Blending Concepts with Text-to-Image Diffusion Models

Lorenzo Olearo[1*], Giorgio Longari[1*], Alessandro Raganato[1], Rafael Peñaloza[1], Simone Melzi[1]

[1*]University of Milano-Bicocca, Milan, Italy.

*Corresponding author(s). E-mail(s): lorenzo.olearo@unimib.it; giorgio.longari@unimib.it;
Contributing authors: alessandro.raganato@unimib.it; rafael.penaloza@unimib.it; simone.melzi@unimib.it;

Abstract

Text-to-image diffusion models can generate high-quality images from individual prompts, yet it remains unclear how reliably they can fuse *two* distinct concepts into a single coherent image in a strictly zero-shot setting. Inspired by human combinational creativity, we operationalize visual concept blending here as conditioning a diffusion model on a pair of text prompts and ask whether the resulting image preserves recognizable traits of both concepts while remaining visually coherent. We study four inference-time blending strategies that intervene at different stages of the diffusion process, including embedding-space interpolation, mid-denoising prompt switching, timestep-wise prompt alternation, and layer-wise conditioning within the denoising network. We evaluate these strategies across diverse blend categories, spanning object–object blends, compound concepts, style–content combinations, and architectural landmarks, and we complement the analysis with a user study involving 100 participants. Results indicate that no single strategy dominates across all categories: timestep-scheduling approaches are often preferred for producing recognisable and well-integrated hybrids, while embedding- and layer-based interventions exhibit characteristic strengths and failure modes in specific settings. Finally, we analyse practical control factors such as prompt ordering, blend ratio, and random seed, showing how they can substantially alter outcomes and providing concrete levers for steering blends more predictably in creative applications.

1 Introduction

Human cognition is grounded in our ability to form, manipulate, and recombine concepts—abstract representations that help us interpret and navigate the world. This cognitive flexibility underlies our creativity, allowing us to generate novel ideas by combining (or *blending*) existing concepts in unexpected or previously unseen ways. Thus, even from a limited number of "base" concepts, it is possible to produce arbitrarily many new constructs. The idea of concept blending, which originates from a theory of cognition [1], has recently been studied in more detail within the areas of knowledge representation, computational creativity, and others [2], using elements of conceptual spaces [3], generalisation and specialisation [4], feature combination [5], and other symbolic transformations. As generative AI advances, and in particular text-to-image diffusion models, examining these architectures' ability to combine distinct concepts into a cohesive visual output without relying on specialized prompts, additional fine-tuning, or explicit symbolic engineering becomes increasingly important.

This paper presents several methods on how the learned representations of these models, under a zero-shot paradigm, can unify multiple ideas into a single coherent image, showing their degree of creative synthesis that can arise in diffusion-based pipelines without further intervention. We investigate the capabilities of diffusion-based text-to-image generative AI models to produce compelling images representing the blend of two concepts. Specifically, to generate visual syntheses of two or more concepts or ideas into one coherent element, relying only on their learned latent representations. In other words, we examine: (i) whether these models are able to produce blends as a zero-shot task, without additional training, fine-tuning, or prompting strategies; (ii) how effective are different blending strategies based on the diffusion architecture to yield coherent and visually meaningful combinations; and (iii) what is the role of factors like the semantic similarity of the prompts, random seed selection, and specificity of the prompts employed in the ultimate success of the produced blend. We do not claim a formal metric of conceptual distance; instead, we study semantic similarity qualitatively via categories (same vs. different classes, etc.) and via observed outcomes in the blending process, and quantify prompt similarity (proxy) via CLIP cosine similarity. Through this lens, we highlight important limitations in the compositional capacities of diffusion models, even as we uncover the surprising potential for creative synthesis that emerges from their training. Our work adopts text-to-image diffusion models [6–8] as a testbed for this phenomenon by treating each concept as a (typically single-word) textual prompt, for instance, "lion" or "cat", and leveraging the model to produce their visual representations. By doing so, we bridge the gap between mental concept representations and their image-based instantiations, examining whether the recent generative pipelines capture a form of compositional creativity akin to what can be observed in human cognition. To our knowledge, this is the first comprehensive investigation that systematically compares multiple zero-shot blending techniques using a user-study–backed analysis, relying purely on the learned latent structure of large diffusion models.

In our recent work [9, 10], we proposed a preliminary approach to concept blending in diffusion models, showing initial evidence of their zero-shot blending capabilities. The present paper significantly expands that investigation in several key directions:

(i) we clearly describe and motivate the blending methods employed, providing a comprehensive analysis of their fundamental properties; (ii) we introduce and motivate new categories of blends, broadening our experimental scenarios; (iii) we enhance our user study with a larger participant base and additional evaluation metrics; (iv) we extend the blending framework to multiple new domains beyond straightforward tangible prompts; and (v) we present novel insights into the role of seed variability in the diffusion pipeline.

To provide a thorough account, we conducted a user study with 100 participants who evaluated 22 concept pairs across five categories (namely, same class, different classes, compound words, choice of style, and architectural landmarks), controlling for both prompt ordering and random seeds. This systematic evaluation reveals important patterns around which methods excel in specific scenarios and how sensitive the blending process can be to seemingly minor input changes. Our main contributions are fourfold:

1. **Comprehensive Comparison of Blending Methods.** We present four zero-shot strategies, ranging from simple interpolation in the text embedding space to different U-Net conditioning, and detail each method's motivations, implementation, and practical trade-offs.
2. **Systematic Analysis of Dependencies.** We show how the quality of a blend depends on factors including prompt design and random seed selection. Our results reveal patterns, biases, and constraints in how diffusion models represent composite concepts.
3. **Novel Blending Tasks and Explorations.** We extend concept blending to broader scenarios, such as combining a specific painting with a general concept, or merging a concept with an emotion or style, evaluating diffusion models well beyond concrete object-object combinations.
4. **Empirical Findings and User Study.** We validate each approach through both qualitative illustrations and comparative user feedback. Our evaluation highlights differences between methods, offering insights into when (and why) certain blends are more successful.

Our experiments show that, although diffusion models can indeed generate compelling conceptual blends in certain cases, no single method universally outperforms the others, although some tend to behave better in general. Some strategies excel at merging highly dissimilar prompts, while others produce subtler merges for closely related concepts. Factors like the chosen seed or prompt ordering critically influence outcomes, revealing that concept blending in text-to-image systems remains sensitive and, at times, unpredictable; although this could be seen as a positive from a creativity point of view. Despite these limitations, we show that zero-shot blending, without fine-tuning, can be harnessed to produce imaginative visual hybrids, indicating a degree of embedded creativity in modern diffusion pipelines. Beyond its relevance for understanding how these models encode compositional knowledge, such blending capabilities could prove valuable for downstream AI tasks like creative design support or interactive brainstorming tools, which rely on the dynamic combination of concepts to spark new ideas [11].

While our primary focus is on single-word prompts for clarity and tractability, we acknowledge that more complex or multi-word descriptions may further challenge (or improve) blending outcomes. We also emphasize that, while our methods systematically produce blended concepts, they do not replicate human conceptual integration. The deeper cognitive processes underlying genuine creativity remain an open area of research, and it is not our claim that these techniques capture or explain them fully. Nevertheless, by investigating diffusion-based blends in a controlled setting, we offer insights into both the potential and inherent constraints of state-of-the-art generative models as tools for AI-driven creativity. To ensure reproducibility, we have made code, user-study materials, and prompt configurations publicly available on GitHub.[1]

The remainder of this paper is organized as follows. Section 2 reviews related work on concept blending from both logical and visual perspectives followed, in Section 3, by a technical overview of diffusion models. In Section 4, we describe our experimental setup and introduce the four zero-shot blending methods that we analyse. Section 5 then presents a detailed analysis of blending dependencies and limitations, followed by Section 6, where we reformulate the blending task to include styles and emotions. Section 7 discusses our user study methodology, results, and key findings. Section 8 concludes with a summary of contributions and future directions in this rapidly evolving field.

2 Related Work

We first recall conceptual blending, then review recent foundation-model approaches to blending, before focusing on visual blending.

2.1 Conceptual Blending

Conceptual blending arose as a theory of cognition that proposes that elements from diverse scenarios are continuously blended in our subconscious [12]. In the context of Artificial Intelligence, the task is often interpreted as trying to "invent" new concepts by combining the identifying features or properties of other existing concepts. A thorough study of this idea in different knowledge domains is available in [2].

Since the original description based on mental spaces [13] is not fully specified, different approaches have been proposed, targeting specific characteristics of the spaces or the resulting blend [14, 15]. For instance, methods based on image schemas [16] or description logics [5].

The task of concept blending is also closely connected to that of building *analogies* between concepts, in that one must understand the commonalities and differences between the underlying concepts. Results in analogical reasoning through word embeddings [17] suggest that embedding spaces behave, up to a point, like conceptual spaces *à la* Gardenfors [3]. Hence, it is reasonable to explore the possibility of constructing blends through an exploration of embedding spaces. Recent advances in large language models have also shown promising capabilities for *verbal* conceptual blending and analogy-driven ideation [18]. Nevertheless, translating such blends into consistent *visual* representations introduces additional challenges in maintaining structural

[1] https://github.com/LorenzoOlearo/blending-diffusion-models

coherence and semantic fidelity, and visual synthesis can provide a more direct medium for conveying complex combinations. This brings us to the task of *visual* concept blending.

2.2 Conceptual Blending in Generative Models

Recent work has begun to realize conceptual blending using *foundation models*, most prominently by leveraging Large Language Models (LLMs) to support ideation, analogy-making, and the discovery of cross-domain mappings. For instance, PopBlends [19] uses LLM-driven strategies (together with structured resources) to help users generate creative blend suggestions for pop-culture references, highlighting how language-based systems can facilitate divergent and convergent thinking during concept combination. In a design-oriented setting, Chen and et al [20] propose a foundation model-enhanced approach for combinational creativity in generative design, where an "additive" concept is fused into a "base" concept to support the exploration and refinement of candidate ideas. Similarly, recent LLM-based design work uses analogy-based structured retrieval to stimulate creative conceptual design via knowledge reuse and recombination [18].

Overall, these approaches primarily emphasise *verbal and conceptual* mechanisms for blending (e.g., suggesting analogies, mappings, or design rationales), showing that LLMs can meaningfully support the production and explanation of blends. On the visual side, Cho et al. [21] propose a diffusion adapter for blending real images and text prompts via blended attention, automating creative hybrids. Our work is complementary: rather than generating blend descriptions or assisting ideation, we focus on *visual synthesis* in text-to-image diffusion models. Specifically, we provide a controlled comparison of *inference-time* interventions (embedding-level, schedule-level, and architecture-level) that directly determine how two prompts are fused into a single image, together with an analysis of practical sensitivities such as prompt ordering, blend ratio, and random seed.

2.3 Visual Blending

To better categorize the evolution of visual blending techniques, we group prior work into three complementary directions: (i) prompt-mediated pipelines, (ii) representation-level interpolation, and (iii) diffusion-time or architectural interventions.

Prompt-mediated pipelines

An early contribution to the field of visual blending is presented in [22], where the authors employed Large Language Models (LLMs) to generate prompts that semantically blend two concepts, followed by a Generative Adversarial Network (GAN) to synthesize the image. The approach heavily relies on the construction of prompts, through *prompt engineering* to describe the image expected as output. In other words, this approach blends concepts online *indirectly* through a verbal description of how they should look.

Representation-level interpolation

In contrast, Melzi et al. 2023 approached the problem by investigating whether conceptual blending corresponds to interpolation within the latent space of the prompts. From their results, the authors concluded that the latent space partially encodes conceptual blending without further training. More recently, Leemhuis and Kutz 2024 extended this idea to further notions of *in-betweenness*. Complementary to prompt-side interpolation, Wang and Golland 2023 study how diffusion models can interpolate *between images*, providing additional evidence that diffusion dynamics can support smooth "in-between" trajectories.

Diffusion-time and architectural interventions

Building on this foundation, we conducted a preliminary exploration in [10], which serves as a direct precursor to this work. In that study, we introduced novel blending strategies designed to exploit both the iterative nature of the synthesis process and the architectural properties of the neural noise estimator in diffusion models. This paper extends our prior work by expanding the analysis of the blending methods, conducting a bigger and more comprehensive user study with new and stronger metrics to evaluate the results, exploring and discussing in depth additional applications and implications of the proposed methods.

Independently from us and in parallel to our research, Li et al. 2024 proposed *balance swap-sampling*, a novel approach for generating blended images from two textual concepts. Akin to Melzi et al. 2023, this latter work exploits the structure of the prompt latent space by embedding two concepts in a latent space and then swapping their column vectors. Along a related schedule-based direction, Kothandaraman et al. 2025 propose a Black-Scholes-inspired mixing schedule that continuously modulates the relative influence of two prompts over the denoising trajectory for concept fusion in diffusion models.

Similarly, Park et al. 2020 introduced so−called *swapping autoencoders* by encoding images in two separate latent spaces representing their structure and texture. By swapping the texture latent space between two images, the decoder is able to generate a new image with the structure of one image and the texture of another. This approach is similar to the one proposed in our previous work [10, see Section 4.4], but requires training a new dedicated decoder, which is not necessary in our approach.

A zero−shot framework which aims at transferring the visual appearance of an object to another object by leveraging its semantic knowledge was proposed by Alaluf et al. 2023. This appearance transfer is achieved by exploiting the cross-attention layers [29] of the denoising network during the synthesis process. More recently, Qiyuan et al. 2024 introduce Attention Interpolation for Text-to-Image Diffusion (AID), a training-free strategy that interpolates cross-attention maps (including prompt-guided variants) to improve consistency when interpolating or fusing conditions. Through a different line of work, Xiong et al. 2024 experimented with the similar task of *text−image* fusion by balancing text and image features in the cross-attention layer.

Very recently, Zhou et al. 2025 introduced FreeBlend, a feedback−driven latent interpolation method for concept blending in diffusion models. In contrast to our approach, which represents concepts using textual prompts, FreeBlend leverages image

embeddings as input through unCLIP [33]. Their method involves injecting different embeddings at various stages of the U-Net noise estimation process, similar to us, but incorporating an additional interpolation module and a refinement stage to enhance fine details in the generated samples. However, as their work was published after the completion of our study, a direct comparison falls outside the scope of this paper.

3 Preliminaries

Recent advances in generative AI have redefined the boundaries of creative computing, allowing machines to synthesize novel and diverse outputs across various domains. In particular, diffusion models, exemplified by Stable Diffusion [34], or DALL-E [33], have showed remarkable capacity for generating richly detailed images via iterative denoising processes, given textual prompts. Trained on large-scale image–text datasets, these models learn latent embeddings that capture both broad visual semantics and fine-grained stylistic attributes. In what follows, Section 3.1 details the fundamentals of diffusion models, describing how iterative sampling leverages latent representations to yield high-quality images. We then connect these fundamentals to the concept of blending before showing (in Section 4) how a single diffusion pipeline can seamlessly merge multiple prompts into inventive, zero-shot hybrids, with no additional training.

3.1 Diffusion Models

Diffusion models [6, 7] have rapidly gained prominence as probabilistic generative architectures for producing high-fidelity images. Conceptually, they learn to transform noise into coherent samples through two complementary phases: (i) a forward process that gradually adds noise to an image over multiple steps, and (ii) a reverse (denoising) process that trains a neural network to iteratively remove noise until converging on the data distribution.

