

Imagine for Me: Creative Conceptual Blending of Real Images and Text via Blended Attention

Wonwoong Cho*

Yanxia Zhang†

Yan-Ying Chen†

David I. Inouye*



Figure 1: Visual and textual conceptual blending results of IT-Blender based on FLUX.1-dev.

*Elmore Family School of Electrical and Computer Engineering, Purdue University

†Toyota Research Institute

Abstract

Blending visual and textual concepts into a new visual concept is a unique and powerful trait of human beings that can fuel creativity. However, in practice, cross-modal conceptual blending for humans is prone to cognitive biases, like design fixation, which leads to local minima in the design space. In this paper, we propose a T2I diffusion adapter ‘‘IT-Blender’’ that can automate the blending process to enhance human creativity. Prior works related to cross-modal conceptual blending are limited in encoding a real image without loss of details or in disentangling the image and text inputs. To address these gaps, IT-Blender leverages pretrained diffusion models (SD and FLUX) to blend the latent representations of a clean reference image with those of the noisy generated image. Combined with our novel blended attention, IT-Blender encodes the real reference image without loss of details and blends the visual concept with the object specified by the text in a disentangled way. Our experiment results show that IT-Blender outperforms the baselines by a large margin in blending visual and textual concepts, shedding light on the new application of image generative models to augment human creativity. Our project website is: <https://imagineforme.github.io/>.

1 Introduction

“Conceptual integration is at the heart of imagination” — Fauconnier and Turner [2008]

Conceptual integration/blending [Fauconnier and Turner, 1998, 2008, Coulson, 2001] is a theory in Cognitive Science, which can describe the human’s cognitive process combining a visual and textual concepts into a new idea. It is one of the most essential virtues in the creative industries (e.g., product design, character design, fashion design, interior design, graphic design, art, and advertisement) because conceptual blending can provide inspirational and creative design ideas by creating new combinations or reinventing existing ones [Gabora, 2002].

Prior works Yang [2009], Hyun and Lee [2018], Cai et al. [2023] have shown that exploring the design concepts and space as much as possible can produce better design results especially during the early phase of the design process (e.g., Conceptual Design [Otto, 2003] and the SCAMPER method [Eberle, 1996] in Concept Generation [Ulrich and Eppinger, 2016]).

However, there can be two challenges to perform cross-modal visual and textual conceptual blending in practice. First, human’s creativity easily got stuck in the suboptimal as shown in design fixation (i.e., a tendency of a designer to overly adhere to a limited set of solutions) [Jansson and Smith, 1991] and Einstellung effect (i.e., a cognitive bias from past experiences or familiar solutions to a problem, preventing them from exploring better alternatives) [Luchins, 1942].

Second, cross-modal conceptual blending itself is not a trivial task. It can be achieved by *selective projection* process determining what and where to integrate the given multiple concepts [Fauconnier and Turner, 1998]. It involves the laborious process for identifying features in each condition and comparing the semantic correspondence to find a way to meaningfully blend them together.

Recent significant advances of text-to-image (T2I) diffusion models [Rombach et al., 2022, Saharia et al., 2022] and their applications for adding an image condition led us to the question, “*Can pretrained diffusion models be used for cross-modal conceptual blending to augment creativity?*”

If so, it can be very useful by 1) providing numerous conceptual blending results to explore broader design possibilities and 2) automating the conceptual blending process to minimize the time required to manually illustrate all design ideas. For example, suppose that we want to come up with a creative product design for sneakers. Instead of struggling with imagining what to combine with and how to apply the selective projection, we can simply give a prompt like “a photo of sneakers, creative design.” and give a set of reference images with a target concept and appearance, e.g., a sport car image for “sleek” or any knitted items for “warm” and “cozy” (e.g., Fig. 1). We may also apply the same style to the multiple objects (e.g., bicycle and car) or add multiple visual concepts to the generated results. Even random reference images can be used to provide a serendipitous inspiration.

The question is how to perform selective projection in diffusion models, which must be done to achieve cross-modal conceptual blending. We think the key is the attention module [Vaswani et al.,

2017] (which is one of the most crucial components of modern diffusion models [Rombach et al., 2022, Black Forest Labs, 2024]) because its mechanism, comparing similarity and selectively applying the value, is conceptually close to the selective projection.

Earlier work, such as IP-Adapter [Ye et al., 2023] and BLIP-Diffusion [Li et al., 2023] proposed encoder-based methods to incorporate a reference image into text-guided generation with additional training. Although they show decent performance in blending visual and textual concepts with a fast inference time, their methods are limited in 1) disentangling textual and visual conditions and 2) preserving the detailed visual concept of the reference image due to the dependencies on the text cross-attention module and an external image encoder.

Meanwhile, RIVAL [Zhang et al., 2023a] and StyleAligned [Hertz et al., 2024] have shown the potential of the pretrained self-attention module of the T2I diffusion models in blending cross-modal concepts. Although they showed impressive performance in disentangling cross-modal concepts and applying detailed visual concepts from their own denoising chain to another, their performance is limited when a real reference image is conditioned due to the distribution shift of the inversion chain [Zhang et al., 2023a]. They also have a slower inference time than the encoder-based methods.

Filling the gap between both baseline approaches, we propose a novel image adapter “Image-and-Text Concept Blender” (IT-Blender) that can imagine for us by blending cross-modal concepts with fast inference time. IT-Blender learns to blend visual concepts from a real image without loss of details, in a disentangled manner from the textual concept (i.e., text determines semantics while a reference image determines visual concepts such as texture, material, color, and local shape).

Briefly, instead of using an external image encoder, we leverage the denoising network as an image encoder to maintain the details of visual concepts. As opposed to recent related literature without an external image encoder [Wu et al., 2025, Tan et al., 2024], our proposed method does not have any architectural dependency (i.e., applicable to both UNet-based [Rombach et al., 2022] and DiT-based diffusion models [Black Forest Labs, 2024]). We design a novel *Blended Attention* on top of the self-attention module, where detailed visual concepts can be preserved, and textual concepts are physically separated, encouraging disentanglement of textual and visual concepts. Blended Attention is trained to be specialized in finding a semantic correspondence between two latents; one from the real reference image and the other from the generated image.

Our baseline experiment results on disentanglement, concept preservation, and blending score (in Appendices) demonstrate that IT-Blender outperforms the baselines in cross-modal conceptual blending in both UNet-based (SD 1.5) and DiT-based (FLUX) architectures.

2 Related Works

In this section, we introduce previous studies related to visual-and-textual conceptual blending, based on diffusion models [Ho et al., 2020, Dhariwal and Nichol, 2021, Song et al.].

Applications for spatially aligned control. Prior works [Zhang et al., 2023b, Mou et al., 2024, Hertz et al., 2022, Tumanyan et al., 2023, Liu et al., 2024] have achieved impressive performance in spatially aligned control. However, their methods are mainly designed for local photo editing instructed by text, which is not suitable for our conceptual blending task to augment creativity.

Applications based on text cross-attention module related to conceptual blending. IP-Adapter [Ye et al., 2023], BLIP-Diffusion [Li et al., 2023], and ELITE [Wei et al., 2023] are closely related to conceptual blending task. They proposed an adapter based on text cross-attention module to encode and incorporate reference image information into the text-guided image generation process. Even though decently working for cross-modal conceptual blending, their methods are limited in two aspects. First, the encoder-based methods often fail in disentangling visual and textual concepts. This is because a reference image relies on the text cross-attention module, potentially entangling the cross-modal information. Second, encoder-based methods are limited in blending the detailed visual concepts because of a dependency on an external image encoder, where visual details can be lost.

Applications of self-attention module related to conceptual blending. Self-attention module is shown to be effective in combining two spatial features. RIVAL [Zhang et al., 2023a] and StyleAligned [Hertz et al., 2024] proposed to modify the self-attention module to be a sort of cross-attention form. Starting from the noise corresponding to the real reference image through inversion methods [Song et al., 2020, Mokady et al., 2023], they combine a denoising chain with an inversion

chain to blend the spatial features. Although they can blend cross-modal concepts without training, their methods are inherently limited when a real reference image is given, due to the distributional gap between the latents from the inversion chain and the denoising chain [Zhang et al., 2023a]. Similar ideas are used in [Cao et al., 2023, Alaluf et al., 2024], but their methods are specifically designed for non-rigid image editing or a blending of two visual concepts in a disentangled manner.

Transformer-based applications related to conceptual blending. StyleDrop [Sohn et al., 2023] proposes to finetune LoRA [Hu et al., 2022] to blend cross-modal concepts. However, their method is limited in scalability, as a separate set of LoRA modules needs to be optimized for each visual concept. Recently, UNO [Wu et al., 2025], OminiControl [Tan et al., 2024], and IC-LoRA [Huang et al., 2024] have shown impressive performance in subject-driven image generation by leveraging diffusion transformers as image encoder. However, their sequentially concatenating methods are only applicable to Diffusion Transformers (e.g., FLUX). Moreover, their methods are not suitable for conceptual blending because of the strong subject preservation.

Generative models augmenting human creativity. Previous studies [Franceschelli and Musolesi, 2024, Hwang, 2022] have shown the potential of generative models for augmenting creativity. Cai et al. [2023] proposed a diffusion framework to diversify image generations to provide inspiration for designers. CreativeConnect [Choi et al., 2024] proposed generative AI pipelines that can help graphic designers to have more design ideas by reference recombination process. Creative Blends [Sun et al., 2025] proposed a system that takes multiple textual concepts as input from users and outputs an image with the blended concepts. The conducted user study shows that visualizing these blended concepts can reduce cognitive load for participants and also foster creativity.

3 Method

In this Section, we describe our proposed method (IT-Blender) that adapts the pretrained projection layers of self-attention module to the visual and textual conceptual blending task. In Section 3.1, we first describe the preliminaries of the T2I diffusion models. In Section 3.2, we introduce IT-Blender with a novel blended attention module that can blend the visual concept of a reference image into the text-guided generation process with enhanced semantic correspondence retrieval for the real image.

