CONSISTENT TEXT-TO-IMAGE GENERATION VIA SCENE DE-CONTEXTUALIZATION

Song Tang[1,2], Peihao Gong[1], Kunyu Li[3], Kai Guo[1], Boyu Wang[4,5], Mao Ye[6,*], Jianwei Zhang[2] & Xiatian Zhu[7,*]

[1]University of Shanghai for Science and Technology, [2]Universität Hamburg, [3]Fudan University, [4]Western University, [5]Vector Institute, [6]University of Electronic Science and Technology of China, [7]University of Surrey

tangs@usst.edu.cn; cvlab.uestc@gmail.com; xiatian.zhu@surrey.ac.uk

ABSTRACT

Consistent text-to-image (T2I) generation seeks to produce identity-preserving images of the same subject across diverse scenes, yet it often fails due to a phenomenon called identity (ID) shift. Previous methods have tackled this issue, but typically rely on the unrealistic assumption of knowing all target scenes in advance. This paper reveals that a key source of ID shift is the native correlation between subject and scene context, called *scene contextualization*, which arises naturally as T2I models fit the training distribution of vast natural images. We formally prove the near-universality of this scene-ID correlation and derive theoretical bounds on its strength. On this basis, we propose a novel, efficient, training-free prompt embedding editing approach, called **Scene De-Contextualization** (**SDeC**), that imposes an inversion process of T2I's built-in scene contextualization. Specifically, it identifies and suppresses the latent scene-ID correlation within the ID prompt's embedding by quantifying the SVD directional stability to adaptively re-weight the corresponding eigenvalues. Critically, SDeC allows for per-scene use (one scene per prompt) without requiring prior access to all target scenes. This makes it a highly flexible and general solution well-suited to real-world applications where such prior knowledge is often unavailable or varies over time. Experiments demonstrate that SDeC significantly enhances identity preservation while maintaining scene diversity.

1 INTRODUCTION

Text-to-image (T2I) generation (Shi et al., 2024; Saharia et al., 2022; Ramesh et al., 2021) aims to synthesize visually compelling and semantically faithful images from prompts. From artistic design to personalized media production, T2I models such as GAN (Tao et al., 2022) and Stable Diffusion (Rombach et al., 2022b) have demonstrated remarkable capability in producing novel scenes that align closely with user intent. However, in narrative-driven visual tasks involving recurring characters or entities, such as animation/video (Lei et al., 2025), personalized storytelling (Avrahami et al., 2024), cinematic pre-visualization (Tao et al., 2024), and digital avatars (Wang et al., 2023), mere alignment with scene descriptions is insufficient: The subject's IDentity (ID) must remain consistent across generated images (no ID shift). Against this backdrop, consistent T2I generation has recently emerged as a focal point of growing interest (Höllein et al., 2024; Wang et al., 2024).

Methodologically, existing approaches reduce ID shift in line with the paradigm of transfer learning (Tang et al., 2024): Extracting invariance from given heterogeneous data. This requires prior knowledge of the complete target scenes, enabling the generative model to map different scene prompts into corresponding features (Zhou et al., 2024; Liu et al., 2025) or image pseudo labels (Avrahami et al., 2024; Akdemir & Yanardag, 2024) for constructing such a diversified dataset. In practice, however, target scenes are not always available, for instance, in online cases, rendering this assumption unrealistic and limiting the practical applicability of these methods. Critically, the underlying cause of ID shift with T2I models remains largely unclear.

*Corresponding author

Figure 1: Illustration of scene contextualization with SDXL. **Left**: The attire of the subject varies with the site. **Right**: The subject's clothing changes with the season.

In this work, we consider that the scene plays a context role that would influence the characterization of identity, called *scene contextualization* (Fig. 1), causing ID shift. This arises because a T2I model is predominantly trained on natural images with a specific data distribution (e.g., caws often appear on the green fields but not in the sea). Consequently, the generated images are constrained to satisfy such internalized priors.

To probe the connection between scene contextualization and ID shift, we formulate a theoretical framework, showing that the contextualization, inherently induced by the attention mechanism, is not only the primary source of ID shift but also inevitable for pre-trained T2I models. Moreover, we derive theoretical bounds on contextualization strength. Building on these insights, we propose a **Scene De-Contextualization** (**SDeC**) approach, which effectively realizes the inverse process of scene contextualization. Specifically, SDeC quantifies the directional stability of the subspace spanned by SVD eigenvectors via a forward-and-backward eigenvalue optimization. After that, the latent scene-ID correlation within the ID prompt's embedding is identified through eigenvalue variations and then suppressed with adaptive eigenvalue weighting. The contextualization-mitigated ID prompt embedding is then reconstructed from the reweighted eigenvalues for subsequent generation. It can work in a one-prompt-per-scene setting, removing reliance on full target scenes.

Our **contributions** are: (1) We propose a *scene contextualization* perspective for ID shift with T2I models; (2) We theoretically characterize and quantify this contextualization, leading to a novel SDeC approach for mitigating ID shift per scene without the need for complete target scenes in advance. (3) Extensive experiments show that SDeC can enhance identity preservation, maintain scene diversity, and offer plug-and-play flexibility at per-scene level and across diverse tasks, e.g., integrating pose map and personalized photo, and generative backbones such as PlayGround-v2.5, RealVisXL-V4.0 and Juggernaut-X-V10.

2 Related Work

The study of consistent T2I generation falls into two phases. The early phase focuses on *personalized T2I generation* (Zhang et al., 2024), where one or a couple of reference images are given to define the identity of interest. These methods tackle ID shift by injecting ID semantics of reference images into a pre-trained T2I model, essentially adopting two strategies. The first leverages cross-attention to progressively inject the information, subject to convergence toward the reference image(s) (Ye et al., 2023). The second is a textual token creation strategy: Transforming the reference image(s) into dedicated tokens to differentiate ID and scene. For example, DreamBooth (Ruiz et al., 2023) and variants (Sun et al., 2025; Hsin-Ying et al., 2025) introduce a unique identifier token to represent the ID and inject it by fine-tuning the generative model. PhotoMaker (Li et al., 2024) fuses the reference image(s) and text embeddings to enhance ID tokens. Textual Inversion (Gal et al., 2022) and its variants (Zeng et al., 2024; Wu et al., 2024) directly create a concept token via textual inversion.

The recent phase moves to the more flexible *reference image-free* setting addressed under two approaches. The first involves *pseudo-label-based self-learning*: Generating candidate images with the entire prompt set, then filtering them with obvious ID shift using some metric (e.g., mutual information in ORACLE (Akdemir & Yanardag, 2024) or clustering in Chosen-One (Avrahami et al., 2024) or self-diffusion in DiffDis (Cai et al., 2025)), further retraining the generative models. Clearly, such methods are expensive due to model retraining. More recent attempts thus emerge in a *training-free* fashion. Storytelling methods (Rahman et al., 2023; Zhou et al., 2024; Tewel et al., 2024) leverage the self-attention mechanism over generated images as an adapter, whilst the state of

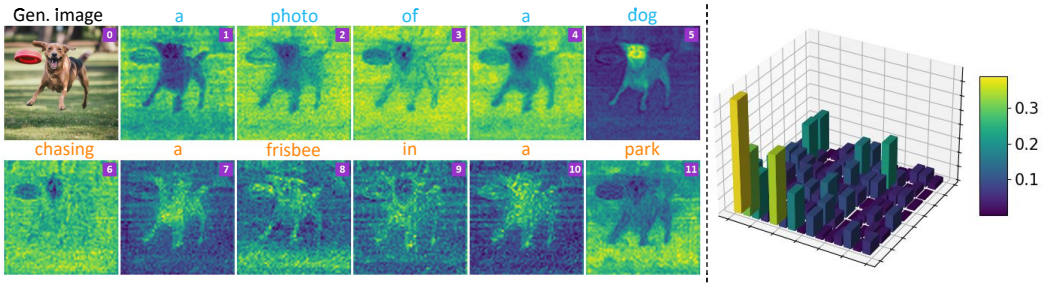

Figure 2: Example of scene contextualization with ID prompt "a photo of a dog" and scene prompt "chasing a frisbee in a park". **Left**: Correlation between tokens and the attention similarity matrix reveals that scene tokens influence ID generation (visualization method follows (Hertz et al., 2023)). **Right**: Similarity between SVD eigenvectors of ID and scene embedding shows they share an overlapping subspace.

the art, 1Prompt1Story (Liu et al., 2025), introduces a prompt re-structuring idea to highlight ID and balance scene's contribution in prompting, with extra need to couple a specific adapter. Commonly, all the methods above assumes the availability of all the scenes which often is not valid in real applications. Importantly, these studies fail to provide an insight on the underlying cause of ID shift with off-the-shelf T2I models. We address these gaps by suggesting a scene contextualization perspective along with theoretical formulation and a flexible *prompt embedding editing* solution.

3 Scene Contextualization

Problem Given an ID text prompt $\mathcal{P}_{\text{id}}$ and K distinct scene text prompts $\{\mathcal{P}_{\text{sc}}^k\}_{k=1}^K$, we form a set of scenario prompts $\mathcal{P} = \{\mathcal{P}^k\}_{k=1}^K$ with $\mathcal{P}^k = \mathcal{P}_{\text{id}} \oplus \mathcal{P}_{\text{sc}}^k$. Feeding each prompt into a T2I model $\mathcal{G}$ produces K generated images $\{I^k\}_{k=1}^K$, where $I^k = \mathcal{G}(\mathcal{P}^k)$. Consistent T2I generation requires that $\{I^k\}_{k=1}^K$ simultaneously (1) preserve the same identity features specified by $\mathcal{P}_{\text{id}}$, and (2) faithfully reflect the scene semantics described in each corresponding $\mathcal{P}^k$.

Interaction between ID and scene The attention mechanism is the key structure in Transformer based T2I models. Denote $Z \in \mathbb{R}^{N\times d}$ as the set of N input token features of dimension d, the self-attention projections are: $K = ZW_K,\ V = ZW_V,\ Q = ZW_Q$. For any query $q \in Q$, the attention output is computed as:

$$O(q) = \alpha^\top V \text{ with } \alpha = \text{softmax}\Big(\big(qK^\top\big)/\sqrt{d_k}\Big). \tag{1}$$

where d_k denotes the feature dimension with W_K. In the prompt embedding space, this attention operation offers an opportunity for scene tokens to inject its context information into ID tokens, potentially leading to ID shift. We name this *scene contextualization*.

For intuitive understanding, we visualize the correlation between scene tokens and ID tokens by their cross-attention similarity matrix, where each row corresponds to α values. As shown in images # 7∼10 of Fig. 2-Left, the bright regions (green/yellow) within the subject (dog) clearly indicate that scene tokens affect the generation of this dog.

3.1 Theorem behind Scene Contextualization

Attention operations are typically executed in a chain-like manner for T2I models. For simplicity, we analyze the first attention block in generative models, without loss of generality.

