MaterialFlow: Attribute-Disentangled Material Transfer via Trajectory-Aware Velocity Modulation

Sung-Lin Tsai[1], Bo-Kai Ruan[1], Yu-Hsuan Chen[1], Wen-Huang Cheng[2,3], and Hong-Han Shuai[1]*

[1] National Yang Ming Chiao Tung University, Hsinchu, Taiwan
[2] National Taiwan University, Taipei, Taiwan
[3] NTU AI Center of Research Excellence (NTU AI-CoRE)
{tsai412504004.ee12,bkruan.ee11,nightdream29479.ee12}@nycu.edu.tw,
wenhuang@csie.ntu.edu.tw,hhshuai@nycu.edu.tw

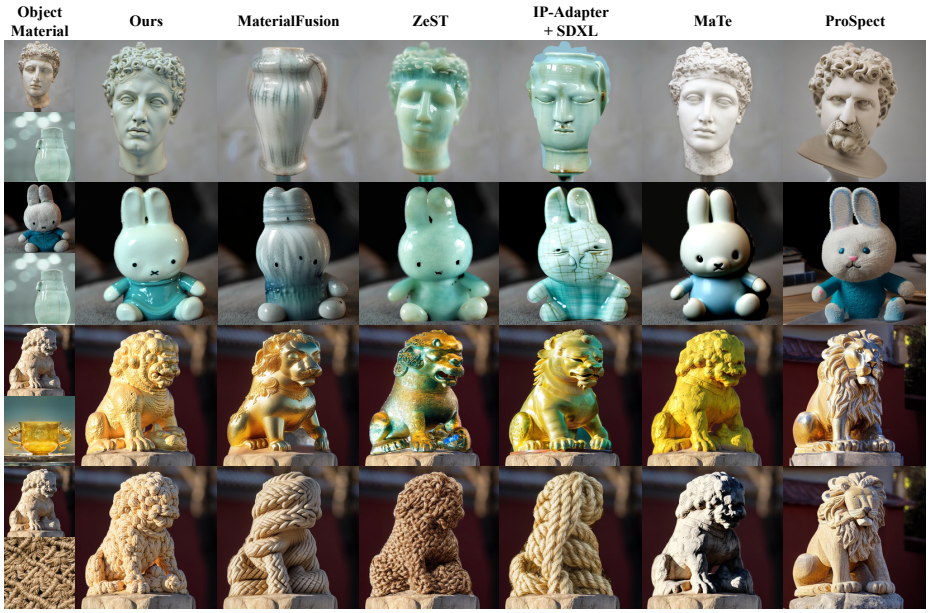

Fig. 1: Qualitative comparison of material transfer results. For each row, the upper-left image is the source and the lower-left image is the material exemplar.

Abstract. Material transfer seeks to re-render an object's surface using a reference exemplar while preserving its original geometry and identity. Current training-free methods often rely on attention-based injection, which entangles attributes like color, texture, and structural patterns into a single representation. This entanglement leads to structural instability and imprecise results. Additionally, inversion-based editing in

* Corresponding author.

flow models is prone to reconstruction-induced trajectory drift, where accumulated errors degrade object details. We propose *MaterialFlow*, a training-free framework for precise material transfer using pre-trained flow models. Our approach introduces a trajectory-aware velocity modulation mechanism that rebalances the generative flow in an inversion-free manner. This ensures stable and semantically consistent editing dynamics without the need for explicit latent reconstruction. We further introduce an attribute-aware disentanglement paradigm that separates reference materials into color, texture, and meso-scale patterns for fine-grained control. Evaluations on the Material Transfer Benchmark (MTB) show that *MaterialFlow* surpasses state-of-the-art methods in transfer quality, inference efficiency, and controllability. Project page: `https://github.com/Sung-Lin/MaterialFlow`.

Keywords: Material Transfer · Image synthesis · Training-free Diffusion models

1 Introduction

Material transfer is defined as the task of re-rendering an object's surface appearance using a reference exemplar while preserving the source object's underlying geometry and semantic identity. Conceptually, it functions as a form of "digital alchemy," allowing designers to decouple an object's "what" (its structure and meaning) from its "how" (its material properties such as reflectance, roughness, and color). The ability to replace wood with marble, leather with fabric, or matte paint with glossy metal benefits numerous domains, including digital content creation [45], augmented reality previews [12], and 3D asset pipelines [26].

Despite the power of modern generative priors, high-quality material transfer remains a daunting challenge. Material cues are inherently and tightly coupled with local illumination and global geometry. Current 2D approaches can generally be grouped into two categories: (1) fine-tuning-based methods like DreamBooth [29] and U-VAP [35], which adapt models to material-specific data but incur high per-task training costs, and (2) training-free feature-injection methods like ZeST [7] and MaterialFusion [11]. While the latter are more efficient, they often rely on complex auxiliary adapters (*e.g.*, ControlNet [43]) or structural cues like depth maps. Even with these additions, existing methods frequently suffer from texture leakage, where material patterns ignore object boundaries, or structural distortion, where the form of the object is warped to match the material characteristics.

The search for more precise control has led to the emergence of flow-based generative models [18, 19]. For instance, unlike traditional diffusion models that follow complex stochastic paths, Rectified Flow (RF) [19] reformulates generation as deterministic ordinary differential equations (ODE) transport. This shift is technically significant for material transfer not merely for its sampling efficiency, but for its trajectory properties. Because RF follows a straight-line latent path, it offers a stable and predictable foundation to directly shape the generation process through velocity field modulation. This allows for the "surgical"

injection of material attributes without the stochastic noise that often triggers structural drift in diffusion-based editing.

However, simply adopting RF is not a "silver bullet." Existing flow-editing frameworks often rely on inversion, where small reconstruction errors accumulate along the deterministic path, leading to significant trajectory drift. In tasks as sensitive as material transfer, these compounded errors frequently result in the loss of original object details. Furthermore, most pipelines encode the reference material as a single entangled representation, causing heterogeneous attributes like color and micro-texture to compete within the editing dynamics. This competition often leads to an "all-or-nothing" transfer that either ignores the reference or destroys the source object's integrity as shown in Fig. 1.

