Peer reviewed Journals

SciChronice

RESEARCH

Cluster Routed Identity Adaptation for Controllable Multimodal Content Generation Under Sparse Brand Supervision and Feedback Constrained Inference

Vo Hung Cuong,[1] Pham Quang Huy,[2] and Dang Duc Long[3]

[1]Faculty of Computer Science, Vietnam-Korea University of Information and Communication Technology, The University of Danang, Urban Area, Hoa Quy Ward, Ngu Hanh Son District, Da Nang 550000, Vietnam
[2]Department of Software Engineering, University of Phan Thiet, 225 Nguyen Thong Street, Phu Hai Ward, Phan Thiet City 77000, Binh Thuan Province, Vietnam
[3]Department of Information Systems, Ha Tinh University, 447 26 Thang 3 Street, Dai Nai Ward, Ha Tinh City 450000, Ha Tinh Province, Vietnam

Abstract

Generative content systems increasingly need to produce images that satisfy both semantic prompts and persistent identity constraints such as brand character, approved palette, recurring layout, and asset-specific visual language. General-purpose text-to-image models often preserve broad prompt semantics while drifting from these finer identity constraints, especially when the target identity is represented by only a few approved examples. This paper studies a cluster routed identity adaptation method for sparse content-identity generation. The method decomposes an identity repository into concept-level clusters, trains compact low-rank adapters for each cluster, routes a user prompt to one or more adapters through a joint text-image embedding space, and supplements prompts with brief-derived and caption-derived identity tokens when cluster support is limited. A second inference pass transfers masked background structure from approved reference assets while retaining generated foreground content. We evaluate the method on a simulated benchmark of 72 content identities, 11,840 approved assets, and 3,600 held-out generation prompts under two-shot, four-shot, eight-shot, and sixteen-shot regimes. Relative to a single identity adapter, the proposed method improves mean identity conformance from 0.673 to 0.761 in the four-shot setting, reduces palette error from 9.8 to 6.1 CIEDE2000 units, and lowers background FID from 38.4 to 29.7. Human preference judgments favor the routed method in 61.4% of pairwise comparisons against retrieval-augmented prompting and 57.2% against DreamBooth-style personalization. The results indicate that routing and low-data prompt supplementation can improve identity stability without fully retraining the base generator.

1. Introduction

Digital content generation has moved from isolated prompt-to-image demonstrations toward production pipelines where repeated visual identity matters. The practical question is not only whether a model can draw a dog, a package, a storefront, or a stylized object, but whether it can generate such content in the recognizable idiom of a specific identity while obeying ordinary instructions about scene, pose, medium, and background. In brand-facing production, this requirement is a technical constraint rather than a preference for aesthetic similarity. A generated image may be semantically accurate and visually plausible while still being unusable because the foreground object carries the wrong proportions, the background conflicts with approved visual grammar, the color palette departs from a narrow range, or the generated layout mixes motifs that are acceptable separately but not acceptable in the same artifact. These failures are visible in current personalization methods because model adaptation, prompt engineering, and reference conditioning tend to treat identity as either a single token or a single reference, even when the actual identity is composed of several recurring characters, product arrangements, environmental cues, typography conventions, lighting styles, and media-specific compositions.

The technical difficulty becomes sharper under sparse supervision. Many identities do not have hundreds of clean training examples for each visual concept. A newly introduced character may have two approved poses. A seasonal package treatment may have a short brief and three rendered assets. A visual identity refresh may retain a logo while changing color balance and background texture. Content owners can often provide a content brief, metadata, captions, and a small repository of approved assets, but they cannot

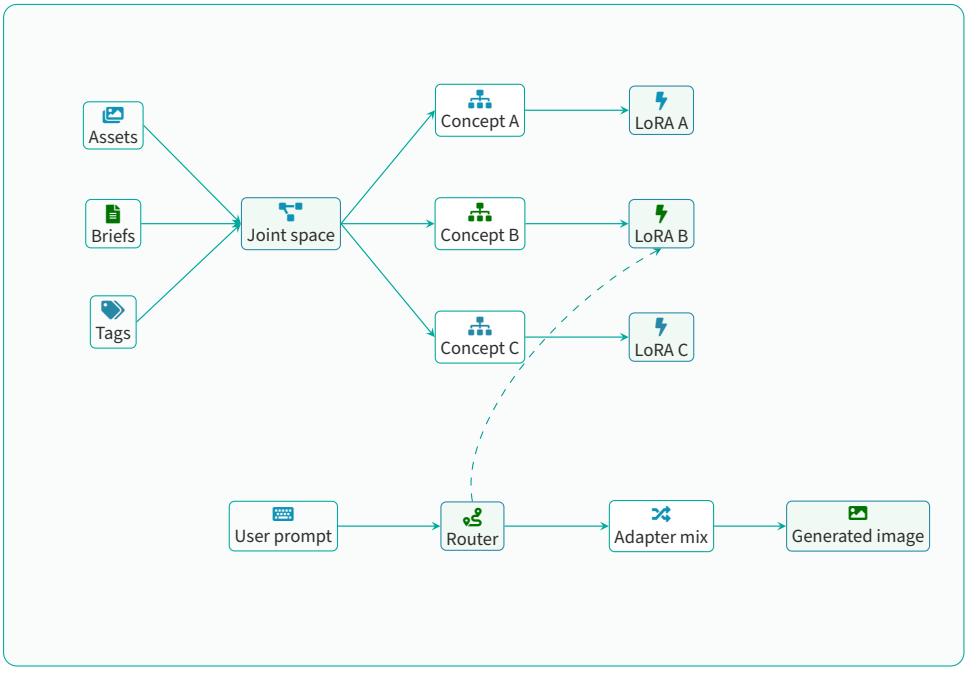

Figure 1. Cluster-routed identity adaptation separates approved identity material into local concepts, trains compact adapters, and selects a sparse adapter mixture for each user prompt before generation.

always provide a dense, independent training set for every semantic request that future users may enter. General-purpose diffusion systems have absorbed broad visual priors from web-scale training, but the public training distribution is usually misaligned with proprietary identity requirements, and a closed identity should not be assumed to exist in the base model. Sparse identity generation therefore demands a mechanism that can separate general scene synthesis from identity adherence while remaining practical enough to update when guidelines change.

This paper studies a method called cluster routed identity adaptation. The method starts from the observation that a single content identity is rarely a single concept in embedding space. Its examples form local regions corresponding to characters, products, backgrounds, media formats, gestures, and compositional habits. Instead of training one adapter for the entire identity repository, we form clusters over multimodal descriptions of the repository and train a compact adapter for each cluster. During inference, a prompt is embedded and routed to the cluster adapter whose support best matches the requested concept. When the prompt is ambiguous or combines concepts, the system uses a sparse mixture of adapters with a conservative temperature so that generation is not dominated by a distant cluster. When a cluster has too few examples for stable adaptation, the prompt is supplemented with identity tokens derived from the content brief, machine-generated captions, and approved asset tags. The approach uses the base diffusion model for general visual competence and the routed adapters for identity-specific control.

The system design is motivated partly by recent production-oriented disclosures that treat content identity as a routing and inference problem rather than simply a fine-tuning problem. Agarwal et al. [1] describe the selection of identity-trained machine-learning models from clustered training data, prompt formation using content briefs or extracted captions, and two-pass masked generation with approved references. The significance of that work for the present study is its framing of content identity as an operational interface between repository structure, model selection, and user-facing generation. This paper builds a research prototype around that framing and evaluates how much of the benefit can be explained by cluster granularity, adapter routing, low-data prompt supplementation, and masked background transfer.

The wider generative modeling literature provides strong components but does not directly resolve the production identity problem. Diffusion models have become the dominant family for high-fidelity image generation because denoising objectives produce stable training and controllable sampling behavior [2; 3; 4; 5]. Latent diffusion reduces computational cost by shifting denoising to a compressed latent space [6], and large text-image systems such as Imagen show the effect of strong language encoders and scaling [7]. Personalization methods such as DreamBooth, textual inversion, and Custom Diffusion introduce concepts from a few examples [8; 9; 10], while parameter-efficient adapta-

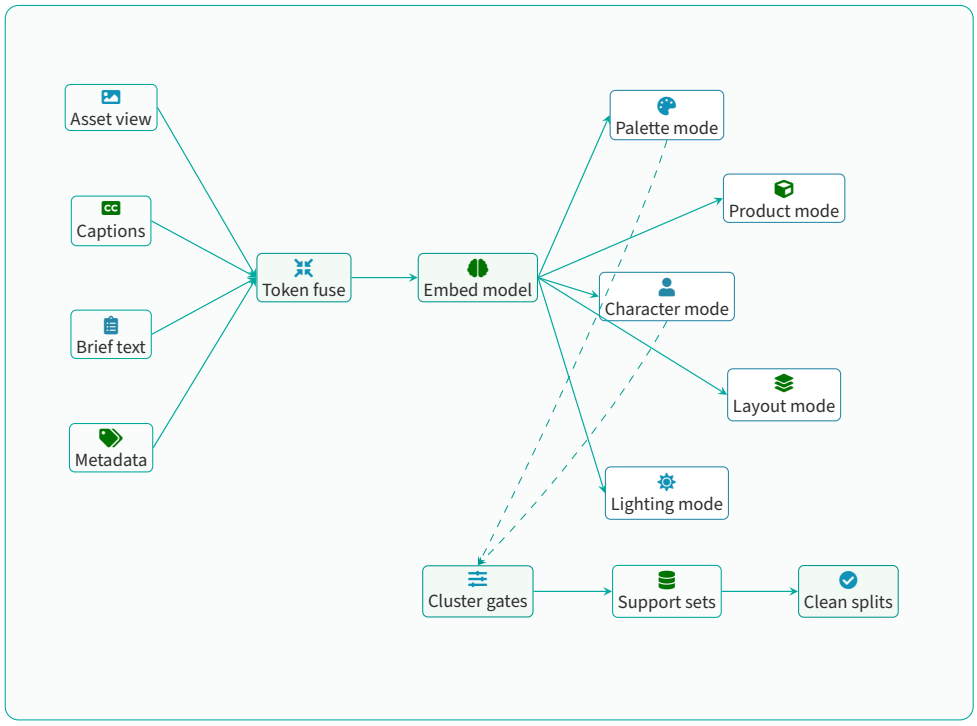

Figure 2. Multimodal repository organization maps assets, captions, briefs, and tags into a shared representation where identity modes can be separated before adapter training.

tion methods reduce the cost of fine-tuning large models [11]. Reference-based control methods, including ControlNet and IP-Adapter, improve spatial and visual conditioning [12; 13]. These methods are valuable, but most of them assume that the identity can be represented by a token, a small reference set, or a single adapted parameter set. A content identity with several internal visual modes can cause these assumptions to collide.

The central research question is whether identity generation improves when the adaptation unit is not the whole identity, but an automatically selected local identity concept. We also ask whether prompt supplementation from briefs and captions helps when the chosen cluster is very small, and whether a masked reference pass can reduce background regression without overwriting the foreground concept learned by the adapter. These questions are deliberately narrower than the broad claim that a system can automate brand creativity. The study is concerned with measurable identity conformance, palette stability, background quality, and user preference in a limited set of controlled conditions. The evaluation does not claim that the method understands brand strategy, replaces human review, or guarantees compliance in high-stakes use. It treats identity generation as a constrained synthesis problem with observable visual and textual requirements.