Theorem 1 *Let $\mathcal{H}_{\text{id}}, \mathcal{H}_{\text{sc}} \subset \mathbb{R}^d$ be the identity/scene semantic subspaces with ideal semantic separation: $\mathcal{H}_{\text{id}} \cap \mathcal{H}_{\text{sc}} = \varnothing$; $\mathcal{Z}_{\text{id}} \subseteq \mathcal{H}_{\text{id}}$ and $\mathcal{Z}_{\text{sc}}^k \subseteq \mathcal{H}_{\text{sc}}$ be the prompt-embedding matrix of $\mathcal{P}_{\text{id}}$ and $\mathcal{P}_{\text{sc}}^k$; the prompt-embedding matrix be partitioned by semantics $\mathcal{Z} = \big[\mathcal{Z}_{\text{id}}; \mathcal{Z}_{\text{sc}}^k\big] \in \mathbb{R}^{n\times d}$. For any query q_{id} associated with the identity, its attention output can be specified to*

$$O(q_{\text{id}}) = \alpha^\top(\mathcal{Z}W_V) = \alpha_{\text{id}}^\top(\mathcal{Z}_{\text{id}}W_V) \ + \ \alpha_{\text{sc}}^\top\big(\mathcal{Z}_{\text{sc}}^k W_V\big), \tag{2}$$

where the attention weights by token index $\alpha = [\alpha_{\text{id}}, \alpha_{\text{sc}}]$ conforming to the row split of $\mathcal{Z}$. Let Π_{id} be the orthogonal projector onto $\mathcal{H}_{\text{id}}$. The projection of $O(q_{\text{id}})$ onto the identity subspace $\mathcal{H}_{\text{id}}$ is

$$\Pi_{\text{id}}\big[O(q_{\text{id}})\big] = \underbrace{\Pi_{\text{id}}\big[\alpha_{\text{id}}^{\top}(\mathcal{Z}_{\text{id}}W_V)\big]}_{\textit{id term: } T_{\text{id}}} + \underbrace{\Pi_{\text{id}}\big[\alpha_{\text{sc}}^{\top}(\mathcal{Z}_{\text{sc}}^{k}W_V)\big]}_{\textit{scene term: } T_{\text{sc}}}. \tag{3}$$

Assume $\Pi_{\text{id}} \circ W_V\big|_{\mathcal{H}_{\text{sc}}}$ denotes an operation where an input from subspace $\mathcal{H}_{\text{sc}}$ is first transformed by W_V and then projected onto subspace $\mathcal{H}_{\text{id}}$. If the conditions, (A) $\alpha_{\text{sc}} \neq \mathbf{0}$ and (B) $\Pi_{\text{id}} \circ W_V\big|_{\mathcal{H}_{\text{sc}}} \neq 0$, hold, then the scene term in Eq. (3) is nonzero: $T_{\text{sc}} \neq 0$.

Theorem 1 suggests that even if $\mathcal{H}_{\text{id}} \cap \mathcal{H}_{\text{sc}} = \varnothing$, the attention could still cause scene contextualization. The two conditions are almost always satisfied with T2I models. Condition (A) is often met for two factors. (i) Keys from scene tokens and queries from ID tokens are rarely strictly orthogonal or sufficiently separated; so the softmax attention weights are unlikely to be exactly zero, leaving scene tokens with non-negligible attention mass. (ii) No enforcement on separating between scene and identity tokens during training, lead scene-to-ID attention positive. For condition (B), it is equivalent to *non-block-diagonality* of W_V w.r.t. the decomposition $\mathcal{H}_{\text{id}} \oplus \mathcal{H}_{\text{scene}}$: No scene vector is mapped with a nonzero component in $\mathcal{H}_{\text{id}}$ — again this condition is not enforced in training.

Theorem 1 assumes an idealized condition of $\mathcal{H}_{\text{id}} \cap \mathcal{H}_{\text{sc}} = \varnothing$. In practice, ID and scene subspaces often exhibit partial overlap. To assess this, we apply SVD to $\mathcal{Z}_{\text{id}}$ and $\mathcal{Z}_{\text{sc}}^{k}$ and compute the similarity between their corresponding eigenvectors. As seen in Fig. 2-Right, the high regions in the similarity matrix reveal nontrivial correlations between $\mathcal{Z}_{\text{id}}$ and $\mathcal{Z}_{\text{sc}}^{k}$ across certain dimensions. Relaxing the disjoint-subspace assumption to this correlated case, we show that scene contextualization still persists below.

Corollary 1 *Assume that $\mathcal{H}_{\text{id}}$ and $\mathcal{H}_{\text{sc}}$ have a nontrivial intersection: $\mathcal{H}_{\cap} := \mathcal{H}_{\text{id}} \cap \mathcal{H}_{\text{sc}}$ with $k_{\cap} := \dim(\mathcal{H}_{\cap}) > 0$ where $\dim(\cdot)$ means space dimensions. If $\alpha_{\text{sc}} \neq 0$, then for a generic linear mapping W_V, which excludes measure-zero degenerate cases, $T_{\text{sc}} \neq 0$ hold.*

The degenerate case above refers to the rare weight setting where W_V maps the scene subspace $\mathcal{H}_{\text{sc}}$ exactly onto a subspace orthogonal to $\mathcal{H}_{\text{id}}$. Such cases form a measure-zero set in the parameter space. With random initialization and continuous optimization, the probability of encountering them is negligible. For a typical W_V, this blocking effect does not arise. Moreover, unlike the idealized assumptions in Theorem 1, in practice $k_{\cap} > 0$, meaning scene tokens always receive some attention. This makes contextualization both easier and stronger, even when W_V only weakly couples $\mathcal{H}_{\text{sc}}$ and $\mathcal{H}_{\text{id}}$. Please see the proof in `Appendix-B`.

Combining Theorem 1 and Corollary 1 yields an insight that, irrespective of whether $\mathcal{H}_{\text{id}}$ and $\mathcal{H}_{\text{sc}}$ overlap, the scene-to-ID projection is generically nonzero, i.e., scene contextualization occurs firmly.

3.2 Bounding the Strength of Contextualization

In this section, we derive a bound on contextualization strength, uncovering the key variables that govern its intensity. Theoretically, we characterize the scene contextualization to T_{sc} (see Theorem 1), thereby its spectral norm $\|T_{\text{sc}}\|_2$ can be a measurement of strength. $\|T_{\text{sc}}\|_2$ can be bounded as below (refer to the proof in `Appendix-C`):

Theorem 2 *Let $P_{\cap}$ be the orthogonal projector onto $\mathcal{H}_{\cap}$; Π_{sc} be the orthogonal projector onto $\mathcal{H}_{\text{sc}}$; $P_{\cap}^{\perp} := \Pi_{\text{sc}} - P_{\cap}$ be the projector onto the orthogonal complement within $\mathcal{H}_{\text{sc}}$. Define $R_{\cap} := \mathcal{Z}_{\text{sc}}^{k}P_{\cap}$, $R_{\perp} := \mathcal{Z}_{\text{sc}}^{k}P_{\cap}^{\perp}$, $T_{\cap} := \Pi_{\text{id}}W_V P_{\cap}$, $T_{\perp} := P_{\cap}^{\perp}W_V\Pi_{\text{id}}$, and $\epsilon := \|\alpha_{\text{sc}}^{\top}\|_2$. The contextualization strength $\|T_{\text{sc}}\|_2$ is bounded as*

$$0 \leq \|T_{\text{sc}}\|_2 \;\leq\; \epsilon \cdot \|R_{\cap}\|_2 \cdot \|T_{\cap}\|_F \;+\; \epsilon \cdot \|R_{\perp}\|_2 \cdot \|T_{\perp}\|_F. \tag{4}$$

In Theorem 2, Π_{id} is formally a subspace operator. In practice, it is often spanned or approximately defined by the ID embedding $\mathcal{Z}_{\text{id}}$ itself. Editing $\mathcal{Z}_{\text{id}}$ is thus equivalent to adjusting the orientation of the subspace, thereby modifying Π_{id}. Following this, we derive a corollary to further specify the upper bound from the ID perspective (see the proof in `Appendix-D`).

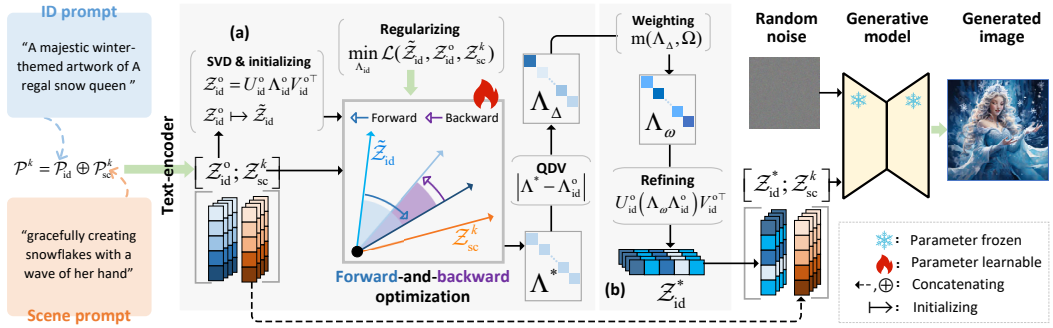

Figure 3: Overview of SDeC. A text prompt $\mathcal{P}^k = \mathcal{P}_{\text{id}} \oplus \mathcal{P}^k_{\text{sc}}$ is encoded into prompt embeddings $[\mathcal{Z}^{\text{o}}_{\text{id}}; \mathcal{Z}^k_{\text{sc}}]$ where $^{\text{o}}$ means "original". SDeC mitigates scene contextualization by **(a)** identifying and **(b)** suppressing latent scene-ID correlation in $\mathcal{Z}^{\text{o}}_{\text{id}}$ (QDV: Quantifying Directional influence Variations). The refined ID embedding $\mathcal{Z}^*_{\text{id}}$ is then concatenated with $\mathcal{Z}^k_{\text{sc}}$ for subsequent generation.

Corollary 2 *$\mathcal{H}_{\text{id}}$ is the subspace spanned by the ID embedding $\mathcal{Z}_{\text{id}}$; U is an orthonormal basis of $\mathcal{H}_{\text{id}}$, i.e., $U = \text{orth}(\mathcal{Z}_{\text{id}})$, where $\text{orth}(\cdot)$ specifies an orthogonalization operation. Writing the projector onto the ID subspace as $\Pi_{\text{id}} = UU^\top$, we have*

$$0 \leq \|T_{\text{sc}}\|_2 \;\leq\; \epsilon \cdot \|R_\cap\|_2 \cdot \underbrace{\|U^\top W_V P_\cap\|_F}_{\sigma_\cap} + \epsilon \cdot \|R_\perp\|_2 \cdot \underbrace{\|W_V^\top P_\cap^\perp U\|_F}_{\sigma_\perp}, \tag{5}$$

In Corollary 2, the term $\sigma_\cap$ measures the energy shared between ID (U) and scene ($\mathcal{Z}^k_{\text{sc}}$) subspaces, while $\sigma_\perp$ denotes the energy of ID projected into the scene-specific subspace. We consider that the majority of scene-ID interaction takes place via $\sigma_\cap$, whilst $\sigma_\perp$ allows for holistic coherence. Thus, we only need to minimize $\sigma_\cap$.

4 Scene De-Contextualization

Overview Grounding on the insights from Corollary 2, we propose the SDeC framework to achieve de-contextualization, including (1) estimating $P_\cap$ and (2) driving $\sigma_\cap \to 0$. SDeC's idea is to quantify the extent to which each direction is influenced by contextualization and then selectively reinforce those that are less affected. Thus, the original ID embedding is edited. Here, the directions that are strongly affected are referred to as the *latent scene–ID correlation subspace*.

Concretely, we approximate $P_\cap$ by identifying the latent scene-ID correlation subspace in $\mathcal{Z}_{\text{id}}$ (Fig. 3(a)), exploiting a learning process. Subsequently, we suppress the correlation subspace (Fig. 3(b)) to reach $\sigma_\cap \to 0$. For clarity, we hereafter denote $\mathcal{Z}_{\text{id}}$ to $\mathcal{Z}^{\text{o}}_{\text{id}}$ and in-training ID embedding is denoted as $\tilde{\mathcal{Z}}_{\text{id}}$.

