OmniColor: A Unified Framework for Multi-modal Lineart Colorization

Xulu Zhang[1], Haoqian Du[1], Xiao-Yong Wei[1,2⋆], and Qing Li[1]

[1] PolySmart Group, The Hong Kong Polytechnic University, HKSAR 999077, China
[2] Sichuan University, Chengdu 610000, China
{compxulu.zhang, haoqian.du}@connect.polyu.hk
{x1wei, qing-prof.li}@polyu.edu.hk

Abstract. Lineart colorization is a critical stage in professional content creation, yet achieving precise and flexible results under diverse user constraints remains a significant challenge. To address this, we propose OmniColor, a unified framework for multi-modal lineart colorization that supports arbitrary combinations of control signals. Specifically, we systematically categorize guidance signals into two types: spatially-aligned conditions and semantic-reference conditions. For spatially-aligned inputs, we employ a dual-path encoding strategy paired with a Dense Feature Alignment loss to ensure rigorous boundary preservation and precise color restoration. For semantic-reference inputs, we utilize a VLM-only encoding scheme integrated with a Temporal Redundancy Elimination mechanism to filter repetitive information and enhance inference efficiency. To resolve potential input conflicts, we introduce an Adaptive Spatial-Semantic Gating module that dynamically balances multi-modal constraints. Experimental results demonstrate that OmniColor achieves superior controllability, visual quality, and temporal stability, providing a robust and practical solution for lineart colorization. The source code is available at https://github.com/zhangxulu1996/OmniColor.

Keywords: Lineart colorization · Multi-modal generation · Image synthesis

1 Introduction

Lineart colorization aims to synthesize high-fidelity visual content from sparse hand-drawn structures [4, 32, 42]. This technology empowers a wide range of applications, including animation production, comic creation, and game design [5, 18, 24, 33]. In this paper, we specifically focus on advancing the capabilities of controllable multi-modal lineart colorization to meet the rigorous demands of real-world creative environments.

The primary challenge in lineart colorization lies in maintaining character identity consistency and stylistic continuity across a sequence of frames [23, 39].

⋆ Corresponding Author

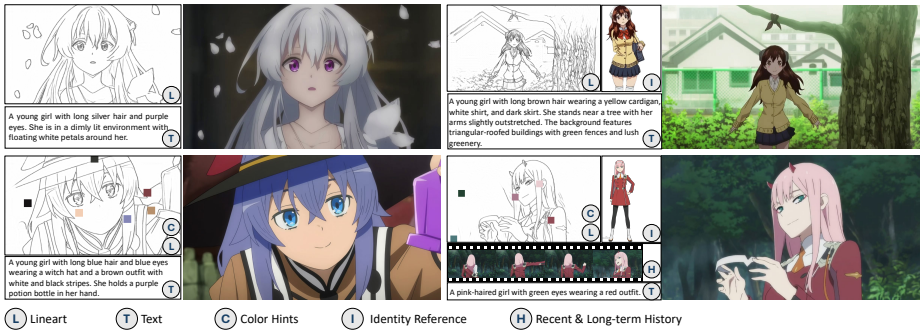

Fig. 1: Demonstration of OmniColor's unified colorization capabilities. From a single lineart (L), our model can generate diverse, high-quality colorized results by integrating different combinations of text (T), color hints (C), identity reference (I), and temporal history (H).

Relying solely on structural lineart guidance is often insufficient to capture a creator's nuanced intent or to ensure global coherence. Recent studies on customized generation further highlight the importance of incorporating user-specific guidance, such as identity, style, or personalized visual concepts, to align generated content with individual creative intent [71–73]. These advances suggest that controllable colorization should move beyond single-condition generation and support flexible combinations of personalized and structural constraints. However, existing lineart colorization methods often suffer from limited flexibility. Most frameworks are designed for one additional control signal, such as text prompts [2, 26, 69], sparse color hints [2, 16, 34, 47, 67, 68], or reference images [14, 23, 31, 32, 39, 60, 61, 70]. This lack of versatility hinders their application in complex, production-level scenes where multiple, potentially overlapping constraints are required. To bridge this gap, we present OmniColor, a unified framework that facilitates multi-modal lineart colorization through flexible combination of various control signals. This approach ensures robust colorization that meets the intricate demands of real-world creative workflows.