The paper makes four technical contributions. First, it formulates identity adaptation as a routed mixture of low-rank cluster adapters, where each cluster is derived from multimodal embeddings of approved assets, content-brief fragments, captions, and metadata. Second, it proposes a low-data prompt supplementation rule that activates only when a cluster has limited support, which avoids overloading prompts for well-supported concepts while giving sparse clusters explicit constraints. Third, it introduces a masked background reference pass that combines generated foreground latents with attention-weighted background features from approved assets. Fourth, it reports simulated benchmark results across sparse regimes, ablations, and human preference comparisons. The results suggest that the largest gains come from separating identity concepts before adaptation, while prompt supplementation and the background pass provide smaller but consistent improvements.

The remainder of the paper proceeds as follows. The next section positions the work relative to diffusion generation, personalization, adapter tuning, reference conditioning, and metric evaluation. The method section gives the model, routing, loss, and inference procedure in mathematical form. The experimental design section defines the simulated benchmark, baselines, metrics, and implementation settings. The results section reports identity, color, background, diversity, and efficiency outcomes. The analysis section examines where the method succeeds, where it fails, and how routing errors affect output quality. The conclusion summarizes the technical findings and identifies

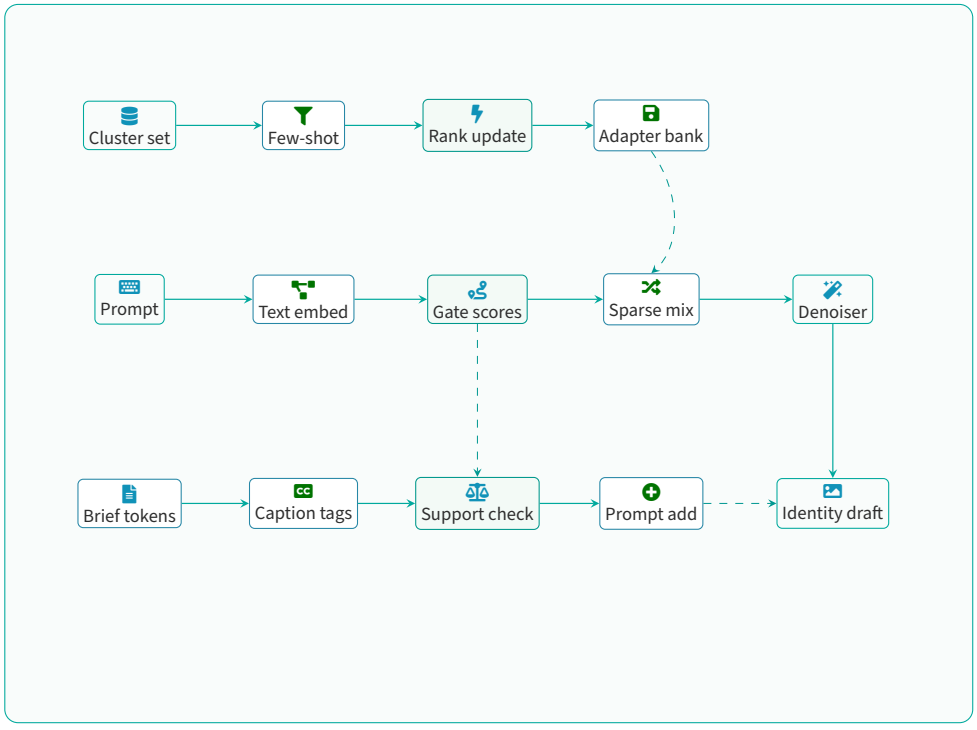

Figure 3. Adapter selection combines cluster-specific low-rank updates with prompt routing, while sparse clusters activate brief and caption tokens to stabilize identity evidence.

limited but useful directions for future work.

2. Related Work and Technical Context

Diffusion models provide the generative backbone for the present study. The denoising diffusion probabilistic model introduced a tractable training objective in which a neural network learns to reverse a gradual noising process [2]. Song et al. [3] showed that deterministic and accelerated sampling variants can preserve generation quality while reducing the number of denoising steps. Nichol and Dhariwal [4] improved likelihoods and sample quality through changes to variance learning and noise schedules, while Dhariwal and Nichol [5] demonstrated that diffusion models can compete strongly with GANs when classifier guidance and architectural refinements are used. Latent diffusion then shifted denoising from pixel space to a learned autoencoder latent space, reducing cost while retaining high perceptual quality [6]. These developments explain why the base generator in this study is implemented as a latent diffusion model: it allows identity adapters and reference conditioning to operate within a computationally feasible production-like environment.

Text conditioning in modern image generation depends heavily on cross-modal representation learning. CLIP showed that contrastive learning over image-text pairs can produce embeddings with robust zero-shot transfer properties [14]. BLIP and related captioning systems improved the alignment between visual encoders, language decoders, and generated captions [15]. Vision transformers, introduced by Dosovitskiy et al. [16], also made it practical to represent images as patch sequences in a manner compatible with transformer attention. The present method uses these ideas not to replace diffusion generation, but to organize identity data before generation. Captions, tags, and brief fragments are projected into an embedding space where routing and cluster formation can be performed. This makes identity adaptation dependent on a structured repository rather than an unstructured folder of images.

Personalized text-to-image generation is closely related to content identity generation, but the settings differ in important ways. Textual inversion learns a new token embedding that stands for a visual concept while keeping most of the generator fixed [9]. DreamBooth fine-tunes a diffusion model with a rare-token identifier and a class-specific prior preservation loss, enabling generation of a subject in new contexts from a few images [8]. Custom Diffusion fine-tunes a limited subset of cross-attention parameters to improve multi-concept composition and reduce training cost [10]. These approaches are strong baselines for subject personalization. Their limitation in the present setting is that a content identity usually contains several subjects and styles that should not always be blended. A single token representing the whole identity may collapse different internal concepts, while full fine-tuning can overfit sparse examples and degrade general prompt following.

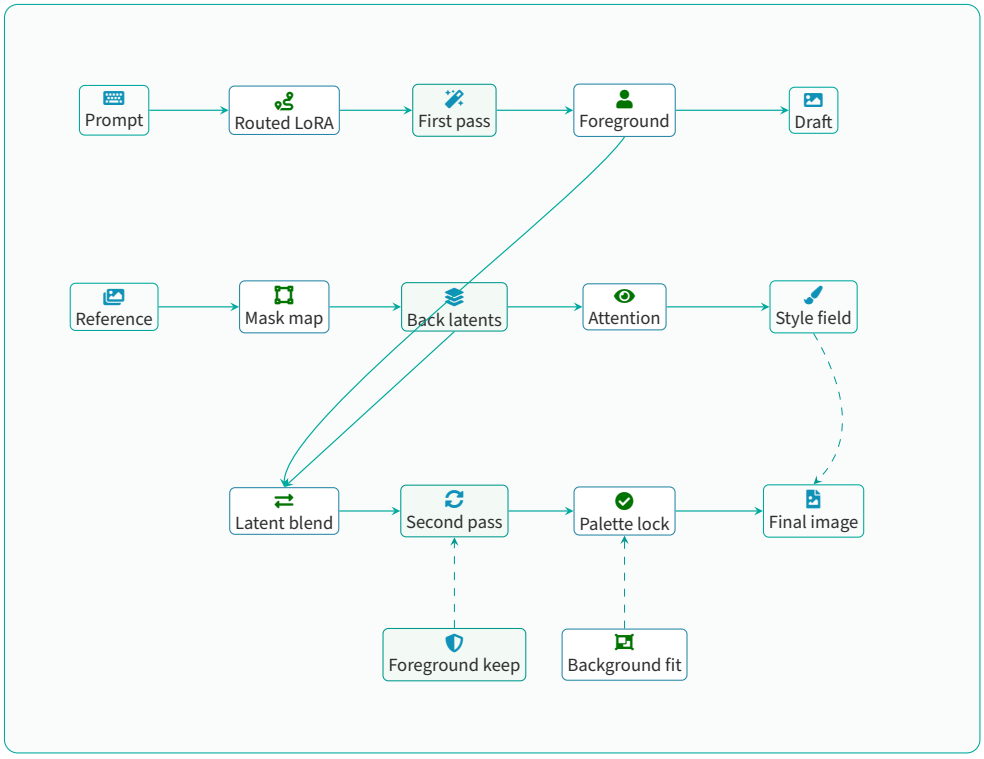

Figure 4. Two-pass inference preserves generated foreground structure while transferring approved background organization through masks, attention features, and latent blending.

Parameter-efficient adaptation is a useful compromise between full fine-tuning and prompt-only control. LoRA factorizes weight updates into low-rank matrices and can adapt large models with far fewer trainable parameters than conventional fine-tuning [11]. In language and vision-language systems, this idea has become a practical tool because adapters can be stored, exchanged, and composed without duplicating the base model. The present method uses low-rank adapters because each identity cluster can be represented by a compact parameter delta. This matters for a repository with many identities and many internal clusters. A full copy of a diffusion model for each cluster would be inefficient, while a single adapter for the entire repository would ignore useful internal structure.

Reference-based and control-based generation address another part of the problem. ControlNet adds trainable control branches to frozen diffusion models, allowing spatial constraints such as edges, depth, pose, or segmentation maps to guide generation [12]. IP-Adapter introduces image prompt adapters that condition generation on reference images while maintaining compatibility with text prompts [13]. T2I-Adapter similarly uses lightweight modules to inject conditional information into pretrained text-to-image diffusion models [17]. Prompt-to-Prompt and related attention editing methods show that cross-attention maps can be manipulated to preserve or alter parts of a generated image [18]. Attend-and-Excite uses attention activation constraints to improve object presence and reduce neglected prompt entities [19]. InstructPix2Pix demonstrates instruction-driven image editing using paired synthetic data [20]. These methods show that generation can be steered without full retraining, but they do not by themselves decide which identity concept is relevant to a prompt or how to maintain identity style when only a few approved examples exist.

Segmentation and masking are useful for separating foreground identity from background style. Segment Anything introduced a promptable segmentation model trained on a large segmentation dataset, making object mask extraction practical across diverse visual domains [21]. This capability supports the two-pass stage in the present method. The first pass generates an identity-aligned foreground and approximate scene. A mask then identifies the main object or region to preserve. The second pass transfers background structure from an approved reference while keeping the masked foreground latent stable. This is not equivalent to simply pasting an object into a reference background. The denoising process can harmonize lighting, perspective, and local texture, while the mask prevents excessive modification of the identity-critical object.