Identifying latent scene-ID correlation subspace in $\mathcal{Z}^{\text{o}}_{\text{id}}$ We first achieve the directional correlation measurement via an "forward-and-backward" optimization (Fig. 3 (a)): First pulling $\tilde{\mathcal{Z}}_{\text{id}}$ closer to the scene $\mathcal{Z}^k_{\text{sc}}$ (forward), followed by restoring back to its original position $\mathcal{Z}^{\text{o}}_{\text{id}}$ (backward). We start with solving $\mathcal{Z}^{\text{o}}_{\text{id}} = U^{\text{o}}_{\text{id}} \Lambda^{\text{o}}_{\text{id}} V^{\text{o}\top}_{\text{id}}$ by Singular Value Decomposition (SVD) (Stewart, 1993). Let $\tilde{\mathcal{Z}}_{\text{id}} = U^{\text{o}}_{\text{id}} \Lambda_{\text{id}} V^{\text{o}\top}_{\text{id}}$ with Λ_{id} initiated as $\Lambda^{\text{o}}_{\text{id}}$; We formulate a two-phase optimization problem:

$$\begin{aligned} \Lambda^* &= \min_{\Lambda_{\text{id}}} \mathcal{L}(\tilde{\mathcal{Z}}_{\text{id}}, \mathcal{Z}^{\text{o}}_{\text{id}}, \mathcal{Z}^k_{\text{sc}}) = \| U^{\text{o}}_{\text{id}} \Lambda_{\text{id}} V^{\text{o}\top}_{\text{id}} - \mathcal{Z}^k_{\text{sc}} \|_2 + \beta \| U^{\text{o}}_{\text{id}} \Lambda_{\text{id}} V^{\text{o}\top}_{\text{id}} - \mathcal{Z}^{\text{o}}_{\text{id}} \|_2, \\ \beta &= 0 \text{ when } iter \leq M, \text{ and } \beta \neq 0 \text{ when } iter > M, \end{aligned} \tag{6}$$

where $iter$ denotes the training iteration index; M denotes the length of the first training phase, correspondingly, iterations $0 \sim M$ correspond to the forward process, while the remaining iterations constitute the backward process.

In the two-phase optimization defined in Eq. (6), the forward phase identifies the directions in $\mathcal{Z}_{\text{id}}$ that align with $\mathcal{Z}^k_{\text{sc}}$, capturing their shared representations. However, some of these directions may

also be essential for representing $\mathcal{Z}_{\text{id}}$ itself. To mitigate potential semantic degradation in $\mathcal{Z}_{\text{id}}$, we introduce a backward phase that progressively removes these ID-associated components. During this phase, to maximally exclude ID-related information, we set $\beta \gg 0$.

After evaluating the directional correlations, we further quantify their strength. In the SVD view, the directions, whose eigenvalues remain nearly unchanged (resistant to both pull and restoration), correspond to robust directions against contextualization. In contrast, those with large variations are the latent scene-ID correlation subspace, which theoretically corresponds to the $P_{\cap}$. In this context, we use *absolute spectral excursion* to quantify the stability in different directions. Assume Λ^* and $\Lambda^{\text{o}}_{\text{id}}$ share the same spectral structure (diagonal with ordered singular values): $\Lambda^{(\cdot)} = \text{diag}(\lambda_1^{(\cdot)}, \ldots, \lambda_r^{(\cdot)})$ with $\lambda_1^{(\cdot)} \geq \cdots \geq \lambda_r^{(\cdot)} \geq 0$ and $r = \text{rank}(\mathcal{Z}^{\text{o}}_{\text{id}})$. The directional correlation of $\mathcal{Z}^{\text{o}}_{\text{id}}$ can be formulated by Eq. (7) where v_i stands for the correlation strength of the i-th direction.

$$\Lambda_\Delta = |\Lambda^* - \Lambda^{\text{o}}_{\text{id}}| = \text{diag}(v_1, \ldots, v_i, \ldots, v_r) \text{ with } v_i = \ |\lambda_i^* - \lambda_i^{\text{o}}|, \tag{7}$$

De-contextualization by suppressing latent scene-ID correlation subspace in $\mathcal{Z}^{\text{o}}_{\text{id}}$ We achieve this through robust subspace filtering, which involves eigenvalue modulation, denoted by $\text{m}(\cdot, \cdot)$, followed by reconstructing the ID prompt embedding using the modulated eigenvalues. This method features (1) relative enhancement on the robust subspace, and (2) soft direction selection without threshold. Let $\mathcal{Z}^*_{\text{id}}$ be the refined ID prompt embedding. The filtering is expressed as

$$\mathcal{Z}^*_{\text{id}} = U^{\text{o}}_{\text{id}} \left(\Lambda_\omega \Lambda^{\text{o}}_{\text{id}}\right) V^{\text{o}\top}_{\text{id}} \text{ with } \ \Lambda_\omega = \text{m}(\Lambda_\Delta, \Omega) = 1 + \Omega \left(\frac{\Lambda_\Delta - \Delta_{\min}}{\Delta_{\max} - \Delta_{\min}} \right), \tag{8}$$

where $\Delta_{\max}$ and $\Delta_{\min}$ are the maximum and minimum entries of Λ_Δ, respectively, and the hyper-parameter $\Omega \geq 1$ controls the weighting strength. In Eq. (8), the weighting values $\Lambda_\omega \in [1, 1 + \Omega]$ are derived from normalized directional influence. Setting $\Omega \geq 1$ ensures that the robust subspace is emphasized while avoiding semantic loss in the shared subspace.

After filtering, we edit the original prompt embedding $[\mathcal{Z}^{\text{o}}_{\text{id}}; \mathcal{Z}^k_{\text{sc}}]$ to $\mathcal{Z}^{k*} = [\mathcal{Z}^*_{\text{id}}; \mathcal{Z}^k_{\text{sc}}]$. We feed $\mathcal{Z}^{k*}$ into the T2I model to produce the final image.

5 Experiments

Benchmark Our evaluation uses the ConsiStory+ (Liu et al., 2025), extending the ConsiStory dataset (Tewel et al., 2024) to 192 prompt sets, generating 1292 images with a wider range of subjects, descriptions, and styles.

Evaluation metrics To assess ID consistency, we employ two metrics: (1) CLIP-I (Hessel et al., 2021), computed as the cosine distance between image embeddings, and (2) DreamSim-F (Fu et al., 2023), better aligned with human judgment of visual similarity closely. As DreamSim, we remove the background using CarveKit (Selin, 2023) and fill random noise, ensuring that similarity measurement focus solely on ID content.

To evaluate the entire scenario (ID + scene), we adopt CLIP-T, the average CLIPScore (Hessel et al., 2021) between each generated image and its corresponding prompt. Note, this metric cannot measure the undesired scene mixture effect (see Fig. 6-Middle in `Appendix-F`). To address this, we introduce a new metric, `DreamSim-B`, specifically designed to quantify inter-scene interference, in the spirit of DreamSim-F, based on foreground masking instead.

Competitors We consider two types of competitors. The first is baseline T2I models, including SD1.5 (Rombach et al., 2022a) and SDXL (Podell et al., 2023). The second includes six state-of-the-art consistent T2I methods: BLIP-Diffusion (Li et al., 2023), Textual Inversion (Gal et al., 2022), PhotoMaker (Li et al., 2024), ConsiStory (Tewel et al., 2024), StoryDiffusion (Zhou et al., 2024), and 1Prompt1Story (1P1S) (Liu et al., 2025). The first three are training based, vs training free for the rest and SDeC. For more extensive test, we introduce (1) 1P1S w/o IPCA, with the attention module IPCA removed to focus on its prompt embedding editing; and (2) SDeC+ConsiStory, to test the complementary effect of our method with existing adapter based method ConsiStory. We exclude IP-Adapter (Ye et al., 2023), as its generated characters are homogeneous in pose and layout, with little ability to follow the scene description instruction. The same setting is applied for fair comparison of all compared methods. (see `Appendix-E.2`).

Table 1: Quantitative comparison. The best and second-best results are marked in **bold** and underlined, respectively. PE: Prompt embedding Editing; POT: Prompt Operation Time (per image); GenT: Generation Time (per image); BL: BaseLine.

	Method	Base model	PE	ID metrics		Scenario metrics		Infer. time↓ (s)		VRAM↓ (GB)	Steps
				DreamSim-F↓	CLIP-I↑	DreamSim-B↑	CLIP-T↑	POT	GenT		
BL	–	SD1.5	✗	0.4118	0.8071	**0.4673**	0.8324	0	2	3.18	50
	–	SDXL	✗	0.2778	0.8558	0.3861	0.8865	0	9	10.72	50
Training	BLIP-Diffusion	SD1.5	✗	0.2851	0.8522	0.3957	0.8187	0	1	3.54	26
	Textual Inversion	SDXL	✗	0.3066	0.8437	0.3919	0.8557	0	10	14.00	40
	PhotoMaker	SDXL	✗	0.2808	0.8545	0.3957	0.8812	0	9	9.74	50
Training-Free	ConsiStory	SDXL	✗	0.2729	0.8604	0.4207	0.8942	0	27	15.58	50
	StoryDiffusion	SDXL	✗	0.3197	0.8502	0.4214	0.8578	0	24	38.44	50
	1P1S	SDXL	✗	**0.2238**	**0.8798**	0.2955	0.8883	0.10	22	13.10	50
	1P1S w/o IPCA	SDXL	✗	0.2682	0.8617	0.3338	0.8637	0.10	19	10.74	50
	SDeC	SDXL	✓	0.2589	0.8655	0.3675	0.8946	0.61	15	12.14	50
	SDeC+ConsiStory	SDXL	✓	0.2542	0.8744	0.4155	**0.8967**	0.67	27	15.56	50

5.1 Results Analysis

Quantitative analysis We draw these observations from Tab. 1: (1) For training-free methods, 1P1S delivers the best result in the ID metrics, whilst suffering serious inter-scene interference (worst DreamSim-B score, also see Fig. 6 in Appendix-F), largely unacceptable for consistent T2I generation. In contrast, our SDeC strikes the best balance between ID consistency and scene diversity. (2) Without the attention IPCA, 1P1S is outperformed by SDeC across all metrics. That indicates that our prompt embedding editing is superior, even without the need for all target scenes in advance. (3) ConsiStory lags behind SDeC in ID metrics, but excels in scene background. (4) SDeC is well complementary with ConsiStory to further push the performance, as they address distinct aspects. (5) Interestingly, training-based methods are even outpaced by most tree-free counterparts in ID consistency, with extra disadvantage in efficiency. (6) In terms of memory and inference time, SDeC introduces negligible overhead on top.

User study To complement those evaluation metrics, we further conduct a user study. We compare with top alternatives: PhotoMaker, ConsiStory, StoryDiffusion, and 1P1S. Specifically, the test images were generated by 30 random prompt sets from ConsiStory+. A total of 20 volunteers were invited to pick which image best balances among ID consistency, scene diversity, and prompt alignment. We measure the performance using the percentage of wins. Tab. 2 shows that SDeC best matches human preference.

Table 2: User study results. Criteria: Best balance in ID consistency, scene diversity, and prompt alignment.

Method	PhotoMaker	ConsiStory	StoryDiffusion	1P1S	**SDeC**
Wins↑	8.17%	20.83%	13.33%	15.00%	42.67%

Qualitative analysis For visual comparison, we show a couple of examples in Fig. 4. For the robotic elephant case, ConsiStory presents varying robotic styles, whilst 1P1S suffers with the scene interference. For the cup of hot chocolate case, the ID shift issue becomes more acute with all previous methods. Instead, SDeC still does a favored job. More qualitative results are provided in Appendix-I.