Such models are well-suited for text-conditioned image (or text-to-image) generation. By injecting semantic cues, usually derived from a text encoder, into the denoising steps, diffusion pipelines can output images that reflect the content and style specified by the user's prompt. Among various text-to-image diffusion approaches, Stable Diffusion [8] introduces a latent-space framework to improve efficiency: images are first encoded into a lower-dimensional latent representation, and the diffusion process occurs there rather than in full-resolution pixel space. A final decoder then maps latent samples back to the visible domain.

Concretely, a common Stable Diffusion pipeline comprises: i) a Variational Autoencoder (VAE) [35] that encodes raw images into a latent space $\mathbb{R}^n$ and decodes them back to pixels, ii) a U-Net [36] with an encoder–decoder architecture, responsible for iteratively denoising latent representations at each timestep, with cross-attention layers that incorporate textual prompts, thus aligning the generated imagery with user-defined concepts, and iii) a text encoder, typically a CLIP model [37], which transforms textual prompts into prompt embeddings. These embeddings modulate the U-Net's denoising trajectory by guiding how noise is removed.

Figure 1 presents a simplified depiction of this pipeline. The latent representation of an image is gradually corrupted in the forward process; then, during sampling,

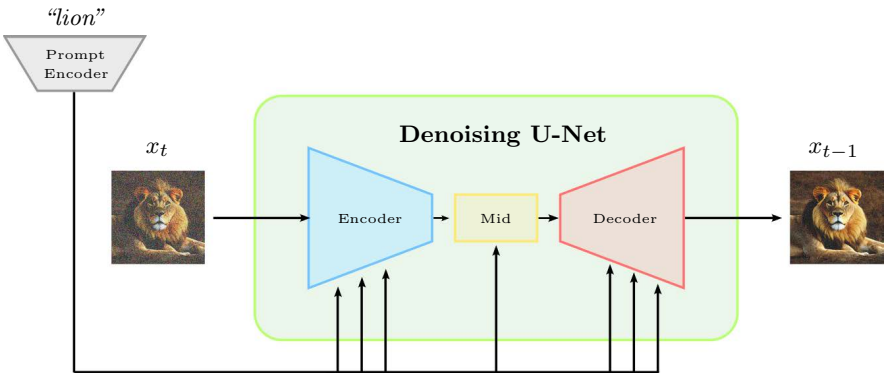

Fig. 1 Simplified overview of the Stable Diffusion pipeline. A latent representation of an image is progressively denoised in the reverse process, with text-encoder embeddings injected at various U-Net layers to guide generation. Images eventually decode back to pixel space via the VAE.

the U-Net removes noise step by step, leveraging the text encoder's guidance. Once sufficiently denoised, the latent is decoded back to pixel space. By operating in a latent domain, Stable Diffusion not only reduces computation but also grants flexible prompt-based control over the generation.

In this work, we adopt Stable Diffusion v1.4 for its open-source availability, community validation, and established hyper-parameters. While more recent variants exist, v1.4 provides a reliable and transparent foundation for investigating how textual concepts can be fused or manipulated without retraining, aligning directly with our exploration of conceptual blending in a generative context.

Diffusion Models and Visual Concept Blending

Our work seeks to operationalize the conceptual blending process using text-to-image diffusion models, which learn to generate images from text prompts. Beyond being a creative application, visual concept blending provides a useful stress test of *compositionality*: the model must combine attributes from two prompts into a single, coherent entity rather than merely retrieving a familiar depiction of either concept. We study this capability in a strictly *zero-shot* setting to probe what diffusion pipelines can achieve *without* additional training, fine-tuning, or explicit concept-specific mappings.

Diffusion models are particularly suitable for this investigation for three related reasons. First, the iterative denoising process provides a temporal axis with multiple intervention points: early steps (high noise) tend to set coarse global structure (layout, overall shape), while later steps refine local details (texture, color, fine attributes). Second, text conditioning is injected throughout the reverse process (e.g., via cross-attention), allowing prompt information to influence generation across timesteps rather than only once at initialization. Third, the separation between prompt embeddings (from the text encoder) and the denoising computation (in the U-Net) enables *inference-time* manipulations that can be applied without retraining, making diffusion pipelines an ideal testbed for controlled comparisons of blending strategies.

Rather than selecting visual features or training specialized networks, we leverage zero-shot methods, manipulating prompt embeddings, latent space, or selecting

components of the diffusion architecture, to merge multiple concepts into a single coherent output, producing a "blended image" that captures essential traits from the input concepts. Building on the intervention points above, Section 4 introduces four complementary inference-time strategies that act at different stages of the pipeline: embedding-level interpolation (`TEXTUAL`), schedule-level prompt control over denoising (`SWITCH`, `ALTERNATE`), and architecture-level conditioning within the denoising network (`USPLIT`).

In the following sections, we describe several diffusion-based blending approaches in detail (Sections 4 and 5) and evaluate them through an extensive user study (Section 7). Our investigation, thus, bridges cognitive theories of conceptual blending with large-scale generative modeling, highlighting how diffusion pipelines might be used for creative blending and where their limitations become apparent.

4 Blending Methods

This section introduces four strategies for performing concept blending within diffusion models, i) `TEXTUAL`, ii) `SWITCH`, iii) `ALTERNATE`, and iv) `USPLIT`, all of which function inside the inference pipeline to avoid any additional fine-tuning. By combining prompts directly in the model's sampling process, each method preserves recognizable features from both original concepts and enables zero-shot compositional blends, without requiring any training. We selected these four strategies as they each exemplify a distinct level or stage at which two prompts can be merged: directly in the text-encoder embedding (`TEXTUAL`), at a specific timestep mid-denoising (`SWITCH`), on alternating timesteps (`ALTERNATE`), or across the encoder–decoder split within the U-Net (`USPLIT`). This variety sheds light on how different points of intervention in the diffusion process can shape the resulting blended image.

Figure 2 provides a qualitative example of how each strategy synthesizes a single image from two distinct textual prompts. In brief:

- `TEXTUAL` merges the concepts in the text encoder's latent space by combining the respective embeddings (e.g., via interpolation) into a single vector.
- `SWITCH` begins denoising with prompt p_1 and switches to p_2 at a chosen timestep.
- `ALTERNATE` alternates the conditioning prompt at each denoising step, interweaving features from each concept across even and odd timesteps.
- `USPLIT` splits the conditioning across different U-Net stages, typically guiding the encoder/bottleneck with p_1 and the decoder with p_2.

The detailed methodology of each blending approach is presented in the following subsections, where we provide a step-by-step explanation of how prompts are injected or combined throughout the denoising process.

Experimental Setup.

In our experimental evaluation, we employ Stable Diffusion v1.4 [8] with the UniPC-MultistepScheduler [38] set at 25 steps. This version uses a fixed pretrained text encoder (CLIP ViT-L/14 [37]). All the images are generated as 512×512 pixels with

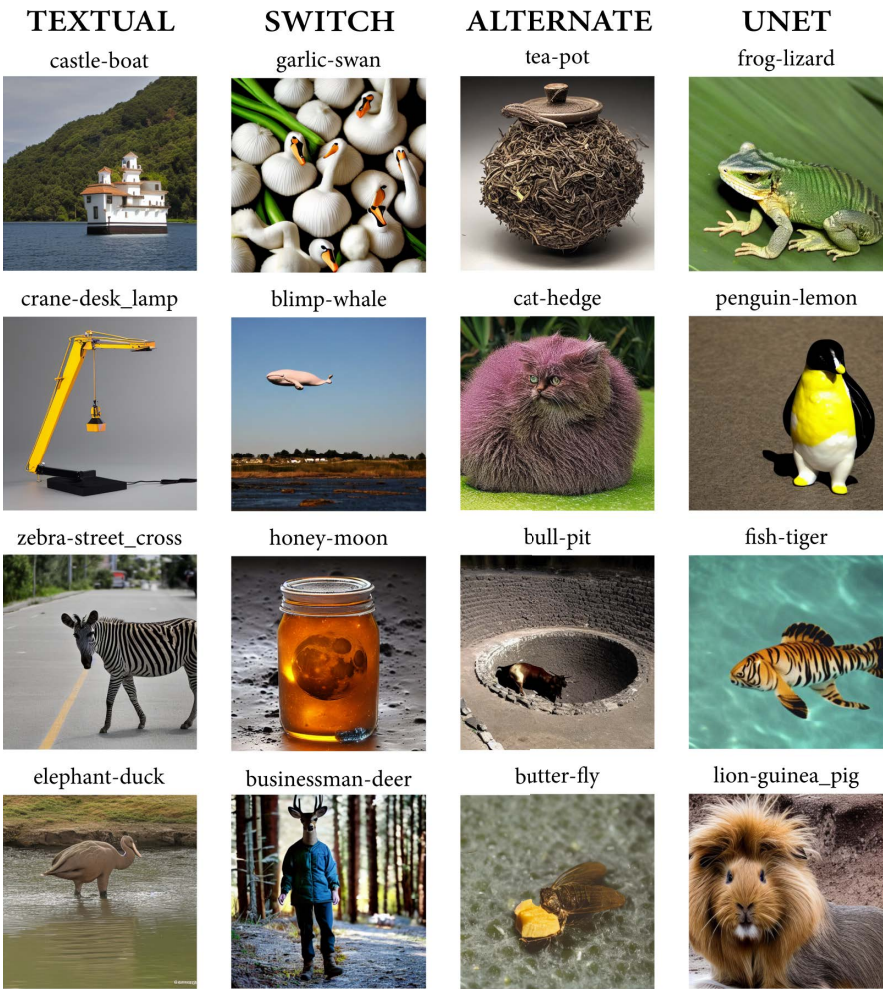

Fig. 2 Examples of the four blending methods, i.e., `TEXTUAL`, `SWITCH`, `ALTERNATE`, and `USPLIT`. Each resulting image is generated from two textual prompts displayed above it.

the diffusion process carried in FP16 precision in a latent space downscaled by a factor of 8. The conditioning signal is provided only in the form of textual prompts and the guidance scale is set to 7.5 across all the experiments. We selected v1.4 over newer variants (e.g., v1.5, v2.1) as it produces high-quality outputs for our short, single-word prompts, while balancing runtime efficiency with image fidelity.

We have integrated all four blending methods under a common codebase, ensuring consistent usage of the model $\mathscr{G}$, seeds, and prompt formatting. Unless otherwise stated, we fix the random seed to isolate the effect under analysis. The repository, complete with sample outputs, prompt configurations, and environment setup details, is publicly available on GitHub.[2]

[2] https://github.com/LorenzoOlearo/blending-diffusion-models

4.1 TEXTUAL: Interpolation in the Prompt Latent Space

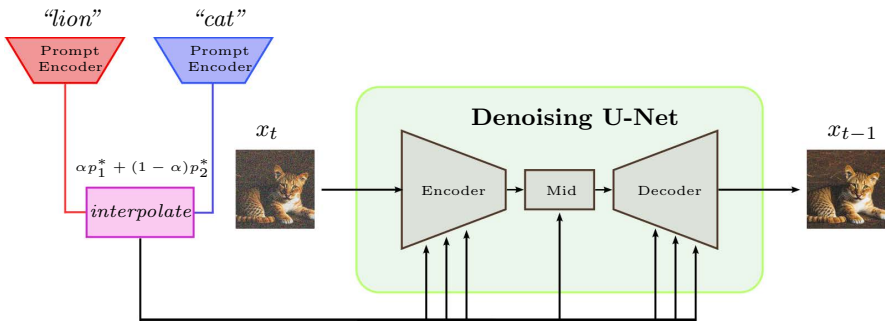

Fig. 3 The TEXTUAL blending pipeline. The two input prompts p_1 and p_2 are blended by linearly interpolating their respective latent representations p_1^* and p_2^*. The sample shown in this figure is generated using the TEXTUAL pipeline, conditioned on the prompts *"lion"* and *"cat"*.

The TEXTUAL blending method merges two prompt embeddings in the text-encoder latent space without any additional fine-tuning. Proposed by Melzi et al. [9], it reflects the idea that vector operations on the embeddings can produce compositional effects. Formally, given two prompts p_1 and p_2, their latent representations p_1^* and p_2^* are first obtained from the text encoder. We then define a weighted interpolation:

$$p^*_{\text{blend}} \;=\; \alpha\, p_1^* \;+\; (1-\alpha)\, p_2^*, \quad \alpha \in [0,1],$$

where $\alpha = 0.5$ yields a simple midpoint (i.e., the Euclidean mean) between p_1^* and p_2^*. This blended vector, p^*_{blend}, is subsequently fed into the diffusion model's cross-attention modules in the same manner as a standard single-prompt embedding would be (see Figure 3).

By adjusting α, one can shift the balance of features in the blended concept. For instance, when combining "*cat*" $\left(p_1^*\right)$ and "*lion*" $\left(p_2^*\right)$, choosing $\alpha \approx 0.3$ emphasizes the cat-like traits more heavily, while $\alpha \approx 0.7$ yields a lion-focused image. Throughout our experiments, we keep α fixed at 0.5, to produce a balanced, symmetrical blend that does not favour either concept.

As with all blending methods in this paper, we use the same random seed across different prompts and blends, ensuring that variations in the output stem primarily from the blending itself rather than the random initialization. We note that the TEXTUAL interpolation in latent space does not strictly guarantee a clear visual hybrid, instead, it captures a point in the semantic domain that might (or might not) correspond to a truly blended image, an inherent uncertainty reflecting the learned representation's complexity.

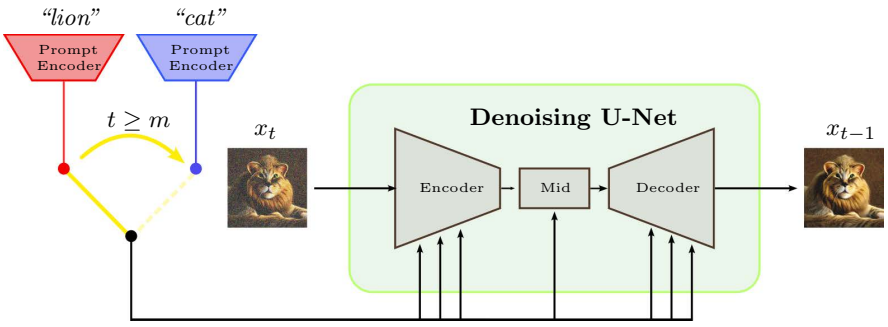

Fig. 4 The `SWITCH` blending pipeline. For the initial m iterations, the U-Net is conditioned on the first prompt p_1, then switched to the second prompt p_2. The sample shown here uses "*lion*" and "*cat*" as prompts, illustrating how the global shape is derived from "*lion*", and the finer details from "*cat*".

4.2 `SWITCH`: Prompt Switching in the Iterative Denoising Process

The `SWITCH` blending method first conditions the model on a prompt p_1 for the initial portion of the denoising schedule, and then, at a specific timestep m, switches to a second prompt p_2. Conceptually, early iterations capture high-level shape and layout from p_1, while later iterations integrate finer details from p_2.