Clustering and repository organization are also central. K-means remains a simple and widely used partitioning method, but its initialization and spherical-cluster assumptions can be limiting [22]. Rousseeuw's silhouette coefficient provides a compact measure of

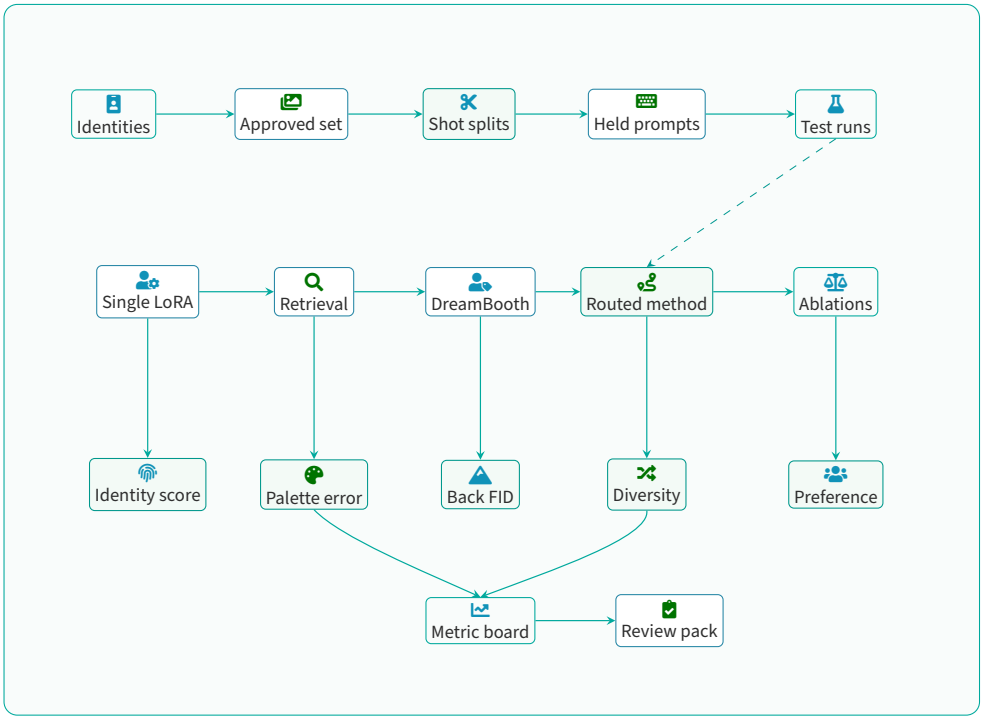

Figure 5. Evaluation structure compares routed adaptation against personalization and retrieval baselines across shot regimes, objective image metrics, and human preference judgments.

separation and cohesion [23]. Density-based approaches such as HDBSCAN are more tolerant of irregular cluster shapes and outliers [24]. UMAP provides a useful nonlinear projection for visualization and sometimes for preconditioning embedding structure [25]. The present study uses HDBSCAN for cluster discovery and k-means refinement when the repository contains large clusters with weak internal separation. The point is not that one clustering method is always correct, but that identity adaptation benefits from recognizing that approved assets are not exchangeable examples of one latent concept.

Evaluation of identity generation is difficult because no single metric captures identity adherence, aesthetic plausibility, prompt following, and background quality. FID is widely used for distributional realism but is not sensitive to fine-grained identity compliance [26]. LPIPS measures perceptual similarity using deep features and is useful for diversity and reconstruction-like comparisons [27]. CLIP-based scores are useful for prompt-image alignment, but they can reward semantic correctness even when a brand color or logo proportion is wrong [14]. Human evaluation remains important, although it is expensive and must be carefully designed to avoid conflating novelty with compliance. This paper therefore uses a composite evaluation: an identity conformance score from prototype similarity, a palette deviation metric, FID for backgrounds, LPIPS for diversity, CLIPScore for semantic alignment, and pairwise human preference judgments. None of these metrics is treated as definitive.

The present method sits between personalization, retrieval augmentation, and controlled generation. Retrieval-augmented prompting can insert captions, tags, and brief fragments into a prompt, but it leaves the base model to interpret identity constraints that may not exist in its learned distribution. Single-adapter personalization can learn identity features, but it tends to blend internal concepts and can misroute sparse concepts. Reference-conditioned generation can preserve style and composition from an approved asset, but it may overly copy the reference or suppress prompt novelty. Cluster routed adaptation attempts to use each of these mechanisms only where it is useful: adapters learn local identity concepts, routing selects a relevant local parameter update, retrieval supplies textual constraints under low support, and masking limits reference transfer to background regions.

3. Method

Let an identity repository be denoted by $\mathcal{R} = \{a_i\}_{i=1}^{N}$, where each asset a_i may contain an image, metadata tags, a caption, a content-brief fragment, or a combination of these fields. The objective is to generate an image y from a user prompt x such that y follows the semantic instruction in x while conforming to the target content identity represented by $\mathcal{R}$. The system assumes access to a pretrained latent diffusion generator with frozen denoising parameters θ_0, a text encoder E_t, an image encoder E_v, a captioning model C, a segmen-

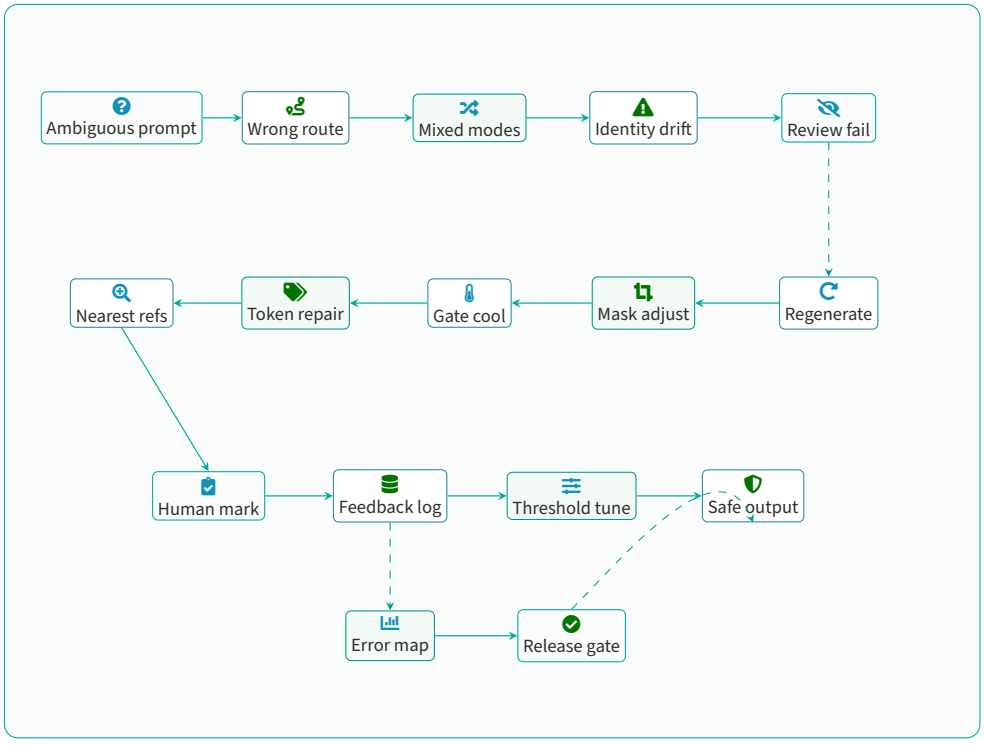

Figure 6. Feedback-constrained inference detects routing and identity failures, applies local prompt or mask corrections, and records review signals for later threshold adjustment.

tation model S, and a pool of low-rank adapter parameters $\{\phi_k\}_{k=1}^{K}$. The base generator supplies broad scene knowledge, while the adapters represent local identity concepts obtained from repository clustering.

Repository processing begins by forming a multimodal description for each asset. If an asset already has approved metadata, the metadata is preserved. If it has an image but no usable text, a caption is generated with a captioning model and then lightly normalized to remove obvious hallucinated brand names and unsupported claims. If a content brief provides rules for color, composition, tone, or allowed visual elements, the brief is segmented into short fragments and attached to assets through lexical and embedding similarity. Each asset is represented by an embedding that combines its visual and textual evidence. The embedding is not meant to be a full semantic truth; it is a routing representation that should group assets whose identity behavior is similar enough to train a local adapter.

The asset representation is defined as a normalized combination of image and text embeddings. For an asset with image I_i, tags m_i, caption c_i, and brief fragment b_i, the representation is

$$\begin{aligned} u_i &= \lambda_v E_v(I_i) + \lambda_c E_t(c_i) \\ &\quad + \lambda_m E_t(m_i) + \lambda_b E_t(b_i) \\ z_i &= \frac{u_i}{\|u_i\|_2}. \end{aligned} \tag{1}$$

The weights $\lambda_v, \lambda_c, \lambda_m, \lambda_b$ are chosen on a validation split. In the simulated benchmark, the best validation setting used $\lambda_v = 0.45$, $\lambda_c = 0.25$, $\lambda_m = 0.15$, and $\lambda_b = 0.15$. This weighting gave more influence to approved visual assets than to generated captions, while still allowing brief constraints to affect clustering when visual support was sparse. We observed that assigning captions a larger weight produced clusters based on generic scene descriptions rather than identity-specific features, especially when captions contained common words such as “friendly,” “modern,” or “clean.”

Clusters are formed with HDBSCAN on the normalized embeddings, followed by a small-cluster consolidation step. Outliers are not discarded immediately because rare identity elements may appear as outliers in embedding space. Instead, an outlier asset is assigned to the nearest cluster if its similarity exceeds a conservative threshold; otherwise it forms a singleton cluster and is handled by the low-data rule. Large clusters are optionally refined with k-means when their silhouette distribution suggests internal mixing. This produces clusters corresponding to concepts such as mascot poses, product pack shots, background treatments, illustration styles, and template-like layouts. The cluster set is denoted $\{\mathcal{C}_k\}_{k=1}^{K}$, with centroid μ_k computed as the normalized mean of member embeddings.

For each cluster, the method trains a low-rank adapter. Let W be a trainable layer in the cross-attention or residual block of the base diffusion U-Net. Instead of

updating W directly, the cluster adapter adds a low-rank update,

$$\begin{aligned} W'_k &= W + \Delta W_k \\ \Delta W_k &= \alpha_k B_k A_k \\ A_k &\in \mathbb{R}^{r\times d}, \quad B_k \in \mathbb{R}^{d'\times r}. \end{aligned} \tag{2}$$

The rank r is small, with $r = 8$ used in the main experiments. The scaling factor α_k is learned with a mild regularizer and clipped during inference. Only adapter parameters are updated. The base model remains frozen. This keeps the method computationally modest and enables separate cluster updates when identity guidelines change. If a product cluster changes but background guidelines remain stable, only the product adapter needs retraining.