3.1 Preliminaries

StableDiffusion and FLUX. StableDiffusion (SD) [Rombach et al., 2022] is widely used open source diffusion models for T2I synthesis. SD is trained with a denoising objective [Ho et al., 2020], and UNet [Ronneberger et al., 2015] is used as denoising networks. FLUX [Black Forest Labs, 2024] is advanced diffusion models based on diffusion transformers (DiT) [Peebles and Xie, 2023], which is trained with a score matching objective. SD 1.5 and FLUX.1-dev are used in our experiment.

Self-Attention module and its application. Self-attention (SA) module [Zhang et al., 2019, Rombach et al., 2022, Black Forest Labs, 2024] is one of the most important components of modern diffusion models. It not only learns to capture long-range dependencies, but also learns to encode spatial representations optimized for similarity comparison; what to aggregate and what to ignore based on semantic correspondence of the input itself. In our paper, the projection layers W_Q , W_K , and W_V are pretrained weights of SA module. For brevity, we omit the layer notation for the projection layers.

As mentioned earlier, Zhang et al. [2023a], Hertz et al. [2024] have shown that visual concept of a reference image can be blended in the generation process of pretrained T2I models without additional training. The detailed methodologies differ, but conceptually they suggested image Cross Attention (imCA) between two latents; Z_{noisy} from a denoising chain and Z_{inv} from an inversion chain, i.e., $\text{imCA}(Z_{\text{noisy}}, Z_{\text{inv}}) = \text{imCA}(Z_{\text{noisy}}, Z_{\text{inv}}; W_Q, W_K, W_V)$. This indicates the key and value of the SA module from Z_{inv} are combined with the query of the SA module from Z_{noisy} , i.e., $\sigma((Z_{\text{noisy}} W_Q)(Z_{\text{inv}} W_K)^T / \sqrt{d_k}) Z_{\text{inv}} W_V$. Note that the operation of imCA is essentially a cross-attention mechanism, but used in a distinct way, i.e., cross-attention in SA layers with W_Q, W_K, W_V .

3.2 Image and Text Blender (IT-Blender)

Setup and overview. We aim to generate an image where cross-modal concepts from a given real image and a text prompt are naturally blended without loss of details, in a disentangled manner. As mentioned in Section 1, attention module can be a key to implement the conceptual blending process.

One key observation for the inversion-based imCA approaches is that: they have the advantage in applying the details of the visual concepts in a disentangled manner, while the performance is degraded when real images are given as input due to the distribution shift of the inversion chain. Hence, our goal is to have a real image adapter that is trained to incorporate a given reference image into the pretrained projection space of the SA module. Since textual concepts are constantly provided through the text CA modules (which are physically separated from the SA modules), IT-Blender aims to blend visual concepts from the reference image with the text-guided generation process.

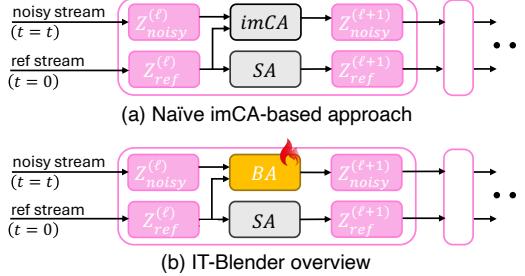


Figure 2: IT-Blender overview

Our method only trains the newly introduced adapter parameters while freezing all the pretrained weights, similar to prior works [Mou et al., 2024, Zhang et al., 2023b, Ye et al., 2023, Tan et al., 2024, Wu et al., 2025]. The denoising objective is used for SD1.5 [Rombach et al., 2022] and the denoising score matching objective is used for FLUX [Black Forest Labs, 2024].

The challenges are 1) how to encode a real image without loss of details, and 2) how to blend the encoded real image feature into the projection space of the pretrained SA module.

Native image encoding. Interestingly, diffusion models already know how to encode a real image X_{ref} into the denoising networks. It can be simply achieved by forwarding a clean version of X_{ref} with $t = 0$. This provides a sequence of latent representations across the L layers of the denoising networks: $(Z_{\text{ref}}^{(1)}, Z_{\text{ref}}^{(2)}, \dots, Z_{\text{ref}}^{(L)})$. This representation has some similarities to the inversion methods in which there is a latent representation at every layer of the network for each denoising step. However, it is fundamentally different because its timestep is set to 0 for all denoising steps, i.e., the clean latent representations can be used at every timestep. We hypothesize that these clean representations are more helpful for conceptual blending because they encode the details of the clean image rather than noisy images as in inversion-based methods. Furthermore, our approach does not require image inversion, which is computationally expensive.

Despite the benefits of this clean representation, it is unclear how to incorporate a set of clean latent features per layer from the denoising networks into the regular denoising process. One simple naïve approach inspired by prior works is to simply use an imCA module to blend the clean reference latent Z_{ref} into the noisy latent Z_{noisy} , i.e., replace SA(Z_{noisy}) modules with imCA($Z_{\text{noisy}}, Z_{\text{ref}}$), as shown in Fig. 2 (a). While in theory this could be done without retraining by using the pretrained self-attention module weights as in Hertz et al. [2024], Zhang et al. [2023a], the performance would be poor because of a significant distribution shift; the reference latents are from a clean image with $t = 0$ while the noisy latents are from noisy images with a $t \geq 0$. Fig. 8 (a) shows the empirical verification of the hypothesis. Thus, a new blending module and finetuning method is needed that can use the clean latents but seamlessly blend the visual concept information into the noisy latents.

IT-Blender. To bridge the gap, we design IT-Blender to have our novel blended attention (BA) module with trainable parameters that can learn how to map the clean Z_{ref} to the Z_{noisy} in the projection space.

As shown in Fig. 2 (b), IT-Blender has two streams; noisy stream and reference stream. The noisy stream refers to the regular denoising chain from $t = T$ to $t = 0$ during sampling or randomly sampled t during training. The reference stream is for encoding a reference image without any noise. Along this stream, $t = 0$ is constantly given for both training and sampling. The same text prompt is used for both streams. The training objective is applied only to the noisy stream.

Blended Attention (BA). As shown in Fig. 3, we design blended attention to have a residual structure with two terms; the first term on the left is the original pretrained self-attention module, which can keep the

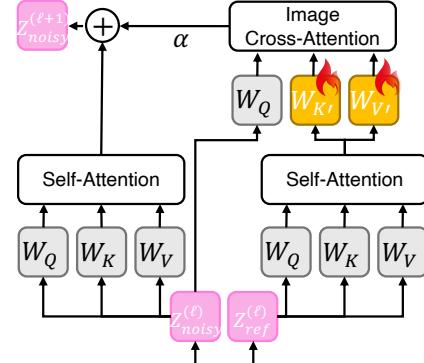


Figure 3: Blended attention at ℓ -th layer.

estimation on the original trajectory. The second imCA term on the right is the key to blended attention, which enables a blending of visual and textual concepts by bridging the clean reference stream with the noisy stream in the output space of the SA module. The ℓ -th self-attention layers of the denoising networks are changed to our blended attention as shown in the equation below:

$$\text{BA} = \text{SA}(Z_{\text{noisy}}^{(\ell)}) + \alpha \text{imCA}(Z_{\text{noisy}}^{(\ell)}, \text{SA}(Z_{\text{ref}}^{(\ell)}); W_Q, W_{K'}, W_{V'}), \quad (1)$$

where $W_{K'}$ and $W_{V'}$ are trainable parameters. The layer notation for the projection layers is omitted for the brevity purpose. α is set to be 1 during the training while set to be a constant < 1 during sampling. In our experiments, we empirically used $\alpha = 0.25$ for SD and $\alpha = 0.6$ for FLUX (the visualization of varying α 's is shown in Fig. 17). For training, both $W_{K'}$ and $W_{V'}$ are randomly initialized. The imCA term in Eq. 1 plays a role in dynamically aligning $\text{SA}(Z_{\text{ref}}^{(\ell)})$ to $\text{SA}(Z_{\text{noisy}}^{(\ell)})$ in the output space of SA by optimizing $W_{K'}$ and $W_{V'}$ to fetch the useful visual information to denoise from the reference stream, driven by the query from the noisy stream.

4 Experiments

Detailed experiment settings and implementation details are provided in Section A and B.

Data. For training and testing SD 1.5 and FLUX, we used a squared subset of LAION2B-en-aesthetic dataset [OpenDiffusionAI, 2025, Schuhmann et al., 2022], which contains around 300k squared images (with at least a resolution of $1,024 \times 1,024$) and their paired text prompt.

Metrics for baseline comparison. We mainly evaluate how well the textual and visual concepts are disentangled. In our cross-modal blending task, semantics (i.e., object) must be determined by a text prompt, and visual concepts (e.g., texture, ingredient, material, color, and local shapes) need to be determined by a reference image. If visual and textual concepts are disentangled well, each of them should maintain high consistency after being blended with different combinations. Therefore, the key to the evaluation is to measure set consistencies for visual concept and the textual concept, respectively. To measure the textual set consistency, we compare a set of generated samples with a fixed text prompt but with different reference images. The generated object must be consistent, and thus we used CLIP [Radford et al., 2021] to measure the semantic similarity between all pairs of the generated images with a fixed prompt. To measure the visual set consistency, we compare the generated samples with a fixed visual prompt but with different text prompts. DINO is used to focus more on pure visual similarity, not semantics, following previous studies [Ruiz et al., 2023, Hertz et al., 2024]. Next, we measure the correct class predictions to measure whether the generated results preserve the textual concept. ChatGPT4.1 [OpenAI, 2023] is used. For SD evaluation, 200 unseen samples with 30 text prompts are used (6,000 samples per baseline in total). For FLUX evaluation, 200 unseen samples with 20 text prompts are used (4,000 samples per baseline in total). We also report the blending score and analysis in Section D.1.