Let the denoising process span T steps (defaulting to 25 in our experiments), and let U-Net$(\mathbf{z}_t, p)$ represent the noise estimation for the sample x at timestep t. The subsequent sample x_{t-1} is then computed by a scheduler, such as the UniPCMultistepScheduler [38]. We implement `SWITCH` as the following four-step procedure:

1. **Initialization.** Sample an initial latent $\mathbf{z}_T$ from a noise distribution using a seed s.
2. **Denoising with p_1.** For timesteps $t = T, T-1, \ldots, m$, condition the U-Net on p_1.
3. **One-time prompt switch.** Immediately before the U-Net evaluation at $t = m-1$, replace the conditioning prompt with p_2.
4. **Denoising with p_2.** Continue denoising for timesteps $t = m-1, m-2, \ldots, 0$, now conditioning on p_2.

Implementation and notation (scheduler-aware).

To avoid ambiguity in step indexing and in the role of the U-Net, we make explicit that the U-Net predicts a noise parametrization while the sampler/scheduler performs the latent update. Let $\hat{\boldsymbol{\epsilon}}_t = \text{U-Net}(\mathbf{z}_t, t, p)$ denote the noise prediction at timestep t, and let $\textsc{Step}(\cdot)$ denote one scheduler update (e.g., UniPCMultistepScheduler [38]) producing $\mathbf{z}_{t-1}$ from $\mathbf{z}_t$ and $\hat{\boldsymbol{\epsilon}}_t$. Under `SWITCH`, the conditioning prompt is selected once per denoising step as $p_t = p_1$ if $t \geq m$, and $p_t = p_2$ otherwise, implementing a

Algorithm 1 `SWITCH`: one-time prompt switching at timestep m

Require: prompts p_1, p_2, switch timestep m, total steps $T > m > 0$, seed s
$\mathbf{z}_T \leftarrow \textsc{InitNoise}(s)$
for $t = T, T-1, \ldots, 1$ **do**
  if $t \geq m$ **then** $p \leftarrow p_1$
  else $p \leftarrow p_2$
  end if
  $\hat{\boldsymbol{\epsilon}}_t \leftarrow \text{U-Net}(\mathbf{z}_t, t, p)$
  $\mathbf{z}_{t-1} \leftarrow \textsc{Step}(\mathbf{z}_t, \hat{\boldsymbol{\epsilon}}_t, t)$
end for
return $\mathbf{z}_0$

single, permanent handover at the boundary between $t = m$ and $t = m - 1$:

$$p_t = \begin{cases} p_1 & \text{if } t \geq m, \\ p_2 & \text{if } t < m. \end{cases}$$

We apply the replacement *at the beginning of the step* (i.e., immediately before the U-Net evaluation at $t = m - 1$) so that each denoising step uses exactly one prompt embedding, avoiding any partial-step mixing. This formulation ensures a clean handover, avoiding the multi-prompt mixing that could arise in partial steps.

This design deliberately exploits the coarse-to-fine nature of the reverse diffusion process: earlier denoising steps (high noise) largely establish the global layout and shape, whereas later steps refine higher-frequency details such as texture, color, and small attributes. `SWITCH` therefore induces an explicit main–modifier hierarchy: p_1 anchors the global structure, while p_2 acts as a modifier that refines the realization of that structure. The switch point m controls how long the "base" concept dominates before the "modifier" takes over; swapping the order of (p_1, p_2) (or choosing a different m) reverses this emphasis.

A key challenge is determining the optimal switch point m. Our experiments test a few values, such as 6 and 18 out of 25 (see Section 5.2). We find that if p_1 and p_2 produce visually dissimilar base images, switching too early can overwrite p_1 almost entirely, while switching too late may fail to incorporate enough of p_2. In some cases, partial blends arise if the global structure diverges significantly between the two prompts' intermediate images.

Although `SWITCH` can yield compelling blends (e.g., combining a "*lion*" shape with "*cat*" textures), abrupt changes in prompt conditioning sometimes lead to unstable transitions. Figure 4 shows how `SWITCH` merges coarse features from p_1 with subtle details from p_2. Still, the best setting of m can be highly prompt-dependent, highlighting the model's sensitivity to semantic similarity of the prompts and underlying training biases.

4.3 ALTERNATE: Prompt Alternating in the Iterative Denoising Process

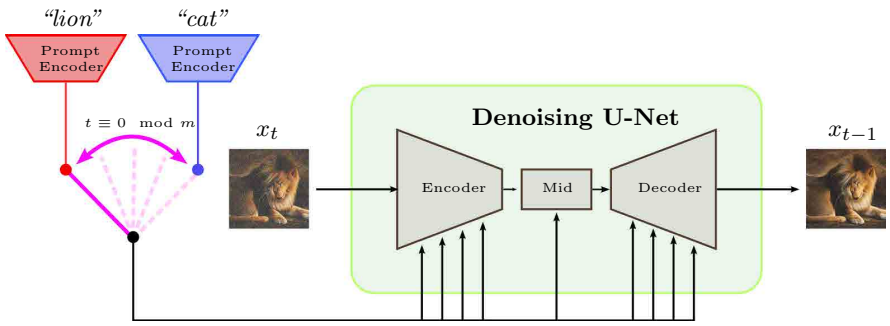

Fig. 5 The ALTERNATE blending pipeline. At the current timestep t, if $t \equiv 0 \pmod{m}$, the U-Net is conditioned on p_1; otherwise, it is conditioned on p_2.

In diffusion-based architectures, the U-Net predicts the noise parametrization at each timestep t, conditioning on the timestep and the prompt. The ALTERNATE blending method modifies this mechanism by presenting two different prompts in alternating order every m timesteps. As illustrated in Figure 5, this approach weaves features from both prompts into the final image, often producing recognizable traits of each concept.

In contrast to SWITCH, which performs a one-time prompt replacement and creates a sequential "base → modifier" dependency, ALTERNATE repeatedly alternates the conditioning prompts across the denoising steps. This repeated interleaving allows both concepts to influence the generation process throughout the trajectory, rather than committing to a single concept for all early (coarse) steps and the other for all late (fine) steps.

Formally, let $\hat{\boldsymbol{\epsilon}}_t = \text{U-Net}(\mathbf{z}_t, t, p)$ denote the noise prediction at timestep t given prompt p, and let $\textsc{Step}(\cdot)$ denote one scheduler update producing $\mathbf{z}_{t-1}$ from $\mathbf{z}_t$ and $\hat{\boldsymbol{\epsilon}}_t$. The ALTERNATE method with modulo parameter m selects the conditioning prompt at each denoising step as:

$$p_t = \begin{cases} p_1 & \text{if } t \equiv 0 \pmod{m}, \\ p_2 & \text{otherwise,} \end{cases} \qquad \mathbf{z}_{t-1} = \textsc{Step}\big(\mathbf{z}_t,\ \text{U-Net}(\mathbf{z}_t, t, p_t),\ t\big).$$

By default, we set the modulo interval to $m = 2$, i.e., we alternate prompts at every timestep: we assign p_1 to even timesteps ($t \equiv 0 \pmod 2$) and p_2 to odd timesteps ($t \equiv 1 \pmod 2$). This yields an approximately balanced allocation of denoising steps between the two prompts (the counts can differ by at most one step depending on the total number of timesteps; in our experiments we use 25 steps). This balanced scenario helps us isolate ALTERNATE's baseline behaviour. In Section 5.2, we discuss how altering this ratio, e.g., 60–40 or 70–30, can shift the emphasis between the two concepts. As with other methods, the denoising process relies on the same random seed, ensuring

that variations in output primarily reflect the degree of prompt alternation rather than chance.

In cases where p_1 and p_2 share a reasonably similar structure (e.g., "*lion*" and "*cat*"), `ALTERNATE` often produces blended images that highlight features from both prompts.

A qualitative and quantitative evaluation of `ALTERNATE` appears in Sections 5.2 and 7, where we compare multiple blending ratios and prompt pairs. There, we also discuss user feedback on how effectively `ALTERNATE` merges two concepts and analyse the circumstances under which it excels or fails to produce a satisfactory blend.

4.4 `USPLIT`: Leveraging the U-Net Conditioning

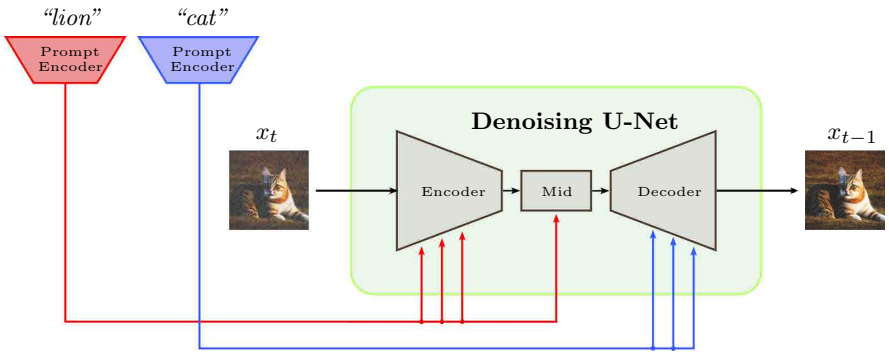

Fig. 6 `USPLIT` pipeline illustration. Encoder+bottleneck blocks (red) receive p_1^*, while decoder blocks (blue) receive p_2^*. This often retains the shape of the first concept and overlays stylistic details from the second.

The `USPLIT` blending method provides a novel paradigm for combining two textual prompts within the U-Net's architecture itself, rather than simply scheduling prompt embeddings outside the U-Net (as in `TEXTUAL`, `SWITCH`, and `ALTERNATE`). In those earlier strategies, either the text-encoder embedding or the timestep conditioning changed, but the underlying U-Net remained unaltered, conditioned on the same latent prompt. Here, however, we exploit the internal cross-attention blocks directly, injecting different prompts at different stages of the network. From a conceptual blending standpoint, the encoder and bottleneck blocks serve to compress an image latent toward the target concept's features (akin to forming a "base structure"), while the decoder reconstructs the refined latent into an output (applying "surface" or stylistic traits). By supplying one prompt to the encoder-bottleneck and another to the decoder, we effectively create a "split" that can yield images combining the form of the first concept with the style or details of the second.

In Stable Diffusion v1.4 [8], there are three cross-attention blocks in the encoder, one in the bottleneck, and three in the decoder. A single prompt embedding p^* would normally pass through all these blocks. Our method selectively switches from p_1^* to

p_2^* partway through the network. This means the U-Net's internal representation can draw on concept p_1 early on, and concept p_2 later, yielding a blended outcome.

Unlike the prior methods, which operate "outside" the U-Net (e.g., interpolating embeddings or swapping prompts across timesteps), `USPLIT` intervenes *inside* the denoising network by controlling which prompt embedding is injected into each cross-attention block. Figure 6 illustrates our default configuration: encoder and bottleneck blocks (red) receive p_1^*, while decoder blocks (blue) receive p_2^*. Intuitively, this split encourages p_1 to dominate the coarse structural representation built in early U-Net stages, while p_2 contributes finer appearance cues (e.g., texture or stylistic traits) during decoding.

To assess how sensitive this approach is to the exact split location, Figure 7 provides a qualitative ablation over all possible switch points across the seven cross-attention blocks. Columns labelled $n{-}m$ isolate the effect of injecting p_1^* into the first n blocks and p_2^* into the remaining m blocks, with all samples generated using the same random seed to control for initialization effects. This analysis shows that splits around the bottleneck or late encoder often preserve the global shape of p_1 while better transferring stylistic details from p_2, although the optimal split can vary across concept pairs.

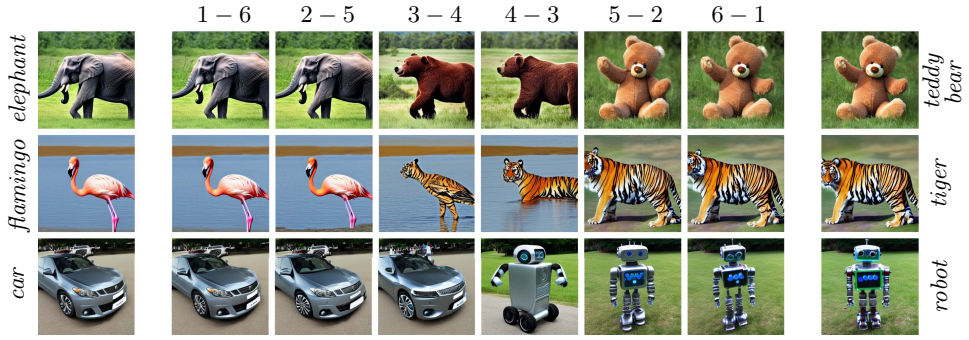

Fig. 7 Qualitative results for the `USPLIT` method. Each row uses different prompt pairs (e.g., "*elephant*" vs. "*teddy bear*"), and columns labelled $n - m$ show different `split` choices, where the first n blocks use p_1^* and the last m blocks use p_2^*. The analysis provides insights into which `split` generates better results. All variants are generated with the same random seed to isolate the effect of the split.

We now formalize the `split` operator used to implement these U-Net conditioning variants.

Formally, let $\mathbf{z}_0$ be an initial latent and let the cross-attention blocks be $\{E_0, E_1, E_2, B, D_0, D_1, D_2\}$. We denote by `split` a partition of the blocks above, indicating which of them receive p_1^* and which receive p_2^*. Then:

We found that guiding the encoder and bottleneck blocks with p_1^* and the decoder blocks with p_2^* (Figure 6) yields stable and visually appealing blends for many concept pairs. Thus, we present our main results with this `split`. However, in Figure 7, we illustrate that not all concept pairs behave uniformly, some pairs benefit from switching earlier or later. Furthermore, while this analysis highlights the architectural sensitivity of this approach, the best split may also depend on seed variance and the semantic

Algorithm 2 A simplified pseudo-code snippet for the `USPLIT` approach, showing how each cross-attention block uses p_1^* or p_2^*.

$\mathbf{z} \leftarrow \mathbf{z}_0$
for $i \in \{E_0, \ldots, D_2\}$ **do**
    if i in `split for` p_1 **then**
        $p^* \leftarrow p_1^*$
    else
        $p^* \leftarrow p_2^*$
    end if
    $\mathbf{z} \leftarrow \text{CrossAttentionBlock}_i(\mathbf{z}, p^*)$
end for
$\mathbf{z}_T \leftarrow \mathbf{z}$

distance between prompts. As we discuss in Section 5, these factors can cause the output to lean more heavily toward one prompt.

In contrast to `TEXTUAL`, `SWITCH`, or `ALTERNATE`, where we alter embeddings outside the denoising network or at the timesteps, `USPLIT` inherently depends on the internal layout of the model's cross-attention blocks. We show a systematic comparison of `USPLIT` with the other three methods in Section 5, where both qualitative examples and user-study feedback help assess the relative strengths and weaknesses of this internal conditioning approach.

5 Dependencies in Concept Blending Procedures

Having introduced four distinct strategies for concept blending with text-to-image diffusion models in Section 4, we now examine the key dependencies that shape how these methods perform in practice. Throughout this section, we discuss with qualitative examples and user-study observations why certain prompt pairs or parameter modifications lead to coherent blends, whereas others result in unexpected or one-sided images. Specifically, we discuss how swapping the order of two prompts (e.g., "*lion*" vs. "*cat*" or "*cat*" vs. "*lion*") can favour one concept visually, depending on the method's internal biases or scheduling approach (Section 5.1), how the blend ratio between two concepts can emphasize one concept more heavily than the other (Section 5.2), how random seed variations can produce radically different results, even for the same prompt pair and blending parameters (Section 5.3), and the limitations (Section 5.4) of this work that can alter the outcome of zero-shot blending.