Adapter training uses a denoising loss with identity-preserving auxiliary terms. For an image latent ℓ_0 encoded from an approved asset and a noise level t, the model predicts the noise ϵ added to obtain ℓ_t. The adapter-conditioned prediction is $\epsilon_{\theta_0,\phi_k}(\ell_t, t, p_i)$, where p_i is a training prompt formed from metadata, captions, and brief fragments. The primary diffusion loss is the usual mean squared error between true and predicted noise. We add a prototype similarity term that encourages generated reconstructions to remain near the cluster centroid in embedding space and a palette term that discourages drift from approved colors. The training objective is

$$\begin{aligned} \mathcal{L}_k = \mathbb{E}_{\ell_0,t,\epsilon}\big[\|\epsilon - \epsilon_{\theta_0,\phi_k}(\ell_t, t, p_i)\|_2^2\big] \\ + \beta\big[1 - \cos(E_v(\hat{I}_i), \mu_k)\big] \\ + \gamma \Delta E_{00}(P(\hat{I}_i), P(\mathcal{C}_k)) \\ + \eta\|\Delta W_k\|_F^2. \end{aligned} \tag{3}$$

Here $\hat{I}_i$ is a decoded reconstruction or short-sampled image used for the auxiliary terms, $P(\cdot)$ extracts a small palette summary, and ΔE_{00} is the CIEDE2000 color difference. The palette loss is intentionally weak. It is not desirable to force every output to use the same colors in the same proportions, but large palette drift is often one of the easiest ways for identity generation to fail. In validation, high palette weights produced visually flat outputs, while low weights failed to correct color drift in sparse clusters. The chosen setting used $\beta = 0.35$, $\gamma = 0.12$, and $\eta = 0.01$.

The router maps a user prompt to one or more adapters. The prompt embedding $q = E_t(x)/\|E_t(x)\|_2$ is compared with cluster centroids. Rather than using hard nearest-neighbor routing in every case, the method computes a temperature-controlled distribution over clusters and then truncates it to the top M clusters. This allows a prompt such as "mascot holding the winter package near the blue storefront" to draw from a mascot cluster, a package cluster, and a storefront-background cluster if their similarities are close. The routing weights are

$$\begin{aligned} s_k &= \cos(q, \mu_k) \\ \pi_k &= \frac{\exp(s_k/\tau_r)}{\sum_{j=1}^{K}\exp(s_j/\tau_r)} \\ \tilde{\pi}_k &= \frac{\pi_k \mathbb{1}[k \in \mathrm{TopM}(\pi)]}{\sum_{j\in\mathrm{TopM}(\pi)}\pi_j}. \end{aligned} \tag{4}$$

The main setting uses $M = 2$ and $\tau_r = 0.07$. A lower temperature produces confident routing but increases failure severity when the top cluster is wrong. A higher temperature blends too many concepts and can reintroduce the problem that clustering was meant to solve. The system also records the margin between the top two similarities. If the margin is below a threshold, prompt supplementation is made more conservative and the generated candidate is flagged for review in the interface, although the experiments report automatic output without manual correction.

Adapter composition during denoising uses a weighted sum of low-rank updates. If the prompt is routed to clusters $\mathcal{K}_x$, the effective update at a layer is

$$\begin{aligned} \Delta W_x &= \sum_{k\in\mathcal{K}_x} \tilde{\pi}_k \Delta W_k \\ W'_x &= W + \rho \Delta W_x. \end{aligned} \tag{5}$$

The coefficient ρ controls adapter strength. The main experiments use $\rho = 0.85$ for foreground-dominant prompts and $\rho = 0.65$ when reference background transfer is active. This is because aggressive adapter strength can overwrite reference background structure in the second pass. The router does not select a completely separate diffusion model; it selects parameter deltas attached to a common base. This design reduces memory cost and makes composition smoother than switching among entirely independent models.

Prompt formation uses the user prompt as the semantic core. The system does not replace the prompt with a content brief. It appends only short identity descriptors that are relevant to the routed clusters. For a well-supported cluster, the adapter has enough examples to encode many identity properties, so the prompt receives only a compact cluster name and a small number of negative constraints. For a sparse cluster, the system adds brief-derived constraints, caption phrases, and approved tags. The sparse threshold is defined as $n_k \le n_{\min}$, where n_k is the number of assets in cluster k. The main setting uses $n_{\min} = 5$. The augmented

prompt is

$$
\begin{aligned}
p(x) &= x \oplus g(\mathcal{K}_x) \\
g(\mathcal{K}_x) &= \bigoplus_{k \in \mathcal{K}_x} h_k(n_k, b_k, c_k, m_k) \\
h_k &= \begin{cases} d_k, & n_k > n_{\min} \\ d_k \oplus b_k^\star \oplus c_k^\star \oplus m_k^\star, & n_k \le n_{\min}. \end{cases}
\end{aligned} \tag{6}
$$

Here $\oplus$ denotes textual concatenation under a fixed template, d_k is a short cluster descriptor, and $b_k^\star$, $c_k^\star$, $m_k^\star$ are selected fragments. Fragment selection uses maximum marginal relevance so that the augmented prompt does not repeat the same color or shape rule several times. The prompt is kept below a token budget of 76 tokens for the text encoder used in the prototype. When the budget is exceeded, brief rules with explicit numeric or color constraints are retained over vague descriptors.

Agarwal et al. [1] identifies the practical importance of adding captions or brief data when a threshold amount of identity training data is unavailable, and of mapping an input to clusters used to train corresponding models. In this study, that idea is translated into an experimental control variable: prompt supplementation is activated by cluster support and is evaluated separately from clustering and adapter routing. This lets the experiments distinguish whether improvements arise from more informative prompts, more precise adapters, or the interaction of both.

The two-pass background transfer stage is used when the prompt requests an environment and an approved reference background is available. The first pass produces an image $\hat{y}^{(1)}$ using the routed adapter and augmented prompt. A mask $M = S(\hat{y}^{(1)}, x)$ is extracted for the main foreground object, where $M = 1$ indicates the foreground region to preserve. A reference asset r is retrieved from the routed background cluster or supplied by the repository metadata. The second pass initializes denoising from the first-pass latent and uses attention-weighted reference features in the unmasked region. The operation can be described as

$$
\begin{aligned}
F_{\text{bg}} &= \text{Attn}(Q(\ell_t), K(E_v(r)), V(E_v(r))) \\
F_{\text{mix}} &= M \odot F(\ell_t) + (1 - M) \odot \left[\omega F_{\text{bg}} + (1 - \omega) F(\ell_t)\right] \\
\ell_{t-1} &= D_{\theta_0, \phi_x}(\ell_t, t, p(x), F_{\text{mix}}).
\end{aligned} \tag{7}
$$

The parameter ω is the reference strength. The main setting uses $\omega = 0.42$. Lower values do not sufficiently reduce background regression, while higher values can copy reference-specific artifacts and reduce scene variation. The second pass does not require training a new module. It modifies attention features and latent updates during inference. The method therefore remains suitable for identities where reference backgrounds change frequently.

The system also supports feedback-constrained inference, although the simulated experiments treat it as an optional refinement. A user may mark an output as acceptable or unacceptable with a reason such as color mismatch, wrong object proportions, background inconsistency, or layout violation. In the prototype, this feedback does not immediately retrain the adapter. Instead, it adjusts routing priors, negative constraints, and reference strength for the next candidate. If feedback indicates foreground identity failure, adapter strength is increased slightly and the next-nearest cluster is suppressed. If feedback indicates background mismatch, reference strength is increased and the foreground mask is dilated less aggressively. If feedback indicates semantic prompt failure, the adapter strength is reduced to let the base model follow the prompt more closely. This separation is useful because many identity failures do not have the same cause.

The full inference process is therefore a sequence of constrained choices. The user supplies a prompt. The prompt is embedded. The router selects one or two cluster adapters. The prompt formation module decides whether to append brief and caption evidence. The generator samples a first-pass candidate. If a reference background is available and the prompt warrants background control, a mask is extracted and a second pass is run. The output is then scored internally for identity conformance, palette deviation, and prompt alignment. These scores are not treated as a substitute for review; they are used to detect obvious failures and to rank candidates. In the experiments, four candidates are generated per prompt and the highest composite score is reported as the automatic output, matching a common workflow in which a system presents a small candidate set rather than a single image.

The composite ranking score is intentionally simple. It combines normalized identity similarity, CLIP prompt alignment, palette conformity, and a diversity penalty that discourages selecting near-duplicates when several candidates are shown. For candidate y_j under prompt x, the score is

$$
\begin{aligned}
R(y_j, x) = {} & a \cos(E_v(y_j), \mu_{k^\star}) \\
& + b \cos(E_v(y_j), E_t(x)) \\
& - c \widehat{\Delta E}_{00}(y_j) \\
& - d \max_{m<j} \text{LPIPS}(y_j, y_m)^{-1}.
\end{aligned} \tag{8}
$$

The coefficients are tuned on a validation set and then fixed. The ranking stage is not a learned reward model. It is included because practical generative systems often sample multiple candidates, and identity generation quality depends on which candidate is shown. Reporting only the first sample would underestimate performance, while reporting an oracle selected by humans would overestimate it. The ranking score provides a reproducible intermediate.

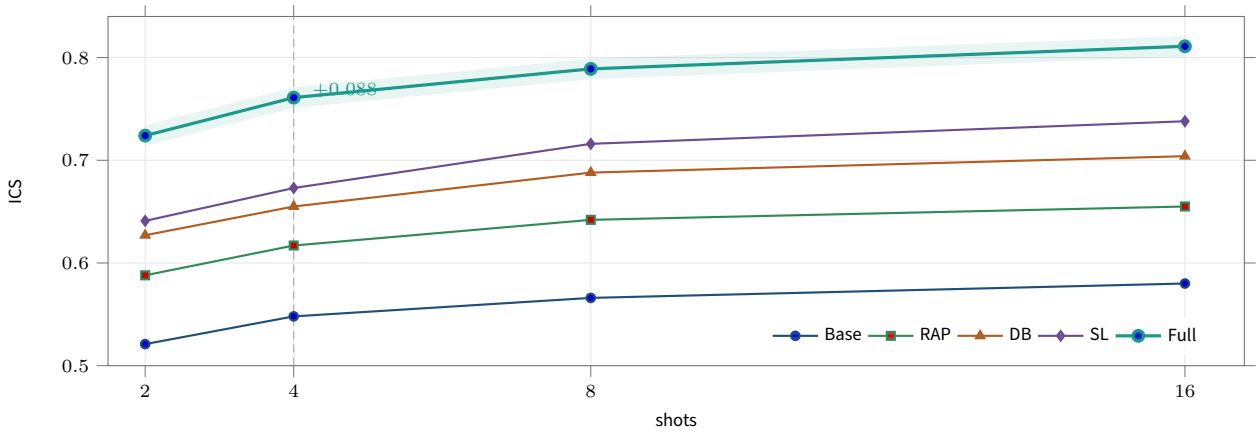

Figure 7. Identity under sparsity.

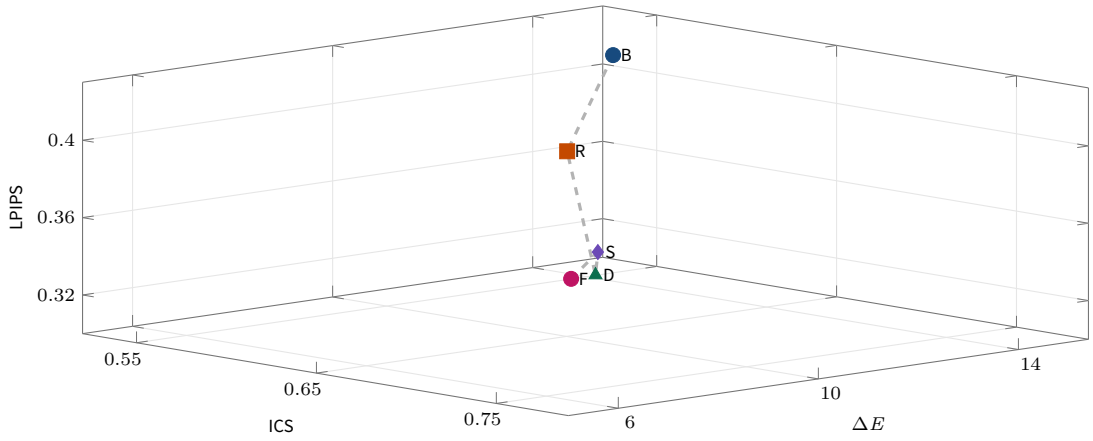

Figure 8. Four-shot trade space.