4.1 Baseline Comparison (SD)

Baselines. To compare the performance of cross-modal conceptual blending in SD, we use two encoder-based methods (BLIP-Diffusion [Li et al., 2023] and IP-Adapter [Ye et al., 2023]) and two inversion-based methods (RIVAL [Zhang et al., 2023a] and StyleAligned [Hertz et al., 2024]). We used SD 1.5 for all the baselines while SDXL [Podell et al., 2023] is used in StyleAligned [Hertz et al., 2024] as their performance in SD 1.5 is worse by a large margin.

Results. Both encoder-based baselines (IP-Adapter and BLIP-Diffusion) show similar patterns. First, the visual concept frequently dominates the generation process, and thus the generated images sometimes do not look like the object given as a text prompt (e.g., the flower train of IP-Adapter and the robot of BLIP-Diffusion in Fig. 5). The same pattern is observed in quantitative evaluations. The encoder-based baselines show the lowest visual and textual set consistencies (Fig. 4 top). This is because they frequently miss the textual concept, yielding inconsistency of the textual and visual sets.

Similarly, the classification results in Fig. 4 bottom show that IP-Adapter and BLIP-Diffusion often miss the target object; out of 200 samples, on average over the prompts, only around 100 samples are classified as a target object. We believe this is because their methods rely on the text CA module, which can inherently limit the disentanglement of visual and textual concepts.

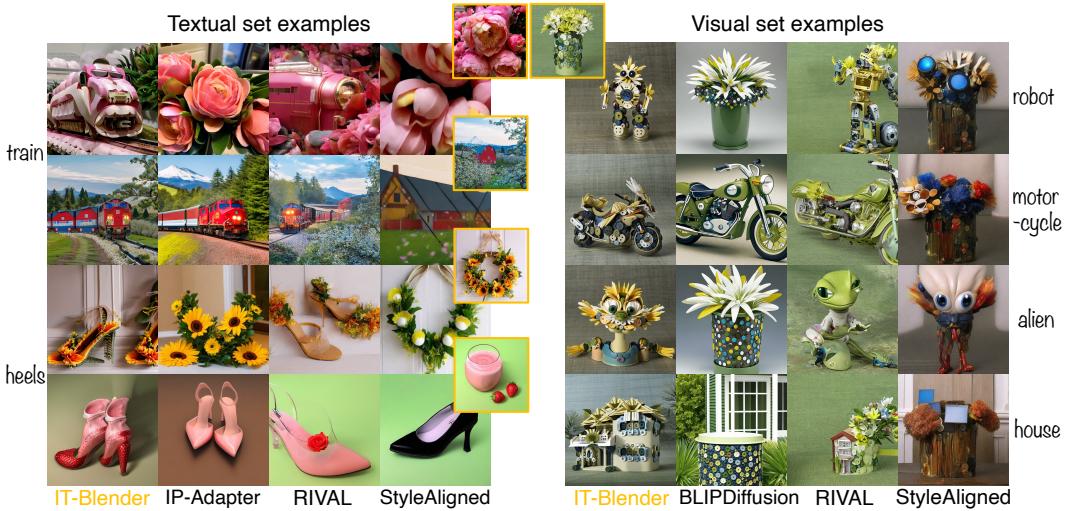


Figure 5: Qualitative comparisons with the baselines in StableDiffusion. For each column of the textual set examples, every two row with the same text prompt need to be semantically consistent. Each column of the visual set examples need to be visually consistent.

Second, when the textual concept is properly applied, the generated results from IP-Adapter and BLIP-Diffusion often lose the details of the visual concept (e.g., the strawberry heels of IP-Adapter and the motorcycle of BLIP-Diffusion in Fig. 5). Additional DINO similarity experiments between a generated image and a reference image (IT-Blender (**0.837**), IP-Adapter (0.812), and BLIP-Diffusion (0.821)) support the observations. This is because IT-Blender does not rely on an external image encoder, while natively encodes images with the denoising networks, retaining visual details better.

As for inversion-based baselines, StyleAligned frequently misses the textual concept, as shown in the motorcycle and house examples in Fig. 5. The lowest classification score in Fig. 4 bottom also quantitatively supports the observation. RIVAL shows worse performance than IT-Blender in both textual and visual set consistencies. This is because their inversion-based method is not specialized in retrieving semantic correspondence between the reference and the generated images, and thus the visual concepts are inconsistently applied to the generated images given varying inputs.

IT-Blender shows good performance in blending visual and textual concepts in a disentangled manner, as shown in the second-best visual set consistency and the best textual set consistency. The highest class prediction also supports the strong performance of IT-Blender in rigidly applying textual concepts. The superior disentanglement performance of IT-Blender is attributed to 1) self-attention-based design, which separates the visual and textual concepts, and 2) strong semantic correspondence retrieval by blended attention, with which the given visual concepts can be consistently applied given varying inputs.

Additional baseline comparisons are provided in Section D, e.g., “blending score” by ChatGPT and occasional unrealistic generations of inversion-based baselines in SD.

4.2 Baseline Comparison (FLUX)

Baselines. To compare the cross-modal conceptual blending performance in FLUX, we used three open-source baselines; IP-Adapter [Ye et al., 2023], OminiControl [Tan et al., 2024], and UNO [Wu et al., 2025]. For IP-Adapter, among two popular open source implementations, we used InstantX implementation [Team, 2024] as it is much better in blending visual and textual concepts. Both OminiControl and UNO are designed for subject-driven image generation by training additional lora modules on top of the pretrained FLUX. The experiment results of IP-Adapter and UNO are based on FLUX.1-dev while OminiControl is based on FLUX.1-schnell.

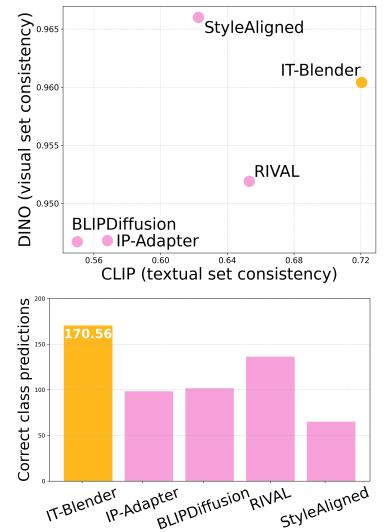


Figure 4: Visualizations of the quantitative comparison with the SD 1.5 baselines.

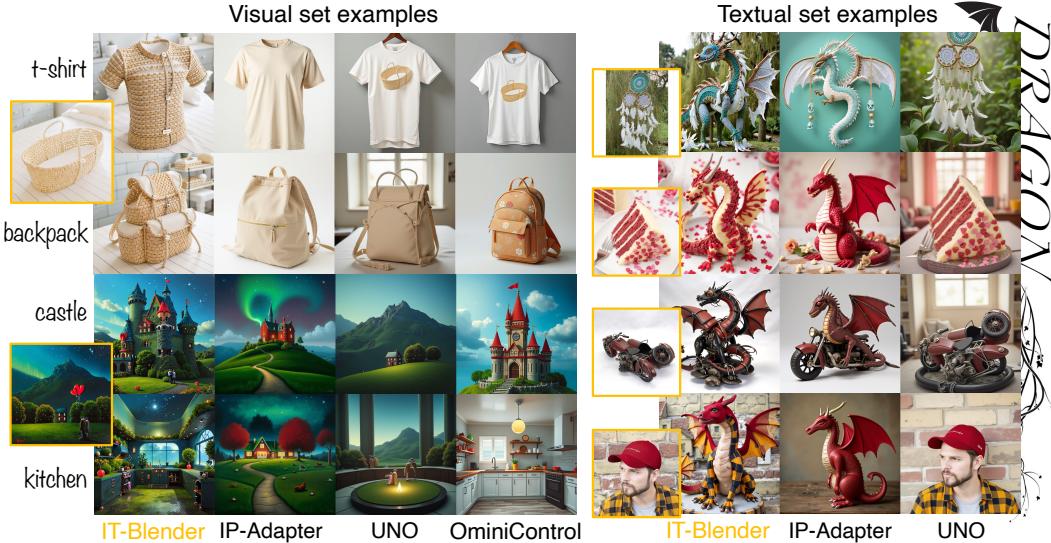


Figure 7: Qualitative comparisons with the baselines in FLUX.

Results. As UNO and OminiControl are specifically trained for subject-driven image generation with paired data, their models are not suitable for blending visual and textual concepts, especially when given visual and textual conditions are not highly correlated. As can be seen in Fig. 7, UNO and OminiControl show strong reference preservation, as shown in the basket-printed t-shirts. However, OminiControl often fails in incorporating the visual concept from the reference image (e.g., the backpack and kitchen examples), while UNO often fails in incorporating the textual concept (e.g., the castle and kitchen examples). IP-Adapter decently blends the visual and textual concepts, but they miss the details of the visual concepts (e.g., the dragons in the second and the fourth rows).

We also observe the similar patterns in the quantitative experiment results. OminiControl shows strong text guidance effect (e.g., the highest textual set consistency and classification in Fig. 6) while inconsistent reference image effect (e.g., the lowest visual set consistency). UNO shows relatively robust performance in preserving the given object in our task, as shown by the high visual set consistency score. However, the given text prompt is often ignored, which is shown by the lowest textual set consistency. IP-Adapter demonstrates lower visual and textual set consistencies compared to ours, similar to the SD experiment results. Compared to the baselines, IT-Blender shows the second-best textual set consistency and the best visual set consistency, showing superior performance in cross-modal conceptual blending.

4.3 Ablation Study and New Applications

In this section, we present interesting applications and visualize the attention mask to better understand what IT-Blender learns. More results are provided in the Appendices (e.g., applying multiple visual concepts in Section C and more interesting results in Section E).

Effects of the blended attention module. To intuitively understand what blended attention learns, we visualize the self-attention mask of BA modules in FLUX. Fig. 8 (a) shows the results. The attention masks of IT-Blender captures the visually corresponding texture area from the reference image. For example, the yellow star from the whale example mostly captures the fur area of the bird in the reference image, while the pink star mostly captures the feather area. However, the attention mask of the naive imCA-based approach (Fig. 2 (a)), does not capture a meaningful area, and thus the generated results are also significantly degraded. This verifies our hypothesis that the distribution shift between clean Z_{ref} and Z_{noisy} is significant, and therefore training $W_{K'}$ and $W_{V'}$ of blended attention is needed to bridge Z_{ref} and Z_{noisy} .