5.2 Further Analysis

Ablation study In SDeC, there are two key designs: (1) Estimate $P_\cap$ in a soft manner by a learning process, and (2) employ the absolute excursion of SVD eigenvalues to identify the latent scene-ID correlation subspace. To isolate their effect, we tailor two variations of SDeC: (1) SDeC w/o soft-estimation, where we operate the shared subspace in a hard way: By constructing a correlation matrix to estimate $P_\cap$ (its implementation details are provided in Appendix-E.3), and (2) SDeC w/o abs-excursion where the corresponding eigenvalue normalizes the eigenvalue variation.

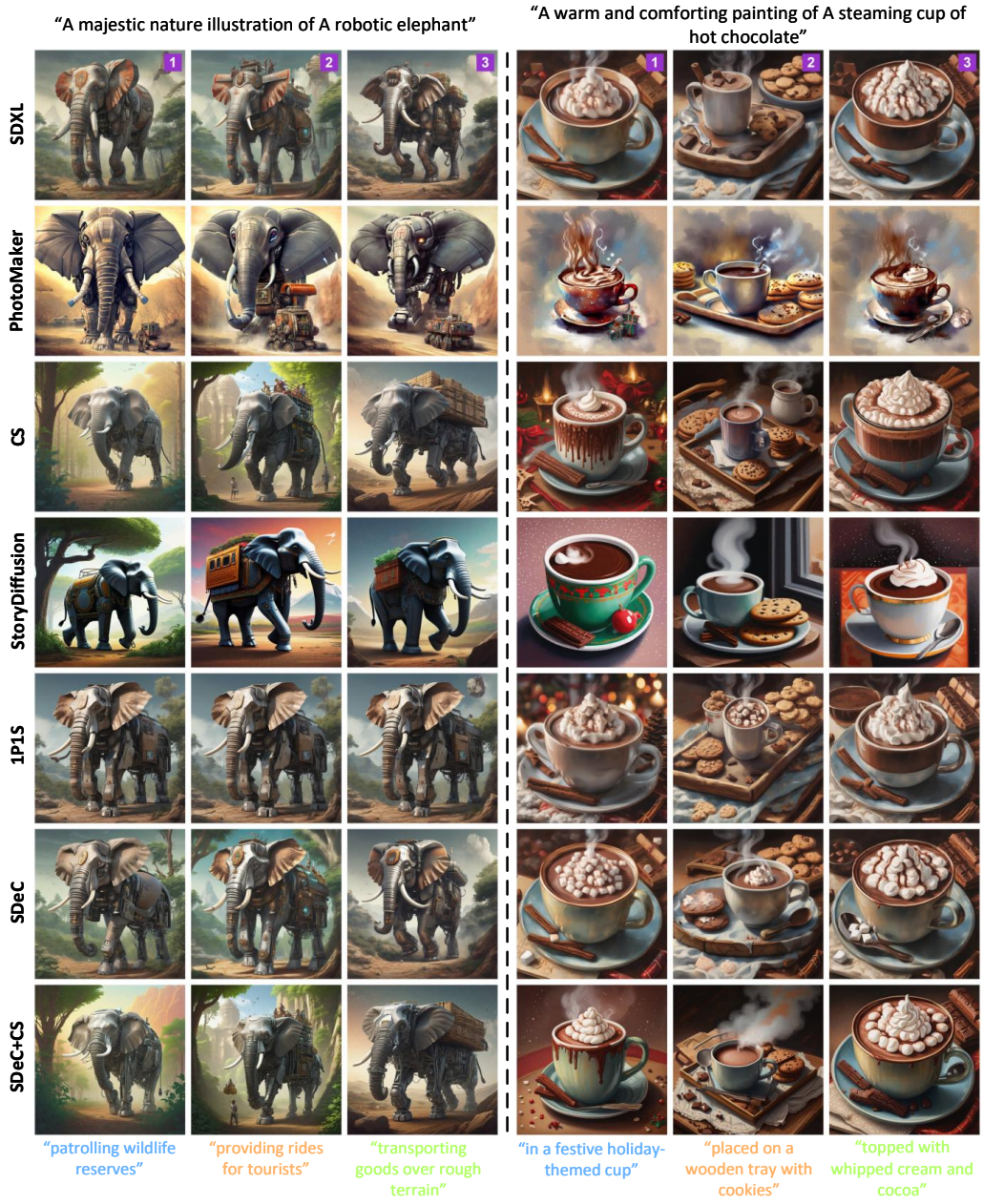

Figure 4: Qualitative results. CS: ConsiStory.

Tab. 3 presents the ablation study results. SDeC ranks first in terms of ID metrics and scenario alignment (see CLIP-T), indicating that the two designs positively affect the final performance. These results are understandable. In reality, the relationship between ID and scene subspaces is complicated. Thus, explicitly constructing $P_{\cap}$ is nearly infeasible. A tractable approach is by approximation using high-dimensional matrix decomposition, which, however, is numerically unstable under limited samples (Yang et al., 2023).

Table 3: Ablation study. Best results are in **bold**.

Method	ID metrics		Scene metrics	
	DreamSim-F↓	CLIP-I↑	DreamSim-B↑	CLIP-T↑
SDeC w/o soft-estimation	0.3351	0.8320	0.4254	0.8755
SDeC w/o abs-excursion	0.3576	0.8190	**0.4440**	0.8569
SDeC	**0.2589**	**0.8655**	0.3675	**0.8946**

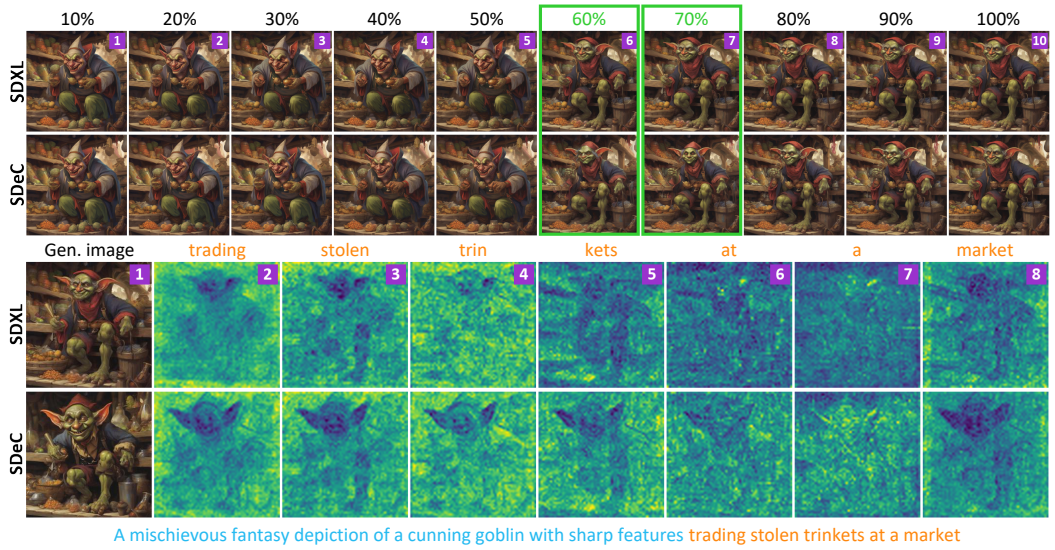

Figure 5: Validation of scene contextualization control. **Top**: PCA-based analysis where the image pairs blocked by the green-highlighted box have noticeable discrepancy. **Bottom**: Correlation between scene token embedding and attention similarity matrix.

Accordingly, SDeC w/o soft-estimation suffers from a performance decrease. As for SDeC w/o abs-excursion, its relative excursion strategy enforces the stability quantifying on a unified scale, disregarding the intrinsic importance of each eigen-direction.

Additionally, relative to SDeC, the $P_{\cap}$ estimated by these two variant approaches is inherently less reliable. When applied to contextualization carrier suppression, it introduces additional and unpredictable interference into the ID embedding. Through the coupling effect of attention, this interference propagates into the scene generation process, ultimately amplifying discrepancies across scenes. This is why the variant approaches have a higher DreamSim-B score than SDeC. The typical qualitative comparison results can be found in `Appendix-G.1`.

Validation of scene contextualization control Our validation is based on a scenario with ID prompt: "A mischievous fantasy depiction of A cunning goblin with sharp features" and scene prompt: "trading stolen trinkets at a market" where "trinkets" is sliced to "trin" and "kets" by text-encoder. Employing SVD to the original ID prompt embedding, we have $\mathcal{Z}_{\text{id}}^{\text{o}} = U_{\text{id}}^{\text{o}} \Lambda_{\text{id}}^{\text{o}} V_{\text{id}}^{\text{o}\top}$. For the refined one $\mathcal{Z}_{\text{id}}^{*}$ obtained by our SDeC, the weighting coefficients is Λ_{ω} (see Eq.(8)).

The key to SDeC is that the detected scene-ID correlation directions can modulate identity traits. To verify this, we apply PCA-based suppression to $\mathcal{Z}_{\text{id}}^{\text{o}}$ using criteria $\Lambda_{\text{id}}^{\text{o}}$ and Λ_{ω}, respectively. Sweeping the cumulative energy threshold in 10% increments yields the ten image pairs shown in Fig. 5-Top. In pairs #6–7, the subject produced by SDeC noticeably diverges from its SDXL counterpart, indicating that SDeC indeed identifies overlapping directions that drive identity adjustment.

Additionally, our prompt editing design operates through the attention module. Accordingly, the most direct evidence of contextualization control is to examine the correlation between token embeddings and the attention similarity matrix. Fig. 5-Bottom visualizes it as $\mathcal{Z}_{\text{id}}^{*}$ drives the generation. With bright green denoting high correlation, we observe that, except for image #7, SDeC yields darker regions with more pronounced subject silhouettes than SDXL. These observations suggest that SDeC reduces the scene contextualization and provides an intuitive explanation for SDeC's effectiveness. The more model analyses are provided in `Appendix-H`.

Analysis of a proprietary commercial product Google's latest flagship T2I product, Nano Banana (Google & DeepMind, 2025), has demonstrated impressive ID-preservation ability. While we cannot integrate our method, we design an interesting test to inspect how this system might work, which may reveal additional insights for future work. The results and speculative analysis are presented in `Appendix-K`.

6 CONCLUSION

In this paper, we identify scene contextualization as a key source of ID shift in T2I generation and conduct a formal investigation. Our analysis shows that this contextualization is an inevitable attention-induced phenomenon and the primary driver of ID shift in T2I models. By deriving theoretical bounds on its strength, we provide a foundation for mitigating this effect. Building on these insights, we introduce SDeC, a training-free embedding editing method that suppresses latent scene-ID correlation subspace through eigenvalue stability analysis, yielding refined ID embeddings for more consistent generation. Extensive experiments validate both the effectiveness and generality of the proposed approach. The limitations and future work are elaborated in Appendix-J.

REPRODUCIBILITY STATEMENT

The code and data will be made available after the publication of this paper.

REFERENCES

Kiymet Akdemir and Pinar Yanardag. Oracle: Leveraging mutual information for consistent character generation with loras in diffusion models. *arXiv preprint arXiv:2406.02820*, 2024.

Omri Avrahami, Amir Hertz, Yael Vinker, Moab Arar, Shlomi Fruchter, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. The chosen one: Consistent characters in text-to-image diffusion models. In *ACM SIGGRAPH 2024 Conference*, pp. 1–12, 2024.

Shengqu Cai, Eric Ryan Chan, Yunzhi Zhang, Leonidas Guibas, Jiajun Wu, and Gordon Wetzstein. Diffusion self-distillation for zero-shot customized image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18434–18443, 2025.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023.

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. *arXiv preprint arXiv:2208.01618*, 2022.

