BrandFusion: Aligning Image Generation with Brand Styles

Parul Gupta Varun Khurana Yaman Kumar Singla
Balaji Krishnamurthy Abhinav Dhall

Adobe Media and Data Science Research, Monash University

Abstract

While recent text-to-image models excel at generating realistic content, they struggle to capture the nuanced visual characteristics that define a brand's distinctive style—such as lighting preferences, photography genres, color palettes, and compositional choices. This work introduces BrandFusion, a novel framework that automatically generates brand-aligned promotional images by decoupling brand style learning from image generation. Our approach consists of two components: a Brand-aware Vision-Language Model (BrandVLM) that predicts brand-relevant style characteristics and corresponding visual embeddings from marketer-provided contextual information, and a Brand-aware Diffusion Model (BrandDM) that generates images conditioned on these learned style representations. Unlike existing personalization methods that require separate fine-tuning for each brand, BrandFusion maintains scalability while preserving interpretability through textual style characteristics. Our method generalizes effectively to unseen brands by leveraging common industry sector-level visual patterns. Extensive evaluation demonstrates consistent improvements over existing approaches across multiple brand alignment metrics, with a 66.11% preference rate in human evaluation study. This work paves the way for AI-assisted on-brand content creation in marketing workflows.

1. Introduction

"If this business were split up, I would give you the land and bricks and mortar, and I would take the brands and trademarks, and I would fare better than you."

- John Stuart, co-founder, Quaker Oats

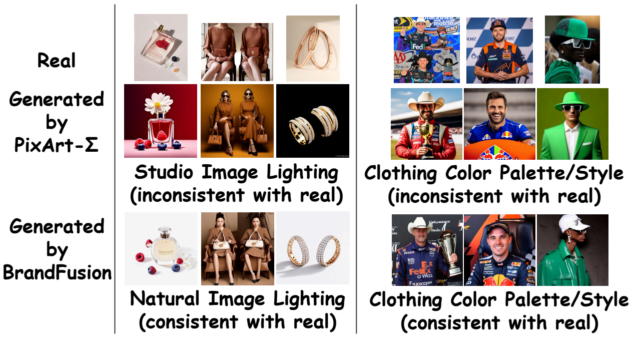

Real
Generated by PixArt-Σ
Generated by BrandFusion
Studio Image Lighting (inconsistent with real)
Clothing Color Palette/Style (inconsistent with real)
Natural Image Lighting (consistent with real)
Clothing Color Palette/Style (consistent with real)

Figure 1. We propose BrandFusion, a method which can generate images *aligned* to the nuanced visual characteristics that define a brand's visual identity, such as lighting preferences, photography genres, color palettes, and compositional choices. Here, the first row shows real images and the second and third rows show corresponding images generated by PixArt-Σ [2] and BrandFusion respectively. Note how BrandFusion images capture brand-specific characteristics such as natural image lighting (first 3 columns), brand-consistent clothing color palette (next 2 columns) and casual clothing style (last column), all of which make them more aligned to the corresponding real brand images.

The AI-powered content creation market is projected to reach $7.9 billion by 2033, with a remarkable CAGR of 7.7% [3]. This explosive growth reflects a fundamental shift in how brands approach content creation, driven by the need to produce high-quality promotional materials at unprecedented scale and speed. Businesses using AI content creation tools publish 42% more content monthly than those relying on traditional methods [20]. Yet despite these advances, a critical gap remains: current AI-driven content generation tools cannot sufficiently capture the nuanced visual characteristics that define distinctive brand identities.
The American Marketing Association defines a brand as "a name, term, sign, symbol, or design, or a combination of them, intended to identify the goods or services of one seller or group of sellers and to differentiate them from those of competitors." A credible brand signals a certain level of quality, fostering consumer trust and repeat purchases. Hence, to firms, brands are an enormously valuable piece of legal property that can influence consumer behavior, be traded, and ensure sustained future revenues [18]. To expand their audience reach, firms not only prioritize positive customer experiences, but also use advertising, espe-

cially on social media platforms, which improves customer awareness about the brand, helps in differentiation from its competitors and eventually drives sales. A brand's social media presence provides it with an owned channel, enabling it to influence narratives, engage with its audience directly, and generate positive media coverage. Therefore, it is imperative that the *elements* which can be used to identify and differentiate a brand are consistently visible in its promotional advertisements.

However, while existing text-to-image methods can generate realistic content consistent with textual instruction semantics, they are unable to capture the distinctive brand *elements* through textual input alone [12]. This limitation is demonstrated through the E-commerce brand FedEx[1] example in Figure 2, where providing brand-relevant characteristics (such as foreground banner with purple and orange gradients, close-up shot) to the prompt fails to properly reflect these elements in generated images, and instead leads to loss of other characteristics such as photorealism.

Unlike artistic style learning [9], which deals with easily identifiable textures and brush strokes, brand visual identity encompasses nuanced, multi-dimensional characteristics. Prior work identifies that brand identity emerges from complex combinations of typography, color palettes, photography genres, lighting preferences, compositional choices, and human representation styles [23, 24]. These elements create subtle but distinctive patterns that differentiate brands within the same market sector, as illustrated in Figure 2(B), where consulting and telecom companies show employees with distinctly different facial expressions and clothing styles, while fashion brands like Gucci and Tommy Hilfiger employ contrasting photographic approaches despite similar subject matter.

Additionally, while artistic style learning techniques typically rely on *latent* embeddings derived from style reference images, this approach limits their practical use for brand style learning. For practical on-brand content creation workflows, it is crucial that marketers can understand and directly control brand characteristics rather than relying only on opaque latent representations. Moreover, current personalization methods [8, 27] require separate fine-tuning for each new brand, making them impractical for large-scale content generation.

This fundamental challenge raises the core research question: *How can we enable automated generation of images that are aligned with a brand's nuanced visual identity characteristics?*

To this end, we introduce BrandFusion, a novel two-component framework that decouples brand style learning from image generation to address these limitations. Our approach consists of: (1) BrandVLM, a vision-language model that predicts both textual brand characteristics and visual style embeddings from marketer-provided context (brand name, sector, campaign details), and (2) BrandDM, a diffusion model trained to generate images conditioned on these dual style representations. This architecture enables scalable brand-aligned content creation without per-brand fine-tuning while maintaining interpretability through textual characteristics that marketers can understand and control. Our work makes the following key contributions:

1. We propose BrandFusion, a novel two-component architecture that decouples brand style learning from image generation, enabling scalable brand-aligned content creation without per-brand fine-tuning.
2. Our framework enables marketers to retain control over visual elements while ensuring brand consistency in the generated images, by learning brand characteristics in the textual domain.
3. We demonstrate effective cross-brand generalization, successfully generating images for unseen brands within the same industry sectors by leveraging sector-level visual patterns learned during training.
4. We show consistent improvements over existing text-to-image and personalization methods across multiple brand alignment measures, achieving a remarkable 66% preference in human evaluation.

2. Related Work

In this section, we situate our work amid the existing literature related to automatic advertisement generation, learning from groups of images to generate similar media, personalized media generation; and using foundational Vision-Language Models to guide Diffusion Models.

Automatic Advertisement Generation: Existing approaches to automatically generate advertisements such as [11, 21, 28, 34–36] take the textual description of the image (background), the image of the product to be advertised, and tagline (text) as input, and generate advertisements showcasing the product and the tagline in design-optimal layouts. [6] uses Diffusion Model to also learn a target *style* from a group of input images for the background generation in the advertisement. [47] proposes a reinforcement learning based framework to continually improve the advertisement generated for a product image by using a reliable feedback network. While these works are able to generate aesthetically pleasing advertisements, they do not consider the *visual identity* of the brand [24]. Even [6], which learns the background visual style, has a limitation – a different model needs to be learnt for each new visual style which requires compute and hyperparameter tuning, thereby reducing its ease of use.