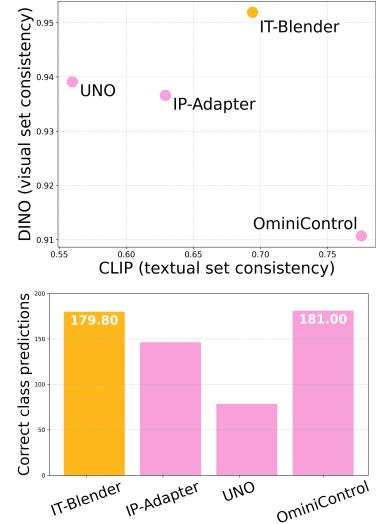


Figure 6: Visualizations of the quantitative comparison with the FLUX baselines.

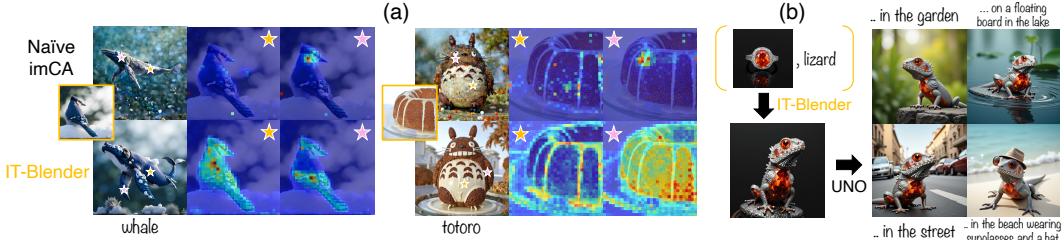


Figure 8: (a) attention mask visualization of IT-Blender and naïve imCA (Fig. 2 (a)). (b) our blended results can be applied to subject-driven generative models to create interesting novel scenes.



Figure 9: Feasible design examples when the given visual and textual concepts are semantically close.

Fesible design. As shown in the owls with diverse desserts in Fig. 1, IT-Blender can create experimental design in a realistic way, which can inspire humans. Interestingly, we also observe that IT-Blender can generate feasible design outcomes as well, especially when a reference image is semantically close to the object given in the text prompt. For example, as shown in Fig. 9, given an indoor or outdoor reference image, IT-Blender can generate the target room with surprisingly coherent visual concepts with the given reference image. Furniture or apparel could be another example.

Additional results. Given a fixed visual and textual concepts, IT-Blender can generate diverse images with varying random noise, as shown in Fig. 10. The creative object generated by IT-Blender can be synthesized in novel scenes with subject-driven models, as shown in Fig. 8 (b).



Figure 10: The results are generated with varying noise.

5 Conclusion

In this paper, we propose IT-Blender that can augment human creativity by automating the cross-modal conceptual blending process of a real image and text. First, IT-Blender uses native denoising networks to encode a real reference image to minimize the loss of visual details, with fast inference time. Second, the encoded visual feature is fed into our novel blended attention modules, which are trained to bridge the distribution shift between the clean reference image and the noised generated image. Third, our blended attention modules are built upon the self-attention module, which can disentangle the textual concept and the visual concept by design. In both SD and FLUX, the experiment results demonstrate that IT-Blender outperforms the baselines in blending cross-modal concepts in terms of disentangling cross-modal concepts and preserving textual and visual concepts. The blending score further verifies the superior performance of IT-Blender in cross-modal conceptual blending. Further discussion of future directions, limitations, and societal impact is provided in Section F. We hope that our research will be able to draw attention to the potential of image-generative models to augment human creativity.

References

- Gilles Fauconnier and Mark Turner. *The way we think: Conceptual blending and the mind's hidden complexities*. Basic books, 2008.
- Gilles Fauconnier and Mark Turner. Conceptual integration networks. *Cognitive science*, 22(2):133–187, 1998.
- Seana Coulson. *Semantic leaps: Frame-shifting and conceptual blending in meaning construction*. Cambridge University Press, 2001.
- Liane Gabora. Cognitive mechanisms underlying the creative process. In *Proceedings of the 4th conference on Creativity & cognition*, pages 126–133, 2002.
- Maria C Yang. Observations on concept generation and sketching in engineering design. *Research in Engineering Design*, 20:1–11, 2009.
- Kyung Hoon Hyun and Ji-Hyun Lee. Balancing homogeneity and heterogeneity in design exploration by synthesizing novel design alternatives based on genetic algorithm and strategic styling decision. *Advanced Engineering Informatics*, 38:113–128, 2018.
- Alice Cai, Steven R Rick, Jennifer L Heyman, Yanxia Zhang, Alexandre Filipowicz, Matthew Hong, Matt Klenk, and Thomas Malone. Designaid: Using generative ai and semantic diversity for design inspiration. In *Proceedings of The ACM Collective Intelligence Conference*, pages 1–11, 2023.
- Kevin N Otto. *Product design: techniques in reverse engineering and new product development*. 2003.
- Bob Eberle. *Scamper on: Games for imagination development*. Prufrock Press Inc., 1996.
- Karl T Ulrich and Steven D Eppinger. *Product design and development*. McGraw-hill, 2016.
- David G Jansson and Steven M Smith. Design fixation. *Design studies*, 12(1):3–11, 1991.
- Abraham S Luchins. Mechanization in problem solving: The effect of einstellung. *Psychological monographs*, 54(6):i, 1942.
- 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.
- 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.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- Black Forest Labs. Flux.1 [dev]. <https://huggingface.co/black-forest-labs/FLUX.1-dev>, 2024. Accessed: 2025-04-27.
- 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.
- Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36:30146–30166, 2023.
- Yuechen Zhang, Jinbo Xing, Eric Lo, and Jiaya Jia. Real-world image variation by aligning diffusion inversion chain. *Advances in Neural Information Processing Systems*, 36:30641–30661, 2023a.
- Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4775–4785, 2024.
- Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-more generalization: Unlocking more controllability by in-context generation. *arXiv preprint arXiv:2504.02160*, 2025.
- Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. *arXiv preprint arXiv:2411.15098*, 2024.

- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.
- Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*.
- 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, 2023b.
- 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*, volume 38, pages 4296–4304, 2024.
- Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.
- 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.
- Bingyan Liu, Chengyu Wang, Tingfeng Cao, Kui Jia, and Jun Huang. Towards understanding cross and self-attention in stable diffusion for text-guided image editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7817–7826, 2024.
- 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15943–15953, 2023.
- Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.
- 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.
- Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yingqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22560–22570, 2023.
- Yuval Alaluf, Daniel Garibi, Or Patashnik, Hadar Averbuch-Elor, and Daniel Cohen-Or. Cross-image attention for zero-shot appearance transfer. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024.
- Kihyuk Sohn, Nataniel Ruiz, Kimin Lee, Daniel Castro Chin, Irina Blok, Huiwen Chang, Jarred Barber, Lu Jiang, Glenn Entis, Yuanzhen Li, et al. Styledrop: Text-to-image generation in any style. *arXiv preprint arXiv:2306.00983*, 2023.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.
- Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Huanzhang Dou, Chen Liang, Yutong Feng, Yu Liu, and Jingren Zhou. In-context lora for diffusion transformers. *arXiv preprint arXiv:2410.23775*, 2024.
- Giorgio Franceschelli and Mirco Musolesi. Creativity and machine learning: A survey. *ACM Computing Surveys*, 56(11):1–41, 2024.
- Angel Hsing-Chi Hwang. Too late to be creative? ai-empowered tools in creative processes. In *CHI conference on human factors in computing systems extended abstracts*, pages 1–9, 2022.
- DaEun Choi, Sumin Hong, Jeongeon Park, John Joon Young Chung, and Juho Kim. Creativeconnect: Supporting reference recombination for graphic design ideation with generative ai. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–25, 2024.
- Zhida Sun, Zhenyao Zhang, Yue Zhang, Min Lu, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Creative blends of visual concepts. In *CHI*, 2025.

- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings, part III* 18, pages 234–241. Springer, 2015.
- 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.
- Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International conference on machine learning*, pages 7354–7363. PMLR, 2019.
- OpenDiffusionAI. laion2b-en-aesthetic-square. https://huggingface.co/datasets/opendiffusionai/laion2b_en-aesthetic-square, 2025.
- 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.
- 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.
- 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.
- OpenAI. Gpt-4 technical report, 2023. URL <https://arxiv.org/abs/2303.08774>. arXiv:2303.08774.
- Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.
- InstantX Team. Instantx flux.1-dev ip-adapter page. <https://huggingface.co/InstantX/FLUX.1-dev-IP-Adapter>, 2024.
- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- Dongmin Park, Sebin Kim, Taehong Moon, Minkyu Kim, Kangwook Lee, and Jaewoong Cho. Rare-to-frequent: Unlocking compositional generation power of diffusion models on rare concepts with llm guidance. *arXiv preprint arXiv:2410.22376*, 2024.

Appendices

Contents of Appendices



Figure 11: Stylized brand logos by IT-Blender with FLUX.