Google and DeepMind. Introducing gemini 2.5 flash image, our state-of-the-art image model. Google Developers Blog / Google AI Studio, 2025. URL https://developers.googleblog.com/en/introducing-gemini-2-5-flash-image/. "Gemini 2.5 Flash Image" (nano-banana) model.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *Proceedings of the Eleventh International Conference on Learning Representations*, 2023.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. ClipScore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

Lukas Höllein, Aljaž Božič, Norman Müller, David Novotny, Hung-Yu Tseng, Christian Richardt, Michael Zollhöfer, and Matthias Nießner. ViewDiff: 3d-consistent image generation with text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5043–5052, 2024.

Lee Hsin-Ying, Kelvin CK Chan, and Ming-Hsuan Yang. Consistent subject generation via contrastive instantiated concepts. *arXiv preprint arXiv:2503.24387*, 2025.

Guojun Lei, Chi Wang, Rong Zhang, Yikai Wang, Hong Li, and Weiwei Xu. Animateanything: Consistent and controllable animation for video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 27946–27956, 2025.

Dongxu Li, Junnan Li, and Steven Hoi. Blip-Diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36:30146–30166, 2023.

Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8640–8650, 2024.

Tao Liu, Kai Wang, Senmao Li, Joost van de Weijer, Fahad Shahbaz Khan, Shiqi Yang, Yaxing Wang, Jian Yang, and Ming-Ming Cheng. One-Prompt-One-Story: Free-lunch consistent text-to-image generation using a single prompt. *arXiv preprint arXiv:2501.13554*, 2025.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Tanzila Rahman, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, Shweta Mahajan, and Leonid Sigal. Make-a-story: Visual memory conditioned consistent story generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2493–2502, 2023.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pp. 8821–8831. Pmlr, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, June 2022a.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022b.

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*, pp. 22500–22510, 2023.

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

Nikita Selin. Carvekit: Automated high-quality background removal framework, 2023.

Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8543–8552, 2024.

Gilbert W Stewart. On the early history of the singular value decomposition. *SIAM review*, 35(4): 551–566, 1993.

Kaiyue Sun, Xian Liu, Yao Teng, and Xihui Liu. Personalized text-to-image generation with auto-regressive models. *arXiv preprint arXiv:2504.13162*, 2025.

Song Tang, Wenxin Su, Mao Ye, and Xiatian Zhu. Source-free domain adaptation with frozen multimodal foundation model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23711–23720, 2024.

Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, and Changsheng Xu. Df-gan: A simple and effective baseline for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16515–16525, June 2022.

Ming Tao, Bing-Kun Bao, Hao Tang, Yaowei Wang, and Changsheng Xu. Storyimager: A unified and efficient framework for coherent story visualization and completion. In *European Conference on Computer Vision*, pp. 479–495. Springer, 2024.

Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. Training-free consistent text-to-image generation. *ACM Transactions on Graphics*, 43(4):1–18, 2024.

Jiahao Wang, Caixia Yan, Haonan Lin, Weizhan Zhang, Mengmeng Wang, Tieliang Gong, Guang Dai, and Hao Sun. OneActor: Consistent character generation via cluster-conditioned guidance. In *Advances in Neural Information Processing Systems*, 2024.

Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4563–4573, 2023.

You Wu, Kean Liu, Xiaoyue Mi, Fan Tang, Juan Cao, and Jintao Li. U-VAP: User-specified visual appearance personalization via decoupled self augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9482–9491, 2024.

Yihe Yang, Hongsheng Dai, and Jianxin Pan. Block-diagonal precision matrix regularization for ultra-high dimensional data. *Computational Statistics & Data Analysis*, 179:107630, 2023.

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

Yu Zeng, Vishal M Patel, Haochen Wang, Xun Huang, Ting-Chun Wang, Ming-Yu Liu, and Yogesh 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*, pp. 6786–6795, 2024.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023.

Xulu Zhang, Xiaoyong Wei, Wentao Hu, Jinlin Wu, Jiaxin Wu, Wengyu Zhang, Zhaoxiang Zhang, Zhen Lei, and Qing Li. A survey on personalized content synthesis with diffusion models. *arXiv preprint arXiv:2405.05538*, 2024.

Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. StoryDiffusion: Consistent self-attention for long-range image and video generation. *Advances in Neural Information Processing Systems*, 37:110315–110340, 2024.

A PROOF OF THEOREM 1

Restatement of Theorem 1 *Let $\mathcal{H}_{id}, \mathcal{H}_{sc} \subset \mathbb{R}^d$ be the identity/scene semantic subspaces, which have ideal semantic separation: $\mathcal{H}_{\mathrm{id}} \cap \mathcal{H}_{\mathrm{sc}} = \varnothing$. Let $\mathcal{Z}_{id} \subseteq \mathcal{H}_{id}$ and $\mathcal{Z}^k_{sc} \subseteq \mathcal{H}_{sc}$ be the prompt-embedding matrix of $\mathcal{P}_{id}$ and $\mathcal{P}^k_{sc}$, respectively; the prompt-embedding matrix be partitioned by semantics $\mathcal{Z} = \left[\mathcal{Z}_{\mathrm{id}}; \mathcal{Z}^k_{\mathrm{sc}}\right] \in \mathbb{R}^{n\times d}$. For any query q_{id} associated with the identity, its attention output formulated by Eq. (1) can be specified to*

$$O(q_{\mathrm{id}}) = \alpha^\top (\mathcal{Z} W_V) = \alpha_{\mathrm{id}}^\top (\mathcal{Z}_{\mathrm{id}} W_V) \; + \; \alpha_{\mathrm{sc}}^\top \left(\mathcal{Z}^k_{\mathrm{sc}} W_V\right), \tag{9}$$

where the attention weights by token index $\alpha = [\alpha_{\mathrm{id}}, \alpha_{\mathrm{sc}}]$ conforming to the row split of Z. Let Π_{id} be the orthogonal projector onto $\mathcal{H}_{\mathrm{id}}$. The projection of $O(q_{\mathrm{id}})$ onto the identity subspace $\mathcal{H}_{\mathrm{id}}$ is

$$\Pi_{\mathrm{id}}\big[O(q_{\mathrm{id}})\big] = \underbrace{\Pi_{\mathrm{id}}\big[\alpha_{\mathrm{id}}^\top (\mathcal{Z}^k_{\mathrm{id}} W_V)\big]}_{\textit{id term: } T_{\mathrm{id}}} \; + \; \underbrace{\Pi_{\mathrm{id}}\big[\alpha_{\mathrm{sc}}^\top (\mathcal{Z}^k_{\mathrm{sc}} W_V)\big]}_{\textit{scene term: } T_{\mathrm{sc}}}. \tag{10}$$

Assume $\Pi_{\mathrm{id}} \circ W_V\big|_{\mathcal{H}_{\mathrm{sc}}}$ denotes an operation where an input from subspace $\mathcal{H}_{\mathrm{sc}}$ is first transformed by W_V and then projected onto subspace $\mathcal{H}_{\mathrm{id}}$. If condition (A) $\alpha_{\mathrm{sc}} \neq \mathbf{0}$ and (B) $\Pi_{\mathrm{id}} \circ W_V\big|_{\mathcal{H}_{\mathrm{sc}}} \neq 0$ hold, then the scene term in Eq. (10)) is nonzero. That is, $T_{\mathrm{sc}} \neq 0$.

Proof 1 *To obtain the result presented above, we need to firstly prove (1) $O_{\mathrm{sc}}(q_{\mathrm{id}}) = \alpha_{\mathrm{sc}}^\top \left(\mathcal{Z}^k_{\mathrm{sc}} W_V\right) \neq 0$ (the second term in Eq. (9)), and (2) its projection onto space $\mathcal{H}_{\mathrm{sc}}$ is non-zero.*

(1) Proving $O_{\mathrm{sc}}(q_{\mathrm{id}}) \neq \mathbf{0}$. *Due to $\alpha_{\mathrm{sc}} \neq \mathbf{0}$, O_{sc} is a non-trivial linear combination of the rows of $Z_{\mathrm{sc}} W_V$; hence, it is nonzero unless $\mathcal{Z}_{\mathrm{sc}} W_V$ vanishes row-wise, which is not the case in functioning models. This establishes that a* nonzero scene term O_{sc} *is present in* $\mathrm{Att}(q_{\mathrm{id}})$.

(2) Proving the projection of $O_{\mathrm{sc}}(q_{\mathrm{id}})$ onto $\mathcal{H}_{\mathrm{id}}$ is non-zero. *From The condition $\Pi_{\mathcal{H}_{\mathrm{id}}} \circ W_V\big|_{\mathcal{H}_{\mathrm{sc}}} = \mathbf{0}$, for any scene vector $z^{(s)} \in \mathcal{H}_{\mathrm{sc}}$, we have $\Pi_{\mathcal{H}_{\mathrm{id}}}\big(W_V z^{(s)}\big) \neq \mathbf{0}$. Then, the scene contribution has a nonzero component along $\mathcal{H}_{\mathrm{id}}$. Because α is a softmax, all its entries are non-negative and at least one scene weight is strictly positive (since $\alpha_{\mathrm{sc}} \neq \mathbf{0}$). Therefore,*

$$T_{\mathrm{sc}} = \Pi_{\mathrm{id}}\big[\alpha_{\mathrm{sc}}^\top (\mathcal{Z}_{\mathrm{sc}} W_V)\big] = \sum_{j \in \mathrm{scene}} \alpha_j \, \Pi_{\mathrm{id}}\big[(z_j^{(s)})^\top W_V\big] \neq \mathbf{0}. \tag{11}$$

B PROOF OF COROLLARY 1

Restatement of Corollary 1 *Assume $\mathcal{H}_{\mathrm{id}}$ and $\mathcal{H}_{\mathrm{sc}}$ have a nontrivial intersection: $\mathcal{H}_\cap := \mathcal{H}_{\mathrm{id}} \cap \mathcal{H}_{\mathrm{sc}}$, $k_\cap := \dim(\mathcal{H}_\cap) > 0$ where $\dim(\cdot)$ means space dimensions. If $\alpha_{\mathrm{sc}} \neq 0$, then for a generic linear mapping W_V, which excludes measure-zero degenerate cases, $T_{\mathrm{sc}} \neq 0$ hold.*

Proof 2 *Pick any nonzero $u \in \mathcal{H}_\cap$. Since $u \in \mathcal{H}_{\mathrm{sc}}$ and $\mathcal{Z}_{\mathrm{sc}} \supseteq \mathcal{H}_{\mathrm{sc}}$, there exists a scene row $z^{(s)}$ such that $z^{(s)} = \beta u + z^\perp$ with $\beta \neq 0$, $z^\perp \perp \mathcal{H}_{\mathrm{id}}$. For any W_V, we further have*

$$\Pi_{\mathrm{id}}(W_V z^{(s)}) = \beta \, \Pi_{\mathrm{id}}(W_V u) + \Pi_{\mathrm{id}}(W_V z^\perp). \tag{12}$$

Consider the set $\mathcal{U} := \{\, W_V : \Pi_{\mathrm{id}}(W_V u) = 0 \,\}$, which is a proper linear subspace of $\mathbb{R}^{d\times d}$ and hence has Lebesgue measure zero. Therefore, for almost every W_V, we have $\Pi_{\mathrm{id}}(W_V u) \neq 0$, implying $\Pi_{\mathrm{id}}(W_V z^{(s)}) \neq 0$. Since $\alpha_{\mathrm{sc}} \neq 0$ has at least one strictly positive entry,

$$T_{\mathrm{sc}} = \Pi_{\mathrm{id}}\big[\alpha_{\mathrm{sc}}^\top (\mathcal{Z}_{\mathrm{sc}} W_V)\big] = \sum_{j \in \mathrm{scene}} \alpha_j \, \Pi_{\mathrm{id}}\big(W_V z_j^{(s)}\big) \neq 0. \tag{13}$$