Artistic style learning: The existing methods to learn artistic styles base their learning on either language guidance or image guidance. We discuss them below:

[1] https://www.fedex.com

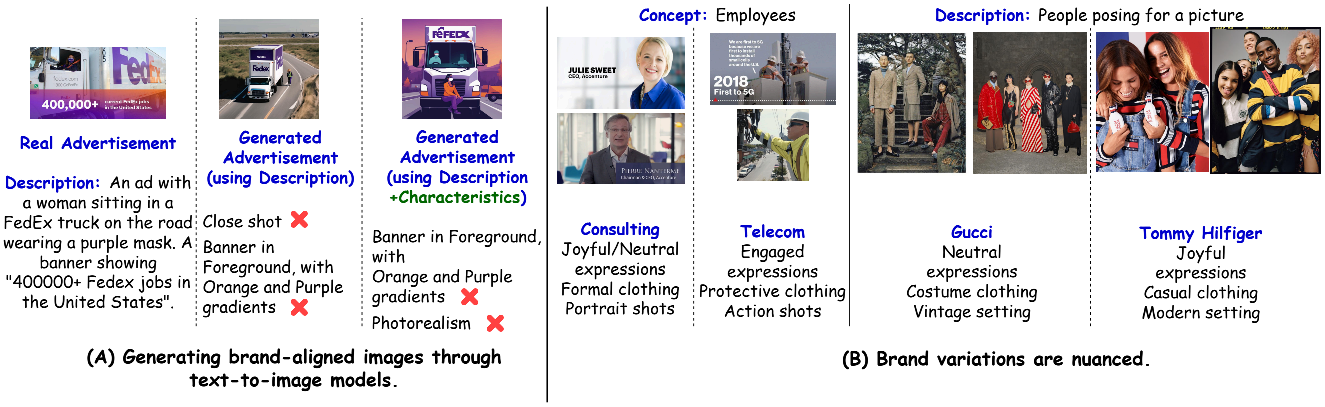

Real Advertisement
Description: An ad with a woman sitting in a FedEx truck on the road wearing a purple mask. A banner showing "400000+ Fedex jobs in the United States".
Generated Advertisement (using Description)
Close shot ✗
Banner in Foreground, with Orange and Purple gradients ✗
Generated Advertisement (using Description +Characteristics)
Banner in Foreground, with Orange and Purple gradients ✗
Photorealism ✗
(A) Generating brand-aligned images through text-to-image models.
Concept: Employees
Consulting
Joyful/Neutral expressions
Formal clothing
Portrait shots
Telecom
Engaged expressions
Protective clothing
Action shots
Description: People posing for a picture
Gucci
Neutral expressions
Costume clothing
Vintage setting
Tommy Hilfiger
Joyful expressions
Casual clothing
Modern setting
(B) Brand variations are nuanced.

Figure 2. (A) Here, we show a real advertisement from the E-commerce brand FedEx, and the corresponding image generated by the SDXL model using its description. This image lacks some brand relevant characteristics, such as *close shot* and *foreground banner with orange and purple gradients*, which are then added to the description to generate another image. As we pass these characteristics, the adherence of the generated image to the prompt reduces even further, with *photorealism* gone. Thus, current text-to-image models are unable to generate brand-aligned images. (B) The variations among brand images are at a very nuanced level, e.g. *Employees* represented in Consulting and Telecom sector images tend to have different facial expressions, clothing style and photography genres. Such differences are evident even among brands from the same sector, e.g. here, models posing for two fashion brands Gucci and Tommy Hilfiger have different expressions and clothing styles. BrandFusion enables alignment of generated images to the brand by capturing these nuances.

- **Natural language based style generation/transfer:** CLIPstyler [19], Styler-DALLE [38] and FastCLIPstyler [31] are some of the recent works which devise ways to transfer the style mentioned in a text prompt onto an image, so the style guidance is applied purely in the linguistic space. This is achieved through the rich CLIP [26] embedding space, by ensuring the correspondence between the output style image and the input style prompt. However, most of the text prompts explored in these works involve only *atomic* level of changes such as colours and textures (e.g. *sunset style* changes the colour palette of the input image to red-orange hues). Another work titled SGDiff [30] provides flexible style guidance through either language or reference image and transfers it onto images of clothing items. Again, the style attributes used for guidance mostly involve colors and texture based changes. On the other hand, reflecting a brand's style requires consideration of a much richer level of features such as human expressions, camera perspective with respect to the subject in the image and many more.
- **Image based style transfer:** ArtBank [46], LSAST [45], StyleDrop [29], DreamStyler [1], Inversion based Style Transfer [43] and RIVAl [44] involve getting style based guidance through one or more reference images which are very similar to each other, i.e. the common features are very easily evident, such as *anime*, *illustration* or *sketch*. The primary focus of these approaches is to reduce the number of images required to infer the style (just one image is sufficient). Also, the styles involved belong only to artistic categories. However, brand promotional content contains real world marketing images whose similarity to each other is on a more abstract level, as evident in Figure 2 (B). Moreover, these methods often require fine-tuning of the base model for each new style, while our approach doesn't require such fine-tuning because the extraction of the brand's visual identity (through VLM) is decoupled from the brand-aligned image generation process (through DM). This has the added advantage of being able to add customizations as per the marketer's preference for each new image that is to be generated.

Personalized Image Generation: A lot of research has been done about manipulating Diffusion Models to generate images that fulfill certain specified criteria. Earliest works aimed towards easier adaptation of large-scale Diffusion Models include ControlNet [42], where image guidance is provided by adding extra conditions through zero convolution layers. While these works need considerable amounts of training data, later works aim to reduce the number of training images to learn a particular concept (object/animal/person's identity), and coin it as *personalized* image generation. These works either fine-tune the weights of the entire model (as in Dreambooth [27]) or *invert* the limited number of samples into the conditional space (as in Textual inversion [8]) to learn representations that can be used to generate the same concept in other environments specified by the text prompts. The major drawback of these approaches is the need to fine-tune repetitively for each new concept. Follow up methods such as PaintByExample [39], CustomNet [41], Mix-of-Show [10] and Infinite-ID [37] introduce new capabilities such as training-free generation, variable viewpoints generation, generating multiple objects in a single image, generating humans in different environments using few images and so on; but all of these works essentially use learnable modules to map the concept representations into the conditional latent space of Diffusion models, or manipulate the cross-attention maps. Since all these representations are learnt in latent space, it is difficult to control specific attributes of the generated images, an aspect that is highly important for marketers.

Vision-Language models based guidance: With the advent of Large-Language Models (LLMs) [33] and their assimilation with Vision Transformer (ViT) [5] models which gave rise to foundational Vision-Language Models (VLMs) [22]; these models are being used extensively in conjunction with generative Diffusion Models to automate several visio-lingual tasks, e.g. Customized Manga Generation [16], User-friendly/smart Image editing, photo optimization [7, 15, 48], etc. Particularly, Customization Assistant [48] enables editing of any input image even on the basis of vague text prompts provided by the user who might not have a clear idea of what they would like to change in the image. Therefore, we utilize a foundational VLM in our approach, to automatically infer the style characteristics relevant to the target image, using just the metadata about the social media post, such as post caption, image description, brand name and sector, date of posting, etc., all of which is provided by the marketer.

3. Data

Collection: We start with a dataset of approx **246K images** belonging to **183 brands** from a wide variety of sectors such as automobiles, fashion, footwear, electronics, airlines, and more (complete list is provided in the Appendix D). These images are scraped from the content posted by these brands on their official twitter handles between Jan 2018 and Feb 2023. We also collect the number of likes received by the post, the post caption and the date of posting. **90%** of the total images of each brand are retained in the **train set**, and rest are used for testing.