• Appendix A: Experiment Settings	14
• Appendix B: Implementation Details	
• Appendix C: Multiple Visual Concepts	15
• Appendix D: Additional Baseline Comparisons	
– D.1: Comparison of Blending Score by ChatGPT (SD and FLUX)	16
– D.2: Qualitative Observation Report (SD)	22
(A limitation of training-free inversion-based method)	
• Appendix E: Additional Results and Analysis	
– E.1: Effect of α of Blended Attention	23
– E.2: Softmax Temperature Control (heuristic for multiple reference images)	
– E.3: Additional Results	24
• Appendix F: Discussion	30
– F.1: Interesting Future Directions	
– F.2: Limitations	
– F.3: Societal Impact	

A Experiment Settings

SD setting. To evaluate the performance with the baselines in SD, we sample 200 samples per prompt. The 30 prompts that we used are as follows:

car, bus, bicycle, chair, truck, tank, lamp, handbag, backpack, heels, train, rabbit cartoon character, owl cartoon character, mouse cartoon character, castle, headphone, motorcycle, kettle, vacuum, toy airplane, robot, sneakers, dragon cartoon character, reindeer cartoon character, alien cartoon character, living room, bathroom, bedroom, kitchen, house

FLUX setting. We sample 200 samples per prompt. The 20 prompts that we used are as follows:

car, bicycle, chair, lamp, headphone, truck, sneakers, handbag, backpack, t-shir", lizard, fish, owl cartoon character, monster cartoon character, dragon, living room, kitchen, castle, 3D apple logo, 3D toyota logo

B Implementation Details

To train IT-Blender with SD 1.5, we use 1 NVIDIA RTX 6000 with a batch size of 16. To train IT-Blender with FLUX, we use 4 NVIDIA L40S GPUs with a total batch size of 16. IT-Blender training and sampling require two streams, as shown in Fig. 2. We simply concatenate them in the batch dimension so that the key-value injections from the reference stream can be easily achieved in each Blended Attention processor.

We train IT-Blender for 5 epochs with a learning rate of $1e-5$ in SD 1.5. We train IT-Blender for 1-2 epochs with a learning rate of $2e-5$ in FLUX. AdamW [Loshchilov and Hutter, 2017] is used in both settings with `betas = [0.9, 0.99]` and `weight_decay = 0.01`.

C Multiple Visual Concepts

IT-Blender can apply multiple visual concepts from multiple reference images.

The naive way is to add additional imCA terms in Eq. 1, e.g.,

$$\begin{aligned} \text{BA} = \text{SA}(Z_{\text{noisy}}^{(\ell)}) + \alpha \text{imCA}(Z_{\text{noisy}}^{(\ell)}, \text{SA}(Z_{\text{ref}_1}^{(\ell)}; W_Q, \textcolor{orange}{W}_{K'}), \textcolor{orange}{W}_{V'}) \\ + \alpha \text{imCA}(Z_{\text{noisy}}^{(\ell)}, \text{SA}(Z_{\text{ref}_2}^{(\ell)}; W_Q, \textcolor{orange}{W}_{K'}), \textcolor{orange}{W}_{V'}), \end{aligned} \quad (2)$$

where Z_{ref_1} and Z_{ref_2} mean two reference images. However, we empirically observe that the results naively mingles the visual features for each query coordinate, which makes the generated image less conspicuous where the visual feature comes from.

To tackle this problem, we came up with a simple idea; concatenating the multiple reference images in sequence dimension before applying softmax of Attention, e.g., $Q \in \mathbb{R}^{HW \times D}$ and $\{K, V\} \in \mathbb{R}^{2HW \times D}$, when two reference images are used. In this way, BA module can exclusively (not strictly though as it is softmax, not hardmax) fetch the visual features from the multiple reference images. We used this approach to blend multiple visual concepts. More examples are provided below in Fig. 12.

Another possible way would be to concatenate multiple reference images in height or weight dimension of the reference image, similar to Huang et al. [2024].



Figure 12: Examples by IT-Blender with FLUX, generated with multiple reference images.

D Additional Baseline Comparisons

D.1 Comparison of Blending Score by ChatGPT (SD and FLUX)

To further measure the blending performance, we use ChatGPT 4.1 [OpenAI, 2023] with a detailed rubric, inspired by the high correlation between human and state-of-the-art LLMs in measuring text and image alignment [Park et al., 2024].

To evaluate, the same samples with the main experiments are used, i.e., the 6000 samples in SD and 4000 samples in FLUX. The results are as shown below:

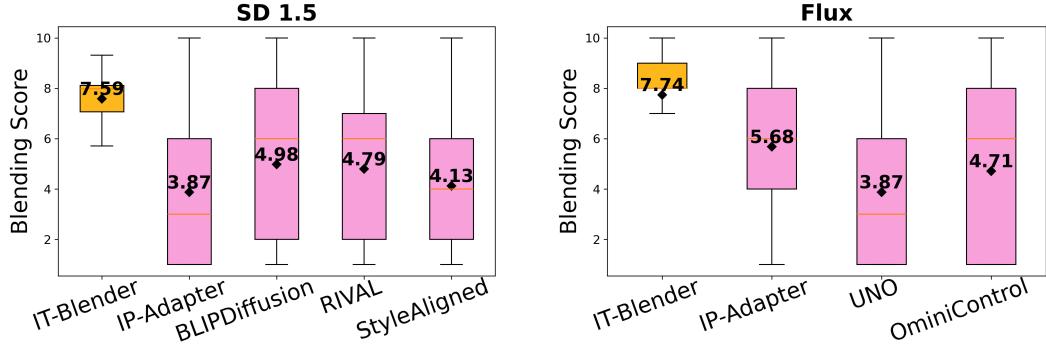


Figure 13: Visualizations of the blending score comparisons with the baselines in SD (left) and FLUX (right).

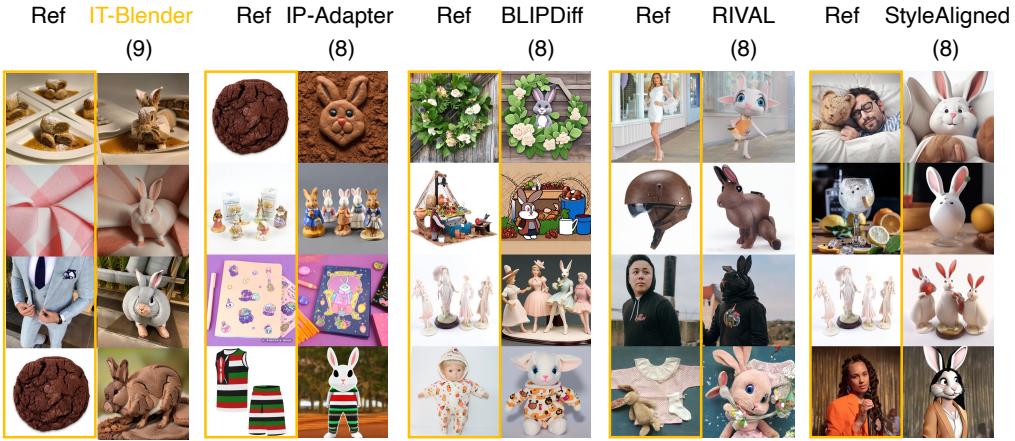
Fig. 13 shows the blending score measure by ChatGPT, given a specific rubric. As shown in the SD-based and Flux-based results, IT-Blender shows the rigid and best performance with the highest mean and lowest variance.

According to the rubric, the highest mean around 8 indicates that our blending results have most elements from both inputs, and they are well integrated.

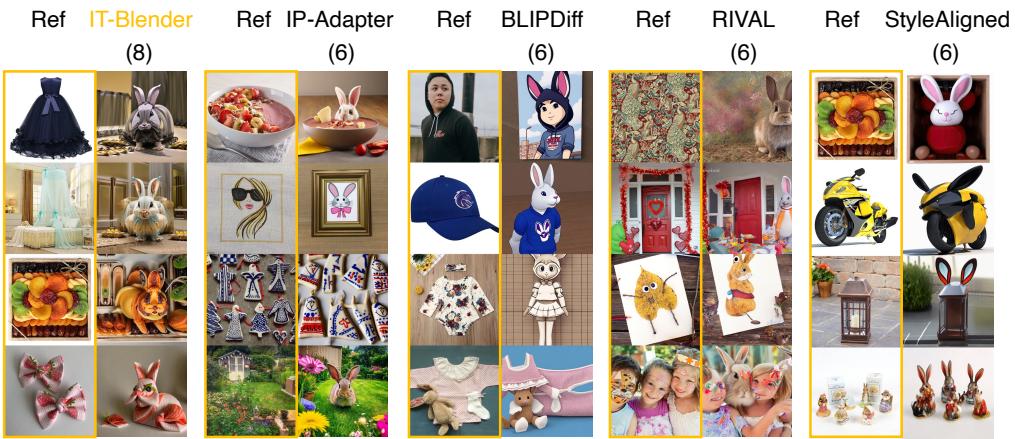
The low variance of IT-Blender indicates that both concepts are consistently blended in a plausible way, without failed or unbalanced integration.

We further visualize the top 10%, 50% (median), and 90% samples in terms of the blending score in Fig. 14 and Fig. 15. The high blending scores around 8-9 show decent performance in blending visual and textual concepts while the low blending scores around 1-2 show poor performance, e.g., only applying one concept or blending cross-modal concepts weakly.

SD15 Top 10 % samples in terms of blending score



SD15 Top 50 % samples in terms of blending score (median)



SD15 Top 90 % samples in terms of blending score



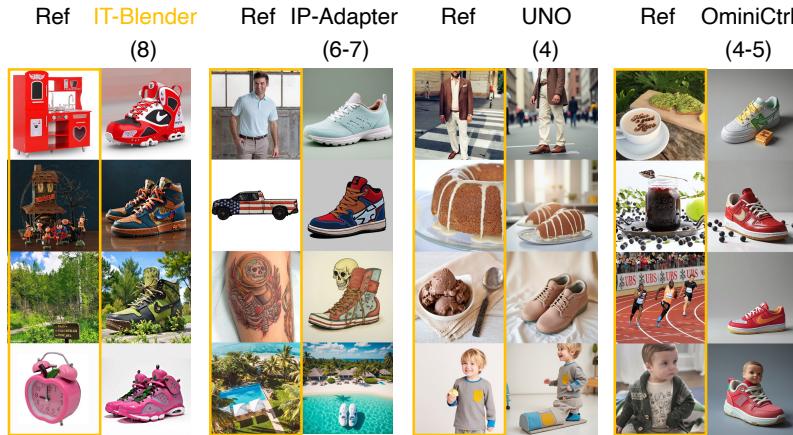
“A photo of a rabbit cartoon character”

Figure 14: Visualization of top 10%, 50%, and 90% samples in terms of blending score (SD). The numbers below each baseline name indicate the blending scores the displayed samples got.

