Mind-the-Glitch: Visual Correspondence for Detecting Inconsistencies in Subject-Driven Generation

Abdelrahman Eldesokey **Aleksandar Cvejic** **Bernard Ghanem** **Peter Wonka**
KAUST, Saudi Arabia
first.last@kaust.edu.sa

Figure 1: **Mind-the-Glitch** is the first pipeline that enables computing *visual correspondences* based on the backbone features of pre-trained diffusion models. The pipeline separates backbone features into semantic and visual components, allowing for *visually matching* keypoints across images, analogous to the well-established *semantic correspondence* task. This provides the first empirical framework for evaluating and localizing visual inconsistencies in subject-driven image generation.

Abstract

We propose a novel approach for disentangling visual and semantic features from the backbones of pre-trained diffusion models, enabling *visual correspondence* in a manner analogous to the well-established *semantic correspondence*. While diffusion model backbones are known to encode semantically rich features, they must also contain visual features to support their image synthesis capabilities. However, isolating these visual features is challenging due to the absence of annotated datasets. To address this, we introduce an automated pipeline that constructs image pairs with annotated semantic and visual correspondences based on existing subject-driven image generation datasets, and design a contrastive architecture to separate the two feature types. Leveraging the disentangled representations, we propose a new metric, *Visual Semantic Matching (VSM)*, that quantifies visual inconsistencies in subject-driven image generation. Empirical results show that our approach outperforms global feature-based metrics such as CLIP, DINO, and vision–language models in quantifying visual inconsistencies while also enabling spatial localization of inconsistent regions. To our knowledge, this is the first method that supports both quantification and localization of inconsistencies in subject-driven generation, offering a valuable tool for advancing this task. Project Page: https://abdo-eldesokey.github.io/mind-the-glitch/

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

1 Introduction

Recent advances in diffusion models have transformed the landscape of image generation, delivering exceptional quality, fidelity, and diversity [8, 5, 32, 36, 27, 30]. These capabilities have revolutionized several domains, including artistic creation, advertising, content production for video games, and storyboards for movies. In many of these applications, maintaining visual consistency of a subject, whether a character or an object, across different generations is crucial. For example, in cinematic and advertising workflows, preserving the identity and appearance of a subject across multiple scenes is essential for narrative coherence and branding. However, most diffusion models operate in a latent space where semantic and visual concepts are entangled, making it challenging to control or preserve specific content using text prompts alone.

To address these limitations, a growing research direction has focused on *subject-driven* generation, which aims to guide diffusion models to produce consistent images of a given subject in different scenes while preserving fine-grained visual details [13, 33, 28, 3, 49]. Nonetheless, a major bottleneck in this line of research has been the lack of reliable evaluation metrics for subject consistency. Since subjects may appear in varying poses and spatial configurations, traditional image similarity metrics such as LPIPS [53], or structural similarity (SSIM) are not well-suited for this task. As a result, many works have relied on feature-based similarities using models like CLIP [29] or DINO [4] as approximations of overall appearance. However, these metrics capture global semantics and tend to overlook subtle visual inconsistencies in object details [26].

Recent works [41, 26] have explored the use of Vision-Language Models (VLMs), such as ChatGPT, to assess consistency between a reference and a generated image. While these approaches are promising, they remain limited to global assessments, and it is often unclear which visual cues or criteria the models rely on to judge consistency. Furthermore, VLM-based metrics lack the ability to localize inconsistent regions within the subject. To enable robust evaluation of subject-driven image generation, it is essential to develop methods that can reliably match the visual appearance of a subject across images, irrespective of variations in pose, scale, or context, and accurately localize inconsistencies for potential correction through post-processing.

This challenge bears similarities to the task of *semantic correspondence*, where corresponding points are matched across image pairs of objects with varying poses, scales, and contexts. Diffusion models have been shown to encode semantically-rich features that have demonstrated great success in computing semantic correspondences across images [39, 42, 52, 22, 51]. However, as illustrated in Figure 1, these semantic features are typically insensitive to appearance-level variations, as they focus primarily on structural or categorical semantics rather than fine-grained visual details. Given that diffusion models are trained for image synthesis, it is reasonable to assume that their internal representations should encode both semantic and visual information, yet only the semantic component has been extensively leveraged so far.

Building on this hypothesis, we propose a novel framework for disentangling semantic and visual features from the backbone of pre-trained diffusion models. Due to the absence of datasets with annotated visually similar or dissimilar regions of a given subject, we introduce an automated dataset generation pipeline that constructs image pairs with annotated semantic and visual correspondences, derived from existing subject-driven generation datasets. Using this dataset, we propose an architecture that disentangles semantic and visual features in a contrastive manner. We then leverage the disentangled representations to derive a metric, the *Visual Semantic Matching (VSM)* metric, that quantifies the degree of visual inconsistency in subject-driven generation. Empirical results show that our approach outperforms existing feature-based metrics such as CLIP and DINO, as well as the Vision-Language Model (VLM) metric, in quantifying visual inconsistencies, while allowing for localizing the inconsistent regions as well. We believe our framework offers a valuable step forward in the evaluation and development of subject-driven image generation.

Our contributions can be summarized as follows:

- We propose a novel framework comprising an automated dataset generation pipeline and an architecture for disentangling semantic and visual features in diffusion models.
- Based on the disentangled features, we present a new metric for evaluating subject-driven generation methods that both quantifies and localizes visual inconsistencies.
- We empirically demonstrate that the proposed metric outperforms existing feature-based metrics commonly used in subject-driven generation.

2 Related Work

In this section, we first provide a brief overview of the *semantic correspondence* task, which serves as a key inspiration for our work. Then we review existing subject-driven image generation approaches, and how consistency is currently evaluated in these methods.