5.1 Symmetry

In conceptual blending theory [1], a main concept typically provides the core structure, which is then enriched by modifier concepts that introduce additional attributes. The resulting blend captures much of the main concept's essence while integrating features from the secondary one. In our setting, `TEXTUAL` is the only purely symmetric method: it averages (or interpolates) two embeddings, so reversing the prompt order does not affect the outcome. By contrast, the other three methods, i.e., `SWITCH`, `ALTERNATE`,

and USPLIT, exhibit inherent asymmetry, meaning that swapping the order of two prompts often produces notably different images.

This asymmetry becomes particularly visible with compound words like pitbull, whose modern usage (a dog breed) differs from its historical roots ("bull in a pit"). For instance, when blending "*pit*" and "*bull*" via SWITCH, which always begins denoising with the first prompt, or USPLIT, which conditions the encoder on the first prompt and the decoder on the second, selecting which prompt is principal determines the final appearance. Even compound words such as "rain bow" or "jelly fish" can show similar order-sensitivity, where the "main concept" might dominate shape and colour. As illustrated in Figure 17, TEXTUAL and ALTERNATE tend to prioritize the second prompt, while SWITCH and USPLIT tend to privilege the first. This unexpected divergence led us to assign the first prompt as the main concept and the second as the modifier to maintain consistency. Consequently, we generate "*bull-pit*" rather than "*pit-bull*".

Figure 8 further shows this phenomenon with the concepts "peacock" and "eagle". Reversing the order can lead to a peacock adopting an eagle's beak or an eagle adopting peacock plumage. Such outcomes highlight a main–modifier dynamic shaped by each method's architecture: TEXTUAL's vector averaging is order-invariant, ALTERNATE partially shifts focus to the second prompt, and SWITCH or USPLIT systematically bias the image toward whichever prompt is introduced first. Although this may be beneficial for practical creative tasks, for example, a user intentionally favouring one concept, unaware designers could be surprised by how drastically prompt order alters results. Our user study (see Section 7) corroborates that participants notice these asymmetries, often judging images based on which concept they perceive as the core. Moreover, the effect is generally more pronounced with single-word prompts, where reversing the main–modifier assignment fully flips the blended outcome. More complex or multi-word prompts may compound this asymmetry even further, introducing additional variability in which concept takes precedence.

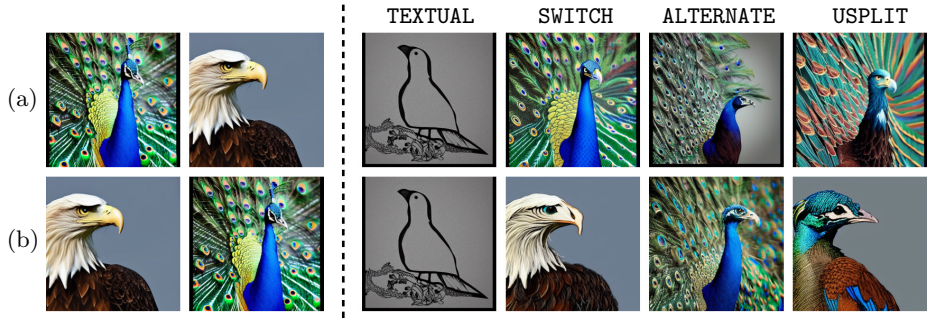

Fig. 8 Comparison between (a) blending "*peacock*" with "*eagle*" and (b) reversing the order to "*eagle*" with "*peacock*". All images share the same initial noise.

5.2 Blend Ratio

When blending two concepts into a single image, it is often desirable to control the relative emphasis each concept exerts. This blend ratio can be viewed as the proportion

of conceptual features drawn from each prompt. Although each method presented in Section 4 offers some mechanism to adjust this ratio, they differ in how that adjustment is realized and how the resulting effect can be interpreted. Figure 9 shows a few examples in which the ratio between "*corgi dog*" and "*duck*" is varied.

From a conceptual blending perspective, it can be defined as selectively adjusting how much of each concept is depicted into the blend projection, where only certain features of each concept are retained, emphasizing the "base" concept while subtly weaving in attributes from the "modifier", or vice versa.

TEXTUAL.

Of all the methods, `TEXTUAL` provides the most direct ratio control, since it uses linear interpolation between prompt embeddings. Specifically, it defines a blended vector p^*_{blend} as $p^*_{\text{blend}} = \alpha\, p^*_1 + (1-\alpha)\, p^*_2$, where $\alpha \in [0,1]$. Setting $\alpha = 0.5$ weights both prompts equally, while more extreme values favour one concept almost entirely.

SWITCH.

In the `SWITCH` method, a single timestep m determines the switch from the first prompt to the second. Shifting m closer to the start yields a blend dominated by p_2, whereas delaying the switch gives more weight to p_1. This is straightforward to tune in principle but can lead to *cartoonification* if the switch is abrupt. As seen in row (c) of Figure 9, excessively late switching sometimes induces a loss of high-frequency details.

ALTERNATE.

The `ALTERNATE` method interleaves each prompt on alternating timesteps of the diffusion process. Controlling the ratio thus involves allocating more timesteps to p_1 or p_2. For instance, dedicating 60% of the timesteps to p_1 and 40% to p_2 yields a correspondingly skewed blend.

USPLIT.

The `USPLIT` approach is the most complex, as it manipulates which layers of cross-attention see each prompt. In principle, a $25-75\%$ blend can be achieved by conditioning one or two early layers on $\mathbf{p}^*_1$ and the remaining layers (including the bottleneck) on $\mathbf{p}^*_2$. However, an odd number of cross-attention blocks requires choosing the middle block, i.e., the bottleneck, to be used for one prompt or the other. As discussed in Section 4.4, we opt to encode the "base" concept in the encoder+bottleneck, then overlay features from the second prompt during the decoder. This design echoes the conceptual blending idea of a "main structure" with added "modifier" details.

Figure 9 shows how each method's ratio-control mechanism can produce distinct outcomes. `TEXTUAL` yields a smooth interpolation in embedding space, `SWITCH` enforces a sharp change at a chosen timestep, `ALTERNATE` distributes timesteps between the two concepts, and `USPLIT` changes which cross-attention blocks are conditioned on each prompt. While all methods allow ratio adjustments, their internal processes, and potential pitfalls (later discussed in Section 5.4), require careful consideration when trying to balance two concepts precisely.

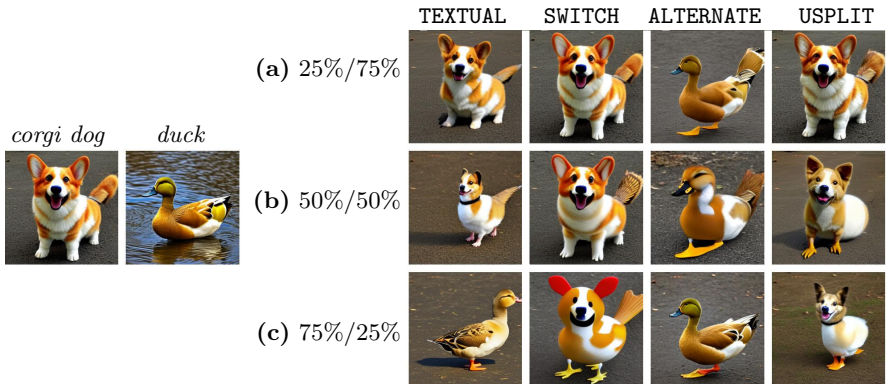

Fig. 9 Comparison between the blending of the concepts *"corgi dog"* and *"duck"* with different blending ratios. Row **(a)** 25% *"corgi dog"* and 75% *"duck"*, row **(b)** both concepts are equally weighted, and row **(c)** 75% *"corgi dog"* and 25% *"duck"*. The same seed is used across all samples.

5.3 The Seed Dependency

A notable complexity of diffusion models is their reliance on the random noise from which images are sampled [39]. In a standard text-to-image diffusion pipeline with classifier-free guidance [40], the initial noise strongly influences colour palettes, object poses, and other low-level traits that persist in the final image, regardless of the semantic content of the prompt. Blending methods are not exempt from this effect: even though they merge two concepts, the seed-specific features can still dominate certain aspects of the generated outcome.

While our user study (see Section 7) does not explicitly address seed variability, we highlight here how consistent each blending method is across different seeds. In practice, we tested ten distinct random seeds for each concept pair. Methods like `USPLIT` tend to produce subtle yet stable blends across seeds, whereas `ALTERNATE` sometimes yields no discernible fusion at all but can, on rare occasions, generate highly appealing hybrids. These observations align with a broader consensus in classifier-free guided diffusion, where similar prompts produce visually similar images when initialized with the same noise, not necessarily in semantic details, but in attributes like pose, colour gradients, or compositional structure.

In our blending context, this implies that if we generate a "*lion*" and a "*cat*" image from the same seed separately, the two images often share broad visual or spatial similarities, factors that then carry over into a subsequent blend. For consistency and to focus on the semantic aspects of blending (rather than seed-induced variance), we fixed the same seed for each prompt pair. Figure 10 illustrates how the blending output changes when we vary the seeds for the individual prompts, for the blend itself, or both. Figure 17 provides additional examples generated from the same initial noise across multiple prompt pairs, illustrating that the methods exhibit consistent qualitative behaviours (including main–modifier asymmetries for the asymmetric schedules) beyond a single example.

Practical implications of control factors

Beyond diagnosing failure modes, the dependencies above can be used as simple *control knobs* when producing blends. (i) *Prompt ordering* is not neutral: in methods with an explicit main–modifier asymmetry (notably `SWITCH` and `USPLIT`), placing as p_1 the concept whose *global shape/layout* should dominate and as p_2 the concept whose *local texture/style details* should be injected yields more predictable outcomes. (ii) *Prompt similarity* provides a rough prior on blend difficulty: pairs with lower semantic similarity tend to require stronger interventions (e.g., later switches / more p_2 allocation, or architectural splitting) and are more prone to collapse into one concept, whereas higher-similarity pairs are often handled adequately by lighter-weight strategies such as `TEXTUAL` interpolation. (iii) *Random seed* controls a substantial portion of low-level structure, regardless to the selected blending method; thus, multiple seeds should be sampled when the goal is creative exploration, while a fixed seed should be used when comparing methods to ensure reproducibility. Taken together, these factors translate into actionable guidance for interactive creative tools: select the order to set the intended dominance, tune the allocation (ratio/switch point/split) to balance influence, and sweep seeds to trade off determinism versus diversity.

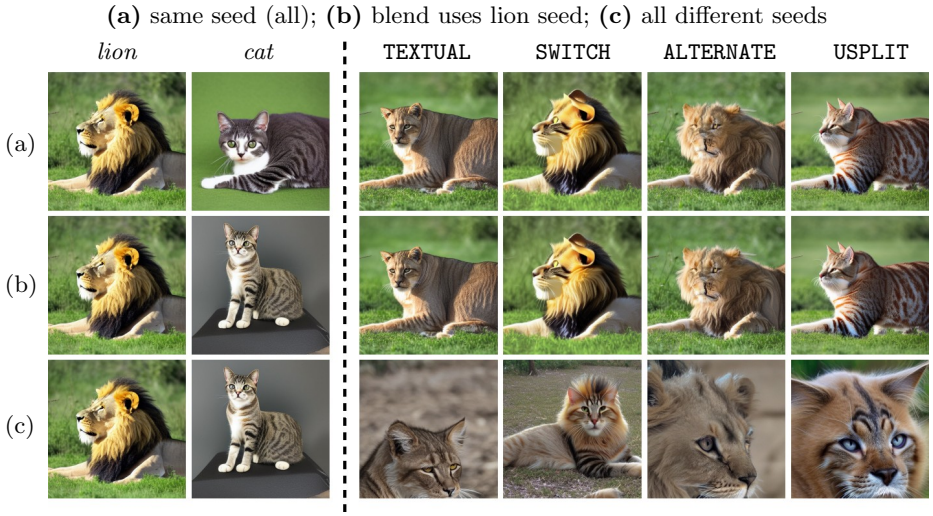

Fig. 10 Comparison of results from blending the prompts "*lion*" and "*cat*": **(a)** all images are generated from the same initial noise, **(b)** individual prompts are generated from different noise, but blending starts from the same noise as the first image, and **(c)** both individual prompts and the blending process begin with different noises.

5.4 Limitations

Each blending method discussed in this paper inevitably comes with methodological and architectural constraints that can affect the quality, consistency, and interpretability of the resulting images. Beyond these specific issues, the broader challenge lies in

ensuring that conceptual blending via neural embeddings aligns well with cognitive expectations derived from classical theories such as Fauconnier and Turner's [1].

`TEXTUAL` relies on linear interpolation in the text-encoder's geometric space, yet we have no guarantee that such interpolation mirrors how humans conceive a smooth transition between concepts. Figure 9, for instance, shows cases where different ratios depict near–single-concept images, reflecting the mismatch between embedding-space geometry and user intuitions. `SWITCH` can cause cartoonification when the prompt changes late in the sampling process, as illustrated in row (c) of Figure 9 and further discussed in Section 7. Determining precisely when to flip the prompt remains a non-trivial task. `ALTERNATE` occasionally yields results that lack aesthetic appeal: highly skewed split points can fragment the final image rather than integrate it, especially for simple or short prompts (e.g. "*cat*", "*lion*"). `USPLIT`, while powerful, is challenging to control because assigning each prompt to different cross-attention blocks can result in subtle or incomplete blends, and is sensitive to architectural specifics like the number of cross-attention layers (Section 4.4 and Figure 7).

Beyond the methods themselves, our reliance on Stable Diffusion v1.4 imposes further constraints. We selected this model for its relative simplicity and manageable computational demands, but newer pipelines or fine-tuned architectures may handle short, general prompts (the kind tested here) more reliably. Likewise, the CLIP text encoder [37], trained on large-scale image-caption pairs, can struggle with minimal prompt descriptions.

An additional limitation emerges from the subjectivity of judging blended images. What one user sees as an elegant synergy, another may perceive as awkward. Our user-study findings (Section 7) hint at how personal aesthetic preferences influence whether a given blend is considered "successful". This effect also interacts with the number and complexity of prompts tested: while we focus on short, single-word prompts (often compound-word splits) to highlight core blending behaviours, more elaborate or multi-word prompts might expose different limitations or require more complex methods of ratio and seed control.

A further limitation is that our experimental blends largely rely on concepts with stable visual referents (concrete objects) or on well-defined stylistic descriptors (e.g., named artistic styles). Fully abstract or mixed blends (e.g., "love+chair", "time+love") remain unexplored: they likely require an additional grounding step that maps abstract notions to *visual proxies* (e.g., representing "love" via culturally dependent symbols such as a heart), or alternative prompting/conditioning mechanisms. Extending our framework to truly abstract only inputs and systematically evaluating different grounding strategies is an important direction for future work.