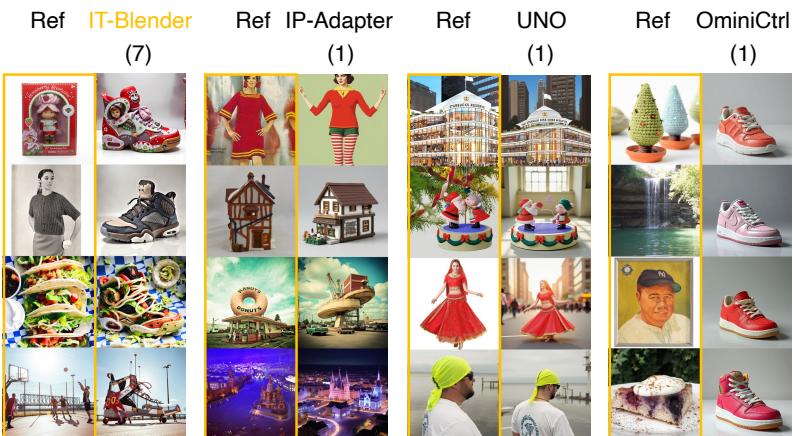
FLUX Top 10 % samples in terms of blending score



FLUX Top 50 % samples in terms of blending score (median)



FLUX Top 90 % samples in terms of blending score



"A photo of sneakers, imaginative, creative, design"

Figure 15: Visualization of top 10%, 50%, and 90% samples in terms of blending score (FLUX). The numbers below each baseline name indicate the blending scores the displayed samples got.

Query for measuring blending score. The prompt we used to measure the blending score is as follows:

You are a helpful assistant who evaluates how well textual and visual concepts are blended in the image generation process. The object in the given first image is conceptually blended result given the text prompt and the second reference image. Text determines semantics while the reference image determines visual concepts such as texture, material, color, and local shape. Evaluate how closely the visual concept in the provided image aligns with the textual concept in the text prompt and the visual concept from the second image. Identify significant overlaps or discrepancies in terms of global object shape, local shape, appearance, texture, material, color, and all the detailed visual components. Analyze the conceptual similarity between the first provided generated image and the text prompt: [PROMPT]. You also need to consider the conceptual similarity between the first provided generated image and the second provided reference image. Provide a concise explanation for your evaluation. Note that we are evaluating cross modal conceptual blending, and thus if one of the crossmodal concepts does not present in the generated image, it has to be considered as failed, even though the first image perfectly matches the second image.

First image: [GEN_IMAGE]

Second image: [REF_IMAGE]

The object in the given first image is conceptually blended result given the text prompt and the second image. Evaluate how closely the visual concept in the provided image aligns with the textual concept in the text prompt and the visual concept from the second image. Identify significant overlaps or discrepancies in terms of shape, appearance, composition, and overall impression. Provide a concise explanation for your evaluation.

Give a score from 1 to 10, according to the following criteria:

10 Perfect conceptual integration: The generated image seamlessly incorporates all core semantic and stylistic elements from both the text and the visual concept. There's no ambiguity in the fusion; it reflects a deep, coherent synthesis of the two modalities.

9 Near-perfect integration: Strong conceptual blending with only extremely minor details or subtleties missing from either modality. The result is still fully coherent and creatively unified.

8 Excellent with minor trade-offs: Most elements from both inputs are present and well-integrated, but one or two key aspects may be simplified. The conceptual overlap is still meaningful.

7 Very good blend, slightly unbalanced: Clear depiction of both concepts with small discrepancies-e.g., one modality slightly dominates the fusion. Still communicates a unified concept.

6 Mostly present, but noticeable gaps: Both modalities are represented, but some important attributes (e.g. color, pose, key terms, or symbolic features) are missing or only vaguely suggested.

5 Moderate representation: Some elements from both text and image are depicted, but several key parts are ignored or distorted. The blend may feel partial or underdeveloped.

4 Unbalanced or sparse blend: One modality is clearly underrepresented or the blend feels superficial. Visuals may include token features from one source without meaningful synthesis.

3 Weak conceptual integration: Few recognizable aspects from both text and image appear; blending feels incomplete or accidental rather than intentional.

2 Minimal blending: Image mostly reflects one modality, with only token or confused reference to the other. Viewers may struggle to infer any deliberate fusion.

1 Failed integration: Generated image does not meaningfully reflect either the textual concept or the visual input. No clear blending is achieved.

Provide your score and explanation (within 20 words) in the following format: ### SCORE: score ### EXPLANATION: explanation

Explanations provided by ChatGPT for blending score. The explanations for the lowest row for each of top 10%, 50%, and 90% are as follows:

- SD15
 - 1. IT-Blender
 - (a) (Top 10%) "The rabbit cartoon's form is clear and well-blended with the cookie's texture and color, though small stylization remains."
 - (b) (Top 50%) "The image blends a rabbit cartoon character with fabric and color from the bows, but lacks full cartoon stylization."
 - (c) (Top 90%) "The rabbit shows garden background and chef attire (from the reference), but lacks strong cartoon character cues from text."
 - 2. IP-Adapter
 - (a) (Top 10%) "Rabbit cartoon is fully integrated; clothing shows reference outfit's colors and stripes, but lacks emblem and exact shape."
 - (b) (Top 50%) "The rabbit matches the text but is realistically rendered, not cartoon-like; the garden environment strongly reflects the reference image."
 - (c) (Top 90%) "The generated image contains only the visual reference style, with no trace of the "rabbit cartoon character" concept."
 - 3. BLIP-Diffusion
 - (a) (Top 10%) ""
 - (b) (Top 50%) "The generated image uses the reference's color and texture but misses key elements of "rabbit cartoon character" in pose and style."
 - (c) (Top 90%) "The generated image reflects only the bread visual, lacking any "rabbit cartoon character" elements from the text prompt."
 - 4. RIVAL
 - (a) (Top 10%) "Strong rabbit cartoon character blending, large expressive eyes and pose, but clothing details lack reference's color and pattern accuracy."
 - (b) (Top 50%) "The generated image shows a real rabbit and kids with cartoonish face paint, but lacks full cartoon character integration."
 - (c) (Top 90%) "Image matches the rabbit cartoon text prompt, but shows almost no visual or conceptual blending with the milk bottles."
 - 5. StyleAligned
 - (a) (Top 10%) "Strong blend; rabbit cartoon matches text while outfit, pose, and background clearly reflect the reference image. Minor stylization only."
 - (b) (Top 50%) "Rabbit forms and some cartoon stylization are present, but features are indistinct and visuals are muddled."
 - (c) (Top 90%) ""
- FLUX
 - 1. IT-Blender
 - (a) (Top 10%) "The sneaker perfectly incorporates the reference flowers' color, material, and shape, with only minor textural differences from the reference."
 - (b) (Top 50%) "Sneakers integrate clock's pink color, shiny texture, and green leaf elements, but clock face and apple shape are abstracted."

(c) (Top 90%) “Strong sneaker form integrates basketball court elements and colors, but local sneaker details and textures are somewhat abstracted.”

2. IP-Adapter

(a) (Top 10%) “The sneaker integrates floral elements—shape and details—from the bouquet while retaining clear sneaker form, with only minor detail loss.”

(b) (Top 50%) “Sneakers (text) are clearly integrated into the villa pool scene (reference), but sneakers’ material/style don’t borrow villa textures.”

(c) (Top 90%) “No sneakers are present; the image depicts buildings and cityscape, failing both text and visual blending criteria.”

3. UNO

(a) (Top 10%) “Sneaker shape is clear and main structure matches “sneakers”, but donut texture dominates, slightly stylizing the footwear concept.”

(b) (Top 50%) “Only the clothing from the reference is blended; no real sneaker shape from the text is present, making fusion superficial.”

(c) (Top 90%) “The generated image only depicts a man with a yellow headscarf, not sneakers; it fails cross-modal blending.”

4. OminiControl

(a) (Top 10%) “The sneaker adopts the Buddha statue’s ivory color, material, and some smooth texture, but lacks significant Buddha-specific shapes.”

(b) (Top 50%) “The sneaker incorporates a doll’s head, referencing the baby, but lacks deeper integration of baby features, mainly merging objects.”

(c) (Top 90%) “The generated image is a sneaker, matching only the text prompt, with no visual or conceptual blending of the cake reference.”

D.2 Qualitative Observation Report (SD)

we observe that the training-free inversion-based baselines sometimes lie off the manifold, so the results are not realistic when cross-modal concepts are blended. We think this is an inherent limitation of training-free methods (in exchange for the benefit of “training free”), which intervene in the sampling trajectory. As shown in Fig. 16, The training-free methods RIVAL and StyleAligned sometimes unrealistically blend the results. The encoder-based baselines IP-Adapter and BLIP-Diffusion often miss the text prompt while the generated results are realistic. IT-Blender combines the benefits, consistently and realistically blending both concepts.

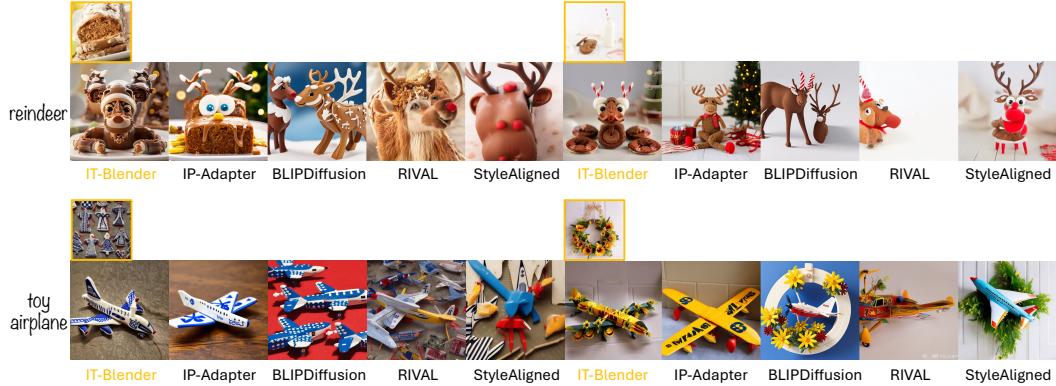


Figure 16: Additional qualitative comparisons (SD).