Annotation: Next, we take a set of N attributes (denoted by i, $1 \leq i \leq N$) related to diverse aspects of images, which are generally used by art directors to describe brand images, inspired from [23, 24]. Some example attributes are image lighting, background, camera perspective, photography genre. Also, since our dataset has humans in 57.02% of the total images, 7 out of these attributes are about the humans in the image (human-related attributes), such as facial expressions, clothing style, clothing color palette, gaze. Each of these attributes can have different possible labels, e.g. the photography genre can be Architectural, Product, Live-stage, Abstract, Candid, Group, etc. Thus, we consider a total of 15 attributes (i.e. $N = 15$). A complete list of the attributes and their corresponding labels is provided in the Appendix D.

For each image in our dataset, we use a Vision Language Model (LLaVA-v1.6-34B [22]) to extract ground truth style characteristics by prompting it to predict at most three labels corresponding to each of our 15 attributes. This creates the textual *verbalization* describing each image's visual characteristics. Additionally, we compute CLIP-Image embeddings [26] as ground truth visual style representations. It is important to distinguish this annotation process from our BrandVLM training: the LLaVA-v1.6-34B model here serves as a ground truth extraction tool that analyzes real images to generate style characteristic labels, whereas our BrandVLM (fine-tuned LLaVA-1.5-13B) learns to predict these same style characteristics directly from textual marketer context without any visual input. In addition, we use a deterministic extractor model to obtain the color palette of each image, out of a set of 40 base colors.

4. Image Generation Aligned with Brand Styles

Our BrandFusion framework generates images aligned to brand visual styles through a two-component architecture trained in sequential phases.

Phase 1: BrandVLM Training. We finetune a Vision-Language model (LLaVA-1.5-13B [22]) to predict both textual style characteristics and visual style representations from marketer-provided context. While no images are provided as input to BrandVLM during training or inference, we choose a VLM over a pure LLM (like LLaMA) because VLMs possess superior understanding of visual concepts and image-related terminology acquired during their multimodal pretraining, making them better suited for understanding about visual style attributes even from textual context alone.

The BrandVLM learns to map marketer context $\mathbf{x}$ (brand name, sector, post text, posting date, image description, and engagement metrics) to two complementary outputs: (1) a textual *verbalization* $\mathbf{y}$ describing the brand's visual characteristics across our 15-attribute taxonomy, and (2) a CLIP global image embedding $\mathbf{e}$ capturing fine-grained visual style information that complements the discrete textual attributes.

This dual prediction approach addresses a key limitation: while textual characteristics provide interpretable control over major style elements (lighting, composition, clothing), visual embeddings capture nuanced details (specific color gradients, subtle textures, photographic quality) that are difficult to verbalize but crucial for brand alignment. The BrandVLM output is: $\hat{\mathbf{e}}, \hat{\mathbf{y}} = \text{BrandVLM}(\mathbf{x})$, with training objective:

$$\mathcal{L}_{\text{BrandVLM}} = \mathcal{L}_{\text{clip}} + \lambda \mathcal{L}_{\text{verb}}, \tag{1}$$

$$\text{where } \mathcal{L}_{\text{clip}} = \|\mathbf{e} - \hat{\mathbf{e}}\|^2, \tag{2}$$

$$\text{and } \mathcal{L}_{\text{verb}} = \mathcal{L}_{\text{LM}}(\mathbf{y}, \hat{\mathbf{y}}) \tag{3}$$

Here, $\mathcal{L}_{\text{LM}}$ denotes the language modeling loss, $\lambda = 0.1$ is a

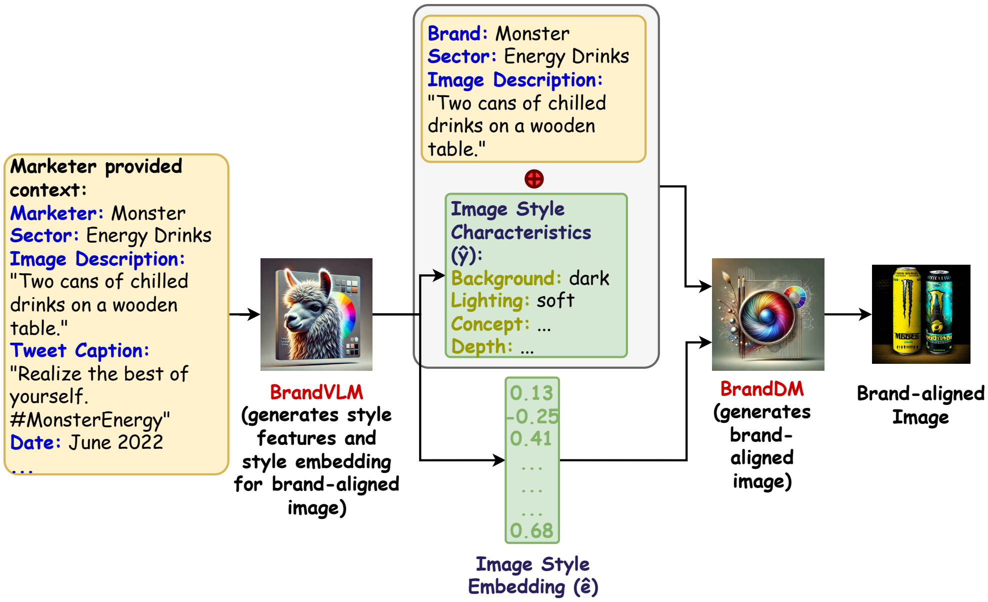

Marketer provided context:
Marketer: Monster
Sector: Energy Drinks
Image Description:
"Two cans of chilled drinks on a wooden table."
Tweet Caption:
"Realize the best of yourself. #MonsterEnergy"
Date: June 2022
...
BrandVLM
(generates style features and style embedding for brand-aligned image)
Brand: Monster
Sector: Energy Drinks
Image Description:
"Two cans of chilled drinks on a wooden table."
Image Style Characteristics (ŷ):
Background: dark
Lighting: soft
Concept: ...
Depth: ...
0.13
-0.25
0.41
...
...
...
0.68
Image Style Embedding (ê)
BrandDM
(generates brand-aligned image)
Brand-aligned Image

Figure 3. The overall process of generating brand-aligned images: (A) Given the brand name & sector, image description, social media post caption, date of posting; the BrandVLM is finetuned to generate the characteristics (or style features) and CLIP Embedding (or style embedding) for the corresponding real image. (B) Given the style characteristics, along with the image caption, brand name, sector and the style embedding; the BrandDM is trained to produce the real image of the brand. (C) While sampling, given the marketer-provided context for the social media post, the BrandVLM generates the style characteristics and the style embedding of the brand-aligned image, which are then ingested by the BrandDM to generate a brand-aligned image.

hyperparameter balancing the two objectives, and $\mathcal{L}_{\text{clip}}$ uses mean squared error between ground truth CLIP embeddings ($\mathbf{e}$) and BrandVLM predictions ($\hat{\mathbf{e}}$), inspired from [17, 48]. We choose MSE over cosine similarity because MSE is more sensitive to the magnitude differences in the embedding space, ensuring that the model learns not only which visual concepts are relevant but also their relative importance or intensity. Cosine similarity, being scale-invariant, would ignore magnitude information that often correlates with the visual prominence of style elements.

Phase 2: BrandDM Training. We finetune a foundation diffusion model (PixArt-Σ or SDXL) to generate brand-aligned images conditioned on both textual and visual style representations. The BrandDM learns to map the combined conditioning $(\mathbf{y}, \mathbf{e})$ to corresponding brand images using standard denoising objective:

$$\mathcal{L}_{\text{DM}} = \mathbb{E}\left[\|\epsilon - \epsilon_\theta\left(\mathbf{z}_t, \mathbf{y}, \mathbf{e}\right)\|^2\right] \tag{4}$$

Cross-Component Alignment. To ensure compatibility between BrandVLM prediction outputs and BrandDM expected inputs, we further perform alignment finetuning of BrandDM. We retain it on 70% of training data unchanged but replace ground truth style representations with BrandVLM predictions for the remaining 30%. This mixed training regime allows the BrandDM to adapt to the BrandVLM's output distribution while maintaining performance on ground truth data.