2.1 Semantic Correspondence using Diffusion Models

Diffusion models have been shown to encode semantically rich features within their backbones, enabling computing semantic correspondence across instances of the same object class and even among semantically similar categories. Some approaches leveraged intermediate features from the decoder of diffusion models [42], and others combine them with features from DINO [4] to enhance semantic representation [52]. Several works [22, 51] proposed learnable aggregation networks that automatically select and fuse features from different layers, trained in a supervised manner using semantic correspondence datasets such as SPair-71k [24]. CleanDIFT [39] further introduces a distillation framework to compress diffusion features, allowing for efficient inference. These semantically rich features have been successfully applied to a range of downstream tasks, including segmentation [45, 23], controllable editing [25, 2, 7], and object manipulation [20]. Inspired by the task of semantic correspondence, we aim to learn feature representations that capture visual appearance rather than semantics. These visual features enable matching regions based on appearance and are well-suited for detecting visually inconsistent regions in subject-driven image generation.

2.2 Subject-Driven Image Generation

Early personalized generation techniques using UNet-based diffusion models [32, 27] primarily focused on encoding subject identity through full-model fine-tuning [34, 17] or low-rank adaptation methods [12, 10, 37, 19], and even training-free approaches [43] Other approaches [35, 9, 15, 1] learned subject-specific embeddings, without modifying model weights, by associating subjects with new tokens in an image or text embedding space. To improve the trade-off between efficiency and fidelity, adapter-based methods [48, 44, 13, 28, 50] introduced zero-shot personalization by conditioning on reference image features through lightweight network modules. These methods showed that injecting image-driven signals or selectively updating parameters could effectively preserve subject identity and visual details, even with limited data.

The emergence of diffusion Transformers (DiTs) [8, 5] introduced a new class of architectures that replace UNets with Vision Transformers as the denoising backbone, offering greater scalability and enhanced contextual understanding. These models support more flexible conditioning mechanisms and exhibit strong in-context learning capabilities, enabling them to generate diverse images of a subject without explicit fine-tuning [38, 18]. To further improve fidelity and identity preservation, several works have applied LoRA-based fine-tuning to DiTs architectures [41, 54, 14, 3], achieving more precise control over subject appearance and consistency. More recently, visual autoregressive models (VARs) have been adopted for subject-driven generation due to their inherently strong conditional modeling capabilities [40, 47, 6]. For an exhaustive overview of subject-driven image generation approaches, we refer the readers to [46].

2.3 Evaluating Subject-Driven Image Generation

Visual similarity is often measured using traditional metrics such as Structural Similarity Index (SSIM), Peak Signal-to-Noise Ratio (PSNR), and LPIPS [53], which assume that the object is spatially aligned across images. However, in subject-driven generation, this assumption does not hold, as the subject may appear in different poses, positions, and contexts. As a result, it has become common to use global feature-based metrics, such as CLIP-Image [29] and DINO [4], which compare the similarity of image-level embeddings extracted from different models. While these methods are robust to spatial changes, they are inherently global and often fail to capture fine-grained appearance details of the subject. A recent trend involves leveraging Vision-Language Models (VLMs) to evaluate subject-driven generation by prompting them to score specific criteria of the generated images [41, 26]. However, it remains unclear how VLMs form their judgments, and they lack the ability to localize the source of inconsistency within the image. In this work, we aim to bridge this gap by leveraging visual features extracted from the backbones of diffusion models to assess

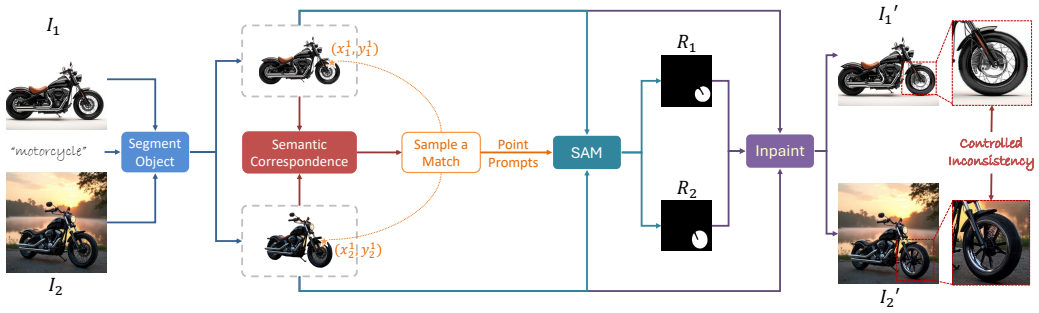

Figure 2: Automated Dataset generation pipeline for producing controlled visual inconsistencies. Access to such pairs of images enables separating visual features from pre-trained backbones in a contrastive manner.

the appearance similarity across images in subject-driven generation. Our approach enables both quantification and spatial localization of visual inconsistencies. By providing a reliable metric that offers fine-grained insight into subject consistency, we take a step toward more robust evaluation and improvement of subject-driven generation.

3 Method

In subject-driven image generation, assessing whether different parts of a subject are visually consistent across two images requires matching visual representations that are robust to changes in pose, scale, and environment. This challenge is similar to the semantic correspondence task, where several approaches [39, 22, 51, 52, 42] were trained on SPair-71k [24], which includes annotated semantic points between image pairs of given subjects. As demonstrated in Figure 1, semantic features are insensitive to appearance changes, making them unsuitable for matching visual appearance. To the best of our knowledge, there exists no dataset with annotated visual correspondences between visually similar or dissimilar regions, similar to SPair-71k.

To bridge this gap, we first introduce a novel automated pipeline for constructing a dataset with visual correspondences, leveraging existing datasets of subject-driven image generation such as Subjects200k [41] (Section 3.1). Using the data generated from this pipeline, we then propose an architecture for disentangling semantic and visual representations from the internal features of diffusion models (Section 3.2). Lastly, we introduce a metric that leverages the disentangled features for empirically evaluating visual consistency between image pairs, which quantifies the degree of consistency and localizes inconsistent regions (Section 3.3).