To bridge this gap, we propose MaterialFlow, an inversion-free material transfer framework designed for pre-trained flow models. Our framework bypasses explicit reconstruction by introducing a trajectory-aware normalization mechanism that rebalances velocity fields to maintain stable and semantically consistent editing dynamics. Moreover, we develop an attribute-aware disentanglement paradigm that explicitly separates the reference material into three factors: color, texture, and structural patterns. By guiding the generative flow with these decoupled representations, MaterialFlow enables precise, training-free material transfer that consistently surpasses state-of-the-art methods in both efficiency and controllability.

Our contributions can be summarized as follows:

- We propose MaterialFlow, a fully training-free framework that operates directly in the latent space of pre-trained flow models, avoiding the trajectory drift common in inversion-based methods.
- We design a normalization mechanism that rebalances exemplar-conditioned velocity fields, ensuring stable editing dynamics and preventing structural distortion. Moreover, a principled paradigm is proposed to explicitly separate reference materials into color, texture, and structural patterns, allowing for fine-grained control over the transfer process.
- Experimental results on the MTB dataset demonstrate that MaterialFlow achieves state-of-the-art overall performance while preserving object details to the greatest extent.

2 Related Work

2.1 Exemplar-guided image editing

Exemplar-guided image editing (EIE) enables precise visual control by modifying source images under the guidance of reference exemplars, allowing users to specify complex appearance attributes that are difficult to describe in language alone. A central challenge in this paradigm is effectively aligning source and exemplar representations while preventing structural distortion. Early works like Paint-by-Example [39] introduce information bottlenecks to integrate features, while subsequent methods explore explicit modulation, such as TF-ICON's [21]

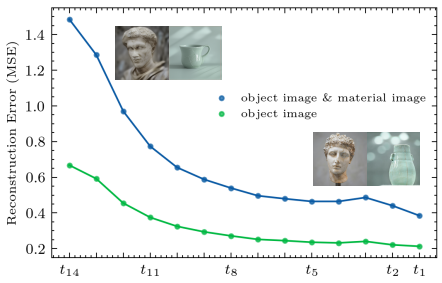

Fig. 2: Reconstruction error across timesteps. Comparison between object-only inversion and joint object–material inversion.

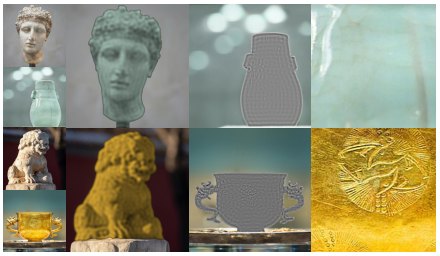

Fig. 3: Visualization of attribute-aware disentanglement. Exemplar representations separated into different material attributes.

self-attention injection, Style Injection's [8] and Z*'s [10] attention reweighting, and AnyDoor's [5] detail extraction. Beyond attention-based designs, MARBLE [6] aligns representations in CLIP space, and MimicBrush [4] employs self-supervised correspondence, with recent efforts focusing on information efficiency [38] and multimodal controllability [42]. However, most existing approaches rely on complex attention modulation or additional trainable modules that increase computational overhead. To enable a more efficient framework, we instead adopt an inversion-free editing paradigm that operates in a fully training-free manner, achieving strong editing performance with low computational cost.

2.2 Material acquisition and transfer

Material acquisition and transfer have evolved from classical studies of illumination and reflectance to recent generative models that support text-driven editing and 3D mesh texturing. While traditional 3D approaches like Text2Tex [3], TEXTure [26], and TextureDreamer [41] rely on explicit geometry estimation, recent works have shifted toward 2D-to-2D transfer to improve practicality. These 2D methods include (1) fine-tuning approaches (*e.g.*, DreamBooth [29], Material Palette [20], U-VAP [35], MARBLE [6], MaTe [13]), which suffer from high costs and overfitting; (2) token-modulation methods like Prospect [45], which are limited by text-space controllability; and (3) feature-injection methods (*e.g.*, MaterialFusion [11], ZeST [7]), which often require auxiliary encoders and structural cues like depth maps. Critically, existing methods typically encode reference materials as homogeneous representations where diverse attributes are entangled within a single latent space, limiting fine-grained control. In contrast, our proposed framework explicitly disentangles complementary attributes, *i.e.*, color, texture, and structural patterns, to guide editing dynamics in a training-free manner, achieving semantically coherent transfer without extra learnable modules or parameter overhead.

3 MaterialFlow

Existing training-free material transfer approaches [7, 11, 45] built upon diffusion and flow models typically depend on inversion-based editing frameworks, where reference material features are extracted and injected via attention manipulation. Despite their effectiveness, existing frameworks are constrained by inversion-induced information degradation, additional attention computations, and entangled material representations. To overcome these limitations, we propose MaterialFlow, an inversion-free material transfer framework. We first treat the reference material image as an exemplar and introduce an exemplar-guided inversion-free editing paradigm in Sec. 3.2. Subsequently, in Sec. 3.3, we further disentangle heterogeneous material attributes into attribute-specific exemplars corresponding to color, texture, and structural patterns. These disentangled representations are then integrated within Algorithm 1 to enable precise and controllable material transfer.

3.1 Preliminaries

Rectified Flow models. Generative flow models [15, 18, 19] learn a time-dependent velocity field to construct a continuous transformation between two probability distributions. Let $\mathbf{Z}_0 \sim p_{\text{data}}$ denote latent samples from the data distribution and $\mathbf{Z}_1 \sim \mathcal{N}(0, \mathbf{I})$ denote samples from a standard Gaussian distribution. The transport between them is defined by the ordinary differential equation

$$d\mathbf{Z}_t = V(\mathbf{Z}_t, t)dt, \quad t \in [0, 1], \tag{1}$$

where $V(\mathbf{Z}_t, t)$ drives samples through continuous latent dynamics from noise to data. Rectified Flow (RF) [19] further simplifies this process by adopting a predefined linear interpolation,

$$\mathbf{Z}_t = (1 - t)\mathbf{Z}_0 + t\mathbf{Z}_1, \tag{2}$$

which avoids complex trajectories and enables efficient latent transport.