Inference. During generation, marketers provide context $\mathbf{x}$ to BrandVLM, which outputs predicted style characteristics $\hat{\mathbf{y}}$ and embeddings $\hat{\mathbf{e}}$. These condition the BrandDM to generate brand-aligned images $\hat{\mathbf{z}}$ (Figure 3).

5. Quantitative Evaluation

5.1. Evaluation Metrics

The performance of the generated images is compared using the following set of metrics –

(1) We calculate the binary classification accuracy of the generated images by training a simple MLP classifier to use (image style labels, brand name) pair as input and predict whether the pair is correct or not. The classifier is trained on equal numbers of correct and incorrect brand-image pairs from our training set. On real data, this classifier achieves 78.05% accuracy, establishing that our 15-attribute taxonomy contains sufficient information to distinguish between brands. We term this metric as **Classifier Accuracy**.

(2) We also finetune the CLIP [26] model (only the final image and text projection layers) over the training set, to get a **BrandCLIP** version of the model, and calculate the image-image and image-text similarities using this version.

(3) For measuring the *brand-alignment* of the generated images, we calculate the JS divergence between the probability distributions of the different style attributes obtained using the real images (brand_{gt}) and the generated ones

Base Model	Brand-style characteristics	Brand-VLM used	Brand-style Embeddings	Cross-component Alignment	Classification Accuracy(↑)	BAS(↓)	BrandCLIP-I(↑) (Image-Image alignment)	BrandCLIP-T(↑) (Image-Prompt alignment)	FID(↓)
PixArt-Σ [2]	-	-	-	-	74.58%	1.9936	0.6228	0.2091	45.93
PixArt-Σ	-	-	-	-	75.44%	1.3098	0.6521	0.2189	27.08
PixArt-Σ	✓	-	-	-	75.91%	1.0206	0.6993	0.2017	24.69
PixArt-$\Sigma^{\ddagger}$	✓	✓	-	-	75.6%	1.3109	0.6875	0.2032	26.31
PixArt-$\Sigma^{\ddagger}$	✓	✓	-	✓	75.9%	1.2422	0.6980	0.2042	26.14
SDV1.5+Dreamstyler [1]	-	-	-	-	74.96%	1.4176	0.6693	0.1818	29.77
SDXL [25]	-	-	-	-	73.16%	3.1084	0.5505	0.1794	62.16
SDXL	-	-	-	-	74.61%	1.2075	0.6569	0.2401	24.61
SDXL	✓	✓	-	-	75.35%	1.2248	0.6424	0.2355	29.85
SDXL+Dreambooth [27]	-	-	-	-	75.45%	1.4509	0.6155	0.2063	28.06
SDXL-IPAdapter [40]	✓	✓	✓	-	73.74%	2.5891	0.5509	0.1814	56.19
SDXL-IPAdapter [40]	✓	✓	✓	✓	75.09%	1.3606	0.7070	0.2083	31.78

Table 1. Performance comparison of different model settings on the test subset where all the brands have been used for training. Blue rows correspond to non-finetuned models. ✓under brand-style characteristics implies they have been added to the text prompt that is input to the Diffusion Model. ✓under brand-style (image) embeddings implies they have been used as input to the Diffusion Model. ✓under Cross-component Alignment implies that the BrandDM has further been finetuned to align with the BrandVLM output space, as described in Section 4. The rows highlighted in green are different versions of our proposed BrandFusion.

Base Model	Brand-style characteristics	Brand-VLM used	Brand-style Embeddings	BAS(↓)	BrandCLIP-I(↑) (Image-Image alignment)	BrandCLIP-T(↑) (Image-Prompt alignment)	FID(↓)
PixArt-Σ [2]	-	-	-	1.3754	0.6472	0.1840	45.94
PixArt-Σ	-	-	-	0.7635	0.6488	0.1886	29.30
PixArt-Σ	✓	✓	-	0.8066	0.6559	0.1956	27.28
SDV1.5+Dreamstyler [1]	-	-	-	0.8843	0.6380	0.1760	36.56
SDXL [25]	-	-	-	4.8071	0.5401	0.1103	142.83
SDXL	✓	✓	-	0.8135	0.6491	0.2135	32.50
SDXL+Dreambooth [27]	-	-	-	0.8786	0.6272	0.1817	28.33
SDXL-IPAdapter [40]	✓	✓	✓	1.3360	0.6363	0.1844	41.92

Table 2. Performance comparison of different model settings on the test subset where the brands in the test set are not observed during training. Blue rows correspond to non-finetuned models. ✓under brand-style characteristics implies they have been added to the text prompt that is input to the Diffusion Model. ✓under brand-style (image) embeddings implies they have been used as input to the Diffusion Model. The rows highlighted in green are different versions of our proposed BrandFusion.

(brand$_{\text{gen}}$). Thus, the brand alignment score, or **BAS** is given by:

$$\text{BAS} = \frac{1}{N}\sum_{i=1}^{N} \text{JSD}\left(i, \text{brand}_{\text{gt}}, \text{brand}_{\text{gen}}\right), \tag{5}$$

$$\begin{aligned} \text{JSD}\left(i, \text{brand}_{\text{gt}}, \text{brand}_{\text{gen}}\right) = \frac{1}{2}\sum_{j=1}^{n_i} r_j^{\text{brand}_{\text{gt}}} \log\left(\frac{r_j^{\text{brand}_{\text{gt}}}}{r_j^m}\right) \\ + (1-r_j)^{\text{brand}_{\text{gt}}} \log\left(\frac{(1-r_j)^{\text{brand}_{\text{gt}}}}{(1-r_j)^m}\right) \\ + r_j^{\text{brand}_{\text{gen}}} \log\left(\frac{r_j^{\text{brand}_{\text{gen}}}}{r_j^m}\right) \\ + (1-r_j)^{\text{brand}_{\text{gen}}} \log\left(\frac{(1-r_j)^{\text{brand}_{\text{gen}}}}{(1-r_j)^m}\right), \end{aligned} \tag{6}$$

$$\text{where} \quad r_j^m = \frac{r_j^{\text{brand}_{\text{gt}}} + r_j^{\text{brand}_{\text{gen}}}}{2}$$

Here, the labels for each attribute are denoted by j, where $1 \leq j \leq n_i$, such that we have total n_i labels for attribute i, and $r_j = \frac{\#\text{images of brand with label } j}{\#\text{total images of brand containing humans}}$ if the attribute is a human-related attribute, and $r_j = \frac{\#\text{images of brand with label } j}{\#\text{total images of brand}}$ otherwise.

(4) Besides these metrics, we also employ the **FID** score [13] to calculate the image quality.

5.2. Setup

In Table 1, we show the performance of images generated in different settings using the above metrics. Note that we use a subset of 50 brands (out of total 183) belonging to 14 different sectors, and sample 80 images per brand from the test set images to get a smaller test set, which has been used in the experiments in Table 1. For the Dreambooth [27] baseline, we finetune the SDXL model separately for each brand in the test set using LoRA [14] (therefore it has 50 models in total), considering all the training set images of the brand as the reference images. In Dreamstyler [1], for each test set image, we search for the closest train set image of that brand (using CLIP image similarity) and use it as the source style image. Next, to observe the effectiveness of our approach over brands not encountered while training, we create a new train-test split, with 26 brands (from 25 different sectors) included in the test set, and the rest in the training set. For each of the 26 brands, we randomly select 160 images to get a test subset of 4,160 images; which has been used to compare the performance of different model settings in Table 2. In both the tables, all the images have been sampled at a resolution of 512×512.