3.1 Data Generation Pipeline with Visual Correspondence

We aim to generate a dataset of image pairs with visual correspondences between *visually* similar and dissimilar regions inspired by the SPair-71k dataset. Given a consistent image pair I_1 and I_2 from a subject-driven dataset, we begin by segmenting the subject in each image using Grounded-SAM [31]. This isolates the subject from the background and makes the computation of correspondences more reliable. Next, we compute semantic correspondences between the two images within the segmented subject regions using the semantic correspondence method CleanDIFT [39]. Specifically, we extract features from the sixth decoder layer of the diffusion model UNet Φ, which has been shown to contain semantically rich information [42, 39], resulting in $F_1^6 = \Phi(I_1)$ and $F_2^6 = \Phi(I_2)$. We then compute the pairwise similarity between the two feature maps to form a similarity matrix $\mathcal{D} = F_1^6 \, {F_2^6}^T$. Corresponding points are obtained by selecting locations with the highest similarity scores using $\arg\max \mathcal{D}$, to obtain the following correspondence mapping:

$$C_1 = \langle (x_1^1, y_1^1), \ldots, (x_1^N, y_1^N) \rangle \ , \qquad C_2 = \langle (x_2^1, y_2^1), \ldots, (x_2^N, y_2^N) \rangle \ , \qquad C_1, C_2 \in \mathrm{R}^{N \times 2} \quad (1)$$

where N denotes the number of correspondences.

Based on these correspondences, we aim to match semantically similar regions across both images and then alter them visually to generate a pair of images with known, localized visual inconsistencies. To

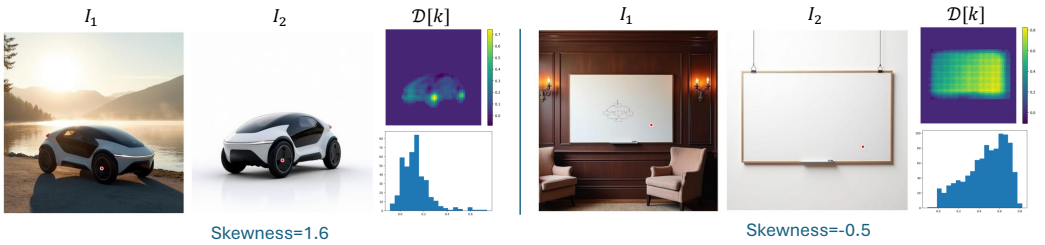

Figure 3: Examples illustrating how the skewness of the matching scores distribution correlates with matching ambiguity. High skewness implies a distinct match, while low skewness indicates diffuse or ambiguous correspondences.

achieve this, we sample a point k from C_1 that has a high similarity score and use the corresponding point pair (x_1^k, y_1^k) and (x_2^k, y_2^k) as prompts to the SAM model [16], which segments a localized region in each image. This produces region masks defined as $R_i = \texttt{SAM}(I_i, (x_i^k, y_i^k))$. We configure SAM to return multiple candidate masks and select the one with the smallest area, as it is more likely to correspond to an isolated semantic part rather than the entire object.

Handling Ambiguous Matches: When segmenting regions, a common challenge arises with flat-textured subjects, such as a whiteboard, where the selected point k tends to match broadly across the object (see Figure 3), leading SAM to segment the entire subject rather than a localized part. To address this, we propose using the skewness of the similarity distribution $\mathcal{D}[k]$ to identify ambiguous matches. The skewness is computed as:

$$\text{Skewness}(\mathcal{D}[k]) = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{\mathcal{D}[k] - \mu}{\sigma} \right)^3 \tag{2}$$

where $n = |\mathcal{D}[k]|$, $\mu = \texttt{mean}(\mathcal{D}[k])$, and $\sigma = \texttt{STD}(\mathcal{D}[k])$. We observe that high skewness corresponds to matches in textured regions, characterized by a long-tailed distribution and low ambiguity. In contrast, low skewness indicates more uniform similarity scores, suggesting ambiguous matches typically associated with flat surfaces as illustrated in Figure 3.

Validating Matched Regions: We apply additional heuristics to alleviate segmentation failures and to ensure that the selected regions in both images correspond to the same semantic part, including constraints on aspect ratio and relative size with respect to the full object. For regions that pass all checks, we perform inpainting using a diffusion-based inpainting model (we use SDXL [27] for efficiency), to obtain two inconsistent images I_1' and I_2'. Specifically, we crop a patch around the region with padding and feed it to the inpainting model to ensure that the model does not observe the rest of the object, promoting the generation of visually distinct content. To confirm that the inpainted region differs from the original, we compute the LPIPS score [53] between the original and inpainted regions in I_i and I_i'. We discard samples with low LPIPS scores to ensure meaningful variation. Further details on the filtering strategy are provided in the supplementary material. We provide an illustration for the data generation pipeline in Figure 2. Eventually, each dataset sample comprises a consistent image pair (I_1, I_2), an inconsistent pair (I_1', I_2') generated with our data generation pipeline, subject masks (O_1, O_2), inpainted region masks (R_1, R_2), subject prompt s and target prompt p.

3.2 Learning Disentangled Semantic and Visual Representations

Given an inconsistent image pair I_1' and I_2', which were generated with our automated pipeline in Section 3.1, where the inconsistent regions are defined by the masks R_1 and R_2, we aim to disentangle semantic and visual features of the diffusion model backbone. We start by extracting multi-layer features for both images using the diffusion model backbone Φ, yielding feature maps $F_1 = \Phi(I_1')$ and $F_2 = \Phi(I_2')$ that include features from multiple decoder layers $l \in L$. To aggregate features from the different layers, we follow the approach of [22], but instead of using a single aggregation network, we employ two separate networks, Ψ_s^l and Ψ_v^l, to aggregate semantic and visual representations separately. Each aggregation network encompasses a ResNet [11] block per decoder layer that are