To effectively manage heterogeneous inputs, we categorize the guidance signals into two distinct types: spatially-aligned and semantic-reference conditions. Spatially-aligned conditions, such as linearts, user-provided color hints, and the most recent history frame provide precise pixel-level constraints on geometry and photometry. In contrast, semantic-reference conditions, including textual descriptions, long-term history frames, and identity reference maps, offer high-level guidance regarding style and character attributes without requiring exact spatial correspondence. To maintain the structural integrity and preservation of high-frequency details in spatially-aligned conditions, we employ a dual-encoder strategy that pairs a VAE [27] for capturing precise color boundaries with a Vision-Language Model (VLM) [1] for localized semantic understanding. This synergy is further reinforced by a Dense Feature Alignment (DFA) loss via DINOv3 [49] supervision, which selectively refines structural correspondences during late denoising stages to ensure artifact-free results that transcend the

limitations of compressed latent spaces. For semantic-reference conditions, we utilize a VLM-only encoding scheme, leveraging compressed semantic tokens to capture essential attributes without the computational burden of dense feature maps. To further optimize efficiency in animation sequences, a Temporal Redundancy Elimination (TRE) mechanism extracts only the unique variations across frames, significantly shortening the token sequence and accelerating inference while maintaining robust long-term consistency.

Another problem in multi-condition generation is the potential for conflicting instructions. For instance, a global style reference may contradict a local color hint. To resolve such ambiguities, we introduce an Adaptive Spatial-Semantic Gating (AS-Gate) mechanism. This module dynamically modulates the influence of semantic tokens based on the local spatial context, allowing the model to prioritize explicit user guidance when necessary. This mechanism effectively prevents artifacts such as color bleeding and ensures that the model remains highly responsive to interactive editing while maintaining a coherent overall style.

Our framework demonstrates superior performance in complex colorization tasks, achieving high-fidelity results that align with both structural constraints and semantic intent, as illustrated in Fig. 1. Through extensive experiments and user studies, we demonstrate that our method not only outperforms state-of-the-art approaches in quantitative metrics but also offers unprecedented flexibility for professional creators. By supporting diverse, mixed-modality controls, our approach provides a practical and scalable solution for the evolving needs of the digital animation and illustration industries.

2 Related Work

2.1 Lineart Colorization

Existing research on lineart colorization primarily focuses on the conditional restoration of colors and textures within the boundaries defined by hand-drawn lineart. To achieve this, modern approaches leverage diverse control signals, including text prompts [2, 26, 69], scribbles [2, 16, 38, 47, 67, 68], and reference images [14, 31, 32, 39, 43, 60, 61, 65, 70, 74, 75]. However, text-based controls typically provide only global semantic guidance and lack the localized precision. Traditional scribble-based methods, while offering chromatic information, often fail to achieve the precise color restoration required for professional-grade results. Furthermore, reference-based approaches are limited by their dependence on dense semantic matching, which hinders the synthesis of elements absent from the reference. Moreover, existing multi-frame or video-based methods [23, 36] typically employ a holistic batch-processing paradigm. This rigid structure lacks scalability, as incorporating additional sketches necessitates re-executing the entire generation pipeline. In contrast, our framework supports decoupled or joint multi-modal control, enabling precise color manipulation while preserving structural integrity. By adopting a sequential generation strategy, our method ensures historical consistency while offering superior flexibility for incremental editing, backtracking, or sequence extension.