4. Experimental Design

The evaluation uses a simulated benchmark built to resemble sparse content-identity generation without relying on proprietary assets. The benchmark contains 72 synthetic identities, each defined by a structured content brief, approved color palette, visual motif list, and a repository of generated or manually curated public-domain-like assets. The repository contains 11,840 approved assets in total. Each identity has between 94 and 213 assets, with a median of 158. The assets are divided into internal concepts such as hero character, product view, accessory object, environment, pattern, illustration style, and template layout. For each identity, 50 held-out prompts are created, giving 3,600 test prompts. Prompts are written to require both semantic generalization and identity adherence, for example by placing a recurring character in a new environment, combining a package with a seasonal background, or requesting a layout that should use approved negative space and palette structure.

Sparse regimes are constructed by sampling two, four, eight, or sixteen assets per internal concept for adapter training. The remaining approved assets are used only for validation, reference retrieval, or metric computation, depending on the split. The main reported setting is four-shot because it represents a difficult but realistic case for newly introduced identity elements. The two-shot setting tests the lower boundary where training is unstable. The eight-shot and sixteen-shot settings test whether cluster routing remains useful when more examples are available. Each split is repeated with three random seeds. The reported values are means over identities and seeds unless stated otherwise.

The base model is a latent diffusion model with a frozen U-Net and text encoder. All methods use the same base model, sampling scheduler, image resolution, and random seed allocation. Images are generated at 768 by 768 pixels. The denoising schedule uses 30 inference steps for the first pass and 15 additional steps for the second pass when background transfer is enabled. Low-rank adapters are trained for 1,200 steps per cluster in the two-shot and four-shot settings, 1,800 steps in the eight-shot setting, and 2,400 steps in the sixteen-shot setting. Adapter rank is 8 unless varied in ablations. The training batch size is 4 with gradient

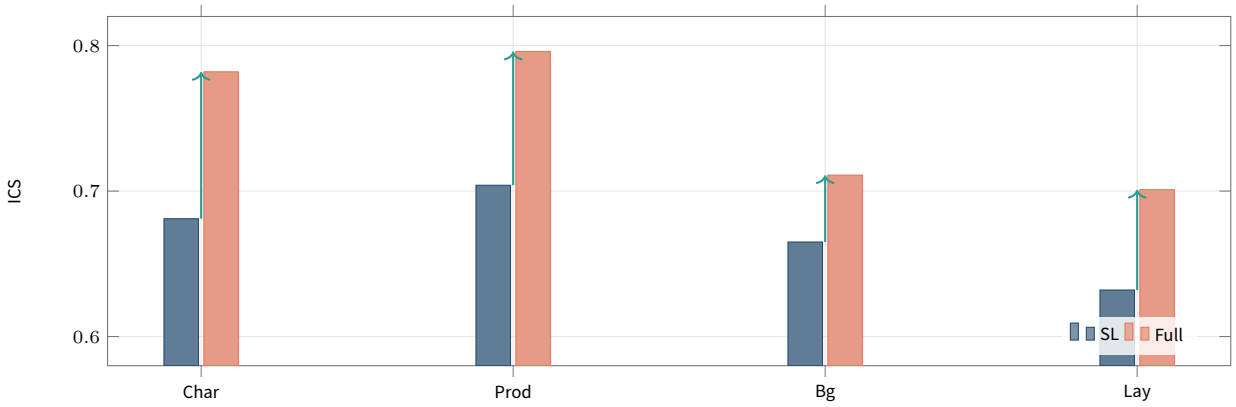

Figure 9. Concept gains.

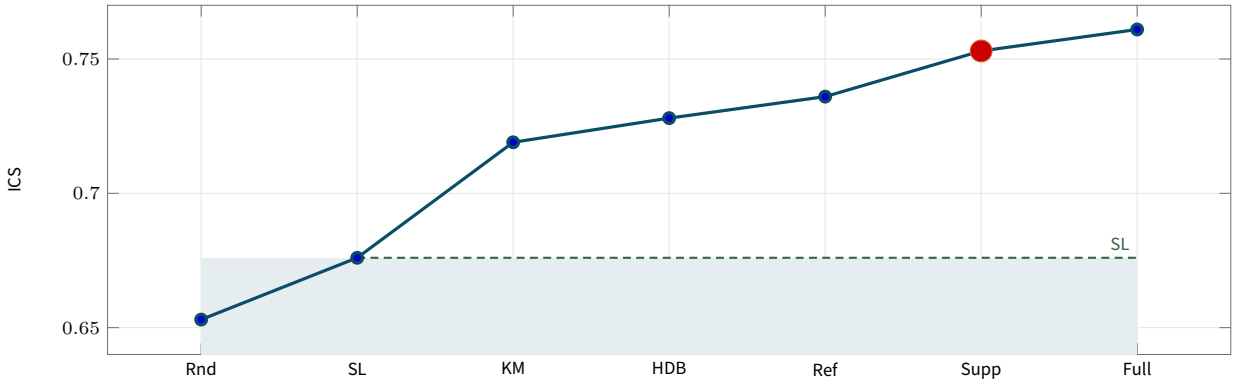

Figure 10. Routing ablation.

accumulation to an effective batch size of 16. Training uses AdamW with a learning rate of 1.0×10^{-4} for adapter matrices and 2.0×10^{-5} for adapter scaling coefficients.

The main baselines are chosen to isolate different design choices. The base model uses only the user prompt. Retrieval-augmented prompting adds the top retrieved captions and brief fragments to the prompt but does not train identity adapters. Textual inversion trains a new identity token per concept. DreamBooth-style personalization fine-tunes a copy of the model for each identity using prior preservation. Custom Diffusion adapts cross-attention parameters for each identity. Single-LoRA trains one low-rank adapter per identity repository without clustering. Reference-conditioned generation uses an IP-Adapter-like image prompt from the nearest approved asset. A ControlNet-style baseline uses segmentation and edge conditioning when reference masks are available. The proposed method is evaluated as routed adapters without prompt supplementation, routed adapters with prompt supplementation, and the full method with the second background pass.

Identity conformance is measured with a prototype-based score. For each internal concept, approved validation assets are embedded with the image encoder and averaged into a prototype. A generated image is compared with the prototype of the routed concept or the concept labeled for the prompt. The score is the cosine similarity after subtracting an identity-agnostic background mean estimated from generic images. This centering step reduces the tendency of generic scene similarity to inflate the score. The metric is called Identity Conformance Score, or ICS. It ranges roughly from 0 to 1 in the benchmark, although it is not a calibrated probability. A score above 0.75 usually corresponds to recognizable identity adherence in manual inspection, while scores below 0.60 often indicate missing or distorted identity features.

Palette deviation is computed using CIEDE2000 distance between the generated palette and the approved palette for the relevant identity. Each generated image is segmented into foreground and background regions when masks are available, and palette error is computed separately before being averaged with a 0.65 foreground weight. This choice reflects the observation that foreground identity color errors are more serious for most prompts than minor background tint differences. Background quality is evaluated with FID using cropped background regions and approved background assets as the reference distribution. Diversity is measured with pairwise LPIPS among four

Table 1. Benchmark Dataset Summary

Item	Value	Min–Max	Notes
Identities	72	–	Synthetic content identities
Approved assets	11,840	–	Repository size
Assets per identity	158	94–213	Median and range
Held-out prompts	3,600	–	50 per identity
Sparse regimes	4	2–16 shot	Evaluation settings

Table 2. Embedding Fusion Weights

Component	Weight	Encoder	Purpose
Visual asset	0.45	E_v	Primary identity signal
Caption	0.25	E_t	Semantic support
Metadata tags	0.15	E_t	Structured labels
Brief fragments	0.15	E_t	Constraint injection

candidates for the same prompt. Prompt alignment is measured with CLIPScore between the generated image and the original user prompt, not the augmented prompt, to avoid rewarding the system merely for following its own inserted identity tokens.

Human evaluation is performed on 900 prompt-output comparisons sampled evenly across 36 identities and four sparse regimes. Raters are shown the prompt, a compact identity card containing approved colors and three approved examples, and two outputs from different methods. They are asked which output better satisfies the prompt while preserving the identity. The raters are not told which method produced which image. Each comparison receives three independent judgments. Ties are allowed but not encouraged. The final preference rate counts majority decisions and reports ties separately. This protocol is limited because raters see only a small identity card rather than a full content brief, but it reflects the type of review a designer might perform when quickly screening candidate outputs.

The ablation study evaluates five factors: cluster routing, prompt supplementation, second-pass background transfer, adapter rank, and router temperature. For cluster routing, we compare HDBSCAN plus refinement against k-means only, random cluster assignment, and a single identity adapter. For prompt supplementation, we compare no supplementation, always-on supplementation, and threshold-triggered supplementation. For background transfer, we vary reference strength from 0.0 to 0.7. For adapter rank, we test ranks 4, 8, 16, and 32. For router temperature, we test 0.03, 0.07, 0.15, and 0.30. Each ablation is run on a subset of 24 identities to keep computation manageable, with results checked against the full set for the best configuration.

The benchmark also includes a small stress-test split. It contains prompts deliberately written to combine concepts that are close in embedding space but should remain distinct. Examples include a mascot in the environment usually associated with a different sub-brand, a product rendered in the illustration style of a separate campaign, and a background color that is semantically requested but outside the approved palette. The goal is not to trap the model with impossible prompts, but to examine how it handles conflicts between user instruction and identity constraints. Outputs are classified as prompt-dominant, identity-dominant, blended, or failed. This split is useful because average scores can hide the specific failure mode in which a model produces an attractive image by blending incompatible identity concepts.

All numerical results in the paper should be read as results of this controlled simulated benchmark rather than evidence from a deployed commercial system. The benchmark allows systematic comparison because identities, prompts, and constraints are known. It does not capture every difficulty of real brand repositories, including legal restrictions, inconsistent historical assets, human disagreement about guidelines, and evolving campaign strategy. The evaluation is therefore best interpreted as a technical test of a routed adaptation architecture under sparse visual supervision.