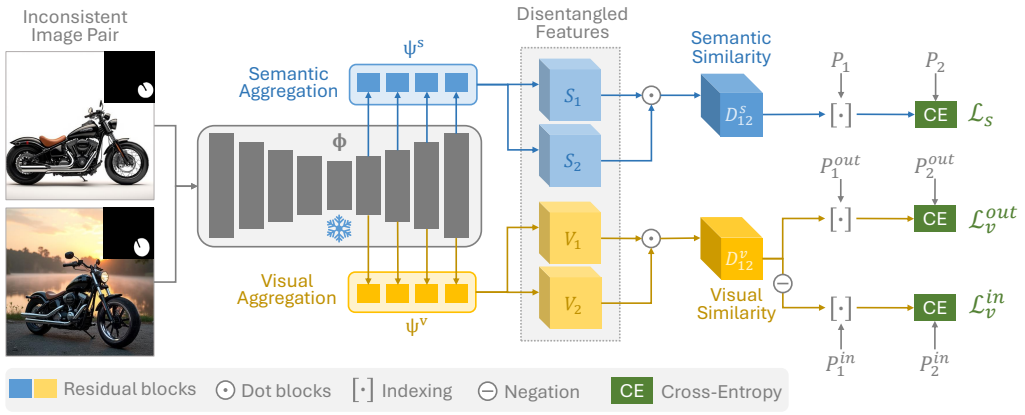

Figure 4: Overview of the proposed architecture for disentangling semantic and visual features from a frozen diffusion backbone Φ. Inconsistent regions between the input image pairs are indicated by binary masks (left). The semantic branch (blue) encourages the features of corresponding semantic points in both images, P_1 and P_2, to align. In the visual branch (yellow), we bring visually consistent points *outside* the inpainted regions, P_1^{out} and P_2^{out}, closer together, while pushing apart features at inconsistent points *inside* the inpainted regions, P_1^{in} and P_2^{in}.

combined using trainable scalar weights w^l:

$$S_i = \sum_l^L w_s^l \, \Psi_s^l(F_i^l) \;, \qquad V_i = \sum_l^L w_v^l \, \Psi_v^l(F_i^l) \;, \qquad S_i, V_i \in \mathrm{R}^{d \times q} \;, \tag{3}$$

where S_i, V_i are the semantic and visual features respectively, $i \in \{1, 2\}$, $d = w \times h$ denotes the flattened spatial dimensions, and q is the feature dimensionality. This architecture allows for flexible selection of features that capture either semantic content or visual appearance. An overview of the architecture is provided in Figure 4.

Next, we compute cyclic similarity matrices between the aggregated semantic and visual features of the two images using dot products:

$$\mathcal{D}_{12}^s = S_1 S_2^T \;, \qquad \mathcal{D}_{21}^s = S_2 S_1^T \;, \qquad \mathcal{D}_{12}^s, \mathcal{D}_{21}^s \in \mathrm{R}^{d \times d} \;, \tag{4}$$

$$\mathcal{D}_{12}^v = V_1 V_2^T \;, \qquad \mathcal{D}_{21}^v = V_2 V_1^T \;, \qquad \mathcal{D}_{12}^v, \mathcal{D}_{21}^v \in \mathrm{R}^{d \times d} \;, \tag{5}$$

Our objective is to encourage semantic features to be similar across all point correspondences, while visual features should be similar only outside the inconsistent regions R_i and dissimilar within them. To achieve this, we adopt a contrastive learning framework that brings similar features closer and pushes dissimilar ones apart. As a first step, we partition the set of pre-computed correspondences C_i from Section 3.1 into two subsets: points that fall inside and outside the inconsistent regions:

$$P_i^{\text{in}} = \left\{ (x_i^j, y_i^j) \;\middle|\; (x_i^j, y_i^j) \in C_i,\; R_i(x_i^j, y_i^j) = 1 \right\} \;, \tag{6}$$

$$P_i^{\text{out}} = \left\{ (x_i^j, y_i^j) \;\middle|\; (x_i^j, y_i^j) \in C_i,\; R_i(x_i^j, y_i^j) = 0 \right\} \;. \tag{7}$$

Then, inspired by the contrastive objective used in CLIP [29], we define a semantic correspondence loss as follows:

$$\mathcal{L}_s = \texttt{CrossEntropy}\big(\mathcal{D}_{12}^s(P_1),\; P_2\big) \;, \quad P_i = P_i^{\text{in}} \cup P_i^{\text{out}} \;. \tag{8}$$

This semantic loss encourages matched points, whether inside or outside the inconsistent region, to exhibit similar semantic feature representations.

To disentangle appearance-specific features, we define a visual loss that explicitly separates consistent from inconsistent regions. For inconsistent regions, we use a negative similarity objective:

$$\mathcal{L}_v^{\text{in}} = \texttt{CrossEntropy}\big(-\mathcal{D}_{12}^v(P_1^{\text{in}}),\; P_2^{\text{in}}\big) \;. \tag{9}$$

This term penalizes similarity in appearance features for points known to be visually inconsistent. For consistent regions, we retain the standard contrastive loss to encourage matching:

$$\mathcal{L}_v^{\text{out}} = \texttt{CrossEntropy}\big(\mathcal{D}_{12}^v(P_1^{\text{out}}),\ P_2^{\text{out}}\big)\ . \tag{10}$$

Additionally, visually corresponding points inside the inpainted regions of the original consistent pairs I_1 and I_2 can be incorporated into Equation (10) to provide additional supervision; however, we omit the formal definition of this part for simplicity. All loss terms are also computed in the reverse direction as well using $\mathcal{D}_{21}^s$ and $\mathcal{D}_{21}^v$, and each loss term is averaged over both directions. The final training objective combines semantic and visual losses as follows:

$$\mathcal{L} = \mathcal{L}_s + \alpha(\mathcal{L}_v^{\text{in}} + \mathcal{L}_v^{\text{out}}) \tag{11}$$

where α is a scaling factor used to prioritize the visual branch, since the semantic branch can already be extracted reliably even without any aggregation, as shown in [42, 39].

3.3 A Metric for Evaluating Subject-Driven Image Generation

Having disentangled visual and semantic features from the backbone of pre-trained diffusion models in the previous section, we now aim to leverage these features to estimate and localize visual consistency in subject-driven image generation. Given two test images I_1 and I_2 generated by a subject-driven generation method, our goal is to evaluate the visual consistency of the subject across the two images.