Although we consider these constraints significant, they do not undercut the overall potential of zero-shot concept blending via text-to-image diffusion. To showcase each method's inherent strengths (rather than flaws from model or prompt limitations), we selected their best outcomes for comparative analysis. Our best outcomes were chosen through a combination of qualitative inspection by the authors and pilot user feedback, ensuring we highlight each technique's characteristic performance under favourable conditions. Ultimately, we expect future refinements, such as adopting next-generation diffusion models, more advanced text encoders, or domain-specific fine-tuning, to

address many of these issues, further bridging the gap between embedding space manipulations and human-like conceptual integration.

Positioning relative to recent blending methods

A strict quantitative comparison with recent zero-shot fusion methods is out of scope for this work, as our goal is diagnostic, to characterize how different inference-time intervention points within a single diffusion pipeline affect the blending task, whereas prior approaches often optimize for different targets and introduce additional inputs, computation, or selection stages beyond prompt-only conditioning. For example, FreeBlend [32] is training-free but relies on unCLIP image conditions (images generated from text) and a staged denoising procedure (initialization, blending, refinement) together with a feedback-driven interpolation mechanism. TP2O / balance swap-sampling [25] generates multiple candidates by swapping elements of text embeddings and then applies CLIP-distance balancing and a segmentation-based selection step to select promising combinations. More recently, AID [30] produces hybrids by interpolating cross-attention maps inside the denoising U-Net (with inner/outer interpolation and a prompt-guided variant), promoting smooth transitions between conditions. Kothandaraman et al. [26] propose a Black-Scholes-inspired rule that adapts prompt mixing over timesteps using CLIP-based scores and scheduler statistics, yielding a continuous, step-wise mixing policy. In contrast, our study focuses on prompt-only inference-time interventions that operate within a single diffusion sampling trajectory (embedding-, schedule-, or architecture-level) and intentionally avoids candidate filtering, so as to characterize how different intervention points affect blending behavior under controlled seeds and ratios.

Qualitative comparison (overlapping prompt-only setting)

To situate our prompt-only blending strategies within recent diffusion-time fusion work, we include a small side-by-side qualitative comparison with two recent methods that intervene through different mechanisms (Figure 11). Qiyuan et al. [30] propose *Attention Interpolation for Text-to-Image Diffusion* (AID), a training-free method that produces hybrids by *interpolating attention maps inside the denoising U-Net* (inner/outer interpolation, with an optional prompt-guided variant), encouraging smooth transitions between conditions. Kothandaraman et al. [26] propose a *Black-Scholes*-inspired prompt-mixing rule that *adapts conditioning over timesteps* via CLIP-based scores and scheduler statistics, yielding a continuous, step-wise mixing policy. The baseline images in Figure 11 come from the respective papers, while the right block reports our four blending methods (`TEXTUAL`, `SWITCH`, `ALTERNATE`, `USPLIT`) on Stable Diffusion v1.4 using the same prompt pairs. Because backbones and reporting protocols differ, this figure is intended to highlight typical behaviors and pipeline trade-offs rather than provide a controlled benchmark. We leave for future work the construction of a full benchmark that unifies inputs, compute budgets, and evaluation targets across methods.

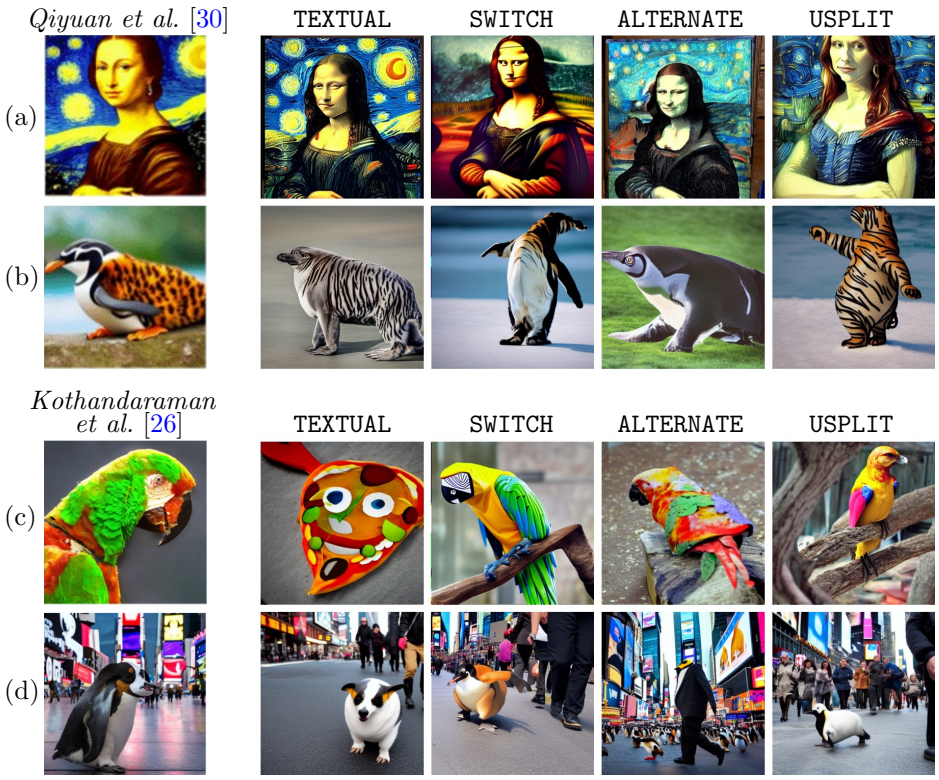

Fig. 11 Qualitative comparison with recent diffusion-based blending methods with the following prompts selected from their results: **(a)** "*A painting of Mona Lisa, extremely detailed, high quality*" and "*A painting of Starry Night, extremely detailed, high quality*"; **(b)** "*A photo of a penguin, extremely detailed, high quality*" and "*A photo of a tiger, extremely detailed, high quality*"; **(c)** "*parrot*" and "*pizza*"; **(d)** "*A corgi walking in Times Square*" and "*A penguin walking in Times Square*". The leftmost column show the output from the corresponding prior work (top block: AID by Qiyuan et al. [30]; bottom block: concept blending with Black-Scholes by Kothandaraman et al. [26]).

6 From Tangible to Intangible: Reimagining Conceptual Blending

In this section, we focus on abstract and subjective prompts, where one concept might represent a specific artistic style, an emotional descriptor, or a cultural theme. This reformulation expands the zero-shot blending paradigm as a style transfer method [28], but without the additional training or specialized architectures that style-transfer approaches often require.

From a conceptual blending perspective, abstract ideas function much like intangible "modifiers", merging their stylistic or emotional essence with the core structure of a more concrete concept. In a text-to-image context, this is highly practical: styles are frequently used via prompt engineering, yet as [28] show, an object's appearance can be transferred to another concept in a purely zero-shot manner, leveraging diffusion models without fine-tuning. We build on this idea using the same four methods

(`TEXTUAL`, `SWITCH`, `ALTERNATE`, and `USPLIT`) discussed in Section 4, simply substituting intangible prompts (e.g., a painting style or an emotional expression) for one of the two concepts.

Sections 6.1, 6.2, and 6.3 explore how intangible concepts, such as paintings, emotions, or architecture, can modify a tangible subject. Although these tasks bear similarities to classical style transfer, our approach maintains the flexibility and compositional properties of zero-shot blending.

6.1 Paintings

One of the most compelling applications of intangible blending involves mimicking the style of a famous painting in a given scene. Notable examples include van Gogh's bold, swirling patterns, Monet's delicate impressions, and Picasso's geometric abstractions. In this setting, we use the painting (identified by its title and author) as the modifier, applying its unique look to a main concept that defines the core content of the image. This choice differs from earlier experiments (e.g., blending lion and cat), which used broader, more general concepts with many possible visual realizations. Instead, each painting has a well-defined style and composition that we expect a diffusion model to have learned during its large-scale training.

We selected five iconic paintings for their visually distinct characteristics, i) "*Starry Night by van Gogh*", ii) "*Water Lilies by Claude Monet*", iii) "*The Scream by Edvard Munch*", iv) "*The Girl with a Pearl Earring by Johannes Vermeer*", and v) "*Guernica by Pablo Picasso*". Thus, the diffusion model's learned representation would be tested against diverse colour palettes and strong stylistic signals. In this setup, we treat each painting (described by its title and artist) as the modifier prompt, blending it with a main concept that defines the subject's semantic content, such as "A portrait of a man". Figure 12 contrasts these blended results with a baseline approach that simply appends "*in the style of [painting]*", i.e. if the main concept is "*A portrait of a man*" and the desired style "*Water Lilies by Claude Monet*", the prompt would be "*A portrait of a man in the style of Water Lilies by Claude Monet*".

Overall, `SWITCH` and `USPLIT` stand out in integrating the painting's style without losing the underlying content. For instance, several methods (except `ALTERNATE`) alter the man's face to resemble Van Gogh's self-portraits, hinting that the model strongly associates Van Gogh's style with a particular facial structure, likely due to extensive training images of his self-portraits. We observed no parallel effect for the other paintings, underscoring how specific artist and subject mappings can emerge from large-scale training data.

6.2 Emotions

A further application of style blending is modulating the mood of an image by mixing a visual subject with an emotion. Following the style blends discussed earlier, we now set the main concept to "*A portrait of a man*" and the modifier to a simple emotion word (e.g., "*anger*", "*happiness*", and "*sadness*").

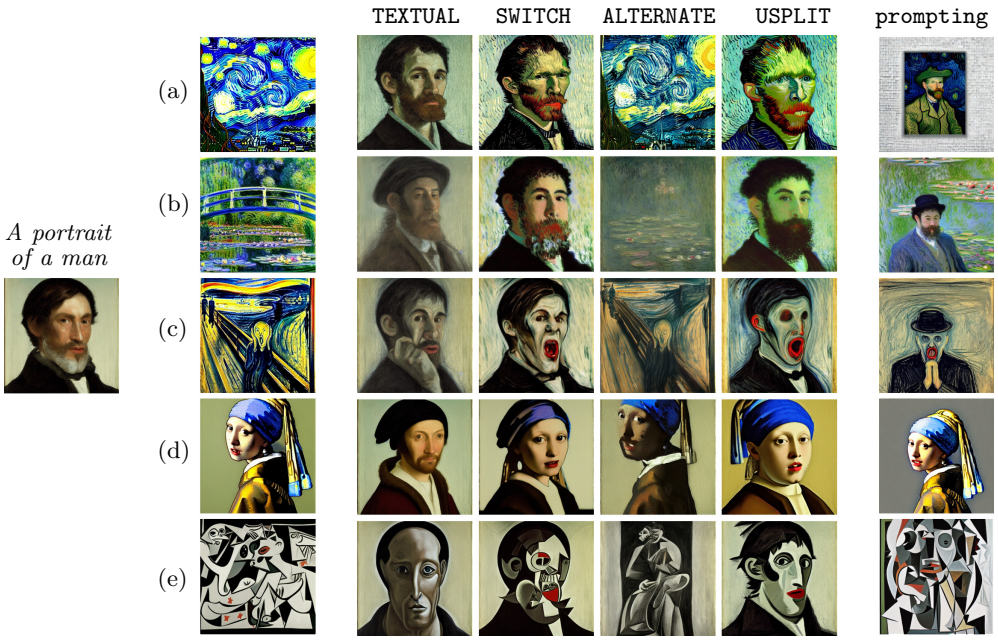

Fig. 12 Style blending examples for the main concept "*A portrait of a man*", combined with (a) "*Starry Night by van Gogh*", (b) "*Water Lilies by Claude Monet*", (c) "*The Scream by Edvard Munch*", (d) "*Girl with a Pearl Earring by Johannes Vermeer*", and (e) "*Guernica by Pablo Picasso*". The last column shows the baseline result using a single "*in the style of*" prompt. All images are generated with the same random seed.

Fig. 13 Samples from blending the main concept "A portrait of a man" with three emotions: (a) "anger", (b) "happiness", and (c) "sadness". All images share the same initial random seed.

As illustrated in Figure 13, each blending method successfully evokes the target emotion, although the "best" outcome may depend on personal preference. Nevertheless, a few distinct behaviours emerge: `SWITCH` and `USPLIT` typically edit the subject's features to reflect the desired emotional state, e.g., furrowing brows or reshaping the mouth, while `TEXTUAL` and `ALTERNATE` may replace the subject entirely, producing a more drastic transformation. In `TEXTUAL`'s case, this follows from the learned latent space of the text encoder (Section 4.1); a single-word prompt like "sadness" can correspond to multiple visual contexts in the large-scale CLIP training set. Yet when paired with a clear prior (like "*A portrait of a man*"), it can guide the pipeline to generate a meaningful emotional display.

Overall, we highlight how zero-shot blending can capture subtle affective cues, despite the abstract nature of emotions compared to well-defined painting styles. Finally, while style-based blending and emotion-based blending share certain similarities, affective transformations are arguably less predictable, given that emotions are inherently more subjective than established artistic styles. The next subsection (Section 6.3) extends intangible blending to architecture, examining how building aesthetics can act as a conceptual modifier for otherwise conventional prompts.

6.3 Architecture

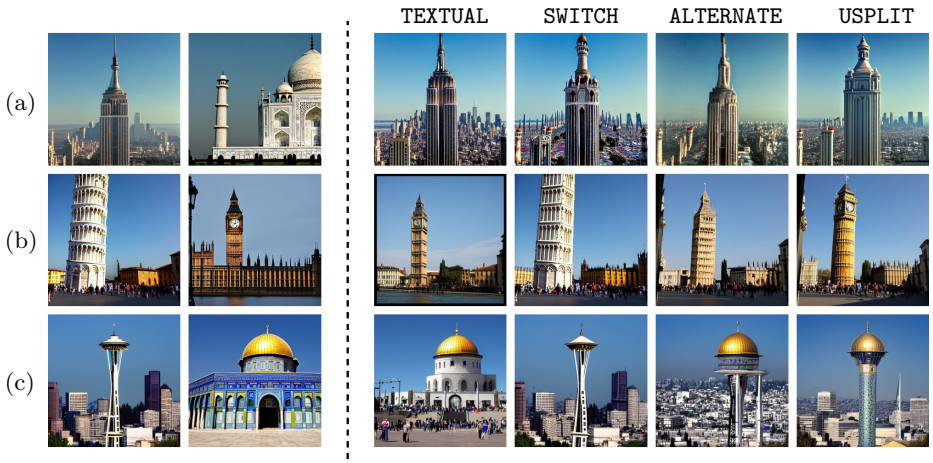

Fig. 14 Results from blending two architectural landmarks. Row (a) shows the blending of "*Empire State Building*" with "*Taj Mahal*", (b) "*Leaning Tower of Pisa*" with "*Big Ben*", and (c) "*Seattle Space Needle*" with "*Dome of the Rock*". All images share the same initial random seed.

Building on our explorations in style (Section 6.1) and emotion (Section 6.2), we now turn to blending two distinct architectural concepts. This task is arguably more challenging than the previous ones, given the difficulty of faithfully reproducing complex geometries and fine architectural details [28]. For this reason, we selected iconic landmarks that feature distinct shapes and enjoy international recognition. From

a conceptual blending perspective, each landmark contributes its defining structural features, which the model attempts to combine into a single, novel architectural entity.