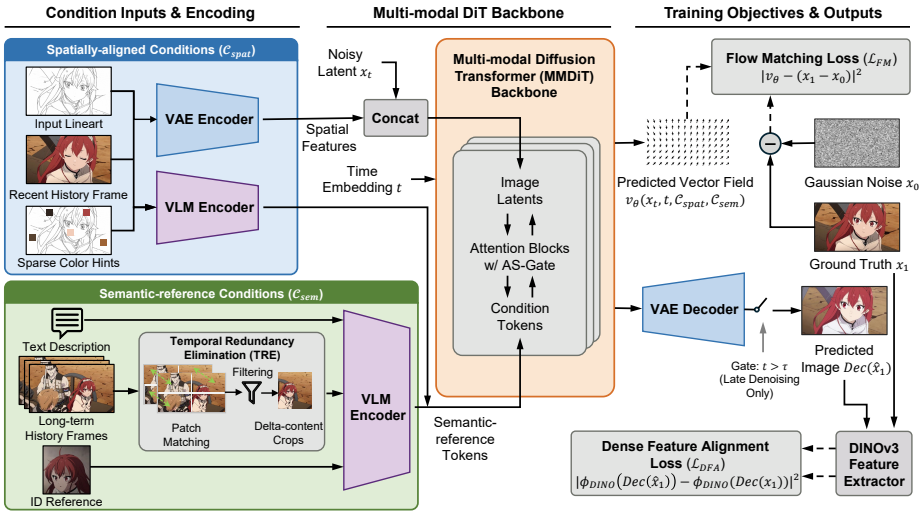

Fig. 2: The architecture of OmniColor. Our framework categorizes auxiliary signals into spatially-aligned conditions (C_{spat}) and semantic-reference conditions (C_{sem}). C_{spat} is processed via a dual-encoder strategy to preserve structural details. C_{sem} is processed via a VLM-only encoder, where a TRE module filters redundant tokens. The MMDiT backbone integrates these features with AS-Gate module to resolve input conflicts. The model is optimized using Flow Matching Loss (L_{FM}) and a DFA loss.

2.2 Unified Multi-modal Generation

Unified multi-modal generation aims to synthesize high-quality visual content by seamlessly integrating diverse input modalities within a cohesive framework. This requires models to possess robust multi-modal comprehension, specifically across textual and visual modalities. To achieve this, existing research primarily follows two distinct paths. The first involves native unified models [11, 13, 15, 41, 52,53,56,59,63,64,76], which achieve early-stage fusion of understanding and generation through large-scale end-to-end training with autoregressive or diffusion objectives. Alternatively, modular frameworks [8–10,30,37,44,50,55,58,62] bridge pre-trained vision-language models [1, 54] with generative backbones [17, 28, 45] via learnable adapters or specialized tokens. However, all frameworks rely on laborious prompt engineering or iterative refinements to resolve ambiguities and balance semantic intent in multiple conditions. In contrast, we categorize the diverse conditions inherent in lineart colorization, employ customized modules for targeted processing, and design an adaptive gating mechanism, ensuring a harmonious integration of structural constraints and semantic creativity.

3 Methodology

In this section, we present OmniColor, a unified framework for controllable and consistent lineart colorization. As shown in Fig. 2, our architecture leverages a

Multi-modal Diffusion Transformer (MMDiT) backbone [45] optimized via flow matching [35] to facilitate seamless integration of diverse control signals.

3.1 Flow Matching with Multi-modal DiT

We formulate the colorization task as a conditional flow matching problem [35]. Let x_1 be the latent representation of the ground truth (GT) image and $x_0 \sim \mathcal{N}(0, \mathbf{I})$ be Gaussian noise. We define a linear probability path $x_t = (1-t)x_0 + tx_1$, where $t \in [0, 1]$. The model v_θ is trained to predict the corresponding velocity field by minimizing the following objective:

$$\mathcal{L}_{FM} = \mathbb{E}_{t\sim\mathcal{U}(0,1),x_0,x_1} \left| v_\theta(x_t, t, \mathcal{C}_{spat}, \mathcal{C}_{sem}) - (x_1 - x_0) \right|^2 . \tag{1}$$

where $\mathcal{C}_{spat}$ and $\mathcal{C}_{sem}$ denote the spatial conditions (e.g., lineart) and semantic conditions (e.g., text), respectively. The backbone utilizes a multi-stream transformer where image latents and conditional tokens interact through joint-attention blocks, enabling bidirectional information flow between modalities.

3.2 Spatially-aligned Condition Injection

Definition of Spatially-aligned Conditions. We define $\mathcal{C}_{spat}$ as the set of conditions that require strict, pixel-level alignment with the target output. In our framework, these conditions primarily encompass the structural constraints of the input lineart, the color distribution of the most recent history frames, and user-provided sparse color hints. Unlike abstract semantic references, $\mathcal{C}_{spat}$ dictates the precise geometric and photometric boundaries that the model must respect to ensure temporal stability and structural fidelity.

Dual-Encoder Feature Extraction. To capture the multi-scale nature of these spatially-aligned signals, we employ a dual-encoder strategy designed to extract both fine-grained textures and localized semantics. Specifically, a pre-trained VAE encoder [1] is utilized to preserve high-frequency details and precise color boundaries from the spatial inputs, ensuring that the model can reconstruct sharp edges. Simultaneously, a VLM encoder [1] extracts dense, localized semantic tokens that provide a higher-level understanding of the scene layout. By fusing these two complementary representations, the MMDiT backbone can maintain rigorous structural integrity while being aware of the underlying semantic context of the spatial constraints.