We begin by passing both images through our architecture to extract semantic and visual feature maps, denoted as F_i^s and F_i^v, respectively, following the aggregation described in Equation (3). As in the previous section, we compute pairwise similarities between features to obtain semantic and visual similarity matrices, denoted by $\mathcal{D}^s$ and $\mathcal{D}^v$, respectively. We then take the maximum similarity score for each point to obtain $\hat{\mathcal{D}}^s = \max(\mathcal{D}^s)$ and $\hat{\mathcal{D}}^v = \max(\mathcal{D}^v)$, representing the best per-point match in the semantic and visual domains.

Semantic correspondences are identified by selecting points whose semantic similarity exceeds a predefined threshold $\mathcal{T}_s$, *i.e.*, $\hat{\mathcal{D}}^s > \mathcal{T}_s$, resulting in a set of confident semantic matches indexed by $\mathcal{J}_s$. This filtering ensures that we only consider regions that are both visible and semantically coherent. We set $\mathcal{T}_s = 0.7$ in all experiments.

To assess visual consistency at these semantically matched locations, we use the same index set $\mathcal{J}_s$ to retrieve the corresponding visual similarity scores from $\hat{\mathcal{D}}^v$. Given a visual similarity threshold $\mathcal{T}_v$, we define the *Visual Semantic Match (VSM)* metric as:

$$\text{VSM}(\mathcal{T}_v) = \frac{1}{|\mathcal{J}_s|} \sum_{j \in \mathcal{J}_s} \delta\left[\hat{\mathcal{D}}_j^v > \mathcal{T}_v\right] \tag{12}$$

where $\delta[\cdot]$ is the indicator function, equal to 1 if the condition holds and 0 otherwise.

This metric captures the proportion of semantically aligned regions that are also visually consistent, providing a quantitative measure of consistency between the two generated images. Inconsistent regions can be identified by subtracting visually matched locations from the semantically matched ones, or by examining regions in $\hat{\mathcal{D}}^v$ with low visual similarity scores.

4 Experiments

In this section, we evaluate the effectiveness of our approach and compare it against commonly used feature-based metrics such as CLIP [29] and DINO [4], as well as VLM-based approaches (*e.g.* ChatGPT-4o). Then, we provide an ablation study to analyze the impact of key architectural design choices and provide further insights into the behavior of our model.

4.1 Implementation Details

Dataset: We use the data generation pipeline described in Section 3.1 to construct the dataset used to train our architecture. The pipeline takes consistent image pairs from the Subjects200k dataset [41] as input, although any subject-driven generation dataset can be used.

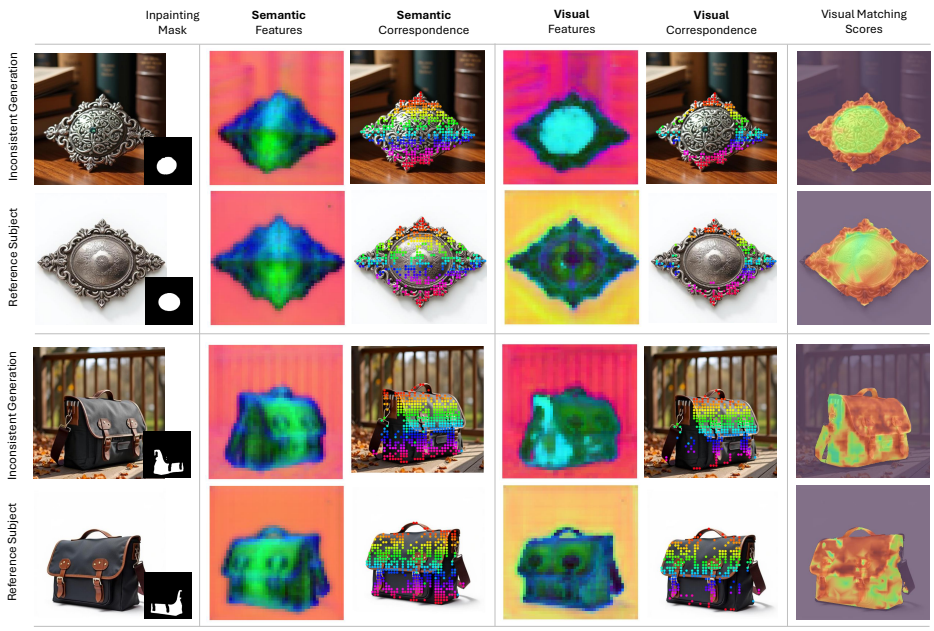

Figure 5: Qualitative examples of semantic and visual features, along with their correspondences, produced by our architecture. Regions that fall within the inpainting mask exhibit visually dissimilar features, enabling the detection of visual inconsistencies based on feature similarity. Dark Red is most consistent and Yellow is least consistent.

During generation, we apply several filtering steps to ensure high-quality training data. We discard samples with matching skewness below 1.3 (see Figure 3), and we enforce that the inconsistent region occupies between 5% and 60% of the subject area. Additionally, we filter out samples where the inpainted region has an LPIPS score below 0.15 to ensure sufficient visual inconsistency. The final dataset consists of 5,000 image pairs for training and 500 image pairs for validation.

Testset: We generate a testset of 100 samples using our dataset, and we manually revise it to ensure the reliability of evaluation.