Inversion-free Editing. Unlike diffusion models, inversion-based editing in Rectified Flow (RF) models [9, 28, 33, 36, 37, 47] suffers from sensitivity to reconstruction errors as indicated by the green curve in Fig. 2. RF models follow a deterministic continuous trajectory governed by an ODE-based velocity field, so small reconstruction deviations introduced during inversion accumulate along the flow path, which causes trajectory drift and reduces controllability during image editing. Recent works [14, 16] propose editing strategies that avoid explicit inversion and better align with the training dynamics of RF models. One representative example is FlowEdit [14], which is an inversion-free and model-agnostic method for text based editing with flow models. Instead of first mapping an image to its latent noise and then solving a separate reverse ODE, FlowEdit directly constructs a transport between the source and target distributions by

defining an ODE that evolves latent states from the source condition to the target condition with lower transport cost than inversion-based editing. To realize this idea within RF, intermediate latents follow the same linear interpolation scheme used in RF training. Specifically, given a source latent $\mathbf{Z}_0^{\text{src}}$, the latent at time t is obtained by

$$\mathbf{Z}_t^{\text{src}} = (1-t)\mathbf{Z}_0^{\text{src}} + t\mathbf{N}_t, \tag{3}$$

where $\mathbf{N}_t \sim \mathcal{N}(0, \mathbf{I})$. A corresponding target latent $\mathbf{Z}_t^{\text{tar}}$ is obtained with the same construction at time t. The editing direction is defined by the difference between the velocity fields of the target and source latents,

$$\Delta V_t = V(\mathbf{Z}_t^{\text{tar}}, t) - V(\mathbf{Z}_t^{\text{src}}, t), \tag{4}$$

and this difference steers the source latent trajectory toward the target distribution, as illustrated in Fig. 4 (a).

During sampling at timestep t_i, the velocity difference is applied to update the source latent using a first-order integration step. Specifically, the source latent is updated as

$$\mathbf{Z}_{t_{i-1}}^{\text{src}} \leftarrow \mathbf{Z}_{t_i}^{\text{src}} + (t_{i-1} - t_i)\,\Delta V_{t_i}. \tag{5}$$

Inversion-free manipulation thus operates directly on the source latent without explicit inversion, which reduces the computational cost from two times numerical function evaluations (NFEs) to one time, and avoids reconstruction-induced drift, leading to more stable and accurate editing. Further analysis of the editing efficiency is provided in the supplementary material.

Problem Formulation. Given a source object image I_O, we first encode it into the latent space of the generative model using a VAE encoder, resulting in the source latent representation $Z^{\text{src}} = \mathcal{E}(I_O)$. The editing process is then performed in the latent space through exemplar-guided inversion-free editing, which will be detailed in Sec. 3.2. In addition to the source object, our method takes a reference material exemplar image I_M as guidance. To enable fine-grained material transfer, we further perform *Attribute-aware Disentanglement* on I_M, decomposing it into three attribute-specific exemplars corresponding to color, texture, and structural patterns, denoted as I_C, I_T, and I_P, respectively. The disentanglement process will be described in Sec. 3.3.

Each attribute exemplar is then encoded into the latent space as $\hat{Z}^C = \mathcal{E}(I_C)$, $\hat{Z}^T = \mathcal{E}(I_T)$, and $\hat{Z}^P = \mathcal{E}(I_P)$, which are subsequently used to compute attribute-specific velocity differences that guide the editing dynamics. The complete editing framework will be presented in Algorithm 1.

3.2 Exemplar-guided Inversion-Free Editing

In inversion-based editing frameworks, the inversion procedure inevitably introduces reconstruction errors that affect the entire editing trajectory. Since the

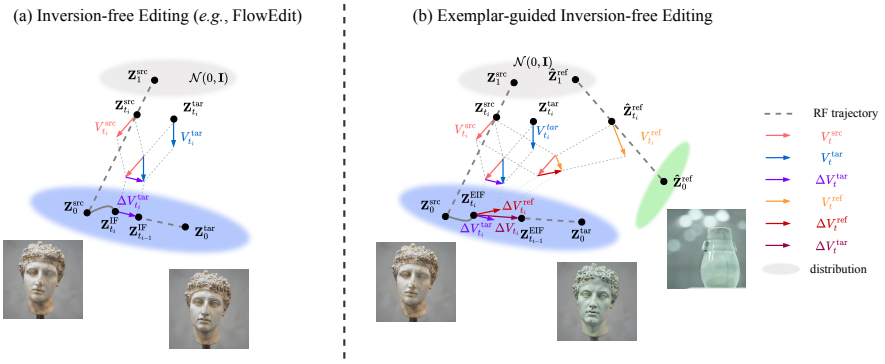

Fig. 4: Comparison between inversion-free editing paradigms. (a) Inversion-free Editing (IF). The latent $\mathbf{Z}_t^{\mathrm{IF}}$ is updated by computing the velocity difference $\Delta V_t^{\mathrm{tar}}$ across timesteps and applying it to the current latent state. (b) Exemplar-guided Inversion-free Editing (EIF). In addition to $\Delta V_t^{\mathrm{tar}}$, an exemplar-conditioned velocity difference $\Delta V_t^{\mathrm{ref}}$ is computed from the reference exemplar. A trajectory-aware normalization mechanism aligns the two velocity fields before aggregation, and the combined velocity is used to update the latent $\mathbf{Z}_t^{\mathrm{EIF}}$.

editing process conditions on the inverted latent representation, any discrepancy between the reconstructed latent and its original counterpart propagates along the generation path and can be amplified through iterative transformations.

This limitation becomes more severe in editing tasks that involve exemplar images, such as material transfer. In addition to inverting the source image, the exemplar image also requires a separate inversion step to obtain its latent representation. When only the object image (the human head sculpture) is inverted, the resulting reconstruction error is illustrated by the green curve in Fig. 2. However, when both the object image and the reference material image (the ceramic exemplar) are simultaneously inverted, the reconstruction errors from the two sources accumulate, as reflected by the blue curve. As a result, an additional reconstruction error is introduced from the exemplar side. These errors jointly influence the velocity field guidance and accumulate throughout the transport process, leading to larger deviations in the editing dynamics. This limitation motivates us to adopt an inversion-free editing framework, which avoids explicit latent reconstruction and provides a more stable and reliable foundation for exemplar-guided material transfer.