Dense Feature Alignment. While VAE provides an efficient latent space for generative modeling, the standard objective $\mathcal{L}_{FM}$ in this space often provides insufficient supervision to guarantee strict spatial alignment and high-frequency structural fidelity. To bridge this gap, we propose to perform optimization in a more dense and semantically rich feature space. Specifically, we utilize DINOv3 [49] to extract dense features and introduce the Dense Feature Alignment (DFA) Loss to enforce this high-fidelity consistency. Recognizing that the early stages of the flow matching process are dominated by noise, we apply this supervision selectively during the late denoising phase ($t \to 1$). When the decoded structures

of predicted latent $\hat{x}_1$ become clear, the DFA loss minimizes the distance between the feature representations of the predicted image and the GT image:

$$\mathcal{L}_{DFA} = \mathbb{1}_{t>\tau} \cdot \left|\phi_{DINO}(\mathrm{Dec}(\hat{x}_1)) - \phi_{DINO}(\mathrm{Dec}(x_1))\right|^2 \tag{2}$$

where $\mathrm{Dec}(\cdot)$ denotes the VAE decoder, ϕ_{DINO} represents the DINO feature extractor, and τ is a time threshold (e.g., 0.7) that restricts the loss to the final refinement stages. This auxiliary objective encourages the model to refine fine-grained structural correspondences and accurate semantic-to-pixel mapping, leading to significantly sharper boundaries and reduced artifacts.

3.3 Semantic-reference Condition Injection

Definition of Semantic-reference Conditions. We define $\mathcal{C}_{sem}$ as the set of conditions that provide high-level semantic guidance without requiring pixel-wise alignment. These conditions are inherently abstract, such as textual descriptions, character identity references, and long-term historical frames. For instance, a text prompt describing a "moonlit night" requires the model to generate a specific cool-toned color palette and low-key lighting, but it does not dictate the exact color values for specific coordinates.

VLM-only Semantic Encoding. Unlike spatial conditions that necessitate a dual-encoder approach, $\mathcal{C}_{sem}$ is processed exclusively via the VLM encoder. We observe that for these abstract references, the compressed semantic tokens extracted by the VLM are sufficient to capture the necessary attributes. Besides, this VLM-only strategy offers significant computational advantages. By avoiding the generation of dense, high-resolution feature maps typically associated with VAE encoders for every reference frame, we drastically reduce the total sequence length processed by the MMDiT backbone.

Temporal Redundancy Elimination. In animation sequences, consecutive frames often contain significant redundant information. To prevent the semantic branch from being overwhelmed by repetitive tokens, we propose a redundancy elimination mechanism. We decompose reference frames into patches and compute color histogram [51] similarities between corresponding patches in adjacent frames. Patches with similarity exceeding a predefined threshold are flagged as redundant. We then perform connected component analysis on the remaining active patches. After filtering out small components, we find the minimal bounding boxes that enclose each remaining active regions. The bounding box is then used to generate a "delta-content" image. This compact representation provides complementary semantic information while further shortening the token sequence, effectively accelerating inference without sacrificing consistency.

3.4 Adaptive Spatial-Semantic Gating

In multi-modal lineart colorization, conflicts often arise between different control signals. For instance, a dense spatial constraint (e.g., a detailed lineart) might provide sufficient structural guidance, rendering certain semantic references (e.g.,

text descriptions) redundant or even contradictory. Conversely, when spatial inputs are sparse, the model must rely more heavily on semantic cues. To resolve these potential conflicts and dynamically balance the contributions of different modalities, we propose the Adaptive Spatial-Semantic Gating (AS-Gate) module. Unlike static fusion methods, AS-Gate adaptively modulates the influence of the semantic branch based on the information density of the spatial branch.

The AS-Gate is integrated into the joint attention mechanism of the MMDiT blocks. Let $\mathcal{H}_{spat} \in \mathbb{R}^{L_i \times D}$ and $\mathcal{H}_{sem} \in \mathbb{R}^{L_t \times D}$ denote the hidden states of the spatially-aligned branch and the semantic-reference branch, respectively. Following the feature recalibration paradigm [22, 45], we first derive a global spatial descriptor $\mathbf{s}_{spat}$ by pooling the image stream to capture the contextual intensity of the spatial constraints:

$$\mathbf{s}_{spat} = \text{MeanPool}\left(W_{up} \cdot \psi(W_{down} \cdot \mathcal{H}_{spat})\right) \tag{3}$$

where $W_{down} \in \mathbb{R}^{\frac{D}{r} \times D}$ and $W_{up} \in \mathbb{R}^{D \times \frac{D}{r}}$ form a low-rank bottleneck with reduction ratio r, and $\psi(\cdot)$ denotes a non-linear activation function. This descriptor is then concatenated with the semantic features to compute the gating score $\mathbf{G}_{sem}$:

$$\mathcal{G}_{sem} = W'_{up} \cdot \psi(W'_{down} \cdot [\mathcal{H}_{sem}; \mathbf{s}_{spat}]) \tag{4}$$