Hyperparameters: We use the backbone of Stable Diffusion 2.1 [32] for all experiments following CleanDIFT [39]. We set the spatial resolution of the aggregated features in Equation (3) to $h = w = 48$, resulting in $d = 48 \times 48 = 2304$, following CleanDIFT [39]. The projected feature dimensionality is set to $q = 384$, as in [22]. We empirically found that setting $\alpha = 10$ in Equation (11) yields the best performance. We train the model for 30 epochs using the AdamW optimizer [21] with a learning rate of 1e-3 that is divided by 10 every 10 epochs. The training is done on 1 A100 GPU (40GB) and takes 12 hours. The source code for the dataset generation pipeline and the proposed architecture are publicly available. [1]

4.2 Evaluating the VSM Metric

To validate our proposed *Visual-Semantic Match (VSM)* metric, described in Section 3.3, we first perform a controlled evaluation where we define an oracle based on the inpainted regions for each image pair in the test set. This oracle is defined based on the ratio of the inconsistent region R_1 to the total object mask O_1 (see Figure 2), computed as:

$$\text{Oracle} = 1 - \left(\frac{\sum R_1}{\sum O_1} \right) \tag{13}$$

This oracle reflects the ground-truth degree of visual consistency between the image pair. An effective metric should produce estimates that strongly correlate with the oracle. We evaluate against commonly

[1] https://github.com/abdo-eldesokey/mind-the-glitch

	Controlled Inconsistency				*Subject-Driven Generation*			
	CLIP	DINO	VLM*	VSM (Ours)	CLIP	DINO	VLM*	VSM (Ours)
Pearson	-0.053	0.087	0.072	**0.448**	0.156	0.164	0.079	**0.405**
Spearman	-0.005	0.120	0.091	**0.582**	0.112	0.146	0.073	**0.369**

Table 1: Average correlation scores across methods. Our VSM achieves significantly higher correlation with the oracle than other metrics both on controlled and realistic settings. *(ChatGPT-4o)

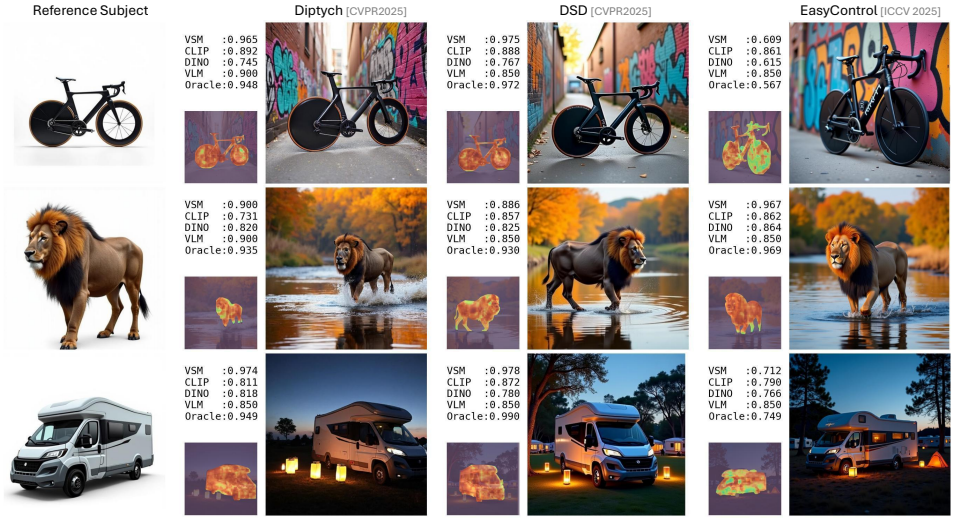

Figure 6: Qualitative examples of evaluating subject-driven image generation approaches using our proposed VSM metric and other existing approaches. Our VSM metric can accurately quantify and localize inconsistency and is more consistent with the oracle. Dark Red is most consistent and Yellow is least consistent.

used metrics in the literature for comparing image pairs: CLIP and DINO image-to-image similarity, as well as ChatGPT-4o as a representative Vision-Language Model (VLM). For VLM, we adopt the prompt used in [26], but modify it to produce a numerical score in the range of 0 to 100.

We report Pearson and Spearman correlations between each evaluated metric and the oracle in Table 1 (left). Our VSM metric shows significantly higher correlation compared to both feature-based similarity metrics and VLM judgments, indicating its greater reliability for measuring visual consistency. Qualitative results in Figure 5 illustrate semantic and visual features with their correspondences. While semantic features align with semantically similar regions regardless of appearance, visual features differ notably within inpainted areas and do not match across image pairs. This is evident in the score heatmaps shown in the rightmost column.

4.3 Evaluating Subject-Driven Image Generation

To evaluate the effectiveness of our VSM metric in a *realistic* subject-driven image generation setting, we evaluate three recent methods: Diptych [38], DSD-Diffusion [3], and EasyControl [54]. Using the reference image I_1 and target prompt p from our test set, we generate 100 images per method and evaluate them using VSM and other metrics as before. To compute the oracle, we manually annotate the generated images to mark visually inconsistent regions and calculate the oracle score as defined in Equation (13).

Table 1 shows that our VSM metric consistently outperforms all other metrics, demonstrating strong generalization to real-world subject-driven generation. We also present qualitative examples in Figure 6, illustrating that VSM aligns most closely with the oracle. This is further supported by the KDE plots in Figure 7, where VSM shows the highest agreement with the oracle, while VLM tends to score most image pairs between 75–95 regardless of visual discrepancies.

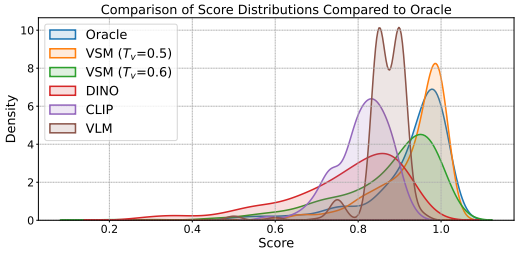

Figure 7: Score Distribution of different metrics compared to the Oracle.

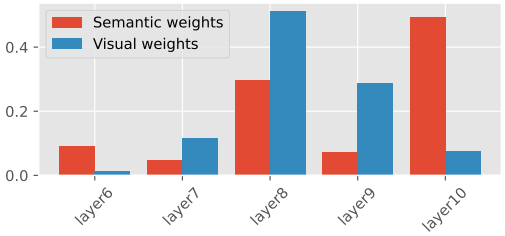

Figure 8: Aggregation weights.

	$\mathcal{T}_v = 0.5$	$\mathcal{T}_v = 0.7$	$\alpha = 1$	Skewness$>$ 1.0	Skewness$>$ 1.5	VSM (Ours)
Pearson	0.465	0.352	0.118	0.232	0.224	**0.448**
Spearman	0.454	0.496	0.104	0.250	0.225	**0.582**

Table 2: Ablation analysis of different hyperparameters.