5.3. Results Discussion

In Table 1, where the test brands are present during training, BrandFusion (BrandVLM + BrandDM) shows robust improvements across nearly all brand-relevant metrics. We experiment with two base diffusion model architectures as BrandDM - PixArt-Σ [2] and SDXL [25]. BrandFusion consistently demonstrates improved classifier accuracy which reflects how well the model generates images that can be classified back into the intended brand, confirming visual alignment with brand identity. With SDXL as BrandDM, BrandFusion achieves 75.35% accuracy, whereas with PixArt-Σ it scores 75.9%. BrandFusion-SDXL gets a BAS of 1.2248, which is an improvement over baseline SDXL (3.1084). However, simple finetuning fares better in aligning the outputs with brand feature distributions. Both Pixart-Σ and SDXL as BrandDM show improvements in BrandCLIP-I over the corresponding baseline models, indicating enhanced fine-grained visual similarity to real brand images. Importantly, BrandCLIP-I captures visual details not reflected in the discrete BAS metric, such as subtle color gradients, texture quality, and photographic nuances that are crucial for brand perception but difficult to categorize into our 15-attribute taxonomy.

To isolate the effect of visual embeddings, we compare SDXL (row 9) with SDXL-IPAdapter (row 12). While BAS scores remain comparable (since both rely on textual characteristics), the significant improvement in BrandCLIP-I (+0.0646) demonstrates that visual embeddings capture complementary brand information beyond discrete textual attributes. This validates our dual-output design: textual characteristics ensure brand alignment across interpretable features, while visual embeddings enhance fine-grained visual fidelity. While absolute CLIP improvements appear modest (0.01–0.05), existing works like StyleDrop [29] demonstrate that such gains correspond to substantial perceptual improvements in style alignment.

Next, we observe a slight drop in BrandCLIP-T scores for BrandFusion over the trained baseline models. This can be attributed to the large prompt (average 223 tokens), which includes image verbalization (average 158 tokens). The large prompt length is likely to reduce prompt adherence. However, this is compensated by the higher perceptual alignment of generated images with target brand styles as can be seen in Section 6. We note that FID trends are inconsistent with other brand-alignment metrics. This discrepancy arises because FID is based on InceptionNet [32] model trained on the ImageNet dataset [4], which lacks representation of marketing-specific visual styles. As a result, FID is not reliable for evaluating brand-coherent generation, since it fails to capture nuanced styling (e.g., depth, lighting, photography genre, etc.). Notably, our SDXL-IPAdapter based BrandFusion with cross-component alignment fares better than the existing Dreambooth personalization and Dreamstyler artistic transfer approaches, in terms of BAS, BrandCLIP-I and BrandCLIP-T scores.

In Table 2, the models are tested on brands never observed during training, however they belong to the same industry sectors whose brands were used while training. We adopt this training regime to investigate our model's cross-brand learning capabilities. This is a more challenging generalization task. BrandCLIP-I scores are consistently improved through BrandFusion for both PixArt-Σ and SDXL baselines, as well as Dreamstyler and Dreambooth baselines, thus signifying higher semantic alignment over unseen brand's images. For both SDXL and PixArt-Σ based BrandDMs, BrandFusion achieves a better Brand-Alignment Score than the respective baseline models. However, simple finetuning strategy leads to an even better BAS for PixArt-Σ, albeit with worse BrandCLIP and FID scores. The FID scores are inconsistent with brand-alignment metrics for unseen brands too, producing better FID for Dreambooth tuned SDXL, even though its Brand-alignment score and BrandCLIP scores are worse than BrandFusion-SDXL. Interestingly, the BrandCLIP-T scores also improve over the baselines (0.1 improvement for SDXL and 0.01 for PixArt-Σ), even with larger verbalization based prompts, signifying increased adherence to brand style characteristics for unseen brands.

5.4. User Study

To further evaluate the adherence of generated images to visual brand styles, we conduct a user study. For the study we ask 25 participants to voluntarily and anonymously answer a questionnaire comprising 20 pairwise image comparisons, sampled from a bank of 300 questions, giving a total of 500 rankings. In each question, the user is shown the real on-brand image for reference, along with two images generated by two different pipelines for comparison. Next, the user is asked to choose which of the two images seems more visually aligned to the reference image. The pipelines chosen for comparison are: (1) Dreambooth-tuned SDXL and (2) our trained BrandFusion based on SDXL. The win rate of BrandFusion is found to be 66.11%, i.e. on an average, a user prefers BrandFusion generated images over Dreambooth generated images 66 times out of 100, thus demonstrating the superior brand-alignment of BrandFusion.

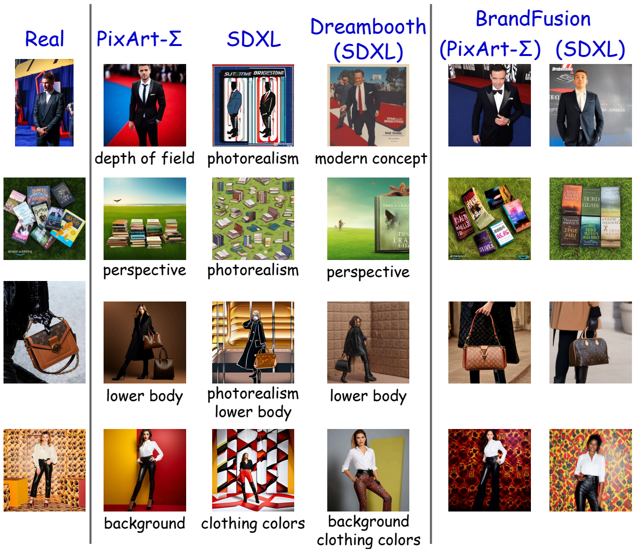


Figure 4. In each row, we show a real image and the corresponding images generated using different methods, specified on the top of each column. Below each row, we mention the style characteristics that are inconsistent between the generated image and corresponding real one. BrandFusion generated images are able to preserve characteristics such as *camera perspective*, *depth of field*, *visible body section*, *clothing colors* and *background pattern*, which are not reflected in the images generated using existing methods.

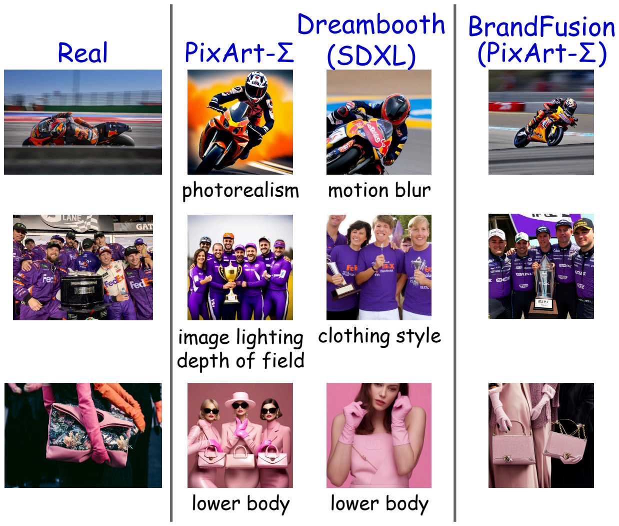


Figure 5. Visualizing images of unseen brands: In each row, we show a real image and the corresponding images generated using different methods, specified on the top of each column. Below each row, we mention the style characteristics that are inconsistent between the generated image and corresponding real one. BrandFusion is able to inculcate brand-relevant characteristics such as *motion blur* and *protective clothing style* for racing and *lower body crop* for handbag accessory images.