We selected iconic landmarks such as the "*Empire State Building*", "*Taj Mahal*", "*Leaning Tower of Pisa*", "*Big Ben*", "*Seattle Space Needle*", and "*Dome of the Rock*", all notable for their recognizable shapes and distinct architectural cues, providing a test of the model's capacity to integrate complex geometry. Figure 14 presents examples of these architectural blends. Here, we can see how `ALTERNATE` and `USPLIT` tend to yield more compelling fusions of structural forms, while `TEXTUAL` typically fails to capture the intended overlap, and `SWITCH` often produces simplified or underspecified outputs. An illustrative case is blending the "*Leaning Tower of Pisa*" with "*Big Ben*". The `USPLIT` method creates a rounded, tilted tower reminiscent of Pisa's hallmark shape, combined with a clock face and colour palette influenced by Big Ben, thereby preserving key features from both landmarks. In contrast, `TEXTUAL`'s latent-space interpolation rarely retains the spatial cues necessary to convey two overlapping architectures, and `SWITCH` can suffer from a lack of detail when switching prompts mid-generation. Overall, these findings show that zero-shot architectural blending remains a difficult challenge for a simple stable diffusion backbone, yet methods like `USPLIT` and `ALTERNATE` can still produce convincing hybrids of iconic structures.

7 Validation and Results

The quality of any conceptual blend often depends on the nature of the concepts being combined. To comprehensively evaluate our blending strategies, we tested each method on five distinct categories of concept pairs, as summarized in Table 1. These categories are intended to highlight divergent properties of the concepts and the resulting blends. Each category represents a different blending scenario. When two concepts belong to the same class, such as pairs of animals (e.g., "*lion-cat*") or foods ("*avocado-pineapple*"), they typically share a visual or semantic common ground that can facilitate a more coherent blend. Conversely, different class objects (e.g., "*turtle-broccoli*") often introduce stronger contrasts in shape, texture, or context, potentially challenging the blending process. Compound words, such as "*butter-fly*" or "*tea-pot*", offer further complexity: many of these compounds merge two distinct concepts into a single expression (e.g., "*butterfly*"), raising the question of whether the blending method will capture the conventional meaning of the compound or produce a literal hybrid of its components. We also examine concept and style, pairing a descriptive prompt like "*a portrait of a man*" with a specific artistic style (e.g., "*Picasso*", "*van Gogh*"), a task reminiscent of style transfer but without any specialized fine-tuning. Finally, we investigate architecture, blending visually iconic structures (e.g., "*Duomo-Taj Mahal*", "*Pisa-Big Ben*") to see how these methods handle complex geometric and structural features.

Because the notion of a "successful blend" is inherently subjective, the evaluation was conducted through a user study. Participants were shown images generated by different blending methods for the same concept pair, and they were asked to rank each image according to how well it captured a coherent hybrid of the two prompts. Figure 15 illustrates a subset of these images and identifies which blending method

Table 1 Full set of concept pairs organized into five categories, each illustrating a distinct type of blending scenario. The author of the painting in the Style section is omitted to better fit the table layout.

Same	
lion	*cat*
owl	*tiger*
avocado	*pineapple*
rabbit	*lion*
apple	*hamburger*
Different	
turtle	*broccoli*
blimp	*whale*
banana	*shoes*
canoe	*walnuts*
rocket	*carrot*
Compound Words	
butter	*fly*
bull	*pit*
kung fu	*panda*
tea	*pot*
man	*bat*
Concept and Style	
A portrait of a man	*Girl with a Pearl Earring*
A portrait of a man	*Guernica*
A portrait of a man	*The Scream*
A portrait of a man	*Starry Night*
Architecture	
Duomo of Milan	*Taj Mahal*
Eiffel Tower	*Sagrada Familia*
Leaning Tower of Pisa	*Big Ben*

generated them. The same set of five categories and prompts (Table 1) was presented, with the main focus on comparing how each method performed under these different blending scenarios.

Quantifying prompt similarity

To provide an operational notion of "concept similarity" aligned with the model we study, we measure similarity directly in the Stable Diffusion v1.4 conditioning space via the cosine similarity of mean-pooled text-encoder embeddings (Table 2). The resulting values broadly match intuition at the category level: same-class object pairs tend to be more similar than cross-domain combinations, whereas Concept+Style pairs

Category	Prompt p_1	Prompt p_2	Cosine sim.
Same	lion	cat	0.766
Same	avocado	pineapple	0.761
Same	owl	tiger	0.750
Same	rabbit	lion	0.736
Same	apple	hamburger	0.707
Different	turtle	broccoli	0.719
Different	rocket	carrot	0.708
Different	blimp	whale	0.693
Different	banana	shoes	0.691
Different	canoe	walnuts	0.646
Concept and Style	A portrait of a man	The Scream	0.565
Concept and Style	A portrait of a man	Girl with a Pearl Earring	0.540
Concept and Style	A portrait of a man	Guernica	0.382
Concept and Style	A portrait of a man	Starry Night	0.378
Compound Words	tea	pot	0.691
Compound Words	man	bat	0.688
Compound Words	kung fu	panda	0.677
Compound Words	bull	pit	0.676
Compound Words	butter	fly	0.656
Architecture	Eiffel Tower	Sagrada Familia	0.553
Architecture	Leaning Tower of Pisa	Big Ben	0.551
Architecture	Duomo of Milan	Taj Mahal	0.486

Table 2 Operationalizing prompt similarity for the user-study concept pairs. For each prompt p, we compute its Stable Diffusion v1.4 text-encoder embedding $e(p)$ and apply mean pooling over token embeddings; we report $\cos\big(e(p_1), e(p_2)\big)$ for each pair (higher means more similar in the model's conditioning space).

(descriptive prompt vs. artwork/style prompt) exhibit substantially lower similarity. Averaged across pairs, this trend becomes explicit: SAME has the highest mean similarity (0.744), followed by DIFFERENT (0.692) and COMPOUND WORDS (0.678), with ARCHITECTURE lower (0.530) and CONCEPT+STYLE the lowest (0.466). Importantly, similarity is not perfectly captured by our coarse categories: several Different-class pairs remain relatively close in embedding space, indicating that "concept similarity" is better treated as a continuous property of the learned text representation rather than a purely categorical notion. This operationalization enables more objective analyses of how blending quality varies with prompt similarity in the same embedding space used for generation, and helps interpret category-wise differences in the user-study results reported below. We emphasize that these values quantify *model-conditioning similarity* rather than an absolute cognitive distance, but they provide a reproducible continuous proxy for contextualizing category-level trends.

7.1 User Study

A user study was conducted with 100 participants, all young adults between 19 and 25 years of age, to evaluate the performance of the four blending methods introduced earlier. Each participant was presented with 22 pairs of prompts, yielding a total of 88 images distributed across the five previously described categories. Before the study, participants received a brief 10-minute introductory session explaining the notion of concept blending. They were then asked to rank each displayed image according to how effectively it conveyed a coherent blend of the two prompts. Specifically, participants were instructed to base their ranking on three criteria: (i) *Concept preservation*—are recognizable visual attributes of *both* prompts present?; (ii) *Integration/balance*—are the concepts fused into a single visual hybrid (rather than two disjoint objects), and does one concept avoid fully dominating the other?; and (iii) *Overall coherence*—does the result look visually plausible and internally consistent (e.g., shape, texture, lighting), without obvious artifacts that undermine the blend? To calibrate judgments, we explained that a successful blend should not merely depict juxtaposition (two objects side-by-side) nor collapse into a near-pure depiction of only one concept; instead, it should integrate features from both concepts into a unified entity (e.g., blending "lion" and "cat" by combining a lion-like mane with a cat-like facial structure). Participants ranked images from 1 (best) to 4 (worst) based on their overall assessment of these criteria. The introduction also provided concrete examples of failure cases (juxtaposition and dominance) and clarified that the task was to judge the quality of *integration* rather than general image beauty. All participants were informed of the prompt pairs under consideration (e.g., "*lion-cat*", "*turtle-broccoli*", "*a portrait of a man-Guernica by Pablo Picasso*"), but the particular blending method

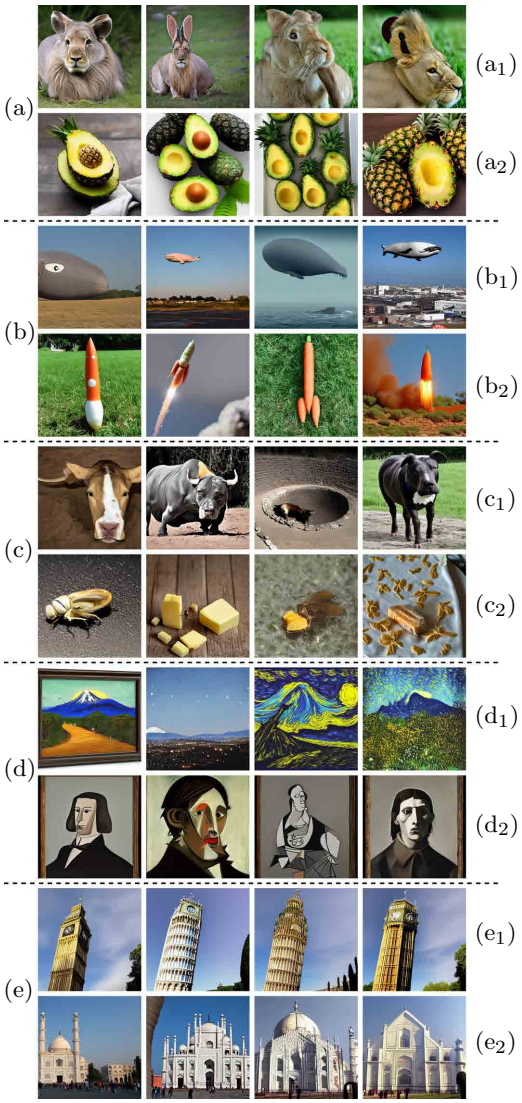

Fig. 15 Two representative blends from each category (a)–(e), showcasing variations within pairs (a_1, a_2) to (e_1, e_2). Highlights key characteristics of the user study materials.

that generated each image was withheld to avoid bias. Moreover, images were presented in a randomized order for each participant and for each concept pair, further minimizing any systematic preference or learning effects over time.

Since there is no single "ground truth" definition of what constitutes a "successful" conceptual blend, subjective human judgment is essential [41]. Conceptual blending inherently involves creativity and can therefore be interpreted differently by different people. By collecting participant rankings, we obtain a direct measure of which blends best capture an intended combination of ideas.

As discussed in Section 5.3, diffusion outputs are highly sensitive to the random seed used during sampling. To control for this variability *without method-specific cherry-picking*, for each prompt pair we predefined a pool of 10 random seeds and selected *one* seed *per pair* (fixed before generating any blends). We then used the same selected seed for all four methods for that prompt pair, so differences in the user study reflect the blending strategy rather than a favorable seed choice. Additionally, each blend was created using an equal ratio of the two prompts (see Section 5.2), preventing the synthesis process from favouring one concept over the other. Although this procedure can, in some cases, yield suboptimal blends, it keeps the experimental conditions balanced across all methods and prompt pairs. To reduce seed-selection bias, two authors independently screened the 10 candidate seeds *at the prompt-pair level* using a simple quality criterion (visual coherence and absence of obvious artifacts), and resolved disagreements by discussion; this internal screening was applied uniformly across methods and was performed prior to running the full study. In future work, automated metrics (e.g., CLIP-based similarity scores [42]) could be used as an objective pre-filter for seed selection.

7.2 Methodology for Evaluation

In total, 100 participants were asked to rank the results of all four blending methods over the 22 pairs of concepts. Our goal is to understand which, if any, of the methods is preferred over this sample. To achieve this, we analyse the relative preference of each method over all others throughout the full survey. In other words, we analyse the frequency at which one method is preferred over another. The rationale behind this choice is the following: if two methods are equally effective, since participants are required to provide relative rankings, it is reasonable to expect that half of the participants "prefer" one of the methods, and half of them "prefer" the other. The preference of a method over another can be modelled as a Bernoulli distribution, where the parameter represents the probability that a method is ranked better than another one. Our goal is to verify whether this parameter is close to 0.5 (i.e., there is no evidence that the subjects show a preference for one method) or away from this middle point, indicating a preference in one direction or the other.

For each possible pair of methods `METHOD-1`, `METHOD-2`, we estimate the Bernoulli parameter by counting the proportion of answers in which `METHOD-1` has a lower rank than (i.e., it is *preferred over*) `METHOD-2`. Table 3 shows the relations where this proportion is above 0.5, suggesting a direct preference among the participants of the survey. This induces a partial order. For instance, the first row of Table 3 expresses that `ALTERNATE` was preferred over `SWITCH`, in 51% of all the comparisons. While this

Table 3 Pair-wise comparison of user preferences of methods, and their p-value. ∼0 expresses that the p-value is below 0.001.

Preference	Proportion	p-value
ALTERNATE < SWITCH	0.51	0.31
ALTERNATE < TEXTUAL	0.60	∼0
ALTERNATE < USPLIT	0.57	∼0
SWITCH < TEXTUAL	0.61	∼0
SWITCH < USPLIT	0.58	∼0
USPLIT < TEXTUAL	0.55	∼0

expresses *some* evidence in favour of ALTERNATE, it is not clear whether this has any statistical power. Hence, the last column of the table presents the p-value for the null hypothesis H_0 : ALTERNATE > SWITCH. We use these p-values to verify whether the preference is statistically significant. In the table, ∼0 expresses that the p-value is below 0.001, which can be interpreted as being *extremely* significant.[3]

In summary, the first row in Table 3 indicates that, although more participants ranked ALTERNATE better than SWITCH, there is no statistical evidence to conclude that ALTERNATE is preferred over SWITCH. However, the subsequent rows respectively demonstrate a clear preference for both ALTERNATE and SWITCH over TEXTUAL and USPLIT. Furthermore, the data shows a preference for USPLIT over TEXTUAL, with all these outcomes exhibiting a high degree of statistical significance (p-value essentially 0). In other words, the data indicates a tie between ALTERNATE and SWITCH, followed by a preference for both over USPLIT, and of the latter over TEXTUAL. These preference relations are summarised in the grey box with bold border of Figure 16: methods depicted lower in the structure are more preferred than those higher above; the lack of connection between ALTERNATE and SWITCH represents the missing evidence in favour of one over the other.

We repeat the same analysis divided by type, as shown in Table 4 and the remaining blocks of Figure 16.

A simple analysis of the figures—where dashed lines express baseline significance, thin lines high significance and thick lines extreme significance—shows that in general ALTERNATE and SWITCH are the most preferred methods, while TEXTUAL is the least preferred with USPLIT remaining in the middle. This trend is, however, reversed for *Same*, where no method is preferred over TEXTUAL. We also see that preferences are indeed dependent on the specific category of blending tested, with no two categories producing the same preference relation.

Moreover, in Table 5, we show the mean (rounded to three significant digits), median, and mode rank assigned to each of the four methods, divided by individual pairs of concepts and by category. The last line also summarises the results globally. The mode tells us the rank that was selected the most, while the median expresses to which rank were at least half of the answers assigned. Hence for instance in the first row (*lion-cat*), SWITCH is ranked most often in first place (mode 1) and half of the

[3]We consider evidence to be *statistically significant* if the p-value is < 0.05; *very significant* if it is < 0.01; and *extremely significant* if below 0.001.

Table 4 Pair-wise comparison of user preferences of methods, and their p-value, grouped by the type of blend executed. ∼0 expresses that the p-value is below 0.001.