4.4 Ablation Analysis

We visualize the learned aggregation weights for the semantic and visual branches in Figure 8. Visual features are primarily derived from decoder layers 8 and 9, while semantic features draw from layers 8 and 10, indicating that certain layers (e.g., layer 8) contribute to both representations. Notably, our learned semantic weights differ from those in [22], likely due to differences in supervision: their model is trained on sparse keypoints, whereas ours uses dense correspondences, favoring later layers with higher spatial resolution.

Table 2 presents an ablation study of key hyperparameters. For the VSM similarity threshold $\mathcal{T}_v$, both low (0.5) and high (0.7) values reduce correlation with the oracle, with $\mathcal{T}_v = 0.6$ yielding the best performance. Setting $\alpha = 1$ in the training loss significantly degrades performance by reducing emphasis on visual features in favor of semantic features, which are easier to learn and already present in diffusion backbones without additional training [42, 39]. In the dataset pipeline, high skewness thresholds overly restrict sample diversity, while low thresholds allow ambiguous matches—both leading to performance drops.

5 Limitations and Future Work

Our feature disentanglement is inherently partial, as visual features may still carry semantic information. Achieving full disentanglement, enabling visual matching across semantically different objects, remains a promising future direction. The quality of the learned features also depends on the reference dataset used in the automated pipeline. The Subjects200k dataset [41], while large-scale, was validated automatically and may include noisy pairs. Manual curation or improved filtering could enhance supervision quality. Additionally, detecting fine-grained inconsistencies is further limited by the spatial resolution of the diffusion features. We plan to address this via multi-scale aggregation and higher-resolution extraction. Lastly, our current framework targets structural and appearance-level inconsistencies. Extending it to handle broader variations, such as artistic style or color, may require further decomposing. We include failure cases in the supplementary to guide future work.

6 Conclusion

We proposed a framework for disentangling diffusion model features into semantic and visual components. Using an automated dataset pipeline with controlled visual inconsistencies, we trained a contrastive architecture to separate the two. This enables visual correspondence across images, complementing semantic matching, and supports fine-grained analysis. Leveraging this, we introduced a metric that quantifies and localizes inconsistencies in subject-driven generation. Our approach outperforms existing metrics, offering a more reliable and interpretable evaluation framework.

Acknowledgment

The research reported in this publication was supported by funding from King Abdullah University of Science and Technology (KAUST) - Center of Excellence for Generative AI, under award number 5940.

References

[1] O. Avrahami, A. Hertz, Y. Vinker, M. Arar, S. Fruchter, O. Fried, D. Cohen-Or, and D. Lischinski. The chosen one: Consistent characters in text-to-image diffusion models. In *ACM SIGGRAPH 2024 conference papers*, pages 1–12, 2024.

[2] Q. Bai, H. Ouyang, Y. Xu, Q. Wang, C. Yang, K. L. Cheng, Y. Shen, and Q. Chen. Edicho: Consistent image editing in the wild. In *arXiv preprint arXiv:2412.21079*, 2024.

[3] S. Cai, E. Chan, Y. Zhang, L. Guibas, J. Wu, and G. Wetzstein. Diffusion self-distillation for zero-shot customized image generation. In *CVPR*, 2025.

[4] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[5] J. Chen, J. YU, C. GE, L. Yao, E. Xie, Z. Wang, J. Kwok, P. Luo, H. Lu, and Z. Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024.

[6] J. Chung, S. Hyun, H. Kim, E. Koh, M. Lee, and J.-P. Heo. Fine-tuning visual autoregressive models for subject-driven generation. *arXiv preprint arXiv:2504.02612*, 2025.

[7] A. Cvejic, A. Eldesokey, and P. Wonka. Partedit: Fine-grained image editing using pre-trained diffusion models. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–11, 2025.

[8] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

[9] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

[10] L. Han, Y. Li, H. Zhang, P. Milanfar, D. Metaxas, and F. Yang. Svdiff: Compact parameter space for diffusion fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7323–7334, 2023.

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[12] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[13] J. Huang, X. Dong, W. Song, Z. Chong, Z. Tang, J. Zhou, Y. Cheng, L. Chen, H. Li, Y. Yan, et al. Consistentid: Portrait generation with multimodal fine-grained identity preserving. *arXiv preprint arXiv:2404.16771*, 2024.

[14] L. Huang, W. Wang, Z.-F. Wu, Y. Shi, H. Dou, C. Liang, Y. Feng, Y. Liu, and J. Zhou. In-context lora for diffusion transformers. *arXiv preprint arxiv:2410.23775*, 2024.

[15] J. Jin, Z. Yu, Y. Shen, Z. Fu, and J. Yang. Latexblend: Scaling multi-concept customized generation with latent textual blending. *arXiv preprint arXiv:2503.06956*, 2025.

[16] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[17] N. Kumari, B. Zhang, R. Zhang, E. Shechtman, and J.-Y. Zhu. Multi-concept customization of text-to-image diffusion, 2023.

[18] Z.-Y. Li, R. Du, J. Yan, L. Zhuo, Z. Li, P. Gao, Z. Ma, and M.-M. Cheng. Visualcloze: A universal image generation framework via visual in-context learning. *arXiv preprint arXiv:2504.07960*, 2025.

[19] C. Liu, V. Shah, A. Cui, and S. Lazebnik. Unziplora: Separating content and style from a single image. *arXiv preprint arXiv:2412.04465*, 2024.

[20] W. Liu, J. Mao, J. Hsu, T. Hermans, A. Garg, and J. Wu. Composable part-based manipulation. In *7th Annual Conference on Robot Learning*, 2023.

[21] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[22] G. Luo, L. Dunlap, D. H. Park, A. Holynski, and T. Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. *Advances in Neural Information Processing Systems*, 36:47500–47510, 2023.

[23] L. Meng, S. Lan, H. Li, J. M. Alvarez, Z. Wu, and Y.-G. Jiang. Segic: Unleashing the emergent correspondence for in-context segmentation. In *European Conference on Computer Vision*, pages 203–220. Springer, 2024.