6. Qualitative Comparison

To qualitatively demonstrate the improved brand-alignment, in Figure 4 we show some example advertisements and their corresponding generated images using three different baselines (PixArt-Σ, SDXL and Dreambooth-SDXL), and compare them with BrandFusion generated images. In row 1, the PixArt-Σ image shows a considerable loss of depth of field (with blurred background), SDXL image is cartoon-like and the Dreambooth image has a vintage concept, unlike the original image (which has modern concept) – none of these differences are present in the BrandFusion images. Similarly, the camera perspective (top view of books) is maintained by BrandFusion in row 2, just like the real image, which does not happen with PixArt-Σ/SDXL/Dreambooth-SDXL. Again, in row 3, since the salient part of the real image is the handbag, only the lower body section of the model is visible, which is consistent in the BrandFusion images, unlike the other baselines. Finally, in row 4, the patterned image background with black-and-white clothing is maintained in BrandFusion, but not in the other baseline images. Next, we also show some examples of images generated for unseen brands in Figure 5, where in row 1, since the image is of a racing motorbike, motion blur image effect is produced in the BrandFusion generated image, just like the real one. This does not happen with Dreambooth or PixArt-Σ, whose image lacks photorealism too. In the middle row, protective style clothing (with caps and jackets) is not generated by Dreambooth, while PixArt-Σ image has diffused image lighting with shallow depth of field (i.e. out-of-focus background). All these inconsistencies are absent in the BrandFusion image, which has natural lighting. In the last row, PixArt-Σ and Dreambooth images show the upper body sections of the models, with handbags at a distance (or absent), while the handbags are highlighted in the BrandFusion image by showing only the lower body section of the models, as is the case with the corresponding real image too.

7. Conclusion

This work introduces BrandFusion, a novel framework that addresses the critical challenge of generating brand-aligned promotional images at scale. Our two-component architecture decouples brand style learning (BrandVLM) from image generation (BrandDM), overcoming the limitations of existing style learning and personalization methods. The key innovation lies in our dual representation approach: textual style characteristics provide marketers with interpretable control, while visual embeddings capture fine-grained perceptual brand nuances. This design ensures generated content adheres to brand guidelines while remaining accessible to marketing professionals. Extensive evaluation demonstrates consistent improvements across brand alignment metrics, achieving a 66.11% preference rate in human study and effective generalization to unseen brands. These results validate our approach's practical utility for large-scale content creation workflows. BrandFusion establishes a foundation for next-generation brand-aware generative models, with potential extensions to video generation and applications beyond marketing. By enabling scalable, interpretable brand-aligned image generation, this work represents a significant step toward practical AI-assisted content creation for professional marketing workflows.

References

[1] Namhyuk Ahn, Junsoo Lee, Chunggi Lee, Kunhee Kim, Daesik Kim, Seung-Hun Nam, and Kibeom Hong. Dreamstyler: Paint by style inversion with text-to-image diffusion models, 2023. 3, 6

[2] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-sigma: Weak-to-strong training of diffusion transformer for 4k text-to-image generation, 2024. 1, 6, 7

[3] Custom Market Insights. Global AI Powered Content Creation Market Size/Share Worth USD 7.9 Billion by 2033 at a 7.7% CAGR, 2024. 1

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 7

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 4

[6] Bi Qi Fong and John See. Branddiffusion: Multimodal personalized marketing visual content generation. In *Proceedings of the 2nd International Workshop on Multimedia Content Generation and Evaluation: New Methods and Practice*, page 72–77, New York, NY, USA, 2024. Association for Computing Machinery. 2

[7] Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan. Guiding Instruction-based Image Editing via Multimodal Large Language Models. In *International Conference on Learning Representations (ICLR)*, 2024. 4

[8] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion, 2022. 2, 3

[9] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. A neural algorithm of artistic style, 2015. 2

[10] Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, WUYOU XIAO, Rui Zhao, Shuning Chang, Weijia Wu, Yixiao Ge, Ying Shan, and Mike Zheng Shou. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. In *Advances in Neural Information Processing Systems*, pages 15890–15902. Curran Associates, Inc., 2023. 3

[11] Shunan Guo, Zhuochen Jin, Fuling Sun, Jingwen Li, Zhaorui Li, Yang Shi, and Nan Cao. Vinci: An intelligent graphic design system for generating advertising posters. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2021. Association for Computing Machinery. 2

[12] Jochen Hartmann, Yannick Exner, and Samuel Domdey. The power of generative marketing: Can generative ai create superhuman visual marketing content? *International Journal of Research in Marketing*, 42(1):13–31, 2025. 2

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 6629–6640, Red Hook, NY, USA, 2017. Curran Associates Inc. 6

[14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 6

[15] Yuzhou Huang, Liangbin Xie, Xintao Wang, Ziyang Yuan, Xiaodong Cun, Yixiao Ge, Jiantao Zhou, Chao Dong, Rui Huang, Ruimao Zhang, et al. Smartedit: Exploring complex instruction-based image editing with multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8362–8371, 2024. 4

[16] Jingbo Wang Yanhong Zeng Xiangtai Li Jianzong Wu, Chao Tang and Yunhai Tong. Diffsensei: Bridging multimodal llms and diffusion models for customized manga generation. *arXiv preprint arXiv:2412.07589*, 2024. 4

[17] Varun Khurana, Yaman Kumar Singla, Jayakumar Subramanian, Changyou Chen, Rajiv Ratn Shah, zhiqiang xu, and Balaji Krishnamurthy. Measuring and improving engagement of text-to-image generation models. In *The Thirteenth International Conference on Learning Representations*, 2025. 5

[18] Philip Kotler and Kevin Lane Keller. *Marketing Management*. Prentice Hall, 14th edition, 2011. 1

[19] Gihyun Kwon and Jong Chul Ye. Clipstyler: Image style transfer with a single text condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18062–18071, 2022. 3

[20] Ryan Law. Marketers Using AI Publish 42% More Content [+ New Research Report], 2024. 1

[21] Jinpeng Lin, Min Zhou, Ye Ma, Yifan Gao, Chenxi Fei, Yangjian Chen, Zhang Yu, and Tiezheng Ge. Autoposter: A highly automatic and content-aware design system for advertising poster generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, page 1250–1260, New York, NY, USA, 2023. Association for Computing Machinery. 2

[22] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, pages 34892–34916. Curran Associates, Inc., 2023. 4

[23] Barbara J. Phillips, Edward F. McQuarrie, and W. Glenn Griffin and. The face of the brand: How art directors understand visual brand identity. *Journal of Advertising*, 43(4):318–332, 2014. 2, 4

[24] Barbara J. Phillips, Edward F. McQuarrie, and W. Glenn Griffin. How visual brand identity shapes consumer response. *Psychology & Marketing*, 31(3):225–236, 2014. 2, 4

[25] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023. 6, 7

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 3, 4, 5

[27] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22500–22510, 2023. 2, 3, 6

[28] Mohammad Amin Shabani, Zhaowen Wang, Difan Liu, Nanxuan Zhao, Jimei Yang, and Yasutaka Furukawa. Visual layout composer: Image-vector dual diffusion model for design layout generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9222–9231, 2024. 2

[29] Kihyuk Sohn, Nataniel Ruiz, Kimin Lee, Daniel Castro Chin, Irina Blok, Huiwen Chang, Jarred Barber, Lu Jiang, Glenn Entis, Yuanzhen Li, Yuan Hao, Irfan Essa, Michael Rubinstein, and Dilip Krishnan. Styledrop: Text-to-image generation in any style, 2023. 3, 7

[30] Zhengwentai Sun, Yanghong Zhou, Honghong He, and P.Y. Mok. Sgdiff: A style guided diffusion model for fashion synthesis. In *Proceedings of the 31st ACM International Conference on Multimedia*, page 8433–8442, New York, NY, USA, 2023. Association for Computing Machinery. 3

[31] Ananda Padhmanabhan Suresh, Sanjana Jain, Pavit Noinongyao, Ankush Ganguly, Ukrit Watchareeruetai, and Aubin Samacoits. Fastclipstyler: Optimisation-free text-based image style transfer using style representations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 7316–7325, 2024. 3

[32] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions . In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, Los Alamitos, CA, USA, 2015. IEEE Computer Society. 7

[33] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. 4

[34] Shaodong Wang, Yunyang Ge, Liuhan Chen, Haiyang Zhou, Qian Wang, Xinhua Cheng, and Li Yuan. Prompt2poster: Automatically artistic chinese poster creation from prompt only. In *Proceedings of the 32nd ACM International Conference on Multimedia*, page 10716–10724, New York, NY, USA, 2024. Association for Computing Machinery. 2

[35] Haohan Weng, Danqing Huang, Yu Qiao, Zheng Hu, Chin-Yew Lin, Tong Zhang, and C. L. Philip Chen. Desigen: A pipeline for controllable design template generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12721–12732, 2024.