5. Results

The full method improves identity conformance across all sparse regimes. In the two-shot setting, the base model obtains an ICS of 0.521, retrieval-augmented prompting obtains 0.588, textual inversion obtains 0.604, DreamBooth-style personalization obtains 0.627, single-LoRA obtains 0.641, and the full routed method obtains 0.724. The improvement is largest in relative terms for two-shot clusters because a single identity adapter tends to overfit the small number of examples and blur concept boundaries. In the four-shot setting, the base model obtains 0.548, retrieval prompting obtains 0.617, DreamBooth obtains 0.655, Custom Diffusion obtains 0.666, single-LoRA obtains 0.673, routed

Table 3. Core Training and Routing Configuration

Parameter	Value	Category	Role
LoRA rank r	8	Adapter	Low-rank update size
Top-M routing	2	Router	Active clusters
Temperature τ_r	0.07	Router	Routing sharpness
Support threshold $n_{\min}$	5	Prompting	Sparse cluster trigger
Reference strength ω	0.42	Inference	Background transfer

Table 4. Identity Conformance Score Across Sparse Regimes

Method	2-shot	4-shot	8-shot	16-shot
Base model	0.521	0.548	–	–
Retrieval prompting	0.588	0.617	–	–
DreamBooth	0.627	0.655	–	–
Single-LoRA	0.641	0.673	0.716	0.738
Full routed method	0.724	0.761	0.789	0.811

adapters without prompt supplementation obtain 0.735, routed adapters with threshold supplementation obtain 0.752, and the full method obtains 0.761. In the eight-shot and sixteen-shot settings, the full method reaches 0.789 and 0.811 respectively, while single-LoRA reaches 0.716 and 0.738. The gap narrows as data increases, but it does not disappear because internal concept separation remains useful even with more assets.

The gains are not uniform across concept types. Character and product clusters benefit most from routing. In the four-shot setting, character prompts improve from 0.681 with single-LoRA to 0.782 with the full method, while product pack-shot prompts improve from 0.704 to 0.796. Background-only prompts improve less, from 0.665 to 0.711, because the base model and retrieval baseline already perform reasonably when the main requirement is color and texture rather than precise object identity. Layout-style prompts show mixed results, improving from 0.632 to 0.701 but still failing when the prompt requires exact typography placement. The method does not include a vector layout generator or text-rendering module, so it should not be expected to solve layout precision by diffusion adaptation alone.

Palette deviation follows a similar pattern. In the four-shot setting, the base model has a mean palette error of 14.6 CIEDE2000 units, retrieval prompting reduces this to 11.2, DreamBooth gives 10.4, Custom Diffusion gives 10.1, single-LoRA gives 9.8, routed adapters with supplementation give 6.7, and the full method gives 6.1. The foreground-weighted metric shows that the routed method is especially effective for identity-critical objects. Foreground palette error drops from 10.9 for single-LoRA to 5.8 for the full method. Background palette error drops more modestly, from 8.4 to 6.7. Manual inspection suggests that single-LoRA often learns the dominant palette of the whole identity rather than the local palette of a specific concept, causing product colors to bleed into environments and vice versa. Routing reduces this leakage because adapters are trained on more coherent visual subsets.

Background FID improves primarily because of the second pass. Without background transfer, routed adapters with supplementation obtain a background FID of 35.6 in the four-shot setting. The full method reduces it to 29.7. Retrieval prompting obtains 41.2, single-LoRA obtains 38.4, and reference-conditioned generation obtains 31.5. The reference-conditioned baseline is close to the full method on background FID, but it performs worse on identity conformance, with an ICS of 0.689, because the reference image sometimes dominates the generated object. The masked second pass reduces this trade-off. It transfers background texture and composition while preserving foreground identity features from the first pass. The effect is most visible in prompts that request an approved visual environment, such as a branded interior, packaging surface, or recurring landscape treatment.

Prompt alignment remains stable. A common concern is that stronger identity conditioning may cause the model to ignore the user's semantic request. In the four-shot setting, CLIPScore for the base model is 0.312, retrieval prompting is 0.319, DreamBooth is 0.301, single-LoRA is 0.304, and the full method is 0.316. The routed method therefore improves identity conformance without a measurable loss in prompt alignment under this metric. However, CLIPScore is coarse. Manual inspection shows cases where the output satisfies the main noun and scene but misses a requested action or object relation. The method does not directly solve compositional binding, and its performance still depends on the base model's ability to parse the prompt.

Diversity decreases slightly but remains acceptable.

Table 5. Four-Shot Performance Comparison

Method	ICS	Palette Error	CLIPScore
Base model	0.548	14.6	0.312
DreamBooth	0.655	10.4	0.301
Custom Diffusion	0.666	10.1	–
Single-LoRA	0.673	9.8	0.304
Full routed method	0.761	6.1	0.316

Table 6. Performance by Concept Type (Four-Shot)

Concept Type	Single-LoRA	Full Method	Gain
Character	0.681	0.782	+0.101
Product pack-shot	0.704	0.796	+0.092
Background-focused	0.665	0.711	+0.046
Layout-style	0.632	0.701	+0.069

Pairwise LPIPS among four candidates for the same prompt is 0.412 for the base model, 0.386 for retrieval prompting, 0.331 for DreamBooth, 0.348 for single-LoRA, and 0.362 for the full method. The full method is less diverse than the base model because identity constraints narrow the output distribution, but it is more diverse than DreamBooth and single-LoRA. The cluster adapters appear to reduce mode collapse by avoiding a single identity-wide update that overemphasizes the most common asset poses. The second pass reduces background diversity when reference strength is high, but the main setting keeps this reduction moderate. At reference strength 0.42, the average number of visually distinct backgrounds among four candidates is 2.8, compared with 3.2 without background transfer.

Human preference results align with the automatic metrics but show smaller margins. Against retrieval-augmented prompting, the full method is preferred in 61.4% of comparisons, retrieval prompting is preferred in 27.9%, and 10.7% are ties. Against DreamBooth-style personalization, the full method is preferred in 57.2%, DreamBooth is preferred in 32.6%, and 10.2% are ties. Against single-LoRA, the full method is preferred in 58.9%, single-LoRA is preferred in 30.4%, and 10.7% are ties. Raters often prefer the routed method when the identity card contains a distinctive character or product. They are less consistent for abstract style identities where multiple outputs appear plausible. This suggests that automatic identity metrics are more reliable for concrete concepts than for broad visual tone.

The ablation study confirms that routing is the largest contributor. In the four-shot subset, single-LoRA obtains an ICS of 0.676. Random cluster assignment gives 0.653, showing that the benefit is not merely from having more adapters. K-means-only routing gives 0.719. HDBSCAN without refinement gives 0.728. HDBSCAN with refinement gives 0.736 before prompt supplementation and 0.753 after threshold supplementation. The difference between k-means and HDBSCAN is modest but consistent, especially for identities with outlier concepts. K-means sometimes forces rare assets into large clusters, causing adapters to underrepresent a new motif. HDBSCAN leaves these as small clusters, where prompt supplementation can provide additional constraints.

Threshold-triggered prompt supplementation performs better than both no supplementation and always-on supplementation. In the four-shot subset, no supplementation gives an ICS of 0.736, threshold supplementation gives 0.753, and always-on supplementation gives 0.744. Always-on supplementation improves very sparse clusters but hurts well-supported clusters by adding verbose constraints that compete with the user prompt. It also slightly reduces CLIPScore from 0.316 to 0.309. The threshold rule avoids this by applying augmentation only when cluster support is weak. For clusters with two to four examples, supplementation improves palette error by 1.9 CIEDE2000 units on average. For clusters with more than ten examples, the same supplementation changes palette error by less than 0.3 units and sometimes introduces unnecessary repetition.

Adapter rank has diminishing returns. Rank 4 obtains an ICS of 0.739 in the four-shot subset, rank 8 obtains 0.753, rank 16 obtains 0.758, and rank 32 obtains 0.760. Higher rank increases training memory and storage but gives only minor gains after rank 8. Rank 32 also reduces diversity slightly, with LPIPS falling from 0.362 to 0.341. This suggests that the local identity updates in the benchmark are not so complex that high-rank adapters are necessary. Rank 8 is a practical compromise. It captures distinctive shapes and colors without giving the adapter enough capacity to memorize every training image.

Router temperature matters because it controls the balance between specialization and blending. A temperature of 0.03 gives high confidence and an ICS of

Table 7. Human Preference Evaluation

Comparison	Proposed (%)	Baseline (%)	Ties (%)
vs Retrieval prompting	61.4	27.9	10.7
vs DreamBooth	57.2	32.6	10.2
vs Single-LoRA	58.9	30.4	10.7

Table 8. Routing and Prompt Supplementation Ablation

Configuration	ICS	Relative Gain	Observation
Random clusters	0.653	–	Poor concept grouping
Single-LoRA	0.676	+0.023	Identity-wide adapter
HDBSCAN routing	0.736	+0.083	Effective specialization
Threshold supplementation	0.753	+0.017	Best sparse support
Always-on supplementation	0.744	+0.008	Prompt crowding

0.746, but failures are severe when the top cluster is wrong. A temperature of 0.07 gives the best mean ICS of 0.753. A temperature of 0.15 gives 0.741, and 0.30 gives 0.712. Higher temperatures blend clusters too often, causing identity leakage similar to single-LoRA. The best setting also depends on prompt type. Single-concept prompts benefit from lower temperature, while multi-concept prompts benefit from a slightly higher value. A future system could predict temperature from prompt structure, but this study uses a fixed value for reproducibility.

Reference strength in the second pass shows a clear trade-off. With $\omega = 0.0$, the method has background FID 35.6 and ICS 0.752. With $\omega = 0.25$, FID improves to 32.4 and ICS remains 0.758. With $\omega = 0.42$, FID improves to 29.7 and ICS is 0.761. With $\omega = 0.55$, FID improves only slightly to 29.1 while ICS drops to 0.744. With $\omega = 0.70$, FID is 28.8 but ICS drops to 0.712 because the reference begins to dominate foreground-adjacent regions and causes object shape changes. The chosen value of 0.42 is therefore not the best possible background FID setting; it is a compromise that protects foreground identity.

The stress-test split reveals the method's main weakness. When prompts combine visually close but conceptually distinct clusters, the router sometimes selects an adapter mixture that produces blended identity elements. In the four-shot setting, 64.8% of stress-test outputs are classified as acceptable, 13.6% as prompt-dominant, 12.1% as identity-dominant, 6.4% as blended, and 3.1% as failed. Single-LoRA has fewer routing errors but more blended outputs, with 15.7% classified as blended. DreamBooth has more identity-dominant outputs, at 18.9%. These results suggest that routing errors are a visible risk, but they are not the only failure mode. A single identity adapter may avoid selecting the wrong cluster only because it has already mixed clusters during training.

Efficiency results are acceptable for a prototype but not negligible. Training all cluster adapters requires 1.28 times the GPU-hours of training a single identity LoRA in the four-shot setting because many clusters are small and training jobs have overhead. However, storage remains modest: the median identity requires 74 MB for all rank-8 adapters compared with 56 MB for a single rank-8 adapter and several gigabytes for a fully fine-tuned model. Inference with routing and prompt supplementation adds less than 80 ms before sampling. The second pass increases total generation time by 38% when enabled. Since background transfer is not required for every prompt, average latency across the full test set increases by 18%. This overhead is meaningful but smaller than the cost of running multiple full candidate generations without ranking.

Failure inspection shows three recurrent error classes. The first is semantic overconstraint, where brief-derived tokens suppress part of the user prompt. For example, if a brief repeatedly says that a character should appear cheerful, outputs may resist prompts asking for a neutral instructional pose. The second is mask boundary instability in the second pass. Thin objects, transparent regions, and patterned accessories can be partially treated as background, causing texture transfer into the foreground. The third is cluster ambiguity. Some identities deliberately use the same palette and motif across products and environments, making local clusters hard to separate. In these cases, the router may select a background adapter when the prompt asks for a foreground product, especially if the prompt mentions environmental nouns more prominently than product nouns.