C PROOF OF THEOREM 2

Restatement of Theorem 2 *Let $P_\cap$ be the orthogonal projector onto $\mathcal{H}_\cap$; Π_{sc} be the orthogonal projector onto $\mathcal{H}_{\mathrm{sc}}$; $P_\cap^\perp := \Pi_{\mathrm{sc}} - P_\cap$ be the projector onto the orthogonal complement within $\mathcal{H}_{\mathrm{sc}}$.*

Define $R_\cap := \mathcal{Z}_{\text{sc}}^k P_\cap$, $R_\perp := \mathcal{Z}_{\text{sc}}^k P_\cap^\perp$, $T_\cap := \Pi_{\text{id}} W_V P_\cap$, $T_\perp := P_\cap^\perp W_V \Pi_{\text{id}}$, *and* $\epsilon := \|\alpha_{\text{sc}}^\top\|_2$. *For contextualization strength* $\left\|T_{\text{sc}}\right\|_2$, *it is bounded by*

$$0 \le \|T_{\text{sc}}\|_2 \;\le\; \epsilon \cdot \|R_\cap\|_2 \cdot \|T_\cap\|_F \;+\; \epsilon \cdot \|R_\perp\|_2 \cdot \|T_\perp\|_F. \tag{14}$$

Proof 3 *Since* $\text{col}(\mathcal{Z}_{\text{sc}}^k) \subseteq \mathcal{H}_{\text{sc}}$, *for any vector* x, $\mathcal{Z}_{\text{sc}}^k x = \mathcal{Z}_{\text{sc}}^k \Pi_{\text{sc}} x = \mathcal{Z}_{\text{sc}}^k (P_\cap + P_\cap^\perp) x$. *Thus*

$$\begin{aligned} T_{\text{sc}} &= \Pi_{\text{id}} \left[\alpha_{\text{sc}}^\top (\mathcal{Z}_{\text{sc}}^k W_V)\right] \\ &= \alpha_{\text{sc}}^\top (\mathcal{Z}_{\text{sc}}^k W_V \Pi_{\text{id}}) \\ &= \alpha_{\text{sc}}^\top ((\mathcal{Z}_{\text{sc}}^k (P_\cap + P_\cap^\perp)) W_V \Pi_{\text{id}}) \\ &= \alpha_{\text{sc}}^\top \mathcal{Z}_{\text{sc}}^k P_\cap W_V \Pi_{\text{id}} + \alpha_{\text{sc}}^\top \mathcal{Z}_{\text{sc}}^k P_\cap^\perp W_V \Pi_{\text{id}}. \end{aligned} \tag{15}$$

Based on $R_\cap := Z_{\text{sc}}^k P_\cap$, $R_\perp := Z_{\text{sc}}^k P_\cap^\perp$, $T_\cap := \Pi_{\text{id}} W_V P_\cap$, $T_\perp := P_\cap^\perp W_V \Pi_{\text{id}}$, *Eq. (15) becomes*

$$\begin{aligned} T_{\text{sc}} &= \alpha_{\text{sc}}^\top \mathcal{Z}_{\text{sc}}^k P_\cap P_\cap W_V \Pi_{\text{id}} + \alpha_{\text{sc}}^\top \mathcal{Z}_{\text{sc}}^k P_\cap^\perp P_\cap^\perp W_V \Pi_{\text{id}} \\ &= \alpha_{\text{sc}}^\top R_\cap T_\cap + \alpha_{\text{sc}}^\top R_\perp T_\perp. \end{aligned} \tag{16}$$

Applying the triangle inequality ($||a+b|| \le ||a|| + ||b||$) *to Eq. (16), we have the following inequality.*

$$\|T_{\text{sc}}\|_2 \;\le\; \|\alpha_{\text{sc}}^\top R_\cap T_\cap\|_2 + \|\alpha_{\text{sc}}^\top R_\perp T_\perp\|_2. \tag{17}$$

For any row vector $x^\top$ *and matrix* A, $\|x^\top A\|_2 \le \|x\|_2 \|A\|_F$ *(column-wise Cauchy–Schwarz). Also* $\|YZ\|_2 \le \|Y\|_2 \|Z\|_2$ *(submultiplicativity). Apply these to each term in Eq. (17) to obtain*

$$\begin{aligned} \|\alpha_{\text{sc}}^\top R_\cap T_\cap\|_2 &\le \|\alpha_{\text{sc}}^\top\|_2 \cdot \|R_\cap\|_2 \cdot \|T_\cap\|_F = \epsilon \cdot \|R_\cap\|_2 \cdot \|T_\cap\|_F, \\ \|\alpha_{\text{sc}}^\top R_\perp T_\perp\|_2 &\le \|\alpha_{\text{sc}}^\top\|_2 \cdot \|R_\perp\|_2 \cdot \|T_\perp\|_F = \epsilon \cdot \|R_\perp\|_2 \cdot \|T_\perp\|_F. \end{aligned} \tag{18}$$

Summing the two inequalities above gives the upper bound claim.

D Proof of Corollary 2

Restatement of Corollary 2 $\mathcal{H}_{\text{id}}$ *is the subspace spanned by the ID embedding* $\mathcal{Z}_{\text{id}}$; U *is an orthonormal basis of* $\mathcal{H}_{\text{id}}$, *i.e.,* $U = \text{orth}(\mathcal{Z}_{\text{id}})$, *where* $\text{orth}(\cdot)$ *means performing orthogonalization on the input. Writing the projector onto the ID subspace as* $\Pi_{\text{id}} = UU^\top$, *we have*

$$0 \le \|T_{\text{sc}}\|_2 \;\le\; \epsilon \cdot \|R_\cap\|_2 \cdot \|U^\top W_V P_\cap\|_F \;+\; \epsilon \cdot \|R_\perp\|_2 \cdot \|W_V^\top P_\cap^\perp U\|_F. \tag{19}$$

Proof 4 *Since* $\Pi_{\text{id}} = UU^\top$, *the term* $\|T_\cap\|_F$ *and* $\|T_\perp\|_F$ *in Theorem 2 can be*

$$\|\Pi_{\text{id}} W_V P_\cap\|_F^2 = \|UU^\top W_V P_\cap\|_F^2 \text{ and } \|W_V^\top P_\cap^\perp \Pi_{\text{id}}\|_F^2 = \|W_V^\top P_\cap^\perp UU^\top\|_F^2. \tag{20}$$

We first prove $\|UU^\top W_V P_\cap\|_F^2 = \|U^\top W_V P_\cap\|_F^2$. *Expanding the left-hand side of it, we have*

$$\begin{aligned} \|UU^\top W_V P_\cap\|_F^2 &= \text{tr}\big((UU^\top W_V P_\cap)^\top (UU^\top W_V P_\cap)\big) \\ &= \text{tr}\big(P_\cap^\top W_V^\top UU^\top UU^\top W_V P_\cap\big). \end{aligned} \tag{21}$$

Since $UU^\top$ *is an orthogonal projector, we have* $UU^\top UU^\top = UU^\top$; *then applying the cyclic property of the trace, we obtain:*

$$\|UU^\top W_V P_\cap\|_F^2 = \text{tr}\big(P_\cap^\top W_V^\top UU^\top W_V P_\cap\big) = \text{tr}\big(U^\top W_V P_\cap P_\cap^\top W_V^\top U\big). \tag{22}$$

Because $P_\cap$ *is itself an orthogonal projector, we have* $P_\cap^2 = P_\cap$ *and* $P_\cap^\top = P_\cap$. *Therefore,*

$$\|UU^\top W_V P_\cap\|_F^2 = \text{tr}\big((U^\top W_V P_\cap)(U^\top W_V P_\cap)^\top\big) = \|U^\top W_V P_\cap\|_F^2. \tag{23}$$

In the similar way, we have $\|W_V^\top P_\cap^\perp UU^\top\|_F^2 = \|W_V^\top P_\cap^\perp U\|_F$. *Substituting it and Eq. (23) to the upper bound from Theorem 2, we finish the proof.*

Table 4: Code link of the comparison methods.

Method	Code link
BLIP-Diffusion (Li et al., 2023)	https://github.com/salesforce/LAVIS/tree/main/projects/blip-diffusion
Textual Inversion (Gal et al., 2022)	https://github.com/oss-roettger/XL-Textual-Inversion
PhotoMaker (Li et al., 2024)	https://github.com/TencentARC/PhotoMaker
ConsiStory (Tewel et al., 2024)	https://github.com/NVlabs/consistory
StoryDiffusion (Zhou et al., 2024)	https://github.com/HVision-NKU/StoryDiffusion
1Prompt1Story (Liu et al., 2025)	https://github.com/byliutao/1Prompt1Story

E Experimental Details

E.1 Parameters setting

SDeC involves three parameters: Trade-off parameter β and switching constant M in Eq. (6) and weighting strength Ω in Eq. (8). We set $(\beta,\ M,\ \Omega) = (10,\ 20,\ 1)$ cross all experiments.

E.2 Model setting

We achieve ID-preserving image generation by editing the ID prompt embeddings during the inference phase. There is no extra training or optimization imposed on the generative models. In practice, we adopt Stable Diffusion (SD) V1.5[1] as the backbone model of BLIP-Diffusion, while the pre-trained Stable Diffusion XL (SDXL)[2] is selected as the backbone model for our SDeC and the rest of the comparison approaches.

The comparison methods are implemented using the unofficial or official codes from GitHub website, whose details are listed in Tab. 4. The computation platform adopts the same configuration as SDeC: NVIDIA RTX A6000 GPU with 48GB VRAM.

In addition, among these comparisons, BLIP-Diffusion (Li et al., 2023) and PhotoMaker (Li et al., 2024) rely on an additional reference image. We produce this reference by feeding the identity prompt into their respective base models, the same as 1Prompt1Story (1P1S) (Liu et al., 2025).

E.3 Implementation details of method SDeC w/o soft-estimation

The goal of **SDeC w/o soft-estimation** is to find the overlapping subspace directions between $\mathcal{Z}_{\text{id}}$ and $\mathcal{Z}_{\text{sc}}$, i.e., directions corresponding to a principal angle $\theta \approx 0$. We achieve this by:

We first perform SVD on both $\mathcal{Z}_{\text{id}}$ and $\mathcal{Z}_{\text{sc}}$ to obtain their orthonormal bases:

$$\mathcal{Z}_{\text{id}} = U_{\text{id}}\Lambda_{\text{id}}V_{\text{id}}^{\top}, \quad \mathcal{Z}_{\text{sc}} = U_{\text{sc}}\Lambda_{\text{sc}}V_{\text{sc}}^{\top}. \tag{24}$$

Then, the column subspaces are constructed as

$$B_{\text{id}} := V_{\text{id}}[:, 1 : r_{\text{id}}], \quad B_{\text{sc}} := V_{\text{sc}}[:, 1 : r_{\text{sc}}], \tag{25}$$

where r_{id} and r_{sc} denote the respective subspace ranks.

Subsequently, based on B_{id} and B_{sc}, we can compute their correlation matrix as

$$M = B_{\text{id}}^{\top}B_{\text{sc}} \ \in \mathbb{R}^{r_{\text{id}} \times r_{\text{sc}}}. \tag{26}$$

By performing SVD, we have $M = U\Lambda V^{\top}$, where the singular values $\sigma_i = \cos\theta_i$ correspond to the cosines of the principal angles θ_i. We select $\sigma_i \geq \tau$ (with τ typically chosen 0.98), where the associated singular vectors indicate the intersection directions.