E Additional Results and Analysis

E.1 Effect of α of Blended Attention

We visualize the effect of α of Eq. 1 in Fig. 17. When $\alpha = 0$, no effect is applied as the imCA term in Eq. 1 becomes zero out. From left to right, as α increases, we can see that the visual concepts are more blended into the generated image. We empirically found that $\alpha = 0.6$ is the best way to get the most natural blend results. However, depending on the user’s intention, $\alpha \in [0.5, 0.8]$ is also good to go with. Especially when reference images and the text prompt are semantically close, $\alpha > 0.6$ can be effective, as shown in some of the results in section E.3.

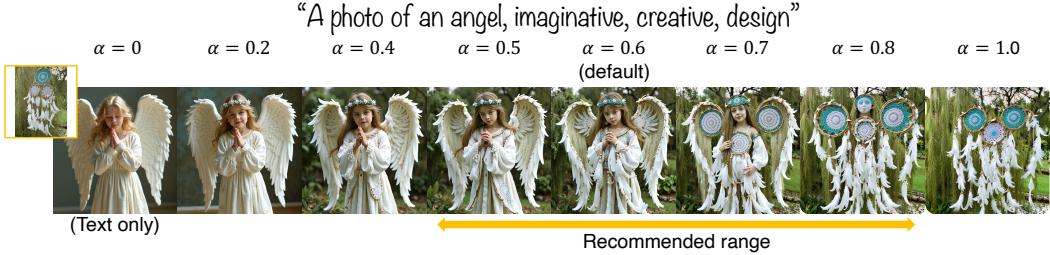


Figure 17: Visualization of the effect of alpha in blended attention with FLUX.

E.2 Softmax Temperature Control (heuristic for multiple reference images)

We empirically observe that applying low temperature to the logits before applying softmax can sharpen the softmax distribution, possibly helping to prevent ambiguous mixtures of visual concepts in exchange for image fidelity. The attention formulation with the temperature can be represented as:

$$\text{Attention}(Q, K, V; \text{temp}) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k} \cdot \text{temp}} \right) V. \quad (3)$$

$1/\text{temp} = 1.0$ indicates the default attention mask while $(1/\text{temp}) > 1.0$ means the attention mask with a sharpened distribution. As shown in the white boxes in Fig. 18, applying lower temperature can make the generated results have more conspicuous visual concept. For example, it is vague to determine whether the owl’s eyes when $1/\text{temp} = 1$ come from the first reference image or the second reference image. On the other hand, when $1/\text{temp} = 1.5$, we can see that the cream texture of the first reference image is drawn more clearly in the generated images. We empirically observe that setting $1 < 1/\text{temp} < 1.5$ can help mitigate ambiguous mixtures of visual concepts when using multiple reference images. However, note that values of $1/\text{temp} > 1.0$ may degrade image fidelity.

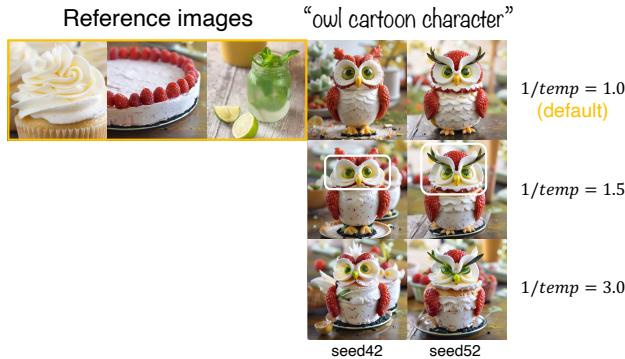


Figure 18: Visualization of the effect of temperature on the attention mask. Lower temperatures result in less ambiguous and more conspicuous application of visual concepts in exchange for the image fidelity. We empirically observe that $1 < 1/\text{temp} < 1.5$ can mitigate the ambiguity when multiple reference images yield ambiguous mixtures of visual concepts.

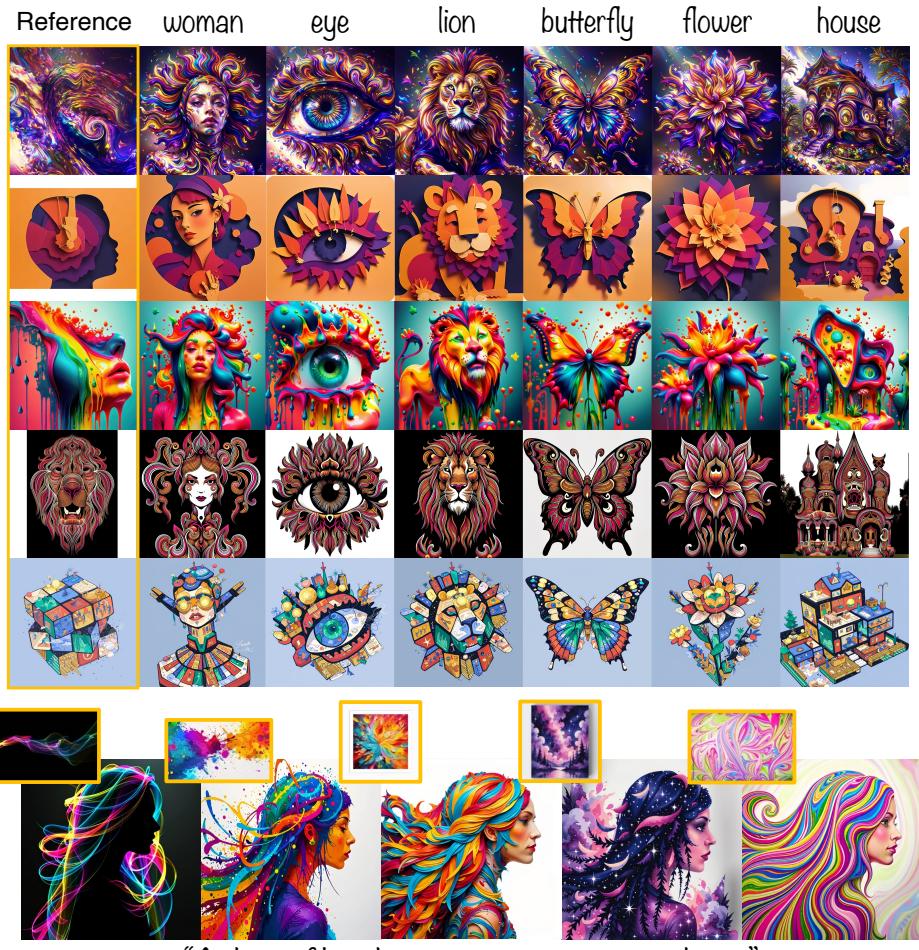
E.3 Additional Results

In this section, we show additional feasible use cases of IT-Blender in diverse design fields. Reference images and a text prompt are semantically close. More additional results with the original resolution can be found on our project page: <https://imagineforme.github.io/>.

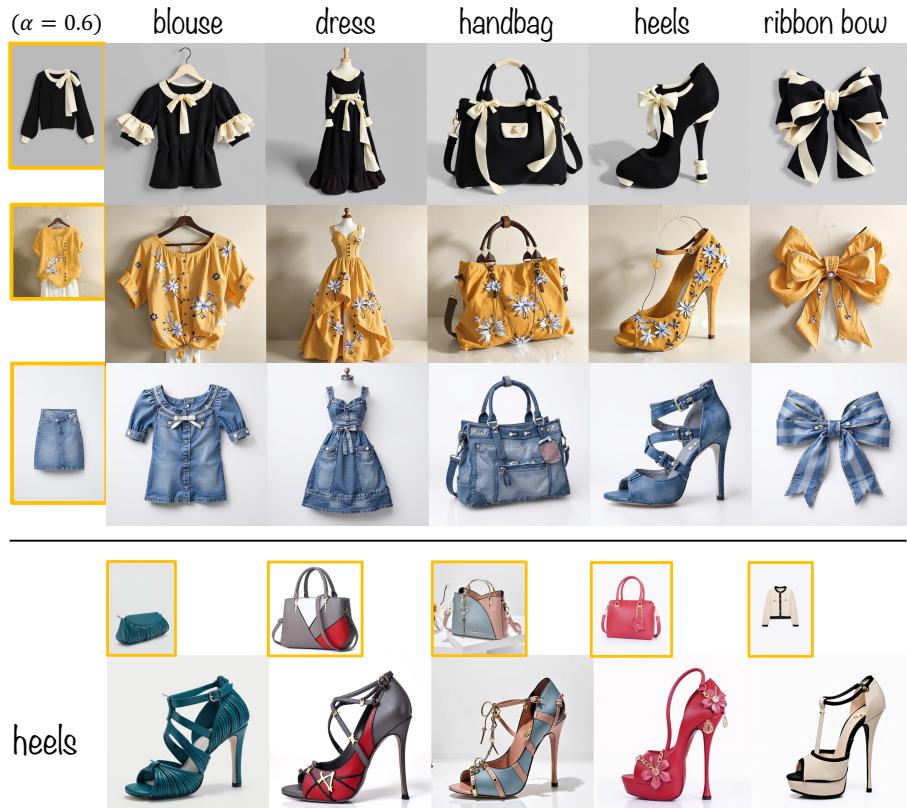


Figure 19: Feasible character design examples by IT-Blender with FLUX.

Graphic Design ($\alpha = 0.7$)



Fashion Design ($\alpha = 0.6, 0.8$)



Reference "A photo of a handbag, imaginative, creative, design" ($\alpha = 0.8$)



Figure 21: Feasible fashion design examples by IT-Blender with FLUX.