Cat.	Test	Proportion	p-value
Same	SWITCH < ALTERNATE	0.57	0.002
	TEXTUAL < ALTERNATE	0.57	0.001
	ALTERNATE < USPLIT	0.60	∼0
	SWITCH < TEXTUAL	0.51	0.33
	SWITCH < USPLIT	0.69	∼0
	TEXTUAL < USPLIT	0.70	∼0
Different	SWITCH < ALTERNATE	0.54	0.047
	ALTERNATE < TEXTUAL	0.62	∼0
	ALTERNATE < USPLIT	0.57	0.001
	SWITCH < TEXTUAL	0.59	∼0
	SWITCH < USPLIT	0.59	∼0
	USPLIT < TEXTUAL	0.54	0.038
Compound words	ALTERNATE < SWITCH	0.51	0.269
	ALTERNATE < TEXTUAL	0.55	0.021
	USPLIT < ALTERNATE	0.54	0.056
	SWITCH < TEXTUAL	0.62	∼0
	SWITCH < USPLIT	0.53	0.079
	USPLIT < TEXTUAL	0.62	∼0
Concept and Style	ALTERNATE < SWITCH	0.67	∼0
	ALTERNATE < TEXTUAL	0.82	∼0
	ALTERNATE < USPLIT	0.69	∼0
	SWITCH < TEXTUAL	0.66	∼0
	USPLIT < SWITCH	0.50	0.44
	USPLIT < TEXTUAL	0.76	∼0
Architecture	SWITCH < ALTERNATE	0.54	0.086
	ALTERNATE < TEXTUAL	0.62	∼0
	ALTERNATE < USPLIT	0.54	0.086
	SWITCH < TEXTUAL	0.69	∼0
	SWITCH < USPLIT	0.57	0.008
	USPLIT < TEXTUAL	0.59	0.002

participants rank it within the first 2 places (median 2). These numbers confirm the previous statistical analysis. Indeed, ALTERNATE and SWITCH tend to be ranked at a better position than USPLIT, and TEXTUAL is typically evaluated last. The exception is in the category *Same*, where TEXTUAL presents mode 1 and median 2 in its rank. We can see that this behaviour is highly influenced by the pair *rabbit-lion* in which more than half of the participants rank TEXTUAL as the best method. It is also evident from the table that neither ALTERNATE nor SWITCH clearly outperforms the other: although ALTERNATE generally has a lower mode (its most frequently assigned rank), SWITCH often shows a lower median, indicating that more participants place it in a higher position overall.

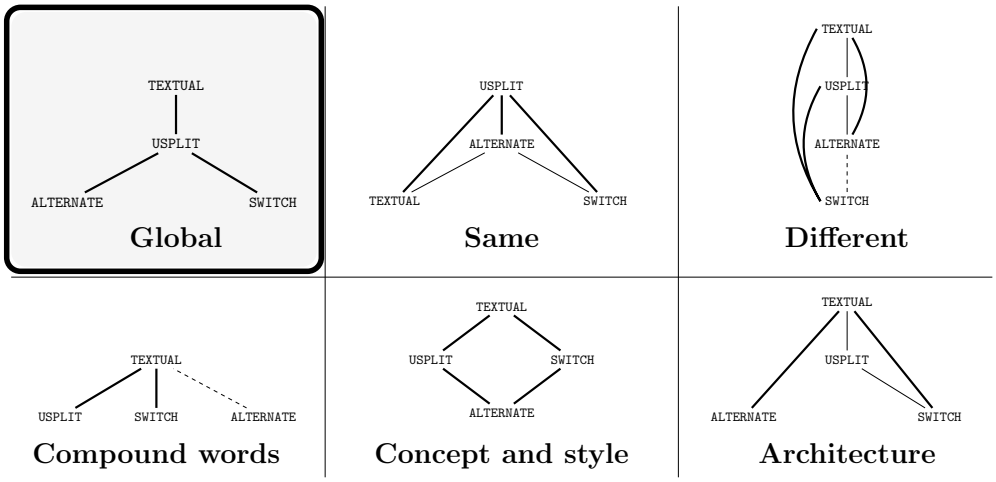

Fig. 16 Ordering preferences by category. In the Hasse diagrams, thick lines mean extremely significant (p-value < 0.001), thin lines mean very significant (p-value < 0.01) and dashed lines mean significant (p-value < 0.05).

7.3 Discussion

The comparison of the methods in the user survey, provides a clear hierarchy of preferences supported by a formal statistical analysis: `ALTERNATE` and `SWITCH` are equally liked, and preferred over `USPLIT`, which is itself preferred over `TEXTUAL`. In this section, we discuss the strengths and weaknesses of these methods explaining the results of the user survey.

Figure 17 highlights method-specific behavior under a controlled setting (same seed, fixed blend ratio 0.5). `TEXTUAL` often behaves like a smooth interpolation that can collapse toward a single dominant concept; `SWITCH` tends to enforce a sequential "base→modifier" dynamic (early structure from p_1, later refinements from p_2); `ALTERNATE` encourages interwoven influence across timesteps and can yield either highly integrated hybrids or stronger competition between structural interpretations; `USPLIT` frequently produces feature-transfer effects (especially texture/appearance) while maintaining a coherent single entity.

- **(a) lion–cat.** `TEXTUAL` is largely cat-dominant; `SWITCH` preserves a lion-like global silhouette while injecting cat-like facial/detail cues; `ALTERNATE` yields the most balanced single-animal hybrid; `USPLIT` mixes lion-like fur/mane cues with a more cat-like head, producing a coherent but slightly stylized hybrid.
- **(b) kung fu–panda.** `TEXTUAL` keeps a panda identity with pose/action cues; `SWITCH` emphasizes an asymmetric base–modifier effect and can preserve prompt artifacts; `ALTERNATE` produces a clearer "panda-as-martial-artist" fusion; `USPLIT` makes the hybridization explicit, combining panda identity with a fighter template.
- **(c) fish–tiger.** `TEXTUAL` tends toward a tiger in an aquatic context; `SWITCH` can lock an early structural interpretation and repaint later with tiger-like details; `ALTERNATE`

	Prompt	ALTERNATE			SWITCH			TEXTUAL			USPLIT		
		mean	mdn	mode	mean	mdn	mode	mean	mdn	mode	mean	mdn	mode
Same	Lion-Cat	2.70	3	4	2.02	2	1	2.61	3	4	2.67	3	3
	Owl-Tiger	2.27	2	1	1.87	2	1	2.30	2	3	3.55	4	4
	Avocado-Pineapple	2.02	2	2	2.53	3	2	2.43	3	1	3.02	3	4
	Rabbit-Lion	3.27	4	4	2.41	2	2	1.59	1	1	2.73	3	3
	Apple-Hamburger	2.43	2	2	2.32	2	2	2.26	2	3	2.99	3	4
	CATEGORY TOTAL	2.54	3	2	2.23	2	1	2.24	2	1	2.99	3	4
Different	Turtle-Broccoli	2.51	3	3	2.48	2	2	1.44	1	1	3.57	4	4
	Blimp-Whale	2.75	3	3	1.89	2	1	3.37	4	4	1.99	2	2
	Banana-Shoes	1.54	1	1	1.86	2	1	2.96	3	3	3.63	4	4
	Canoe-Walnuts	1.72	1	1	3.37	4	4	2.58	3	2	2.33	2	3
	Rocket-Carrot	3.20	3	3	1.77	2	2	3.44	4	4	1.59	1	1
	CATEGORY TOTAL	2.35	2	3	2.27	2	2	2.76	3	4	2.62	3	4
Compound Words	Butter-Fly	2.48	2	2	3.15	4	4	2.24	2	1	2.14	2	1
	Bull-Pit	3.14	4	4	1.61	1	1	2.92	3	3	2.33	2	2
	Kung fu-Panda	1.80	1	1	2.36	2	2	3.40	4	4	2.45	3	3
	Tea-Pot	1.76	2	1	2.85	3	3	2.11	2	1	3.28	3	4
	Man-Bat	3.21	3	3	1.83	2	2	3.26	3	4	1.70	1	1
	CATEGORY TOTAL	2.48	2	1	2.36	2	2	2.79	3	3	2.38	2	1
Concept and Style	Vermeer	1.98	2	1	2.27	2	1	3.16	4	4	2.59	3	2
	Picasso	2.13	2	1	2.05	2	2	3.19	3	4	2.63	3	3
	Munch	1.67	1	1	2.27	2	2	3.70	4	4	2.36	2	3
	van Gogh	1.51	1	1	3.49	4	4	2.86	3	3	2.14	2	2
	CATEGORY TOTAL	1.82	1	1	2.52	2	2	3.23	3	4	2.43	2	2
Architecture	Duomo-Taj Mahal	2.14	2	1	2.20	2	1	2.70	3	3	2.91	3	4
	Eiffel-Sagrada Familia	1.95	2	1	2.28	2	3	3.42	4	4	2.35	2	2
	Pisa-Big Ben	3.04	3	4	2.13	2	1	2.52	3	4	2.31	2	2
	CATEGORY TOTAL	2.38	2	1	2.20	2	1	2.90	3	4	2.52	2	2
	GLOBAL TOTAL	2.33	2	1	2.32	2	2	2.75	3	4	2.60	3	2

Table 5 Statistics (mean; median [mdn]; and mode) about the rank of each method over the whole survey, divided by pairs of concepts and category.

often yields heavy pattern/texture competition; `USPLIT` gives a clean single-object hybrid (fish-like form with tiger-like striping/color).

- **(d) car–lego bricks.** `TEXTUAL` remains close to a realistic car with weak lego cues; `SWITCH` introduces blocky/modular surface hints while staying car-centric; `ALTERNATE` pushes toward a lego-built toy interpretation; `USPLIT` often sits between, preserving car structure while strengthening brick-like geometry/material cues.
- **(e) tea–pot.** `TEXTUAL` leans toward a pot-like object with mild "tea" influence; `SWITCH` and `USPLIT` frequently express tea as material/texture (tea leaves) with weaker pot geometry; `ALTERNATE` yields a particularly legible hybrid (teapot silhouette made of tea leaves).
- **(f) bull–pit.** `TEXTUAL` stays largely bull-like; `SWITCH` preserves bull structure while injecting localized "pit" cues; `ALTERNATE` is more likely to flip toward a competing scene-level interpretation (emphasizing a pit/hole); `USPLIT` returns to a coherent single-subject fusion with strong feature transfer.

Blending in the Textual Latent Space (`TEXTUAL`).

`TEXTUAL` 's interpolation in the latent space of prompts does not always produce meaningful visual hybrids, as evidenced by the method's consistently lower ranking in

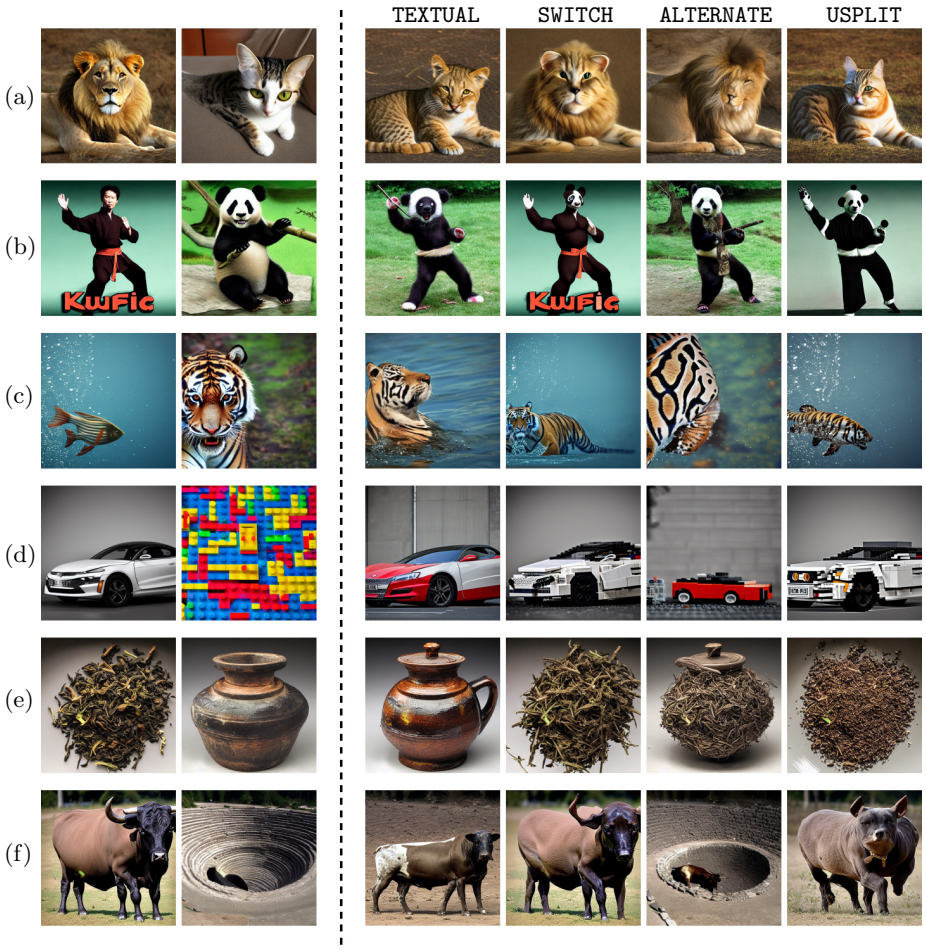

Fig. 17 Comparison of the blending methods with the same seed. On the left, the individual prompts, on the right, the blended images. (a) "*lion-cat*", (b) "*kung fu-panda*", (c) "*fish-tiger*", (d) "*car-lego bricks*", (e) "*tea-pot*", and (f) "*bull-pit*".

the user survey. We attribute this shortcoming to how the text encoder (CLIP in Stable Diffusion v1.4) arranges prompts in its latent space [8, 37]. There is no guarantee that points interpolated between two concept embeddings correspond to semantically valid blends; for instance, *"car"* and *"lego bricks"* in Figure 17 yield a mostly unaltered vehicle with few, if any, Lego-like details. Moreover, we observe substantial seed-to-seed variability, so even if the text prompt embedding remains constant, small differences in the initial noise can affect how visual and semantic elements are merged. Nonetheless, `TEXTUAL` sometimes outperforms other methods for concept pairs in the *Same* category (e.g., closely related items), likely because the overlap in visual or semantic features is easier to capture via simple interpolation.

Blending in the Iterative Reverse Process (SWITCH and ALTERNATE).

The SWITCH and ALTERNATE methods blend two concepts by altering the conditioning prompt during the iterative denoising process of the diffusion pipeline. These two methods are tied as the most effective overall, with ALTERNATE, despite its simplicity, showing a slight advantage in *Different* and *Architecture* contexts. This indicates that while both methods are generally reliable, the choice between them may be influenced by the specific blending context.

As mentioned in Section 4.2, results from SWITCH vary considerably depending on the timestep at which the prompt is switched from one concept to the other. Finding the right timestep is crucial for achieving a good visual blend. This behaviour is particularly evident in the case of "*tea-pot*", as illustrated in Figure 17. While "*tea*" and "*pot*" together form a semantically meaningful compound word, they represent visually distinct concepts. Recall that SWITCH starts by generating an image based on the first prompt (in this case "*tea*"), and then—midway through the diffusion process—switches to the second prompt ("*pot*"). After the switch, the model struggles to change and adjust the existing distribution to align with the prompt *pot*. Hence, the final image predominantly depicts the first concept, failing to blend the two effectively.