[36] Qin Wu and Peizi Zhou. Advertisement synthesis network for automatic advertisement image synthesis. *International Journal of Antennas and Propagation*, 2024(1):8030907, 2024. 2

[37] Yi Wu, Ziqiang Li, Heliang Zheng, Chaoyue Wang, and Bin Li. Infinite-id: Identity-preserved personalization via id-semantics decoupling paradigm, 2024. 3

[38] Zipeng Xu, Enver Sangineto, and Nicu Sebe. Stylerdalle: Language-guided style transfer using a vector-quantized tokenizer of a large-scale generative model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7601–7611, 2023. 3

[39] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. *arXiv preprint arXiv:2211.13227*, 2022. 3

[40] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models, 2023. 6

[41] Ziyang Yuan, Mingdeng Cao, Xintao Wang, Zhongang Qi, Chun Yuan, and Ying Shan. Customnet: Zero-shot object customization with variable-viewpoints in text-to-image diffusion models, 2023. 3

[42] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 3

[43] Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10146–10156, 2023. 3

[44] Yuechen Zhang, Jinbo Xing, Eric Lo, and Jiaya Jia. Real-world image variation by aligning diffusion inversion chain, 2023. 3

[45] Zhanjie Zhang, Quanwei Zhang, Huaizhong Lin, Wei Xing, Juncheng Mo, Shuaicheng Huang, Jinheng Xie, Guangyuan Li, Junsheng Luan, Lei Zhao, et al. Towards highly realistic artistic style transfer via stable diffusion with step-aware and layer-aware prompt. *arXiv preprint arXiv:2404.11474*, 2024. 3

[46] Zhanjie Zhang, Quanwei Zhang, Wei Xing, Guangyuan Li, Lei Zhao, Jiakai Sun, Zehua Lan, Junsheng Luan, Yiling

Huang, and Huaizhong Lin. Artbank: Artistic style transfer with pre-trained diffusion model and implicit style prompt bank. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7396–7404, 2024. 3

[47] Du Zhenbang, Feng Wei, Wang Haohan, Li Yaoyu, Wang Jingsen, Li Jian, Zhang Zheng, Lv Jingjing, Zhu Xin, Jin Junsheng, Shen Junjie, Lin Zhangang, and Shao Jingping. Towards reliable advertising image generation using human feedback. In *European Conference on Computer Vision*, 2024. 2

[48] Yufan Zhou, Ruiyi Zhang, Jiuxiang Gu, and Tong Sun. Customization assistant for text-to-image generation, 2024. 4, 5

Appendices

A. Hyperparameters

All the trainings were done on 8 Nvidia A100 GPUs.

- BrandVLM training: Batch size 1024, Learning rate $2e-5$, Number of epochs 1.5, AdamW Optimizer, Cosine learning rate scheduler with warmup ratio 0.03
- BrandDM training
  1. PixArt-Σ: Batch size 1024, Learning rate $2e-5$, Number of epochs 40, CAME optimizer, Constant learning rate scheduler
  2. SDXL: Batch size 1024, Learning rate $1e-6$, Number of epochs 20, AdamW Optimizer, Constant learning rate scheduler
  3. SDXL-IPAdapter: Batch size 32, Learning rate $1e-6$, Number of epochs 20, AdamW Optimizer, Constant learning rate scheduler
- Brand-Classifier training (Used for evaluation): Batch size 32, Learning rate $1e-3$, Number of epochs 80, AdamW Optimizer, 42K parameters, training strategy – to obtain (brand style attribute labels, brand) pairs during training, for each brand's train images (forming correct pairings), we sample an equal number of images from other brands of the same industry sector and pass the brand as input to get incorrect pairings
- BrandCLIP training: OpenAI's ViT-L-14-336 base model, Batch size 128, Learning rate $1e-3$, Number of epochs 5, AdamW Optimizer, 0.1 weight decay

B. BrandVLM performance comparison

Model version	Verbalisation evaluation			Embedding evaluation	
	BLEU score (↑)	IoU score (↑)	BAS (↓)	MSE (↓)	Cosine similarity (↑)
LLaVA-1.5-13B	0.0788	0.1859	16.1064	-	-
BrandVLM	0.4569	0.5499	3.3752	0.5102	0.5598

Table 3. Different VLM performances for generating image style characteristics (verbalization) and brand-relevant image style embeddings on the full test set. The non-finetuned baseline does not predict the style embeddings. The IoU score is the Intersection over Union of the style labels predicted by the VLM and style labels present in the ground truth verbalizations. BAS refers to the Brand Alignment Score as explained in Section 5.1.

