint arXiv:2303.09522*, 2023. 14

[48] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15897–15907, 2023. 2, 3, 14

[49] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *arXiv preprint arXiv:2305.10431*, 2023. 14

[50] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 2, 3, 14

[51] 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. 3

[52] Yuxuan Zhang, Jiaming Liu, Yiren Song, Rui Wang, Hao Tang, Jinpeng Yu, Huaxia Li, Xu Tang, Yao Hu, Han Pan, et al. Ssr-encoder: Encoding selective subject representation for subject-driven generation. *arXiv preprint arXiv:2312.16272*, 2023. 14

[53] Ruoyu Zhao, Mingrui Zhu, Shiyin Dong, Nannan Wang, and Xinbo Gao. Catversion: Concatenating embeddings for diffusion-based text-to-image personalization. *arXiv preprint arXiv:2311.14631*, 2023. 14

A. Preliminaries for T2I Diffusion Models

As evidenced in numerous contemporary studies regarding T2I models, Stable Diffusion (SD) has emerged as a predominant backbone for text-to-image (T2I) models. SD involves training diffusion models in latent space, reversing a forward diffusion process that introduces noise into the image. A notable feature of SD is its integration of cross-attention, facilitating various conditions like text input. Operating in VQVAE latent space, SD not only achieves exceptional generative performance surpassing that in pixel space but also significantly reduces computational demands.

UnCLIP models (such as DALL-E 2) are very similar to the Stable Diffusion. In contrast, the UnCLIP stack takes the modular approach. UnCLIP first trains the diffusion text-to-image to the image prior (f_θ) to estimate the image embeddings (z_x) from the text embeddings (z_y). In parallel, a UNet-like diffusion image generator (h_ϕ) is trained to generate images (x) conditioned on ground truth vision embeddings (z_x).

Traditionally, T2I prior is modeled to estimate x_0-prediction instead of ϵ-prediction. Given the forward function $z_x^{(t)} \sim q(t, z_x)$, the goal of f_θ is to directly estimate z_x for all timesteps $t \sim [0, T]$ as:

$$\mathcal{L}_{prior} = \mathop{\mathbb{E}}_{\substack{t\sim[0,T],\\ z_x^{(t)}\sim q(t,z_x)}} \Big[||z_x - f_\theta(z_x^{(t)}, t, z_y)||_2^2\Big]. \quad (2)$$

ECLIPSE presents the contrastive learning strategy (Eq. 1) instead of minimizing Eq. 2. The diffusion image generator is trained by following standard ϵ-prediction formulation. Here, h_ϕ will estimate the ground truth added Gaussian noise $\epsilon \sim N(0, I)$, given the noise image $X^{(t)}$ for all timesteps $t \sim [0, T]$ and input conditions (such as z_x, z_y).

$$\mathcal{L}_{decoder} = \mathop{\mathbb{E}}_{\substack{\epsilon\sim N(0,I)\\ t\sim[0,T],\\ (z_x,\, z_y)}} \Big[||\epsilon - h_\phi(x^{(t)}, t, z_x, z_y)||_2^2\Big]. \quad (3)$$

For models like Kandinsky v2.2, we drop the z_y to condition the model on z_x. Therefore, λ-*ECLIPSE* also only conditions the image generation with z_y in the prior stage.

B. Image-Text Interleaved Training Details

Dataset Creation In constructing the dataset, we adhered to the data processing pipeline of Subject Diffusion [27]. We utilized the LAION-5B High-Res dataset, requiring a minimum image size of 1024x1024 resolution. Original captions were replaced with new captions generated by BLIP-2 (flan-t5-xl)[3]. Subjects were extracted using Spacy[4]. For each subject, we identified bounding boxes employing Grounding DINO [24], setting both box-threshold and text-threshold values to 0.2. We retained images with 1 to 8 detected bounding boxes, discarding the rest. Additionally, captions with multiple instances of identical objects were filtered, allowing a maximum of 6 identical objects. Following bounding box detection, individual subject masks were isolated using Segment-Anything (SAM) [19]. To enhance the efficiency of the training process, we pre-processed the dataset by pre-extracting features from CLIP vision and text encoders. During this phase, images predominantly featuring a background (white portion) exceeding 10% of the total area were excluded. We preserved bounding boxes with a width-height ratio ranging from 0.08 to 0.7 and logit scores of at least 0.3. Masks constituting less than 40% of the bounding box area were discarded. For the selection of subjects in images, we constrained the range to 1-4 subjects per image, excluding those with more than 4 subjects. At last, the interleaved image-text examples with respective ground truth images are shown in Figure 9.