However, combining additional exemplar-guidance introduces a fundamentally different requirement from a text-guided setting. Unlike text-guided editing, where the target condition typically induces localized and semantically consistent modifications, exemplar-guided editing aims to inject only attribute-level information from a reference image while preserving the primary semantic content and structural integrity of the source image. In this setting, directly applying the velocity difference derived from the exemplar-conditioned trajectory may lead to an imbalance in flow guidance. Specifically, the exemplar trajectory may

exhibit displacement magnitudes or directional tendencies that dominate the source trajectory, causing the editing process to either (i) insufficiently affect the source image due to a negligible velocity norm, or (ii) overly conform to the exemplar appearance at the expense of structural fidelity. This imbalance is further discussed in Sec. 4.4.

Therefore, rather than assuming implicit trajectory compatibility, it becomes necessary to explicitly align the exemplar-conditioned velocity field with that of the source trajectory. This alignment ensures that exemplar-driven modifications are introduced in a controlled manner, preserving the semantic backbone of the source while enabling effective attribute transfer. To address this issue, we introduce *Exemplar-guided Inversion-Free Editing*, a principled extension that rebalances the exemplar-conditioned velocity field to maintain stable and semantically consistent editing dynamics, as illustrated in Fig. 4 (b). The key idea is to align the displacement magnitudes of the two interpolated trajectories. Specifically, we introduce a trajectory-aware normalization mechanism that dynamically calibrates the exemplar-conditioned velocity field according to the relative trajectory displacement between the source and exemplar conditions. The rebalanced velocity difference is then computed as follows:

$$\Delta V_t^{\text{ref}} = \lambda_t V^{\text{ref}}\left(\hat{\mathbf{Z}}_t^{\text{ref}}, t\right) - V^{\text{src}}(\mathbf{Z}_t^{\text{src}}, t)\,, \quad \lambda_t = \frac{\|\mathbf{Z}_t^{\text{src}} - \mathbf{Z}_0^{\text{src}}\|}{\left\|\hat{\mathbf{Z}}_t^{\text{ref}} - \hat{\mathbf{Z}}_0^{\text{ref}}\right\| + \epsilon}, \tag{6}$$

where $\hat{\mathbf{Z}}_t^{\text{ref}}$ denotes the interpolated latent trajectory under reference exemplar conditioning, and ϵ is a small positive constant introduced for numerical stability. The rebalanced velocity difference is used to guide the source latent along the forward generative trajectory. When both textual and exemplar conditions are present, the complete velocity applied to the source latent is formulated as:

$$\Delta V_t = \Delta V_t^{\text{tar}} + \Delta V_t^{\text{ref}}. \tag{7}$$

Based on this formulation, we derive the MaterialFlow algorithm, which generalizes text-guided inversion-free editing to include additional exemplar-guidance.

Although *Exemplar-guided Inversion-free Editing* introduces richer visual guidance beyond textual descriptions, directly conditioning on a holistic material exemplar remains insufficient for fine-grained material transfer. A material image simultaneously encodes multiple visual factors, including color composition, micro-texture statistics, and structural patterns, which contribute differently to the perceived appearance. When treated as a single entangled latent condition, these heterogeneous factors compete within the editing dynamics and often lead to unintended structural changes or over transfer of irrelevant attributes. As illustrated in Fig. 5, such entanglement frequently produces distorted structures or exaggerated patterns. To address this issue, we explicitly disentangle the internal attribute structure of the exemplar representation to enable more controllable and semantically aligned material transfer.

Algorithm 1 MaterialFlow algorithm

Input: source object latent Z^{src}; material exemplar I_M; $\{t_i\}_{i=0}^{T}$; $n_{\max}$, $n_{\min}$
Output: Z_0^{MF}
Init: $Z_{t_{\max}}^{\text{MF}} = Z_0^{\text{src}}$
$\hat{Z}^C, \hat{Z}^T, \hat{Z}^P \leftarrow \mathcal{E}(\mathcal{D}(I_M))$ ▷ Attribute-aware Disentanglement
for $i = n_{\max}, n_{\max}-1, \ldots, n_{\min}$ **do**
  $N_{t_i}, \hat{N}_{t_i} \sim \mathcal{N}(0, \mathrm{I})$
  $Z_{t_i}^{\text{src}}, \leftarrow (1-t_i)Z_0^{\text{src}} + t_i N_{t_i}$
  $Z_{t_i}^{\text{tar}} \leftarrow Z_{t_i}^{\text{MF}} + Z_{t_i}^{\text{src}} - Z_0^{\text{src}}$
  $\Delta V_{t_i}^{\text{tar}} \leftarrow V(Z_{t_i}^{\text{tar}}, t_i) - V(Z_{t_i}^{\text{src}}, t_i)$
  $\Delta V_{t_i}^{\text{ref}} \leftarrow 0$
  for all $a \in \{C, T, P\}$ **do** ▷ Exemplar-guided Inversion-Free Editing
    $\hat{Z}_{t_i}^a \leftarrow (1-t_i)\hat{Z}_0^a + t_i \hat{N}_{t_i}$
    $\lambda_{t_i}^a \leftarrow \frac{\|Z_{t_i}^{\text{src}} - Z_0^{\text{src}}\|}{\|\hat{Z}_{t_i}^a - \hat{Z}_0^a\| + \epsilon}$
    $\Delta V_{t_i}^a \leftarrow \lambda_{t_i}^a V(\hat{Z}_{t_i}^a, t_i) - V(Z_{t_i}^{\text{src}}, t_i)$
    $\Delta V_{t_i}^{\text{ref}} \leftarrow \Delta V_{t_i}^{\text{ref}} + \Delta V_{t_i}^a$
  end for
  $\Delta V_{t_i} \leftarrow \Delta V_{t_i}^{\text{tar}} + \Delta V_{t_i}^{\text{ref}}$
  $Z_{t_{i-1}}^{\text{MF}} \leftarrow Z_{t_i}^{\text{MF}} + (t_{i-1} - t_i)\Delta V_{t_i}$
end for
Return: Z_0^{MF}