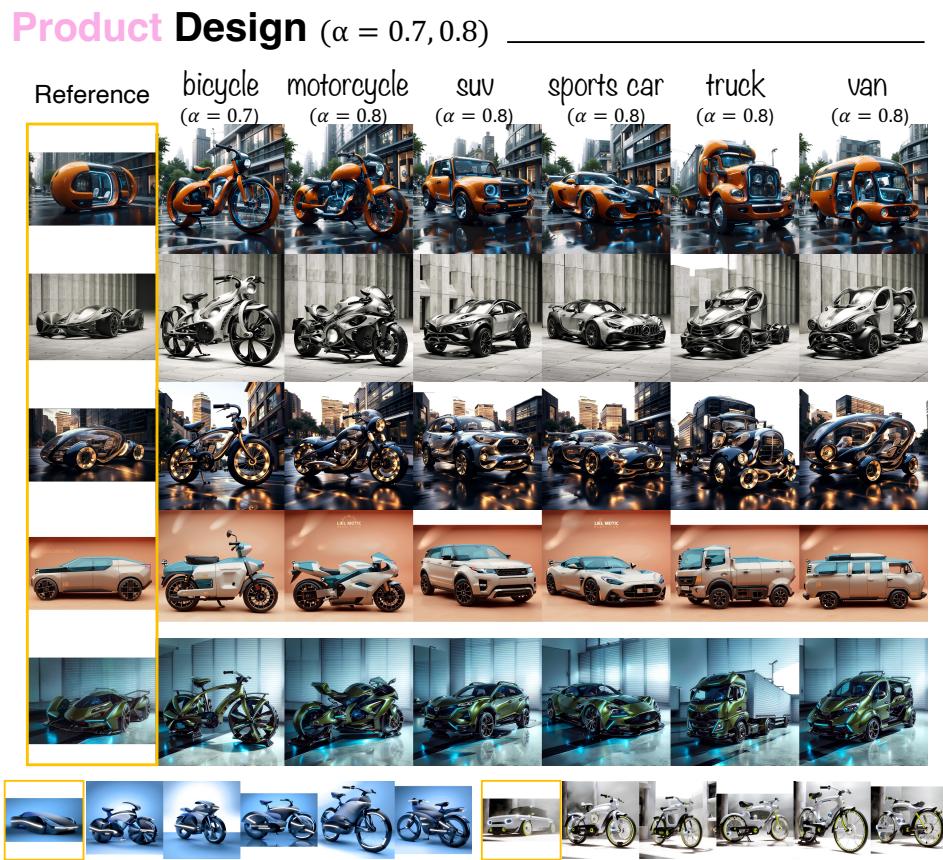
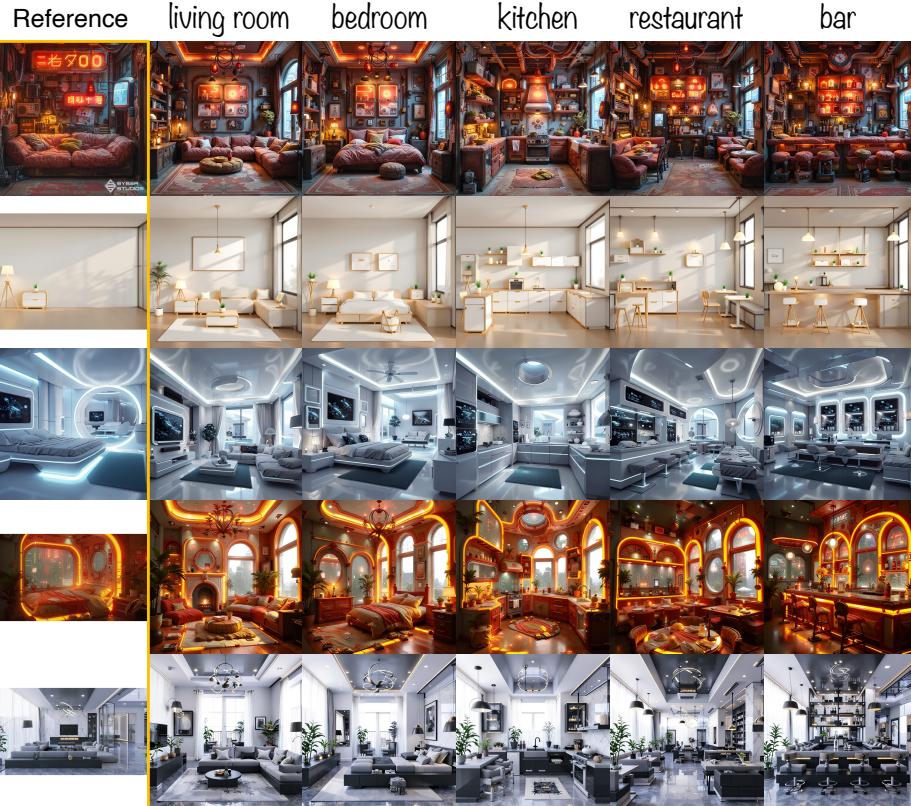


Figure 22: Feasible product design examples by IT-Blender with FLUX.

Interior Design ($\alpha = 0.7$)



Architectural Design ($\alpha = 0.6$)



Figure 23: Feasible interior and architectural design examples by IT-Blender with FLUX.

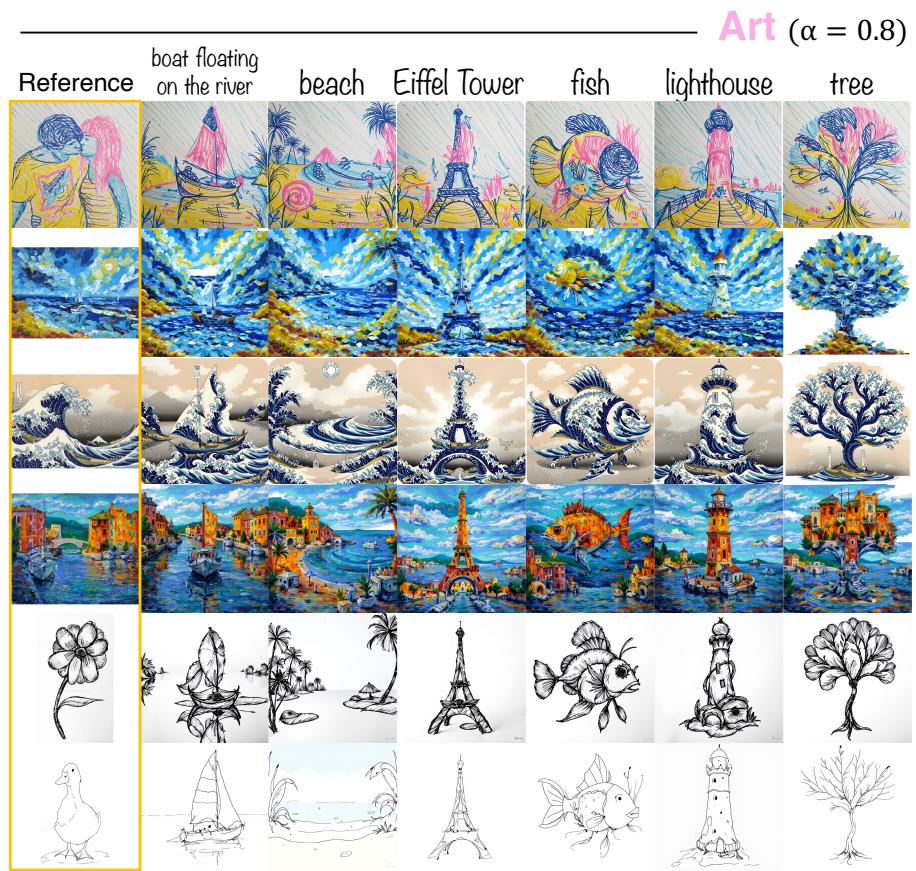


Figure 24: Feasible art examples by IT-Blender with FLUX.

F Discussion

F.1 Interesting Future Directions

Our proposed blended attention module learns to be specialized in retrieving semantic correspondence between the real image and the generated image, and it combines the visual concept with the text-guided generated image in a plausible way. We believe this technique can be useful in other creativity fields as well, such as music, text and video. For example, suppose we have music generative models. Given an arbitrary table tapping sound, the generated music would have the table tapping sound as a central theme in a plausible way. In another case, suppose that we have text generative models. Given a dialogue from a specific target person as input to the BA module, the generated text will be personalized for that individual.

F.2 Limitations

Even though IT-Blender shows impressive performance in cross modal conceptual blending, there can be several limitations. First, visual concept subtraction is not working well. It would be interesting if visual concept subtraction could be achieved.

Second, the global shape variation of the generated objects is limited. In IT-Blender, the semantics of the generated image are determined by a textual condition, and the visual concepts, such as color, texture, local shape, and material, are determined by the reference image. As can be seen in our experiments, the visual concepts can be applied with a large variation. However, the variation of the global shape (i.e., the object) is relatively limited, e.g., given “heels”, the results literally look like “heels”. We believe human designers can imagine global shape as well, which we think can be the gap with IT-Blender.

Third, there is room for fully supporting human designers. The aesthetic (i.e., how it looks) is one of the most important features of design, for which IT-Blender can significantly help human designers. However, a good human designer can consider many other features, such as functionality, usability, durability, affordability, and cultural relevance, for which IT-Blender may not be helpful. Further exploration and research are needed for AI that can consider all the important features in design.

F.3 Societal Impact

Positive societal impact. IT-Blender can augment human creativity, especially for people in creative industries, e.g., design and marketing. With IT-Blender, designers might be able to have better final design outcome by exploring wide design space in the ideation stage.

Negative societal impact. As shown in Fig. 9 and Fig. 19-24, IT-Blender can be used to apply the design of an existing product to the new products. The user must be aware of the fact that they can infringe on the company’s intellectual property if a specific texture pattern or material combination is registered. We encourage users to use IT-Blender to augment creativity in the ideation stage, rather than directly having a final design outcome.