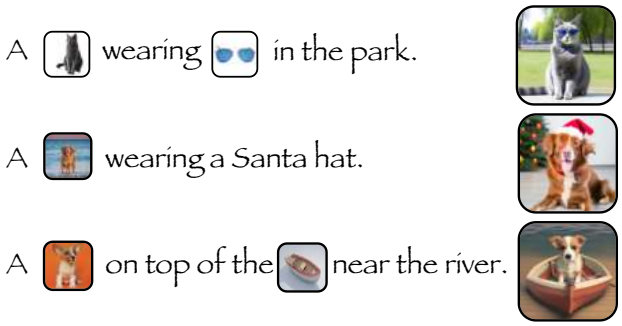

Figure 9. Examples of image-text interleaved training data. Left column shows the input the prior model and right images shows the ground truth corresponding images. Note: these examples are generated from λ-*ECLIPSE* for better interpretability.

Dataset Statistics In the final analysis, our dataset comprised a total of 1,990,123 images. The distribution of subjects per image exhibited a range from 1 to 4, with the following breakdown: 1,479,785 images featuring one subject, 432,831 images with two subjects, 65,597 images containing three subjects, and 11,910 images showcasing four subjects. The overall count of unique subjects acquired from this dataset amounted to 30,358. We partitioned our dataset into an 80:20 split between training and validation, reserving the remaining 1.6 million images for training and the rest for validation.

[3] https://huggingface.co/Salesforce/blip2-flan-t5-xl

[4] https://spacy.io

Table 5. **The detailed overview of subject-driven text-to-image generative methodologies.** * represents the backbone base models listed are subject to potential updates or modifications.

Method	Multi-Concept	Finetuning Free	Base Model	Image Input
Re-Imagen [9]	✗	✓	Imagen	Single
Textual Inversion [12]	✗	✗	SDv1.4	Multiple
DreamBooth [38]	✗	✗	SDv1.4	Multiple
Custom Diffusion [20]	✓	✗	SDv1.4	Multiple
ELITE [48]	✗	✓	SDv1.4	Single
E4T [13]	✗	✗	SD	Single
Cones [25]	✓	✗	SDv1.4	Single
SVDiff [15]	✓	✗	SD	Multiple
UMM-Diffusion [28]	✗	✓	SDv1.5	Single
XTI [47]	✗	✗	SDv1.4	Multiple
Continual Diffusion [44]	✓	✗	-	Multiple
InstantBooth [43]	✗	✓	SDv1.4	Multiple
SuTi [10]	✗	✓	Imagen	Multiple
Taming [17]	✗	✓	Imagen	Single
BLIP-Diffusion [21]	✗	✓	SDv1.5	Single
Cones 2 [26]	✓	✗	SDv2.1	Single
DisenBooth [7]	✗	✗	SDv2.1	Single
FastComposer [49]	✓	✓	SDv1.5	Single
Perfusion [45]	✓	✗	SDv1.5	Multiple
Mix-of-Show [14]	✓	✗	Chilloutmix	Multiple
NeTI [1]	✗	✗	SDv1.4	Mulitple
Break-A-Scene [4]	✓	✗	SDv2.1	Single*
ViCo [46]	✗	✗	SDv1.4	Mulitple
Domain-Agnostic [3]	✗	✗	-	Single
Subject-Diffusion [27]	✓	✓	SDv2	Single
HyperDreamBooth [39]	✗	✗	SDv1.5	Single
IP-Adapter [50]	✗	✓	SDv1.5	Single
Kosmos-G [29]	✓	✓	SDv1.5	Single
Zip-LoRA [42]	✓	✗	sdxl	Multiple
CatVersion [53]	✗	✗	SDv1.5	Multiple
SSR-Encoder [52]	✓	✓	SDv1.5	Single
λ-*ECLIPSE* (ours)	✓	✓	Kv2.2	Single

C. Implementation Details

The λ-*ECLIPSE* transformer prior architecture is significantly more compact compared to other Text-to-Image (T2I) methodologies. Our model employs a configuration of 16 Attention Heads, each with a dimension size of 32, alongside a total of 10 layers. Additionally, the embedding dimension size for our model is set at 1280, supplemented by 4 auxiliary embeddings (including, one for canny edge map). As λ-*ECLIPSE* is not a diffusion prior model, we do not keep time embedding layers. Overall, the λ-*ECLIPSE* model comprises approximately 34 million parameters, establishing it as a streamlined yet effective solution for Personalaized-T2I. Notably, the standard UnCLIP T2I priors contain 1 billion parameters.

D. Extended P-T2I Baselines Comparison

We further expand our comparative analysis of P-T2I methods encompassing a total of 32 approaches including ours and parallel works. Table 5 summarizes them into four crucial aspects: 1) multi-concept support, 2) fine-tuning free, 3) base model types, and 4) the required number of input images. To summarize, λ-*ECLIPSE* is the only methodology built on top of the UnCLIP models while supporting multi-concept personalization, fine-tuning free, and only requires a single reference concept image. We detail the comparison below:

Multi-Concept Generation. Multi-concept generation enables users to integrate multiple personal subjects to generate an image that follows the text prompts and aligns with all the concept visuals. Of the 32 methods analyzed, 14 offer this capability. Ten methods require separate training for

each concept and then an additional fusing step for combining the learned concepts. (Mix-of-Show). While only four methods support fast multi-concept personalization. Only a few can learn auxiliary guided information such as canny edge, depth maps, or open-pose and adapt style variation.