3.3 Attribute-aware Disentanglement

Attribute disentanglement is widely used in image synthesis and editing to enable controllable manipulation of visual factors. Prior works [2, 24] show that separating visual attributes in the embedding space leads to more structured and controllable generation. However, existing training-free material transfer methods [7,11,45] typically encode the reference material as a unified latent representation. As a result, heterogeneous attributes are entangled within a single feature space, limiting the ability to selectively manipulate individual material characteristics. Instead of attribute-level transfer, the model often superimposes the overall appearance of the reference material on the source object, which may cause structural distortion and detail degradation. To address this issue, we propose an attribute-aware disentanglement paradigm for material transfer. Specifically, we decompose the reference material into **three complementary components**, (1) color (I_C), (2) texture (I_T), and (3) structural patterns (I_P), and design dedicated decoupling mechanisms for each attribute. For clarity, Fig. 3 provides visual examples of the corresponding attribute exemplars. By explicitly modeling their distinct roles, our framework enables more precise and controllable material transfer through selective guidance of the editing dynamics.

Color. To achieve effective color disentanglement, we first extract the semantic low-frequency component of the material exemplar I_M via 2D Fourier low-pass filtering:

$$I_M^{\text{low}} = \mathcal{F}^{-1}(\mathcal{H}_{\text{low}} \odot \mathcal{F}(I_M)), \tag{8}$$

where $\mathcal{H}_{\text{low}}$ denotes a radial low-pass mask in the frequency domain. Following prior studies on color representation in generative models [12,30], we summarize the material's semantic color by extracting a representative chromatic vector $\mathbf{c}_{\text{rep}}$ in the CIELab space from I_M^{low}. Finally, to construct the color exemplar I_C we align this representative color with the object's low-frequency luminance distribution:

$$I_C(x) = \mathbf{c}_{\text{rep}} \cdot \mathcal{H}\big(Y_{\text{obj}}^{\text{low}}(x)\big)\,, \tag{9}$$

where $Y_{\text{obj}}^{\text{low}}$ denotes the object's low-frequency luminance and $\mathcal{H}(\cdot)$ represents histogram alignment. This preserves the shading structure while transferring the semantic color of the material.

Texture. In image synthesis, texture attributes characterize surface material properties such as metallicity, roughness, glossiness, and granularity [32]. Unlike color, which primarily conveys chromatic information, texture is predominantly encoded in the high-frequency components of an image, reflecting local spatial variations and fine-scale structural details [46]. To isolate the texture-specific signal from the material exemplar I_M, we apply a 2D Fourier high-pass filter:

$$I_M^{\text{high}} = \mathcal{F}^{-1}(\mathcal{H}_{\text{high}} \odot \mathcal{F}(I_M))\,, \tag{10}$$

where $\mathcal{H}_{\text{high}}$ denotes a radial high-pass mask in the frequency domain, satisfying $\mathcal{H}_{\text{high}} = 1 - \mathcal{H}_{\text{low}}$. The resulting high-frequency component I_M^{high} captures local contrast patterns and micro-structural variations while suppressing global chromatic information. We therefore designate

$$I_T = I_M^{\text{high}} \tag{11}$$

as the texture exemplar, which serves to guide the transfer of surface-level material characteristics independently of semantic color.

Structural Patterns. Beyond color and texture, an often-overlooked yet crucial attribute in material transfer is the presence of *structural patterns*. While texture primarily captures high-frequency micro-variations, structural patterns describe meso-scale spatial organizations such as spots, streaks, stripes, or repeated motifs distributed across a surface. This attribute is particularly prominent in materials exhibiting characteristic spatial arrangements, including wood grain, apple skin speckles, animal fur markings, or other abstract, non-object-centric references. Unlike stochastic texture variations, structural patterns encode structured and spatially correlated configurations that are perceptually salient and semantically meaningful. To disentangle this property from purely chromatic and high-frequency signals, we extract representative material patches from the surface region of the exemplar image I_M. These patches preserve the spatial arrangement and repetitive structural characteristics inherent to the material while minimizing global color bias. Formally, we define the structural pattern exemplar as

$$I_P = \mathcal{P}(I_M), \tag{12}$$

where $\mathcal{P}(\cdot)$ denotes a patch extraction operator that selects spatially coherent material regions from the exemplar image.

Overall Pipeline. Algorithm 1 summarizes the procedure of MaterialFlow. Starting from the source latent Z_0^{src}, we iteratively update the latent state along a reverse-time schedule from $n_{\max}$ to $n_{\min}$ using a trajectory-aware velocity modulation scheme. Here, $\{\hat{Z}_0^a\}_{a\in\mathcal{A}}$ denotes the set of disentangled attribute exemplars in latent space, where $\mathcal{A} = \{C, T, P\}$ corresponds to color, texture, and structural pattern, respectively. Each $\hat{Z}_0^a$ represents the initial latent constructed from the attribute-specific exemplar. At each timestep t_i, we first compute the exemplar-guided velocity difference based on the source trajectory. We then aggregate attribute-wise velocity modulations across all $a \in \mathcal{A}$, followed by a first-order integration step to update the MaterialFlow latent $Z_{t_i}^{\text{MF}}$. The final edited latent Z_0^{MF} is obtained after completing the reverse-time iteration.

4 Experiments

To systematically evaluate the proposed MaterialFlow, we conduct experiments on the **Material Transfer Benchmark (MTB)** [13] under a unified training-free evaluation protocol. Specifically, all compared methods operate without any additional fine-tuning. Pretrained weights of the backbone generative model [15, 23, 27] and optional auxiliary modules (*e.g.*, IP-Adapter [40]) are allowed to be loaded, but no parameter updates are performed during evaluation. This protocol ensures a fair comparison with our approach, which directly operates on the native generative model without introducing any additional modules.

To further enable objective and scalable evaluation, we introduce a VQA-based evaluation metric that assesses whether the generated images correctly reflect the target material attributes while preserving the original object semantics. Together with a complementary user study, these metrics provide both automatic and human-centered perspectives for quantitative comparison across different methods.

4.1 Experimental Setup

Dataset. The MTB dataset is a real-world benchmark constructed from images collected from Unsplash [31]. It contains 30 object images and 30 reference material images, forming diverse object–material pairs for evaluating material transfer in realistic scenarios.