The results should not be read as evidence that the routed method always dominates reference-based or personalization-based methods. Reference-conditioned generation performs well when the desired output should closely resemble a particular approved asset. DreamBooth performs well for a single recurring subject with several clean examples and little need to represent multiple internal concepts. Retrieval prompting is attractive when training adapters is impossible or when iden-

Table 9. Adapter Rank Sensitivity

Rank	ICS	LPIPS	Trend
4	0.739	–	Under-parameterized
8	0.753	0.362	Best trade-off
16	0.758	–	Small gain
32	0.760	0.341	Lower diversity

Table 10. Effect of Background Reference Strength

ω	Background FID	ICS	Outcome
0.00	35.6	0.752	No transfer
0.25	32.4	0.758	Balanced improvement
0.42	29.7	0.761	Best overall setting
0.55	29.1	0.744	Identity degradation begins
0.70	28.8	0.712	Reference dominance

tity constraints are mostly textual. The routed method is most useful when the identity repository has several internal visual concepts, when at least a few approved examples exist for each concept, and when generated content must combine new prompt semantics with recurring local identity features.

6. Analysis and Discussion

The strongest evidence from the simulated benchmark is that identity repositories benefit from being treated as structured rather than homogeneous. A single adapter per identity can learn common colors and broad style, but it often treats rare or local concepts as noise. This is especially problematic when the identity includes both foreground subjects and backgrounds. The adapter update is then pulled in multiple directions: it must preserve a character's shape, a package's color, and a background's texture, even though a given prompt may require only one or two of these. Cluster routing changes the adaptation problem by making each update narrower. The adapter for a character cluster does not need to model every approved background, and the adapter for a background cluster does not need to preserve product geometry. This division appears to explain much of the ICS improvement.

The value of prompt supplementation is more conditional. It helps when the adapter lacks enough examples to internalize a visual rule. In the two-shot setting, a content brief that names palette constraints and allowable shape features can prevent severe drift. Captions can also make sparse visual evidence more legible to the text-conditioned generator. However, supplementation is not always beneficial. Long prompts can dilute the user's instruction, exceed text-encoder capacity, or cause the model to attend to generic adjectives instead of identity-specific nouns. The threshold-triggered rule works because it treats text evidence as a support mechanism, not as a universal solution. When a cluster has sufficient examples, the adapter should carry most identity information.

The background pass addresses a different failure mode. Adapted diffusion models sometimes generate strong foreground identity while producing generic or low-quality backgrounds. This can happen because sparse concept examples focus on the subject and do not provide enough background variation. It can also happen because identity adapters alter attention patterns in ways that reduce the base model's background diversity. A reference background gives the generator a concrete approved structure, but unmasked reference conditioning can overwrite the generated subject. The masked two-pass procedure avoids some of this conflict by separating regions. Its effectiveness depends on mask quality, reference relevance, and the strength of attention transfer. In the experiments, it improves background FID with a small positive effect on overall identity conformance, but only when reference strength is moderate.

One subtle outcome is that human preference gains are smaller than metric gains. This is not surprising. Automatic metrics compare outputs to hidden prototypes, while human raters see only a limited identity card. If a generated image violates a rule that is not visible in the card, raters may not penalize it. Conversely, raters sometimes prefer visually rich images that slightly violate palette constraints. This difference is important for deployment. A system optimized only for measured identity conformance may produce outputs that are technically closer to approved assets but less compelling to human reviewers. A system optimized only for human first impressions may drift from formal guidelines. The prototype therefore uses metrics for ranking and failure detection, not as the sole definition of quality.

The method also raises a question about identity specificity. A cluster adapter can preserve local identity features, but it can also encode accidental details

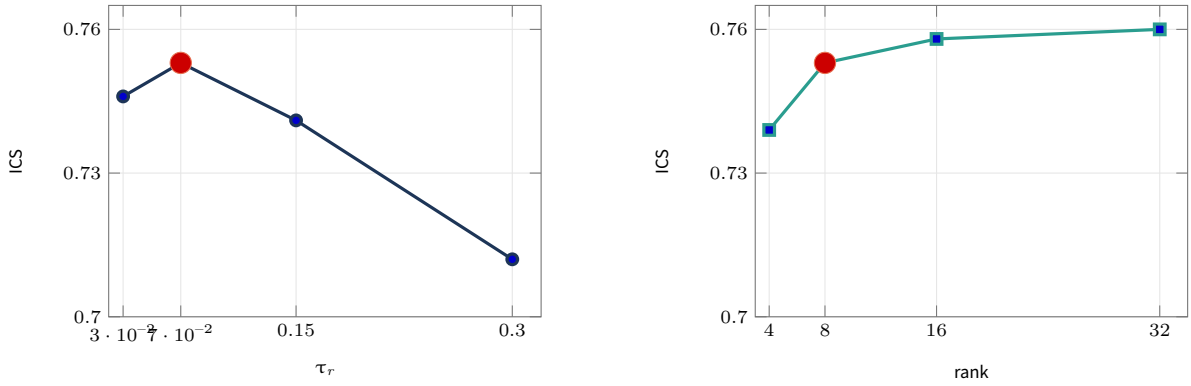

Figure 11. Sensitivity profile.

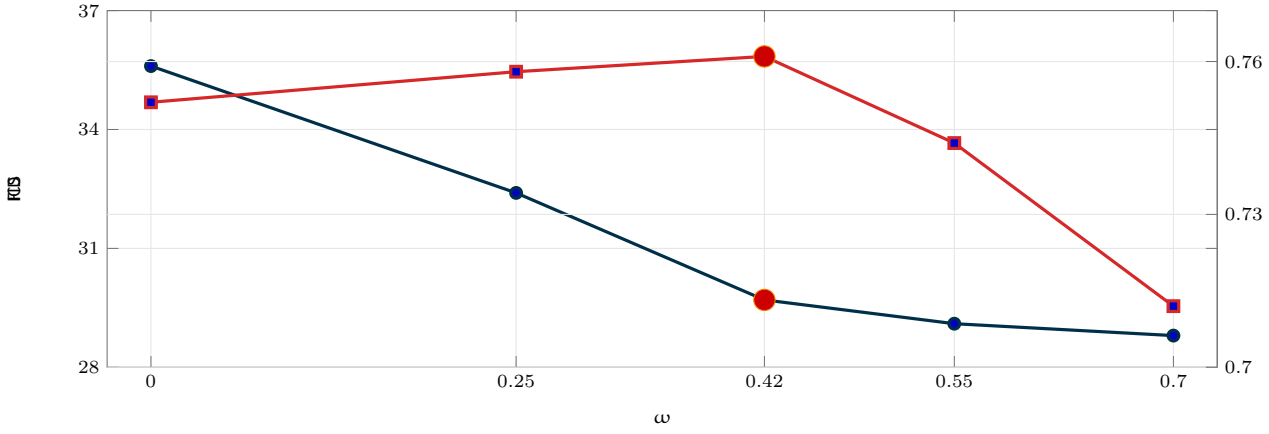

Figure 12. Reference balance.

from the sparse training examples. For example, if all four examples of a package show the same camera angle, the adapter may treat that angle as part of the identity. Prompt supplementation can reduce this by adding brief constraints that distinguish essential rules from incidental details, but the method does not fully solve the problem. A more principled approach would separate invariant identity attributes from contingent presentation attributes. This might require causal data augmentation, explicit attribute labels, or counterfactual training examples, none of which are assumed in the current benchmark.

Routing errors deserve particular attention. A wrong adapter can be worse than no adapter because it confidently imposes the wrong identity concept. The router uses embedding similarity between the prompt and cluster centroids, but prompts and asset clusters are not always expressed in the same language. A user may ask for a "welcome scene" while the repository names the relevant concept "entryway background." A user may ask for "the round snack box" while captions refer to "cylindrical packaging." These vocabulary differences can misroute the prompt. Retrieval of brief fragments partly mitigates the problem, but the router still depends on representation quality. Future versions may need a learned router trained on prompt-cluster pairs, or an interactive router that asks for clarification when margins are low.

Adapter composition is another source of complexity. A prompt may genuinely require several identity concepts, but combining low-rank updates is not guaranteed to produce linear combination of visual behaviors. Two adapters can interfere inside attention layers, especially when both modify cross-attention for overlapping tokens. In the experiments, using the top two clusters works better than hard routing for multi-concept prompts, but using more than two clusters reduces identity conformance. This suggests that adapter composition should be sparse and possibly layer-specific. A background adapter might be applied more strongly in deeper spatial layers, while a character adapter might be applied in cross-attention blocks associated with foreground tokens. The current method uses the same routing weights across layers, which is simple but not necessarily optimal.

The palette results show that color constraints can be learned more reliably when the repository is clustered. A single adapter sees all approved colors and may learn a broad identity palette, but prompts often require a local palette. Product packaging may use a

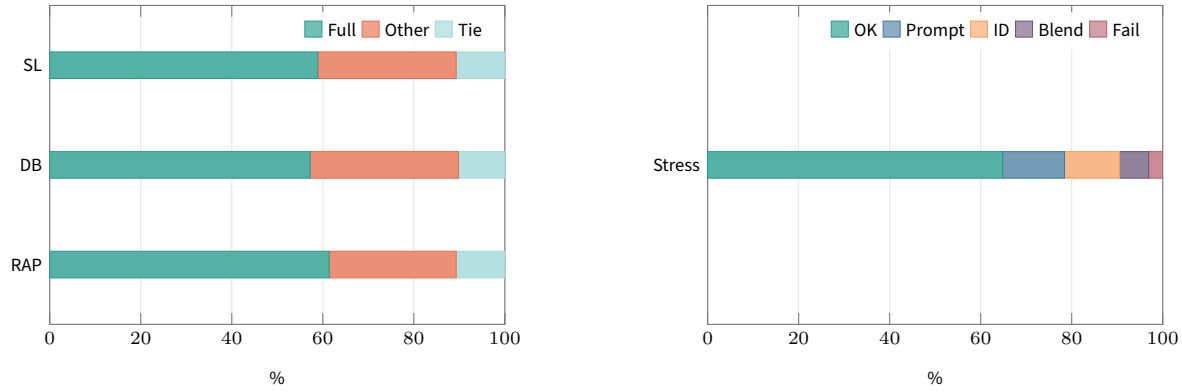

Figure 13. Preference and stress.

saturated accent color, while backgrounds use muted variants. If these are not separated, the generated image may use the accent everywhere. Cluster routing reduces this because each adapter sees a narrower palette distribution. The weak palette loss further discourages large drift. Still, color metrics are imperfect. A low average palette error does not guarantee that a logo color is exact, and a high palette error may be acceptable when the prompt asks for unusual lighting. Palette control may require explicit color management beyond diffusion adaptation.

The method does not solve text rendering. If a content brief requires exact typography, slogans, or legal copy, diffusion image generation remains unreliable. The benchmark includes layout-style prompts, but it does not require exact readable text beyond simple marks. For production workflows, generated backgrounds and illustrations may need to be combined with deterministic layout tools that render text and vector assets exactly. The routed adaptation method is better viewed as a component for generating visual material around fixed brand assets, not as a replacement for document composition systems.