Fine-tuning Free (Fast Personalization). Many models require test-time fine-tuning. Each varies on which part alteration occurs, as early models tend to modify the whole UNet. In contrast, recent models tune a small portion of the cross-attention layers or introduce additional layers performing as adapters. In our analysis of P-T2I models, 13 out of 32 models employ a fine-tuning free approach which enables fast personalization.

Diffusion Independent. A majority of the reviewed models utilize diffusion models, with Stable Diffusion being the predominant choice, spanning versions 1.4, 1.5, 2.1, and XL. Few adapt Imagen (SuTi, Taming) and interestingly Mix-of-show employs ChillOutMix as their pretrained model, known for its adeptness at preserving realistic concepts like human faces. A unique outlier in this landscape is our λ-*ECLIPSE*, the only one that eschews the use of any diffusion prior model.

Easiness to Use (Input and Output). Ease of use, particularly regarding model input and output, significantly impacts user convenience in subject-driven generalization. A more user-friendly model typically requires a single input image, as opposed to multiple images of the same subject. In our study, 18 models offer P-T2I capabilities with just one input image. In contrast, others often require 4 to 5 images of the subject. Additionally, some models necessitate storage space for learned concepts, ranging from a few hundred kilobytes (e.g., Perfusion, HyperDreamBooth) to several megabytes (e.g., Zip-LoRA). Our model stands out by eliminating the need for individual concept pre-learning or storing any artifacts for P-T2I utilization, offering a streamlined, efficient user experience.

E. More Qualitative Results

In this section, we present an array of extended examples of P-T2I generation, which presents complex compositions that are hard to achieve in λ-*ECLIPSE* as well as other models. Figure 10 showcases the complexity level that notably escalates as it goes from top to bottom. As the complexity of the visual concepts increases, we observe that all methodologies perform poorly in maintaining the subject-fidelity. This applies to λ-*ECLIPSE* as well. Interestingly, λ-*ECLIPSE* retains the compositional understanding while baselines perform poorly in all scenarios.

In addition, we demonstrate some examples of how P-T2I methods exhibit inconsistency in their generated results depending on each trial. Figure 11 exhibits consistency in single and multiple concept generation across both models. However, Kosmos-G exhibits inconsistency in generating multiple concepts, sometimes placing the Ironman suit in various parts of the dog or omitting it entirely. This also means that λ-*ECLIPSE* reduces the diversity of the generated images to improve the consistency. However, this behavior is observed across the UnCLIP family of models.

In our final exploration, we examine the preservation of actual human facial characteristics in target scenarios, combined with varying captions. While each model attempts to retain the original face, none can precisely replicate the exact personal facial features. These examples generally fail to accurately represent the intended compositions, except in one instance in IP-Adapter FaceID.

F. Limitations

Our work represents a pioneering effort in harnessing a method that utilizes the latent space of pre-trained CLIP models for P-T2I generation, it is important to acknowledge certain limitations. Firstly, CLIP, despite its capabilities, does not perfectly capture hierarchical representations, and this limitation can occasionally result in suboptimal outcomes. Consequently, the CLIP contrastive loss can cause the model to deviate from the original subject features, making performing P-T2I on complex concepts, such as human faces, more challenging. We anticipate that the enhancement of CLIP representations will correspondingly improve the performance of our framework in the realm of P-T2I mapping. The λ-*ECLIPSE* model has been trained with 34 million parameters and 1.6 million images. However, increasing the data quality and parameters might lead to improvements in the results.

G. Broader Impact

The recent retraction of the LAION-5B dataset, prompted by concerns over the presence of data samples potentially breaching ethical guidelines, has significantly affected various research initiatives. This situation underscores the crucial importance of careful dataset selection and raises a cautionary note for users about the ethical dimensions involved. In line with this, the community needs caution in the application of associated methodologies as well. Subject-driven image generation or Personalized Text-to-Image (P-T2I) methods have the potential to be a transformative tool in numerous domains.

For their positive influence, they enable users to effortlessly generate, modify, and synthesize original subjects into diverse environments, thereby enriching creative expression. On the other hand, the ease of altering and cre-

ating images raises concerns about the responsible use of this technology which requires significant ethical and legal considerations. We recommend developers provide a more secure way (such as image attribution) for end-users to ensure accountability for misuse of such models [18]. Users must be acutely aware of being able to infringe intellectual property rights and create misleading or harmful content. As such, those employing subject-driven image-generation techniques should exercise careful judgment, ensuring that their work adheres to ethical standards and legal boundaries. It is imperative that the broader implications of this technology are considered, and that a commitment to responsible and conscientious use guides its application.

Figure 10. Qualitative examples of increasing complexity of novel visual concepts as we move from top to bottom.

Figure 11. Qualitative examples of showcasing the consistency comparisons between Kosmos-G and λ-*ECLIPSE*.

Figure 12. Qualitative examples of showcasing the failure cases on human faces on Kosmos-G, IP-Adapter (SDXL), IP-Adapter (FaceID), and λ-*ECLIPSE*.