Implementation Details. We implement MaterialFlow based on the FLUX.1-dev model [15]. During inference, the number of diffusion timesteps is set to 28, while editing is performed within a predefined interval corresponding to 10 effective steps ($n_{\max} - n_{\min}$).

For training-free baselines (IP-Adapter+SDXL [23, 40], ZeST [7], and MaterialFusion [11]), we follow the official implementations, adopting recommended settings and loading pretrained auxiliary modules when required. For methods that originally require additional training data for fine-tuning (ProSpect [45] and MaTe [13]), we follow their inference pipelines but disable additional fine-tuning, ensuring that all methods are evaluated under a unified training-free setting.

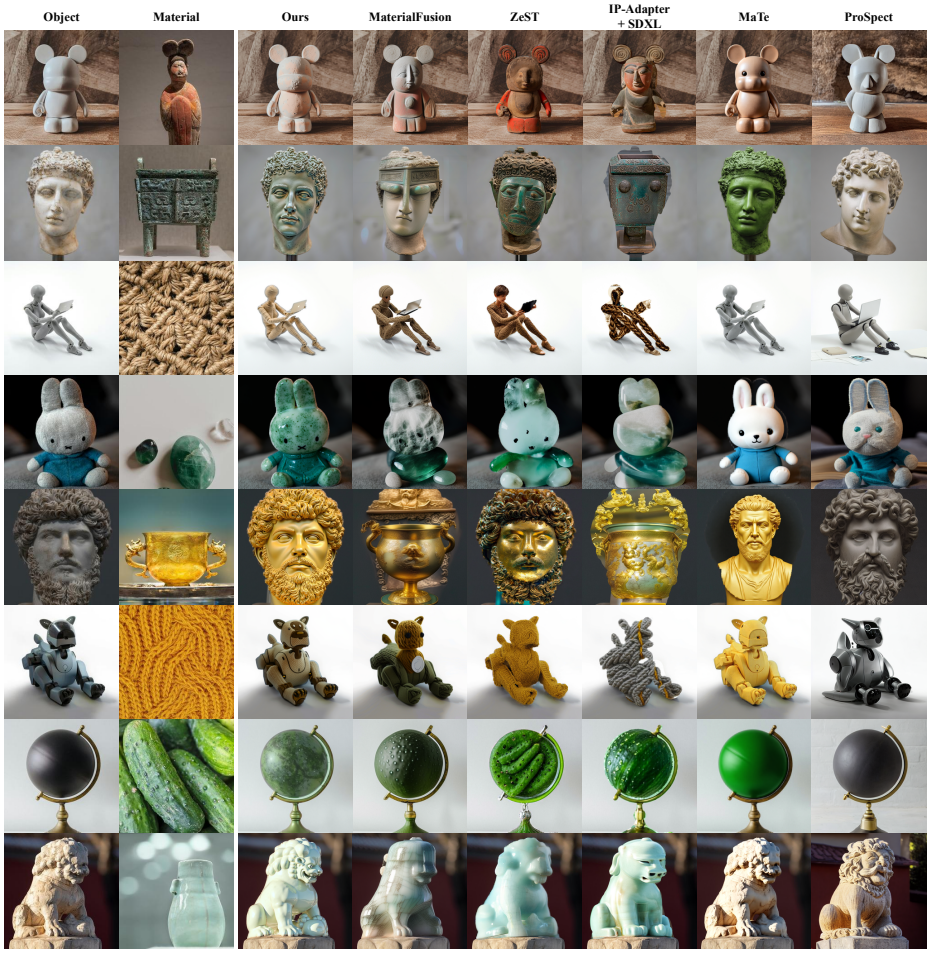

Fig. 5: Qualitative comparisons. The first column shows the source object image, and the second column shows the reference material exemplar.

4.2 Qualitative Results

Figure 5 presents qualitative comparisons between MaterialFlow and several baseline methods, including MaterialFusion [11], ZeST [7], IP-Adapter+SDXL [23, 40], MaTe [13], and ProSpect [45]. As shown in the figure, our approach achieves effective material transfer while preserving fine-grained object details.

Methods that rely on auxiliary conditioning modules, such as MaterialFusion, ZeST, and IP-Adapter, can strongly capture appearance cues from the reference material. However, this often comes at the cost of degrading the structural details of the original object. For instance, the facial features of the human sculpture and the clothing details of the rabbit toy are noticeably distorted. In contrast, methods such as MaTe and ProSpect normally require data-driven optimization (*e.g.*,

Table 1: Quantitative comparisons. The table reports two types of evaluation metrics: (1) VQA-based transfer quality, including Color, Texture, and Pattern alignment, where Avg. denotes their average; (2) user study results, where Object indicates object alignment, Material indicates material alignment, and Overall indicates overall performance. Aux. denotes whether additional auxiliary modules are used.

	Transfer Quality				User Study			Aux.
	Color↑	Texture↑	Pattern↑	Avg.↑	Object↑	Material↑	Overall↑	
IP-Adapter + SDXL [40]	41.11%	39.00%	32.22%	37.44%	9.52%	6.90%	6.19%	✓
ProSpect [45]	21.44%	12.33%	5.78%	13.18%	22.38%	0.71%	1.67%	–
ZeST [7]	58.11%	**57.00%**	**49.56%**	54.89%	54.76%	**22.62%**	48.57%	✓
MaterialFusion [11]	41.33%	38.00%	31.78%	37.04%	33.33%	13.10%	28.81%	✓
MaTe [13]	67.00%	40.00%	20.44%	42.48%	68.81%	2.14%	10.00%	–
MaterialFlow (Ours)	**79.89%**	<u>42.67%</u>	<u>34.78%</u>	**52.45%**	**83.81%**	<u>20.48%</u>	**59.29%**	–

fine-tuning or embedding refinement). Under our unified training-free evaluation setting, their performance becomes unstable, indicating strong dependence on training resources.

MaterialFlow does not rely on attention-based cross-image correspondence to associate object and material images. Instead, our method guides the editing dynamics through velocity-field manipulation in latent space, preventing material patterns from being attached to semantically inconsistent regions. This behavior is evident in the first row of Fig. 5, where other methods impose the pattern onto the bear model, while our approach preserves the object structure and produces more coherent material adaptation.