Thus, the explicit basis for the intersection subspace in the original space can be written as $B_{\cap} = B_{\text{id}}U_{(:,\mathcal{I})}$, where $\mathcal{I} = \{i : \sigma_i \geq \tau\}$. Applying an additional QR orthonormalization on $B_{\cap}$ yields the final intersection basis $\hat{B}_{\cap}$. The projection operator onto the intersection subspace then can be

[1] https://huggingface.co/stable-diffusion-v1-5/stable-diffusion-v1-5
[2] https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0

Figure 6: Demonstration of scene-level interference caused by the prompt integration strategy in 1Prompt1Story (Liu et al., 2025). The visual interference elements are: **Top**: The bent tree in images #3∼6, **Middle**: The oak tree in all images, and **Bottom**: The cake in images #1∼5.

expressed by $P_{\cap} \approx \hat{B}_{\cap}\hat{B}_{\cap}^{\top}$. Corresponding to the proposed scheme in SDeC, we equivalently detect the latent scene–ID correlation subspace. After that, we can realize suppressing this correlation subspace by employing $\mathcal{Z}_{\mathrm{id}}^{*} = \mathcal{Z}_{\mathrm{id}}(I - P_{\cap})$ directly.

F Demonstration of Scene-level Interference in 1Prompt1Story

The previous 1Prompt1Story method (Liu et al., 2025) proposed a prompt strategy that merges the ID prompt with all scene prompts into a single input, followed by calibration across different scenes using SVD-based scene prompt selection. Subsequently, all images are generated from this consolidated prompt with a fixed ID component, thereby reducing ID shift. With the aid of the proposed attention module, serving as an adapter, this strategy further enhances identity preservation.

However, the proposed prompt strategy in 1Prompt1Story inevitably introduces pronounced *scene-level* interference. For an intuitive view, we demonstrate this phenomenon from 1Prompt1Story's results on the ConsiStory+ benchmark. As shown in Fig. 6, in the Top, images #1 and #2 both exhibit similar dense vegetation in the bottom-right corner, while the remaining images consistently feature a bent tree (mentioned in the prompt of image #5); in the Middle, all images are dominated by the visual element of an oak tree (mentioned in the prompt of image #4).

G Further Model Analysis

G.1 Qualitative Comparison of Ablation Study

As a supplement to Tab. 3, in Fig. 7, we present a qualitative comparison between methods SDeC w/o soft-estimation, SDeC w/o abs-excursion, and SDeC. It is seen that the qualitative results are consistent with the data in Tab. 3.

Figure 7: Qualitative comparison results of SDeC and its two variations SDeC w/o soft-estimation, SDeC w/o abs-excursion. The results of SDeC significantly outperform the variant methods. Moreover, SDeC w/o abs-excursion disregards the original importance of the SVD eigen-directions, causing greater semantics loss than both SDeC w/o soft-estimation and SDeC, and consequently ranks last in generation quality.

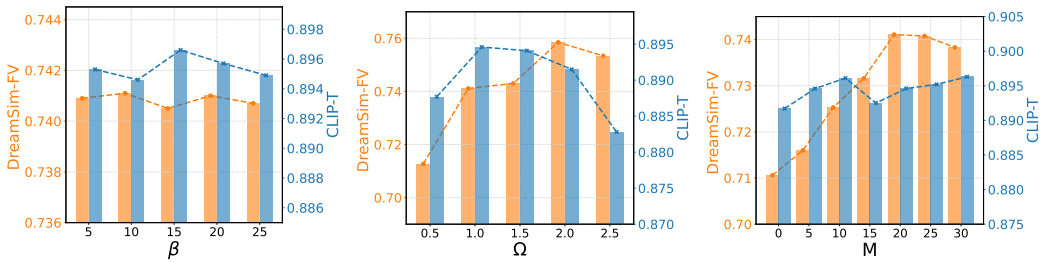

Figure 8: Parameter sensitivity analysis results of $\beta \in [5, 25]$, $\Omega \in [0.5, 2.5]$, and $M \in [0, 30]$. The DreamSim-F score is inverted to DreamSim-FV = 1-DreamSim-F, so higher values indicate better performance, as with CLIP-T score.

In particular, method SDeC w/o abs-excursion performs worse than method SDeC w/o soft-estimation, which supports our previous discussion. Specifically, SDeC w/o abs-excursion disregards the inherent importance of different SVD eigen directions, resulting in greater semantic loss and severe subject deformation (for example, see image #2 in Fig. 7-Right). In contrast, SDeC w/o soft-estimation, which employs a matrix transformation approach, just suffers from less accurate in identifying latent overlapping subspaces, but remain the key semantic information. Thus, it achieves better ID preservation than SDeC w/o abs-excursion.

G.2 Parameter sensitivity analysis

Our SDeC method involves three parameters, including weighting strength Ω in Eq. (8) and trade-off parameter β and switching constant M in Eq. (6). In Fig. 8, we provide the varying curves of DreamSim-F and CLIP-T scores as the three parameters change. As shown in Fig. 8-Left, the DreamSim-FV and CLIP-T scores remain near-steady when a wide range of $[5, 20]$, indicating our method does not rely on the careful selection of trade-off parameter β.

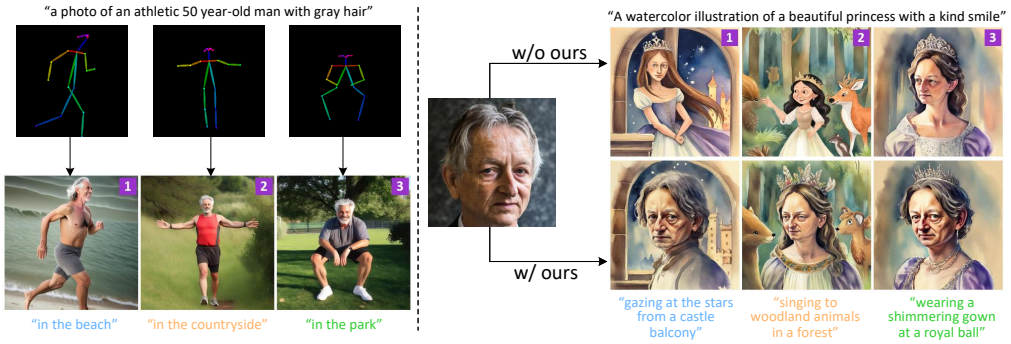

Figure 9: **Left**: Incorporating SDeC with ControlNet under the control of pose map. **Right**: Integrating SDeC with PhotoMaker, where a photo of *Geoffrey Hinton* serves as the reference.

The results in Fig. 8-Middle show that Ω's increase leads to a trade-off phenomenon between the DreamSim-FV and CLIP-T score. This phenomenon is reasonable. In SDeC, to avoid semantics loss, we enhance the robust subspace by imposing larger weights Ω on its corresponding eigen-directions (see Eq. (8)). When the increase of Ω remains within a reasonable range, the native correlation between ID and scene is reduced, improving both ID preservation and scene retention. Once it exceeds a threshold (e.g., > 1), however, the generative model is prone to pay more attention to the stronger ID components, leading to continued gains in ID preservation but convergence in scenes.

As for parameter M, Fig. 8-Right presents that DreamSim-FV curve progressively climbs and reaches the top at $M = 20$, while ClIP-T curve is with only minor fluctuations. The results are consistent with our expectations. In the forward-and-backward process, sufficient time in the forth stage drives the ID prompt embedding closer to that of the scene, enabling the identification of directions in ID prompt embedding's eigen-space most sensitive to contextualization. As this process acts only on ID, scene diversity remains unaffected.

H Supplemental Experiments

H.1 Integrating with other generative tasks

SDeC reduces the ID shift by editing the ID prompt embedding, without modifying the generative models. Consequently, it is compatible with different visual generation tasks. In this part, we incorporate the proposed method with two typical models with different generative goals. One is ControlNet (Zhang et al., 2023), which introduces controllable conditions (e.g., edge maps, pose maps, or depth maps) to enable structured control over image generation. The other is PhotoMaker (Li et al., 2024), which leverages an input reference image to preserve identity features, generating consistent subject images across diverse scenarios. As shown in Fig. 9, our SDeC demonstrates excellent compatibility and ID-preservation.

H.2 Consistent story generation with multiple subjects

In this part, we present the effect of SDeC as the ID prompt involves multiple subjects. Fig. 10 demonstrates a toy case involving an "elderly man" and a "cat". It can be observed that these two subjects are basically consistent, although some ID shift: The shorter cat fur in images #4, #9, and #10; the absence of glasses in images #1, #7; the hat in image #8.

H.3 Generality to base generative model

Extensive experiments with the SDXL base model show that our SDeC enables semantically meaningful editing. However, if its effectiveness were confined to only a subset of generative models, its practical applicability would be greatly limited. To demonstrate its generality, we integrate SDeC with four representative base models and compare performance before and after integration. Specif-

Figure 10: Consistent story generation with multiple subjects. The results show that our SDeC can generate images featuring multiple expected characters, with minor ID shift, for example, the hat (image #8) and glasses (images #1, #7).

ically, in addition to SDXL, the other base models include PlayGround-v2.5-1024px-Aesthetic[3], RealVisXL-V4.0[4], and Juggernaut-X-V10[5].

From the comparison presented in Fig. 11, we draw two main observations. First, within the SDXL group, SDeC substantially alters the subject in image #2, while making only minor adjustments to the others, for example, consistently sharpening the cats' chins. This is reasonable, as SDXL already performs well in ID preservation for the remaining images. Second, in contrast, when the base models exhibit evident identity divergence (see the other three groups), equipping them with SDeC leads to a significant improvement in the quality of the generated images.

I More qualitative results

As a supplement to Fig. 4, we provide more qualitative comparison results in Fig. 12.

J Limitation and Future Work

Despite improving ID consistency while maintaining scene diversity, SDeC, as a prompt-embedding editing approach, cannot fundamentally resolve ID shift. First, Theorem 1 proves that the attention mechanism is the central origin of ID shift, even when ID and scene prompt embeddings occupy disjoint subspaces. Second, in line with Theorem 2, SDeC essentially performs an indirect form of contextualization control by reducing the overlapping energy between ID and scene embeddings.

Therefore, designing attention modules tailored to preserve ID represents a promising direction for future work. In this paper, however, our theoretical contributions, such as Theorem 1 and Corollary 1, are intended to reveal the mechanistic link between scene contextualization and ID shift. Their practical contribution to attention design remains limited. This limitation arises from idealized assumptions (e.g., sharp subspace partitioning and linearity) and the neglect of real data geometry and attention dynamics. These factors render the precise construction of $P_{\cap}$ numerically unstable and difficult to apply in high-dimensional, sample-limited settings. Addressing these challenges will be a central focus for future work.