Another undesired behaviour of SWITCH is the *cartoonification* of the produced blend. When unable to shift the pixel distribution towards the new prompt, the diffusion pipeline corrects the existing noisy image latent by progressively removing the high-frequency details, resulting in a picture with flat colours and simplified shapes. This behaviour can be observed in the "*kung fu-panda*" blend produced by SWITCH in Figure 17. In all our experiments, this kind of behaviour could not be observed in any of the other methods.

The ALTERNATE method, which alternates between the two prompts at each timestep, tends to produce consistent results when the two blended concepts are visually distant. The unique type of visual blend that this technique produces is arguably even more interesting when the two concepts are both visually and semantically very different. This is the case of "*tea-pot*" and "*butter-fly*", where the model creates an image that literally and spatially contains both the first and the second prompt. This is also evident in the "*bull-pit*" blend in Figure 15, where the ALTERNATE generates what could be described as a "*a bull in a pit*".

Blending in the U-Net (USPLIT).

USPLIT occupies a middle ground in the rankings, consistently outperforming TEXTUAL but falling behind both ALTERNATE and SWITCH in most comparisons. Its performance indicates that it is a viable blending method in certain contexts but lacks the robustness and adaptability of the top-performing methods.

Compared to the other blending methods, USPLIT, which encodes the image (latent) conditioned on the first prompt in the bottleneck and then decodes it with the second prompt, produces more subtle blends while maintaining generally consistent results. This subtlety might explain why USPLIT ranks below ALTERNATE and SWITCH in our user study, as the visual blend is less pronounced compared to the more evident combinations achieved by the other methods.

From our experiments, in general, USPLIT seems to be able to modify the shape of the first concept with prominent features of the second. This is particularly evident in the "*fish-tiger*" and "*car-lego bricks*" blends shown in Figure 17. Interestingly, on the "*kung fu-panda*" blend, USPLIT seems to slightly change the visual representation of the first prompt while matching the colours of the second one. This subtle blending is also evident in the "*bull-pit*" blend of Figure 17, where surprisingly the pipeline creates an image that somewhat resembles an actual *pitbull* although nothing in the prompt makes any reference to this breed of dog.

Although the USPLIT method consistently produces more subtle blends, its results depend on the alignment between the visual representation of the first prompt and the embedding of the second. Specifically, the image generated by the diffusion pipeline based on the first prompt must contain visual features that can be effectively leveraged by the cross-attention layers in U-Net when incorporating the embedding of the second prompt. When this alignment is not possible, as in the case of the blend between "*tea*" and "*pot*", the model struggles to identify which elements of the first prompt should be modified by the second, resulting in an unsuccessful blend.

8 Conclusions

This paper investigated how text-to-image diffusion models can blend two distinct concepts, introduced as separate prompts, into a single coherent image. We examined four zero-shot blending strategies, each leveraging a different aspect of the diffusion process, and showed that they can operate effectively without specialized training, fine-tuning, or complex prompt engineering. Although each strategy can produce high-quality hybrids, our results indicate that no single approach dominates in all scenarios; rather, blending outcomes depend heavily on the nature of the concept pair (e.g., animal–animal vs. architectural structures), the initial noise seed, and the user's subjective notion of a "successful" blend. By systematically evaluating the proposed methods across five concept categories, we showed the zero-shot compositional capability of diffusion models. This capability underscores the models' ability to perform creative fusion of ideas, complementing broader AI research on compositional reasoning and generative creativity [43]. Our user study with 100 participants provided clear empirical evidence of where each strategy excels, reinforcing that the iterative approaches (SWITCH and ALTERNATE) are often preferred but remain sensitive to user choices, such as prompt order and the point of switching, while USPLIT and TEXTUAL handle certain cases reliably but can falter when concepts are highly dissimilar or insufficiently anchored. The findings presented here align with ongoing research in knowledge representation, interactive design, and computational creativity [44]. This opens up creative applications, such as rapid design prototyping or cross-domain concept generation, that extend beyond conventional use cases like style transfer [45]. We also discussed the limitations of our work. First, our study mostly focused on short, discrete prompts, thus leaving open questions about how these methods handle longer or more complex descriptions. Second, the inherent variability of diffusion outputs can lead to unpredictable or "cartoonified" images, especially when the model

struggles with the newly introduced concept. Third, the user study sampled general participants; results might differ with domain experts or professional artists.

As future work, several concrete next steps arise from this study. We aim to test our four blending methods within more recent diffusion models, e.g., SDXL [34], which could yield more refined or stable outputs. An interesting avenue is to incorporate symbolic constraints or partial blending into the pipeline, ensuring that only certain features of one concept transfer to the other. Multi-step blending (e.g., merging more than two concepts) may further stretch the boundaries of creative composition. Finally, large-scale user studies with participants from specialized domains, such as art, design, or psychology, could indicate additional nuances in aesthetic preferences and the cognitive interpretation of blended images.

Although we did not explicitly compare to fine-tuned or specialized style-transfer frameworks (e.g., DreamBooth-like methods [46]), our focus on zero-shot methods fills a distinct niche: it shows how well diffusion models as-is handle concept fusion. By showcasing practical, zero-shot blending techniques, we have highlighted diffusion models' latent potential for creative synthesis without specialized training or prompt engineering. We encourage the community to build upon our publicly released codebase, prompts, and user-study protocols to replicate and extend our findings in more diverse application contexts.

Declarations

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Ethical approval

Not applicable. Ethical review and approval were not required for this study because it involved an anonymous questionnaire administered to competent adult university students, posed minimal risk, and did not involve the collection or processing of personal data or special-category data. In accordance with the EU General Data Protection Regulation (Reg. (EU) 2016/679, Recital 26) and the Italian Privacy Code (Legislative Decree 196/2003, as amended by Legislative Decree 101/2018), anonymous information falls outside the scope of data-protection obligations. The study did not fall under healthcare or clinical research frameworks requiring ethics-committee review.

Consent to participate

Informed consent was obtained from all participants. Participation was voluntary and limited to adults ($\geq$ 18 years). Prior to participation, instructors explained the purpose of the study in class and participants were informed through an introductory screen in the questionnaire. Only students who agreed to participate were included. No personal or identifying data were collected, and participants could discontinue participation at any time.

Consent to publish

Not applicable. The study did not collect personal or identifiable data, and all reported results are based solely on anonymous responses.

Competing interests

The authors declare no competing interests.

Data availability

To ensure reproducibility, code, user-study materials, prompt configurations, and all anonymous survey responses are publicly available at https://github.com/LorenzoOlearo/blending-diffusion-models.

References

[1] Fauconnier, G., Turner, M.: The Way We Think: Conceptual Blending And The Mind's Hidden Complexities. Basic Books, New York (2003)

[2] Confalonieri, R., Pease, A., Schorlemmer, M., Besold, T., Kutz, O., Maclean, E., Kaliakatsos-Papakostas, M.: Concept Invention Foundations, Implementation, Social Aspects and Applications: Foundations, Implementation, Social Aspects and Applications, (2018). https://doi.org/10.1007/978-3-319-65602-1

[3] Gardenfors, P.: Conceptual Spaces: The Geometry of Thought. A Bradford book. MIT Press, Cambridge, MA (2004)

[4] Confalonieri, R., Eppe, M., Schorlemmer, M., Kutz, O., Peñaloza, R., Plaza, E.: Upward refinement operators for conceptual blending in the description logic EL++. Ann. Math. Artif. Intell. **82**(1-3), 69–99 (2018) https://doi.org/10.1007/S10472-016-9524-8

[5] Porello, D., Kutz, O., Righetti, G., Troquard, N., Galliani, P., Masolo, C.: A toothful of concepts: Towards a theory of weighted concept combination. In: Simkus, M., Weddell, G.E. (eds.) Proceedings of the 32nd International Workshop on Description Logics (2019)

[6] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning, pp. 2256–2265 (2015). PMLR

[7] Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)

[8] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10684–10695 (2022)

[9] Melzi, S., Peñaloza, R., Raganato, A.: Does stable diffusion dream of electric sheep? In: CEUR Workshop Proceedings, vol. 3511, pp. 1–11 (2023). CEUR-WS

[10] Olearo, L., Longari, G., Melzi, S., Raganato, A., Peñaloza, R.: How to Blend Concepts in Diffusion Models (2024). https://arxiv.org/abs/2407.14280

[11] Sun, Z., Zhang, Z., Zhang, Y., Lu, M., Lischinski, D., Cohen-Or, D., Huang, H.: Creative Blends of Visual Concepts (2025)

[12] Fauconnier, G., Turner, M.: Conceptual blending, form and meaning. Recherches en communication **19**, 57–86 (2003)

[13] Fauconnier, G., Turner, M.: Conceptual integration networks. Cognitive Science **22**(2), 133–187 (1998) https://doi.org/10.1207/s15516709cog2202_1

[14] Eppe, M., Maclean, E., Confalonieri, R., Kutz, O., Schorlemmer, M., Plaza, E., Kühnberger, K.-U.: A computational framework for conceptual blending. Artificial Intelligence **256**, 105–129 (2018) https://doi.org/10.1016/j.artint.2017.11.005

[15] Kutz, O., Bateman, J., Neuhaus, F., Mossakowski, T., Bhatt, M.: In: Besold, T.R., Schorlemmer, M., Smaill, A. (eds.) E Pluribus Unum, pp. 167–196. Atlantis Press, Paris (2015). https://doi.org/10.2991/978-94-6239-085-0_9

[16] Hedblom, M.M., Kutz, O., Neuhaus, F.: Image schemas in computational conceptual blending. Cognitive Systems Research **39**, 42–57 (2016) https://doi.org/10.1016/j.cogsys.2015.12.010 . From human to artificial cognition (and back): new perspectives of cognitively inspired AI systems

[17] Si, M.: Word embedding for analogy making. In: ICCC, pp. 282–286 (2022)

[18] Wang, B., al: From analogy to innovation: A creative conceptual design approach leveraging large language models. Advanced Engineering Informatics **67**, 103427 (2025)

[19] Wang, S., Petridis, S., Kwon, T., Lisle, L., Mo, B.H., Ma, C., Watkins, O., Hearst, M.A.: Popblends: Strategies for conceptual blending with large language models. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, pp. 1–19 (2023)

[20] Chen, L., al: A foundation model enhanced approach for generative design in combinational creativity. Journal of Engineering Design **35**(11), 1394–1420 (2024)

[21] Cho, W., Zhang, Y., Chen, Y.-Y., Inouye, D.I.: Imagine for me: Creative conceptual blending of real images and text via blended attention. arXiv preprint arXiv:2506.24085 (2025)

[22] Ge, S., Parikh, D.: Visual conceptual blending with large-scale language and vision models (2021). https://arxiv.org/abs/2106.14127

[23] Leemhuis, M., Kutz, O.: The boat, the house and the in-between: Conceptual blending as betweenness relation. In: Hedblom, M.M., Kutz, O. (eds.) Proceedings of The Eighth Image Schema Day. CEUR Workshop Proceedings, vol. 3888. CEUR-WS.org, Bozen-Bolzano, Italy (2024). https://ceur-ws.org/Vol-3888/Paper_5.pdf

[24] Wang, C.J., Golland, P.: Interpolating between images with diffusion models. arXiv preprint arXiv:2307.12560 (2023)

[25] Li, J., Zhang, Z., Yang, J.: Tp2o: Creative text pair-to-object generation using balance swap-sampling. In: European Conference on Computer Vision, pp. 92–111 (2024). Springer

[26] Kothandaraman, D., Lin, M., Manocha, D.: Financial models meets generative art: Black-scholes-inspired concept blending in text-to-image diffusion. In: Proceedings of the 33rd ACM International Conference on Multimedia, pp. 12209–12217 (2025)

[27] Park, T., Zhu, J.-Y., Wang, O., Lu, J., Shechtman, E., Efros, A.A., Zhang, R.: Swapping autoencoder for deep image manipulation. In: Advances in Neural Information Processing Systems (2020)

[28] Alaluf, Y., Garibi, D., Patashnik, O., Averbuch-Elor, H., Cohen-Or, D.: Cross-Image Attention for Zero-Shot Appearance Transfer (2023)

[29] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

[30] Qiyuan, H., Wang, J., Liu, Z., Yao, A.: Aid: Attention interpolation of text-to-image diffusion. Advances in Neural Information Processing Systems **37**, 97766–97799 (2024)

[31] Xiong, Z., Zhang, Z., Chen, Z., Chen, S., Li, X., Sun, G., Yang, J., Li, J.: Novel Object Synthesis via Adaptive Text-Image Harmony (2024)

[32] Zhou, Y., Shen, H., Wang, H.: FreeBlend: Advancing Concept Blending with Staged Feedback-Driven Interpolation Diffusion (2025). https://arxiv.org/abs/2502.05606

[33] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)

[34] Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: SDXL: Improving latent diffusion models for high-resolution image synthesis. In: The Twelfth International Conference on Learning Representations (2024). https://openreview.net/forum?id=di52zR8xgf

[35] Kingma, D.P., Welling, M.: Auto-Encoding Variational Bayes (2022). https://arxiv.org/abs/1312.6114

[36] Ronneberger, O., Fischer, P., Brox, T.: 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, pp. 234–241 (2015). Springer

[37] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., *et al.*: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763 (2021). PMLR

[38] Zhao, W., Bai, L., Rao, Y., Zhou, J., Lu, J.: UniPC: A Unified Predictor-Corrector Framework for Fast Sampling of Diffusion Models (2023). https://arxiv.org/abs/2302.04867

[39] Karthik, S., Roth, K., Mancini, M., Akata, Z.: If at First You Don't Succeed, Try, Try Again: Faithful Diffusion-based Text-to-Image Generation by Selection (2023). https://arxiv.org/abs/2305.13308

[40] Ho, J., Salimans, T.: Classifier-free diffusion guidance. In: NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications (2021)

[41] Martins, P., Urbancic, T., Pollak, S., Lavrac, N., Cardoso, A.: The good, the bad, and the aha! blends. In: Toivonen, H., Colton, S., Cook, M., Ventura, D. (eds.) Proceedings of the Sixth International Conference on Computational Creativity, ICCC 2015, pp. 166–173. computationalcreativity.net, Park City, Utah (2015)

[42] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., *et al.*: Learning transferable visual models from natural language supervision. In: Proc. of International Conference on Machine Learning, pp. 8748–8763 (2021). PMLR

[43] Park, D., Kim, S., Moon, T., Kim, M., Lee, K., Cho, J.: Rare-to-frequent: Unlocking compositional generation power of diffusion models on rare concepts with llm guidance. In: International Conference on Learning Representations (2025)

[44] Liu, N., Li, S., Du, Y., Torralba, A., Tenenbaum, J.B.: Compositional visual generation with composable diffusion models. In: European Conference on Computer Vision, pp. 423–439 (2022). Springer

[45] Huang, L., Chen, D., Liu, Y., Shen, Y., Zhao, D., Zhou, J.: Composer: creative and controllable image synthesis with composable conditions. In: Proceedings of the 40th International Conference on Machine Learning, pp. 13753–13773 (2023)

[46] Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: 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, pp. 22500–22510 (2023)