4.3 Quantitative Comparisons

Transfer Quality. Previous works on material transfer often evaluate using generation quality metrics such as PSNR, SSIM [34], LPIPS [44], and CLIP similarity [25], where the edited image is compared with the source object via reconstruction metrics and with the reference material using CLIP similarity. However, these metrics do not fully reflect the objective of material transfer, which emphasizes fine-grained attribute alignment rather than low-level similarity. Additional analysis and representative cases are provided in the supplementary material.

To address this limitation, inspired by prior visual question answering (VQA) based evaluation paradigms [17], we introduce a VQA-based protocol for measuring transfer quality. Specifically, we employ Qwen3-VL-30B-A3B [1] as the evaluation model. Given the source object image, reference material exemplar, and edited result, the model evaluates alignment across three attributes: color, texture, and structural patterns, while verifying that the object structure is preserved. We report the proportion of samples achieving high alignment for each attribute and compute their average as the overall Transfer Quality score. Prompt details and additional evaluation results using the proprietary GPT-5.5 model [22] are provided in the supplementary material.

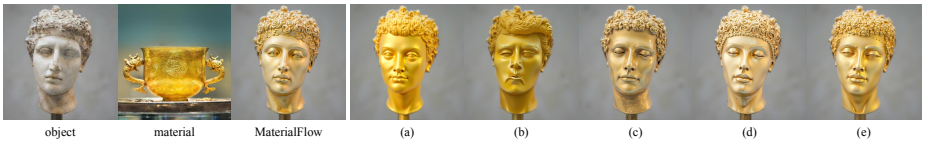

Fig. 6: Ablation study. Visualization of the effects of attribute-aware disentanglement, trajectory-aware normalization, and attribute-specific guidance.

As shown in Tab. 1, our method achieves the best average alignment score across the three attributes, demonstrating its effectiveness in producing faithful material transfer while preserving object details. Conventional generation metrics are also reported in the supplementary material for reference.

User Study. To further evaluate material transfer from a perceptual perspective, we conduct a user study. Participants are shown pairs of object images and reference material images, together with edited results from different methods. They evaluate the results from three aspects: (1) Object alignment, measuring how well object details are preserved relative to the source object image; (2) Material alignment, assessing how well the material appearance matches the reference material image; and (3) Overall performance, reflecting the overall quality considering both aspects. More details of the user study protocol are provided in the supplementary material. As shown in Tab. 1, our method achieves the best overall score of 59.29% among all compared methods, while preserving object details more faithfully than competing approaches.

4.4 Ablation Study

We visualize the effects of different components in MaterialFlow in Fig. 6. As shown in Fig. 6(a), without *Attribute-aware Disentanglement*, the exemplar material is treated as a homogeneous representation, making it difficult for the model to align with the desired attributes and leading to less controllable generation results. Fig. 6(b) shows that removing *Trajectory-aware Normalization* causes the editing dynamics to become biased toward either the object or the material, resulting in structural artifacts such as the curved and dragon-like distortions on the sculpture. A detailed quantitative ablation of these two components is provided in the supplementary material.

Fig. 6(c)–Fig. 6(e) illustrate the effect of increasing the guidance scales for the velocity-field differences of color, texture, and structural patterns, respectively. Strengthening each attribute-specific guidance enhances the corresponding material characteristics in the edited results. A more detailed quantitative analysis of attribute-level Transfer Quality is provided in the supplementary material.

5 Conclusion

We propose MaterialFlow, a training-free framework for exemplar-guided material transfer in flow models. Our approach adopts an *Exemplar-guided Inversion-free Editing* paradigm that stabilizes editing dynamics through trajectory-aware normalization, avoiding reconstruction errors inherent in inversion-based methods. Moreover, *Attribute-aware Disentanglement* separates reference materials into color, texture, and structural patterns, enabling controllable and fine-grained transfer. Experiments on the MTB dataset show that MaterialFlow achieves strong overall performance while preserving object details.

Acknowledgements

This work was partially supported by the National Science and Technology Council, Taiwan (Grants: NSTC-112-2221-E-A49-094-MY3, NSTC-112-2221-E-A49-059-MY3, NSTC-112-2628-E-002-033-MY4, and NSTC-114-2634-F-002-004), the Taiwan Centers of Excellence (TCE), and the Center of Data Intelligence: Technologies, Applications, and Systems (Grants:115L900901/115L900902/115L900903), National Taiwan University, from the Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education, Taiwan.

References

1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
2. Butt, M.A., Wang, K., Vazquez-Corral, J., Van de Weijer, J.: Colorpeel: Color prompt learning with diffusion models via color and shape disentanglement. In: European Conference on Computer Vision. pp. 456–472. Springer (2024)
3. Chen, D.Z., Siddiqui, Y., Lee, H.Y., Tulyakov, S., Nießner, M.: Text2tex: Text-driven texture synthesis via diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 18558–18568 (2023)
4. Chen, X., Feng, Y., Chen, M., Wang, Y., Zhang, S., Liu, Y., Shen, Y., Zhao, H.: Zero-shot image editing with reference imitation. Advances in Neural Information Processing Systems **37**, 84010–84032 (2024)
5. Chen, X., Huang, L., Liu, Y., Shen, Y., Zhao, D., Zhao, H.: Anydoor: Zero-shot object-level image customization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6593–6602 (2024)
6. Cheng, T.Y., Sharma, P., Boss, M., Jampani, V.: Marble: Material recomposition and blending in clip-space. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 13061–13071 (2025)
7. Cheng, T.Y., Sharma, P., Markham, A., Trigoni, N., Jampani, V.: Zest: Zero-shot material transfer from a single image. In: European conference on computer vision. pp. 370–386. Springer (2024)