C. Prompt used for training BrandVLM

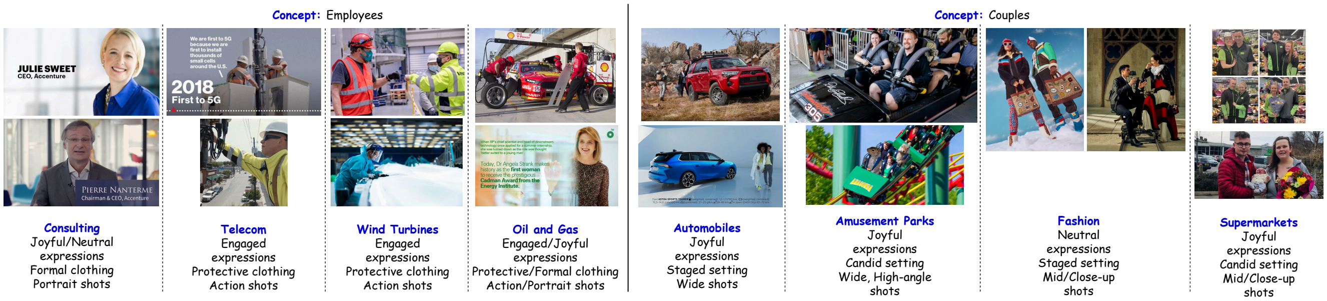
```
A marketer from company company which
    belongs to the sector companySector
    wants to create a social media post
    containing an image for marketing
    purposes. The following information
    about the social media post is available
    :
(1) Social media post text: tweetText
(2) Date of posting: tweetTimeStamp
(3) Number of likes on a scale of 0 to 100
    that the social media post is expected
    to receive: tweetLikesPercentile
(4) Image description: imCaption
(5) Image tags: imKeywords.
Now, considering the company's visual
    identity and the above information,
    predict the colors and tones describing
    the image that the marketer will use in
    the social media post from the lists
    given below. Also predict the spatial
    coverage ratios (with respect to the
    total image area) of the colors and
    tones that will be used.
- Allowed colors: [Beige, Black, Blue,
    Bright_Green, Brown, Cream, Cyan,
    Dark_Blue, Dark_Brown, Dark_Gray,
    Dark_Green, Dark_Pink, Dark_Red, Emerald
    , Gold, Gray, Green, Khaki, Lavender,
    Light_Blue, Light_Green, Lilac, Magenta,
     Maroon, Mud_Green, Mustard, Off_White,
    Olive, Orange, Pink, Plum, Purple, Red,
    Royal_Blue, Silver, Tan, Turquoise,
    Violet, White, Yellow]
- Allowed tones: [warm, neutral, cool]
Now, for each of the following 8 visual
    attributes, predict at most 3 labels
    which should be the most prominent in
    the image out of the given list of
    labels.
(1) Image Lighting: [bright, dark, moderate
    , studio, natural, soft, hard, light
    glare, vignette, colored, light on
    subject]
(2) Perspective: [bird eye view, worm eye
    view, fish eye view, panorama view,
    centered composition, rule of third,
    altered perspective, framed image, high
    angle photo, low angle photo, vertical
    composition, corner shot, point of view
    shot, audience perspective]
(3) Image Background: [solid, pattern,
    gradient, background as frame, textured,
     wood, blurred, transparent, bright,
    dark, light]
(4) Color Palette: [grayscale, monotone,
    two tone, bright colors, pastel colors,
    complementary colors, analogous colors,
    inverted colors, galaxy colors, aquatic
    colors, sunset colors, autumnal colors]
(5) Photography Genre: [architectural,
    candid, staged, portrait, selfie, group,
     product, fashion, beauty, bridal,
    interior, street, landscape, sky, still-
    life, action, underwater, botanical,
    historical, amateur, abstract, live
    stage]
(6) Concept: [illustration, photorealism,
    typography, vintage, graphic design,
    cartoon, incomplete art, wave pattern,
    text heavy]
(7) Depth: [wide angle shot, mid shot,
    close up shot, macro shot, motion blur,
    radial blur, gaussian blur, fully
    focused subject, unfocused subject,
    partly focused subject, bokeh effect,
    isolated focal point, multiple focal
    points, bright focal point, dark focal
    point, shallow depth of field]
(8) Image Effects: [short exposure, long
    exposure, neutral density filter,
    artificial shadow, silhouette, pixelated
     image, vanishing point, negative space,
     motion capture, cut-out, symmetric,
    asymmetry, low saturation, high
    saturation, low contrast, high contrast]
If the main subject of this image contains
    a human, then for each of the following
    7 attributes, predict at most 3 labels
    out of the provided list which should be
     relevant to the subject in the image,
    otherwise predict 'Not applicable' for
    each of these attributes.
(1) Hair Style: [short, covered, wavy,
    loose, varied, straight, neat, ponytail,
     casual, tied back, flowing, curly, updo
    , pulled back, braided, Not applicable]
(2) Facial Expression: [engaged, content,
    focused, neutral, joyful, relaxed,
    contemplative, Not applicable]
(3) Clothing Style: [casual, athletic,
    formal, business, swimwear, business
    casual, traditional, protective,
    beachwear, costume, form fitting, Not
    applicable]
(4) Clothing Color Palette: [neutral,
    colorful, vibrant, monochrome, earthy,
    pastel, muted, Not applicable]
(5) Posing: [standing, seated, holding,
    leaning, active, reclined, walking,
    stretching, dynamic, running, relaxed,
    confident, Not applicable]
(6) Gaze: [forward, downward, sideways,
    away, upward, outward, engaged, Not
    applicable]
(7) Visible Body section: [upper body, full
     body, hand only, lower half, close up,
    midsection, full back, head shot, Not
    applicable]
Answer properly in JSON format with the
    following keys - "colors_and_tones", "
    image_lighting", "perspective", "
    image_background", "color_palette", "
    photography_genre", "concept", "depth",
    "image_effects", "hair_style", "
    facial_expression", "clothing_style", "
    clothing_color_palette", "posing", "gaze
    ", "visible_body_section". The values of
     the JSON should be in a dictionary for
    colors_and_tones and a list for all
    other keys. Do not include any other
    information in your answer.
```

Listing 1. Prompt used while training BrandVLM

Considered Brand Sectors
Aerospace, Agricultural Heavy Equipment, Airline, Amusement Park, Automobile, Beauty, BioTech, Brewery, Car rental, Construction, Consulting, Consumer goods, Cruise, Defense, Drink, E-commerce, Education, Electronics, Entertainment, Eyecare, Fashion, Finance, Fitness, Food, Footwear, Gaming, Gas, Hardware, Healthcare, Home appliances, Homecare, Hospitality, Insurance, Jewelry, MLM, Networking, NGO, Oil, Parcel service, Pet Supermarket, Petrol station, Petroleum, Pharma, RailRoad, Research, Restaurant, Ride sharing, Satellite, Software, Sports, Supermarket, Telecom, Tires, Tourism, Underwater Diving, Watches, Wind Turbines

Characteristic	Labels
Image Lighting	Bright, Dark, Moderate, Studio, Natural, Soft, Hard, Light glare, Vignette, Colored, Light on subject
Perspective	Bird eye view, Worm eye view, Fish eye view, Panorama view, Centered composition, Rule of third, Altered perspective, Framed image, High angle photo, Low angle photo, Vertical composition, Corner shot, Point of view shot, Audience perspective
Image Background	Solid, Pattern, Gradient, Background as frame, Textured, Wood, Blurred, Transparent, Bright, Dark, Light
Color Palette	Grayscale, Monotone, Two tone, Bright colors, Pastel colors, Complementary colors, Analogous colors, Inverted colors, Galaxy colors, Aquatic colors, Sunset colors, Autumnal colors
Photography Genre	Architectural, Candid, Staged, Portrait, Selfie, Group, Product, Fashion, Beauty, Bridal, Interior, Street, Landscape, Sky, Still-life, Action, Underwater, Botanical, Historical, Amateur, Abstract, Live stage
Concept	Illustration, Photorealism, Typography, Vintage, Graphic design, Cartoon, Incomplete art, Wave pattern, Text heavy
Depth	Wide angle shot, Mid shot, Close up shot, Macro shot, Motion blur, Radial blur, Gaussian blur, Fully focused subject, Unfocused subject, Partly focused subject, Bokeh effect, Isolated focal point, Multiple focal points, Bright focal point, Dark focal point, Shallow depth of field
Image Effects	Short exposure, Long exposure, Neutral density filter, Artificial shadow, Silhouette, Pixelated image, Vanishing point, Negative space, Motion capture, Cut-out, Symmetric, Asymmetry, Low saturation, High saturation, Low contrast, High contrast
Hair Style	Short, Covered, Wavy, Loose, Varied, Straight, Neat, Ponytail, Casual, Tied back, Flowing, Curly, Updo, Pulled back, Braided
Facial Expression	Engaged, Content, Focused, Neutral, Joyful, Relaxed, Contemplative
Clothing Style	Casual, Athletic, Formal, Business, Swimwear, Business casual, Traditional, Protective, Beachwear, Costume, Form fitting
Clothing Color Palette	Neutral, Colorful, Vibrant, Monochrome, Earthy, Pastel, Muted
Posing	Standing, Seated, Holding, Leaning, Active, Reclined, Walking, Stretching, Dynamic, Running, Relaxed, Confident
Gaze	Forward, Downward, Sideways, Away, Upward, Outward, Engaged
Visible Body Section	Upper body, Full body, Hand only, Lower half, Close up, Midsection, Full back, Head shot

Appendix D. The Brand sectors, Brand Style Characteristics and their corresponding labels considered in our approach. For details please refer to Section 3.



Appendix E. Here, we show more examples of nuanced Brand variations. Different brands from a particular sector can have some common characteristics, such as *Protective clothing style* of *employees* in *Telecom* sector, or *Joyful facial expressions* and *Candid setting* images of *couples* in *Amusement Parks* sector.