[3] https://huggingface.co/playgroundai/playground-v2.5-1024px-aesthetic
[4] https://huggingface.co/SG161222/RealVisXL_V4.0
[5] https://huggingface.co/RunDiffusion/Juggernaut-X-v10

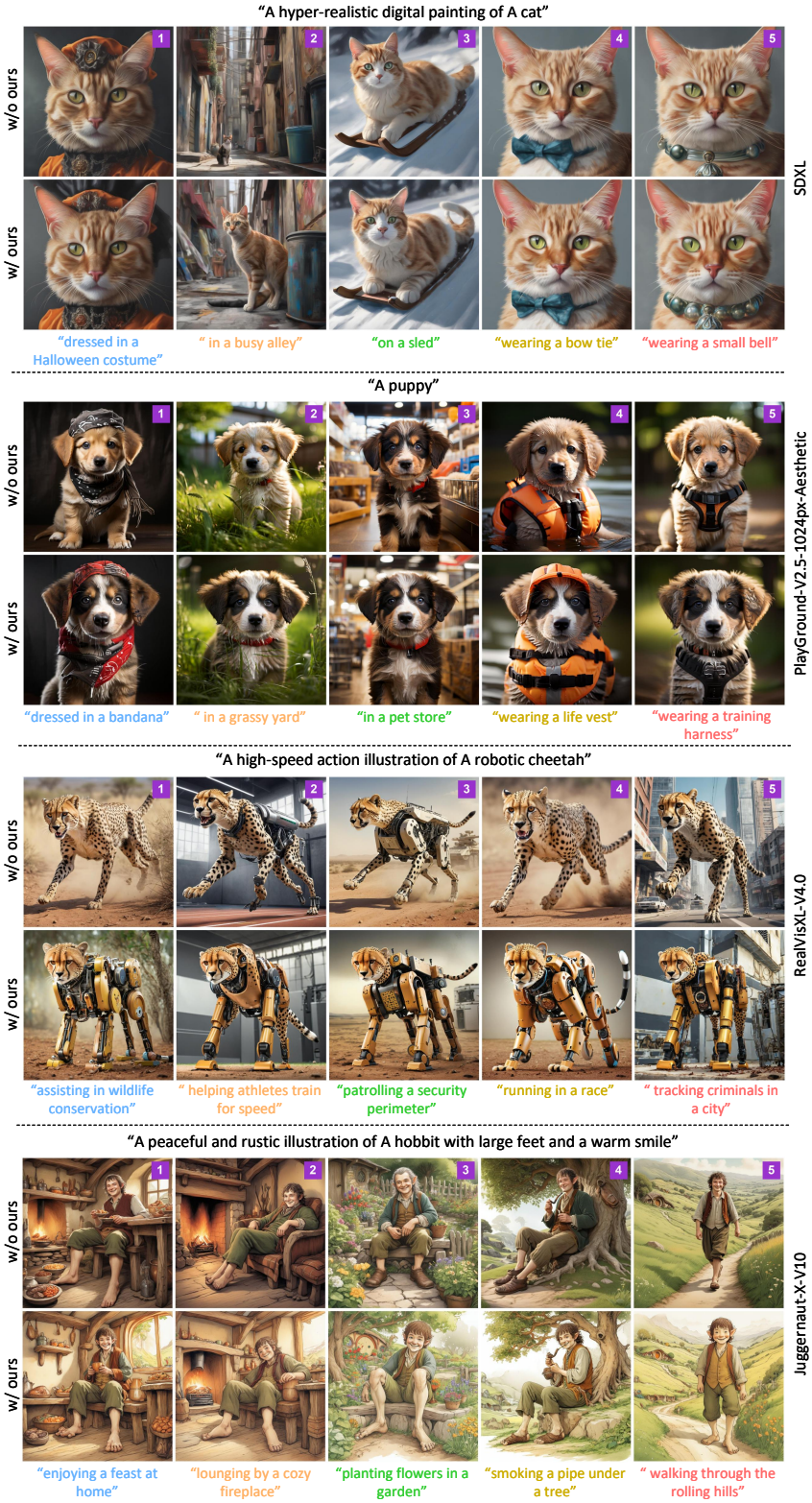

Figure 11: Comparison of combining SDeC with the base models. From **top** to **bottom**, there are results of **SDXL** group, **PlayGround-v2.5-1024px-Aesthetic** group, **RealVisXL-V4.0** group, and **Juggernaut-X-V10** group, respectively.

Figure 12: Supplementary qualitative comparison results.

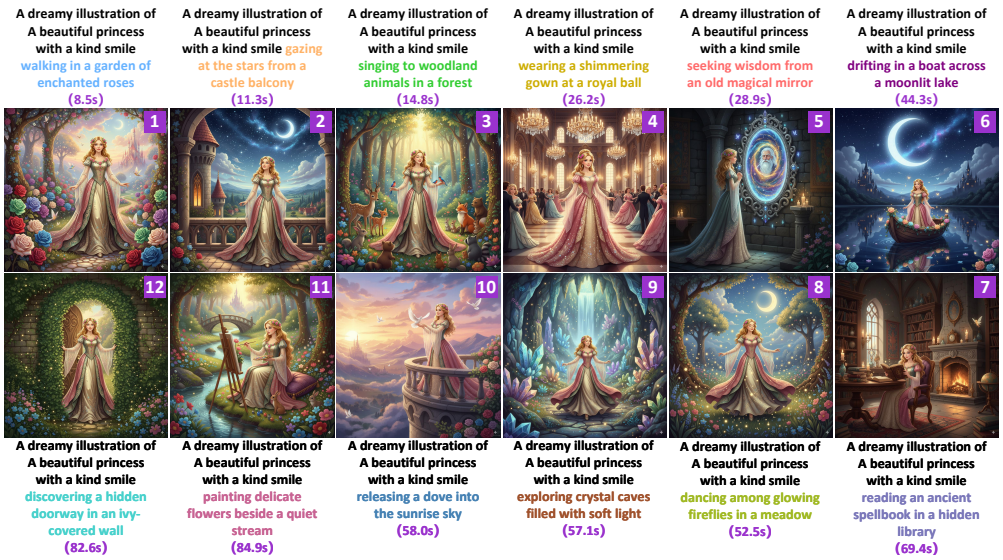

Figure 13: Results of baseline experiment where the images are arranged in a U-shape. All scenarios share the same ID prompt, and the scene prompts are different, as in the setting of this paper.

K An Empirical Study of Nano Banana's ID-preservation

K.1 Building baseline

As the beginning of our exploration, this baseline experiment generates 12 images using Nano Banana[6] under the setting formulated in the `Problem Statement` of Sec. 3. In practice, we start a new session and sequentially input the 12 prompts into Nano Banana via its official interface in dialogue mode. We set the ID prompt to "*A dreamy illustration of a beautiful princess with a kind smile*", and the scene prompts are generated using ChatGPT-5.0[7].

As shown in Fig. 13, the generated images exhibit excellent ID consistency, except for the hairstyle in image #4 and the clothing in image #5. We further observe that the generation time increases steadily from **8.5** seconds for the first image to **82.6** seconds for the last. This phenomenon suggests that Nano Banana might leverage previously generated images as contextual information during subsequent generation. If this speculation holds, Google's method implicitly incorporates a reference image to enforce ID consistency, the same as the personalized T2I generation (see `Sec. 2 Related Work`). Accordingly, the strong ID preservation becomes interpretable. In Nano Banana, the developed image editing technique (a feature also emphasized officially) is employed to integrate the subject from the reference image with the target scene. The subjects in images #1, #2, #4, #6, #8, and #9 provide compelling support: Their clothing and pose are mirrored, which exhibits clear signs of editing.

K.2 Validating leveraging prior images as context

As presented above, our conclusion regarding reference images is drawn under the assumption that contextual information is exploited in Nano Banana. To test this assumption, we selected scenario #1 and #4 from the baseline as the first and last ones, respectively, and inserted 10 intermediate scenarios between them. Based on this, we conducted two perturbation experiments as follows. Of note, to avoid interference between experiments, each experiment is conducted in a newly created session.

In the first experiment, we introduce substantial perturbations to the intermediate scenes by asking ChatGPT-5.0 to produce prompts that are markedly different from the first scenario. The results in Fig. 14 reveal significant differences between the first and last images (e.g., hairstyle, costume, and headdress), suggesting that the Nano Banana model indeed accounts for contextual information.

[6] https://aistudio.google.com/models/gemini-2-5-flash-image
[7] https://chatgpt.com/

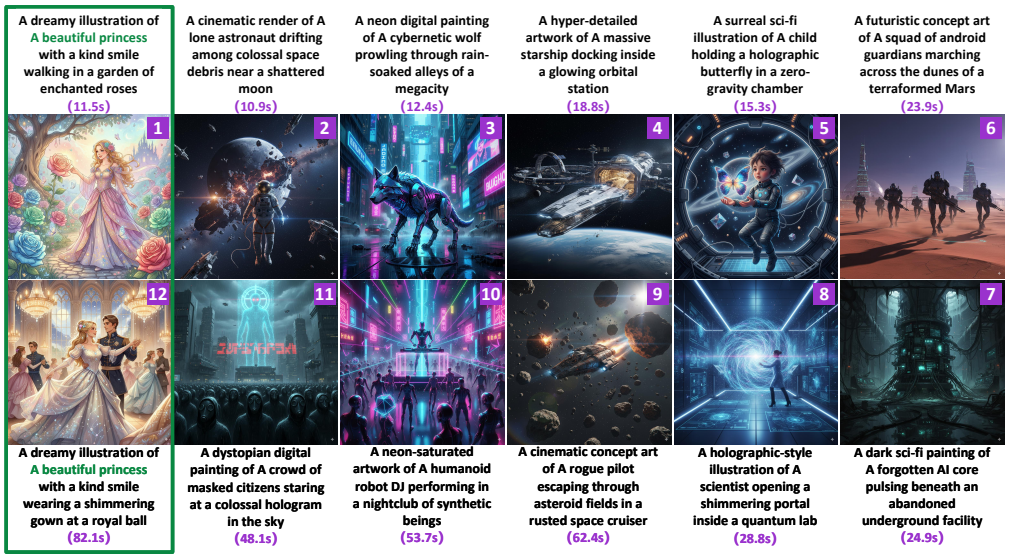

Figure 14: Results of experiment with major perturbation where the images are arranged in a U-shape. The intermediate scenarios (#2∼#11) significantly differ from the first one.

In the second experiment, we perform minor perturbations: The intermediate scenarios adopt a literary style similar to that of the first scenario, achieved by instructing ChatGPT-5.0 to replicate its style literally. Meanwhile, the ID component remains aligned with the first, but is slightly varied by substituting the subject with related terms (e.g., "woman" and "girl"). As shown in Fig. 15, compared with the baseline, the subjects in the first and last images again exhibited notable differences (e.g., hair color, hair length, and costume style). Moreover, the last image and its neighbors displayed higher facial similarity than the first. We further extend the number of scenarios to Nano Banana's maximum support (**22**). The results in Fig. 16 reveal that the differences between the first and last images are further amplified, making it difficult to regard the presented subjects as the same person.

While the observed ID shifts under the two types of perturbations suggest Nano Banana adopts a context strategy, another evidence arises from the side-by-side comparison of experimental results. Specifically, compared with minor perturbations (Fig. 15 and Fig. 16), strong perturbations (Fig. 14) result in smaller ID shifts. A reasonable explanation is that in the strong-perturbation case, the constructed context exhibits much larger semantic differences. This enables the generation process of image #12 to more easily converge attention on the most similar first scenario, thereby achieving ID-preservation. In contrast, in the minor-perturbation case, the attention distribution remains relatively flat. As a result of blending the semantics of multiple scenarios, the final generated image displays a much more pronounced ID shift.

K.3 Comparing with our method

In summary, the ID consistency observed in Nano Banana can be attributed to the implicit use of reference images, where previously generated results serve as contextual input. This allows image editing techniques to enforce high-quality ID consistency. In contrast, our method avoids such strong assumptions: The "one prompt per scene" feature provides an unconstrained usage condition and requires neither intensive computation nor additional data overhead.

Methodologically, Nano Banana falls within the category of personalized T2I generation. It typically relies on a reference dataset, taken as context, to model ID invariance, aligning with the principles of transfer learning. In contrast, SDeC takes a conceptually distinct path by pursuing a novel prompt embedding editing paradigm, derived from a native generative perspective on ID shift: Scene contextualization in each individual image.

Figure 15: Experiment with minor perturbation where the images are arranged in a U-shape. In the intermediate scenarios (#2∼#11), the scene prompts generated by ChatGPT-5.0 follow a literary style similar to that of the first, while the ID prompts remain consistent with the first, differing only in the substitution of the subject with a related concept.

Figure 16: Extended experiment with minor perturbations, under the same setting as Fig. 15, with the number of images increased to the maximum continuous generation times of Nano Banana.