8. Chung, J., Hyun, S., Heo, J.P.: Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8795–8805 (2024)
9. Deng, Y., He, X., Mei, C., Wang, P., Tang, F.: Fireflow: Fast inversion of rectified flow for image semantic editing. In: Forty-second International Conference on Machine Learning (2025), https://openreview.net/forum?id=JFafMSAjUm
10. Deng, Y., He, X., Tang, F., Dong, W.: Z*: Zero-shot style transfer via attention reweighting. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6934–6944 (2024)
11. Garifullin, K., Nikolaev, M., Kuznetsov, A., Alanov, A.: Materialfusion: High-quality, zero-shot, and controllable material transfer with diffusion models. arXiv preprint arXiv:2502.06606 (2025)
12. Hou, Y., Zeng, X., Wang, Y., Yang, M., Chen, X., Zeng, W.: Gencolor: Generative color-concept association in visual design. In: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. pp. 1–19 (2025)
13. Huang, N., Liu, H., Lin, Y., Huang, K., Chen, C., Guo, J., Lee, T.y., Li, X.: Mate: Images are all you need for material transfer via diffusion transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15117–15126 (2025)
14. Kulikov, V., Kleiner, M., Huberman-Spiegelglas, I., Michaeli, T.: Flowedit: Inversion-free text-based editing using pre-trained flow models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19721–19730 (2025)
15. Labs, B.F.: Flux. https://github.com/black-forest-labs/flux (2024)
16. Li, Y., Gao, L., Zhang, J., Zeng, P., Xiang, L., Wen, H., Shen, H.T., Song, J.: Reversible inversion for training-free exemplar-guided image editing. arXiv preprint arXiv:2512.01382 (2025)
17. Lin, Z., Pathak, D., Li, B., Li, J., Xia, X., Neubig, G., Zhang, P., Ramanan, D.: Evaluating text-to-visual generation with image-to-text generation. In: European Conference on Computer Vision. pp. 366–384. Springer (2024)
18. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: International Conference on Learning Representations (2023)
19. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: International Conference on Learning Representations (2023)
20. Lopes, I., Pizzati, F., de Charette, R.: Material palette: Extraction of materials from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4379–4388 (2024)
21. Lu, S., Liu, Y., Kong, A.W.K.: Tf-icon: Diffusion-based training-free cross-domain image composition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2294–2305 (2023)
22. OpenAI: GPT-5.5 System Card. https://openai.com/index/gpt-5-5-system-card/ (Apr 2026)
23. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: SDXL: Improving latent diffusion models for high-resolution image synthesis. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=di52zR8xgf

24. Qin, J., Gomez-Villa, A., Li, S., Yang, S., Wang, Y., Wang, K., van de Weijer, J.: Free-lunch color-texture disentanglement for stylized image generation. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025), https://openreview.net/forum?id=e0ltwNcoIp
25. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
26. Richardson, E., Metzer, G., Alaluf, Y., Giryes, R., Cohen-Or, D.: Texture: Text-guided texturing of 3d shapes. In: ACM SIGGRAPH 2023 conference proceedings. pp. 1–11 (2023)
27. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
28. Rout, L., Chen, Y., Ruiz, N., Caramanis, C., Shakkottai, S., Chu, W.S.: Semantic image inversion and editing using rectified stochastic differential equations. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=Hu0FSOSEyS
29. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22500–22510 (2023)
30. Tsai, S.L., Huang, B.L., Shen, Y.T., Yeo, C.Y., Tseng, C., Ruan, B.K., Lien, W.S., Shuai, H.H.: Color me correctly: Bridging perceptual color spaces and text embeddings for improved diffusion generation. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 10074–10082 (2025)
31. Unsplash Inc.: Unsplash. https://unsplash.com
32. Vainer, S., Boss, M., Parger, M., Kutsy, K., De Nigris, D., Rowles, C., Perony, N., Donné, S.: Collaborative control for geometry-conditioned pbr image generation. In: European Conference on Computer Vision. pp. 339–357. Springer (2024)
33. Wang, J., Pu, J., Qi, Z., Guo, J., Ma, Y., Huang, N., Chen, Y., Li, X., Shan, Y.: Taming rectified flow for inversion and editing. In: Forty-second International Conference on Machine Learning (2025), https://openreview.net/forum?id=uDreZphNky
34. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
35. Wu, Y., Liu, K., Mi, X., Tang, F., Cao, J., Li, J.: 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)
36. Xie, C., Li, M., Li, S., Wu, Y., Yi, Q., Zhang, L.: DNAEdit: Direct noise alignment for text-guided rectified flow editing. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025), https://openreview.net/forum?id=JxBA9OJExP
37. Xu, P., Jiang, B., Hu, X., Luo, D., He, Q., Zhang, J., Wang, C., Wu, Y., Ling, C., Wang, B.: Unveil inversion and invariance in flow transformer for versatile image editing. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 28479–28489 (2025)

38. Yan, Z., Ma, Y., Zou, C., Chen, W., Chen, Q., Zhang, L.: Eedit: Rethinking the spatial and temporal redundancy for efficient image editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17474–17484 (2025)
39. Yang, B., Gu, S., Zhang, B., Zhang, T., Chen, X., Sun, X., Chen, D., Wen, F.: Paint by example: Exemplar-based image editing with diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18381–18391 (2023)
40. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)
41. Yeh, Y.Y., Huang, J.B., Kim, C., Xiao, L., Nguyen-Phuoc, T., Khan, N., Zhang, C., Chandraker, M., Marshall, C.S., Dong, Z., et al.: Texturedreamer: Image-guided texture synthesis through geometry-aware diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4304–4314 (2024)
42. Yu, Q., Chow, W., Yue, Z., Pan, K., Wu, Y., Wan, X., Li, J., Tang, S., Zhang, H., Zhuang, Y.: Anyedit: Mastering unified high-quality image editing for any idea. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 26125–26135 (2025)
43. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: IEEE International Conference on Computer Vision (2023)
44. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: 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)
45. Zhang, Y., Dong, W., Tang, F., Huang, N., Huang, H., Ma, C., Lee, T.Y., Deussen, O., Xu, C.: Prospect: Prompt spectrum for attribute-aware personalization of diffusion models. ACM Transactions on Graphics (ToG) **42**(6), 1–14 (2023)
46. Zhao, C., Cai, W., Dong, C., Hu, C.: Wavelet-based fourier information interaction with frequency diffusion adjustment for underwater image restoration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8281–8291 (2024)
47. Zhou, Z., Deng, Y., He, X., Dong, W., Tang, F.: Multi-turn consistent image editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15792–15801 (2025)