There is also an intellectual property and governance dimension, although it is outside the main technical evaluation. The method assumes that the repository contains approved assets and that adapters are trained only for authorized identities. This assumption matters because identity generation can be used to mimic styles or brands without permission. A production system should include access control, audit logs, provenance metadata, and review workflows. The technical mechanism of routing does not by itself enforce appropriate use. It can make authorized generation more controllable, but it could also make unauthorized mimicry more systematic if misapplied. The paper does not evaluate misuse prevention.

The evaluation being simulated has both advantages and limitations. It allows precise control over sparse regimes, cluster labels, and prompt splits. It also allows results to be compared across baselines without exposing proprietary identity data. However, simulated identities are cleaner than many real repositories. Real repositories often contain obsolete assets, contradictory guidelines, low-resolution images, campaign-specific exceptions, and undocumented preferences held by human reviewers. These issues may reduce clustering quality and adapter reliability. A real-world evaluation should measure how the method handles noisy repositories and guideline changes over time.

One practical implication is that identity generation systems should store intermediate structure. It is not enough to store final adapters. The system should preserve cluster membership, centroid embeddings, prompt fragments, brief-to-asset links, routing logs, and generation scores. These artifacts allow debugging. If an output fails, one can inspect whether the wrong cluster was selected, whether the prompt was over-supplemented, whether the mask leaked, or whether the adapter itself is undertrained. This kind of observability is important because identity failures are often subtle. A user may say “it does not look like us,” and the system needs technical evidence to diagnose what happened.

Another implication is that model updating can be localized. If a content identity changes, cluster-level adapters make it possible to update only affected concepts. A new package design can form a new cluster or replace the old package adapter. A changed background treatment can update background clusters while leaving mascot clusters stable. This is more efficient than retraining a full identity model and less risky than editing a monolithic adapter. The benchmark does not include a temporal update experiment, but the architecture is compatible with incremental maintenance.

The method is deliberately conservative in its claims. It improves measurable identity adherence in a constrained benchmark, but it does not remove the need for review. It works best for visual concepts that have some approved examples and reasonably descriptive metadata. It is less reliable for abstract tone, exact

text, subtle cultural cues, or contradictory prompts. It also depends on the base diffusion model; if the base model cannot generate a requested object or relation, the adapter cannot reliably fix that deficiency. The value of routing is therefore bounded by representation quality, repository quality, and base generator competence.

7. Limitations and Threats to Validity

The first limitation is that the reported results come from a simulated benchmark. Although the benchmark is designed to resemble sparse identity generation, it cannot fully represent the irregularity of real production repositories. Real content identities often include years of accumulated assets with inconsistent lighting, file formats, usage rights, and historical guideline changes. Some assets may be approved only for a specific campaign, region, or medium. The simulated benchmark treats approval as uniform and stable. This simplification makes controlled evaluation possible but may overstate the ease of clustering and adaptation.

The second limitation concerns the definition of identity conformance. ICS uses prototype similarity in an embedding space. This is useful for comparing methods, but it is not a complete measure of identity. Embeddings may miss exact geometry, symbolic meaning, legal constraints, or cultural associations. A generated logo-like shape may be close in embedding space but still unacceptable because one stroke is wrong. Conversely, an image may have lower embedding similarity while satisfying a content brief through composition or tone. The paper therefore reports multiple metrics and human preferences, but none of them fully captures the judgment that a trained reviewer would make.

The third limitation is prompt distribution. The test prompts are more systematic than real user prompts. They are written to cover known internal concepts and to vary scene, pose, and background. Real prompts may be underspecified, colloquial, misspelled, or contradictory. They may refer to internal campaign names that do not appear in captions. They may ask for content that violates guidelines. The router may fail more often under such conditions. A deployed system would need clarification behavior, prompt rewriting, and policy checks that are not evaluated here.

The fourth limitation is that the second-pass background method depends on segmentation quality. Segment Anything and related models are strong general tools, but masks can fail around transparent objects, fine hair-like structures, complex product edges, and overlapping objects [21]. In the current method, mask errors can transfer background texture into foreground regions or preserve foreground artifacts that should be harmonized with the background. Mask dilation and erosion help but do not solve all cases. More robust region control may require joint segmentation and generation rather than a separate mask extraction stage.

The fifth limitation is adapter interference. Weighted composition of low-rank updates is simple and efficient, but it is not guaranteed to preserve the behavior of each adapter independently. The method restricts composition to the top two clusters, yet some prompts require more complex combinations. Layer-specific routing, token-specific routing, or gated attention could improve this, but they add training and inference complexity. The current results should therefore be interpreted as evidence for a practical baseline rather than a final solution to multi-concept identity composition.

The sixth limitation is that the method may encode biases or errors present in the approved repository. If the repository underrepresents certain contexts, the generated outputs may also underrepresent them. If old assets contain undesirable stereotypes or outdated visual practices, clustering and adapters may preserve those patterns. Prompt supplementation from briefs can correct some issues when the brief is explicit, but it cannot infer missing ethical or representational constraints. Human governance remains necessary.

The seventh limitation is computational. Although low-rank adapters are efficient compared with full fine-tuning, training many cluster adapters still creates operational overhead. Small clusters are cheap individually but numerous. The second pass also increases latency. In an interactive design tool, this overhead may be acceptable for high-value generation but not for every quick draft. Caching, asynchronous candidate generation, and selective use of background transfer may be needed.

The final threat to validity is that baseline implementations may not represent every possible tuning strategy. DreamBooth, Custom Diffusion, textual inversion, reference conditioning, and ControlNet-style methods each have many variants and hyperparameters. The experiments use reasonable configurations under a shared compute budget, but stronger tuning of a baseline could narrow some gaps. The main conclusion should therefore be about the observed value of cluster routing and threshold supplementation under controlled conditions, not about universal superiority over all personalization methods.

8. Conclusion

This paper presented a cluster routed identity adaptation method for sparse content-identity generation. The method decomposes an approved identity repository into multimodal clusters, trains compact low-rank adapters for local identity concepts, routes user prompts to the most relevant adapters, supplements prompts with brief and caption evidence when cluster support is limited, and applies a masked second pass to transfer approved background structure. The design responds to a practical issue in generative content systems: iden-

tity adherence is often local, sparse, and operationally constrained. Treating the repository as a homogeneous training set can cause concept blending, color leakage, and background regression.

The simulated evaluation suggests that routing is the main source of improvement. In the four-shot setting, the full method improves identity conformance from 0.673 for a single identity adapter to 0.761, while reducing foreground-weighted palette error and improving background FID. The gains persist in two-shot, eight-shot, and sixteen-shot settings, although the margin narrows as more examples become available. Prompt supplementation helps most when clusters contain five or fewer examples, and always-on supplementation is less effective because it can crowd the prompt with constraints. The masked background pass improves background quality when reference strength is moderate, but high reference strength can damage foreground identity.

The results also show several limits. The method remains dependent on embedding quality, mask quality, and the base generator's semantic competence. It does not guarantee exact text rendering or formal brand compliance. Routing errors can impose the wrong identity concept, and adapter composition can blend concepts when prompts are ambiguous. Human preference judgments favor the routed method over several baselines, but the margins are smaller than automatic metric differences, which indicates that visual appeal and identity compliance are related but not identical.

A useful direction for future work is to learn routers from reviewed generation logs rather than relying only on embedding similarity. Another direction is layer-specific adapter composition, where foreground, background, color, and layout constraints are injected at different parts of the denoising network. Better uncertainty estimates could help a system ask for clarification when two clusters are nearly tied. Repository cleaning and temporal update experiments would also make the evaluation closer to real identity management, where assets change and guidelines evolve.

The broader implication is modest but practical. Content-identity generation benefits from preserving structure between the asset repository and the generative model. Clusters, adapters, brief fragments, masks, references, and routing logs are not incidental engineering details; they are the mechanisms through which a generative model can be made more accountable to a specific visual identity. The present method does not remove human judgment from content creation, but it can reduce avoidable identity drift and make failures easier to diagnose. That is a useful step for controlled generative workflows where semantic novelty and identity consistency must coexist.

References

[1] D. Agarwal, U. Moorarka, S. Agrawal, V. N. Reddy, K. K. Jain, and A. Revanur, "Content identity based digital content generation," Feb. 5 2026. US Patent App. 18/790,506.

[2] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in Advances in Neural Information Processing Systems, vol. 33, pp. 6840–6851, 2020.

[3] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in International Conference on Learning Representations, 2021.

[4] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in Proceedings of the 38th International Conference on Machine Learning, vol. 139 of Proceedings of Machine Learning Research, pp. 8162–8171, PMLR, 2021.

[5] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," in Advances in Neural Information Processing Systems, vol. 34, pp. 8780–8794, 2021.

[6] 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, pp. 10684–10695, 2022.

[7] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, "Photorealistic text-to-image diffusion models with deep language understanding," in Advances in Neural Information Processing Systems, vol. 35, pp. 36479–36494, 2022.

[8] 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, pp. 22500–22510, 2023.

[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," in International Conference on Learning Representations, 2023.

[10] N. Kumari, B. Zhang, R. Zhang, E. Shechtman, and J.-Y. Zhu, "Multi-concept customization of text-to-image diffusion," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.

[11] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," in International Conference on Learning Representations, 2022.

[12] L. Zhang, A. Rao, and M. 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.

[13] 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.

[14] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in Proceedings of the 38th International Conference on Machine Learning, vol. 139 of Proceedings of Machine Learning Research, pp. 8748–8763, PMLR, 2021.

[15] J. Li, D. Li, C. Xiong, and S. C. H. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in Proceedings of the 39th International Conference on Machine Learning, vol. 162 of Proceedings of Machine Learning Research, pp. 12888–12900, PMLR, 2022.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer,

G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in International Conference on Learning Representations, 2021.

[17] C. Mou, X. Wang, L. Xie, Y. Wu, J. Zhang, Z. Qi, and Y. Shan, "T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 4296–4304, 2024.

[18] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross-attention control," in International Conference on Learning Representations, 2023.

[19] H. Chefer, Y. Alaluf, Y. Vinker, L. Wolf, and D. Cohen-Or, "Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models," ACM Transactions on Graphics, vol. 42, no. 4, pp. 1–10, 2023.

[20] T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to follow image editing instructions," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18392–18402, 2023.

[21] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4015–4026, 2023.

[22] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 1027–1035, Society for Industrial and Applied Mathematics, 2007.

[23] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," Journal of Computational and Applied Mathematics, vol. 20, pp. 53–65, 1987.

[24] R. J. G. B. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," in Advances in Knowledge Discovery and Data Mining, vol. 7819 of Lecture Notes in Computer Science, pp. 160–172, Springer, 2013.

[25] L. McInnes, J. Healy, N. Saul, and L. Grossberger, "Umap: Uniform manifold approximation and projection," Journal of Open Source Software, vol. 3, no. 29, p. 861, 2018.

[26] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in Advances in Neural Information Processing Systems, vol. 30, 2017.

[27] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 586–595, 2018.