[24] J. Min, J. Lee, J. Ponce, and M. Cho. Spair-71k: A large-scale benchmark for semantic correspondence. *arXiv preprint arXiv:1908.10543*, 2019.

[25] C. Mou, X. Wang, J. Song, Y. Shan, and J. Zhang. Dragondiffusion: Enabling drag-style manipulation on diffusion models. *arXiv preprint arXiv:2307.02421*, 2023.

[26] Y. Peng, Y. Cui, H. Tang, Z. Qi, R. Dong, J. Bai, C. Han, Z. Ge, X. Zhang, and S.-T. Xia. Dreambench++: A human-aligned benchmark for personalized image generation. In *The Thirteenth International Conference on Learning Representations*, 2025.

[27] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024.

[28] G. Qian, K.-C. Wang, O. Patashnik, N. Heravi, D. Ostashev, S. Tulyakov, D. Cohen-Or, and K. Aberman. Omni-id: Holistic identity representation designed for generative tasks. *arXiv preprint arXiv:2412.09694*, 2024.

[29] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

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

[31] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

[32] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[33] L. Rout, Y. Chen, N. Ruiz, A. Kumar, C. Caramanis, S. Shakkottai, and W.-S. Chu. RB-modulation: Training-free stylization using reference-based modulation. In *The Thirteenth International Conference on Learning Representations*, 2025.

[34] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023.

[35] M. Safaee, A. Mikaeili, O. Patashnik, D. Cohen-Or, and A. Mahdavi-Amiri. Clic: Concept learning in context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6924–6933, 2024.

[36] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

[37] V. Shah, N. Ruiz, F. Cole, E. Lu, S. Lazebnik, Y. Li, and V. Jampani. Ziplora: Any subject in any style by effectively merging loras. In *European Conference on Computer Vision*, pages 422–438. Springer, 2024.

[38] C. Shin, J. Choi, H. Kim, and S. Yoon. Large-scale text-to-image model with inpainting is a zero-shot subject-driven image generator. *arXiv preprint arXiv:2411.15466*, 2024.

[39] N. Stracke, S. A. Baumann, K. Bauer, F. Fundel, and B. Ommer. Cleandift: Diffusion features without noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[40] K. Sun, X. Liu, Y. Teng, and X. Liu. Personalized text-to-image generation with auto-regressive models. *arXiv preprint arXiv:2504.13162*, 2025.

[41] Z. Tan, S. Liu, X. Yang, Q. Xue, and X. Wang. Ominicontrol: Minimal and universal control for diffusion transformer. *arXiv preprint arXiv:2411.15098*, 3, 2024.

[42] L. Tang, M. Jia, Q. Wang, C. P. Phoo, and B. Hariharan. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems*, 36:1363–1389, 2023.

[43] Y. Tewel, O. Kaduri, R. Gal, Y. Kasten, L. Wolf, G. Chechik, and Y. Atzmon. Training-free consistent text-to-image generation. *ACM Transactions on Graphics (TOG)*, 43(4):1–18, 2024.

[44] Q. Wang, X. Bai, H. Wang, Z. Qin, A. Chen, H. Li, X. Tang, and Y. Hu. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024.

[45] Q. Wang, A. Eldesokey, M. Mendiratta, F. Zhan, A. Kortylewski, C. Theobalt, and P. Wonka. Zero-shot video semantic segmentation based on pre-trained diffusion modelss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[46] Y. Wei, Y. Zheng, Y. Zhang, M. Liu, Z. Ji, L. Zhang, and W. Zuo. Personalized image generation with deep generative models: A decade survey. *arXiv preprint arXiv:2502.13081*, 2025.

[47] Y. Wu, L. Zhu, L. Liu, W. Qiao, Z. Li, L. Yu, and B. Li. Proxy-tuning: Tailoring multimodal autoregressive models for subject-driven image generation. *arXiv preprint arXiv:2503.10125*, 2025.

[48] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

[49] Y. Zeng, V. M. Patel, H. Wang, X. Huang, T.-C. Wang, M.-Y. Liu, and Y. Balaji. Jedi: Joint-image diffusion models for finetuning-free personalized text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6786–6795, June 2024.

[50] Y. Zeng, V. M. Patel, H. Wang, X. Huang, T.-C. Wang, M.-Y. Liu, and Y. Balaji. Jedi: Joint-image diffusion models for finetuning-free personalized text-to-image generation. In *CVPR*, 2024.

[51] J. Zhang, C. Herrmann, J. Hur, E. Chen, V. Jampani, D. Sun, and M.-H. Yang. Telling left from right: Identifying geometry-aware semantic correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3076–3085, 2024.

[52] J. Zhang, C. Herrmann, J. Hur, L. Polania Cabrera, V. Jampani, D. Sun, and M.-H. Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *Advances in Neural Information Processing Systems*, 36:45533–45547, 2023.

[53] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[54] Y. Zhang, Y. Yuan, Y. Song, H. Wang, and J. Liu. Easycontrol: Adding efficient and flexible control for diffusion transformer. *arXiv preprint arXiv:2503.07027*, 2025.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Both the abstract and Section 1 (the introduction) summarize the scope and all contributions of the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5 provides a discussion on the limitations of the proposed approach and how it can potentially be improved in future work.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The results in the paper are either empirical or qualitative.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide exhaustive details about the proposed approach in section 3 (Method), and we provide all used hyperparameters in section 4.1 (implementation details).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Yes the source code and the data need to reproduce the results in the paper will be made publicly available.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All these details are provided in section 4.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We provide 1-sigma experiments due to the cost of training and since the gap to the competing methods is significant.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Details are provided in section 4.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Code of Ethics was reviewed an followed.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The proposed approach can benefit the research community in subject-drive image generation.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The proposed method is analytical and does not impose any risk.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All used models, data, and code were referenced and credited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Exhaustive details on the new data were provided, and the data will be made publicly available.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No human subject were used in the paper.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No human subject were used in the paper.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method does not include any use of LLMs

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.