The gating is restricted solely to the semantic branch, recognizing that modulating the scale of a single component is equivalent to recalibrating the relative importance between spatial and semantic constraints. The final modulated semantic output $\hat{\mathbf{H}}_{sem}$ is formulated as:

$$\hat{\mathcal{H}}_{sem} = \mathcal{H}_{sem} \odot (\sigma(\mathcal{G}_{sem}) + 0.5) \tag{5}$$

where $\sigma(\cdot)$ denotes the Sigmoid function and $\odot$ is the element-wise product.

To maintain training stability and preserve the pre-trained capabilities of the foundation model, we implement these projection layers using a LoRA-based structure with zero-initialization [21]. By initializing the up-projection weights (W_{up}, W'_{up}) to zero, the term $\sigma(\mathcal{G}_{sem}) + 0.5$ evaluates to 1.0 at the start of training. This ensures that the AS-Gate acts as an identity mapping initially, allowing the model to gradually learn to scale the semantic contribution within a flexible range of $[0.5, 1.5]$ as training progresses.

4 Experiments

4.1 Dataset

Training Set. To train our unified framework, we curate a large-scale dataset featuring six types of auxiliary signals. The raw data is sourced from 25 high-quality animation series. We extract frames with resolutions exceeding 1280×720 pixels at a frequency of 1 frame per second. We employ *pySceneDetect* [6] for shot boundary detection to define temporal relationships. (1) **Spatially-aligned**

conditions include linearts, recent history frames, and sparse color hints. **Lineart** is extracted using a GAN-based framework [7] specifically optimized for anime style. **Recent history frame** is defined as the immediate preceding frame within the same shot to provide pixel-level temporal guidance. **Color hints** are simulated by randomly sampling uniform pixel blocks (minimum 10×10 pixels) from the ground truth. (2) **Semantic-reference conditions** encompass text descriptions, ID reference maps, and long-term history frames. For **textual descriptions**, we use Qwen3-VL-32B [66] to generate hierarchical captions (shot-level followed by frame-level) to ensure global-local consistency. For **ID maps**, we utilize Qwen3-VL-32B for quality filtering, SAM3 [3] for character segmentation, and CLIP features for cross-frame identity matching. **Long-term history frames** are sampled from adjacent but distinct shots. These frames are processed by the TRE module to remove redundant information.

The final dataset contains 120,000 high-quality pairs. During training, we adopt a stochastic sampling strategy with the following activation probabilities: Lineart (100%), Text (95%), Recent History (60%), Color Hints (50%), Long-term History (30%), and ID Maps (15%).

Test Set. The evaluation set is constructed from 6 unseen animation series. Following the same condition extraction pipeline as the training set, we obtained 305 shots and 900 test samples. Each sample is equipped with a lineart, a text description, and several color hints. Additionally, 136 samples include an ID reference image to evaluate identity-consistent generation. Furthermore, to facilitate the sequential generation workflow, we incorporate both recent and long-term history by utilizing either the ground truth or previously generated frames as inputs, where the immediate preceding frame and a distant earlier frame serve as the corresponding historical conditions.

4.2 Experimental Setup

Training. The parameters of the VAE encoder and the MMDiT backbone in our framework are initialized from Qwen-Image-Edit-2509 [58], and the VLM encoder adopts the pre-trained weights of Qwen2.5-VL-7B [1]. In the TRE module, we adopt a patch size of 28, set the similarity threshold to 0.85 to discard redundant patches with similarity exceeding this value, and employ a minimum connected component size of 10 to filter out small regions. The AS-Gate employs a low-rank projection structure with a LoRA rank of $r = 16$. The DFA Loss is incorporated with a weight of $\lambda_{DFA} = 0.1$ and only activated during the late denoising phase ($\tau = 0.7$ in Eq. (2). We use the AdamW [40] optimizer with a constant learning rate of 1×10^{-5} and an effective batch size of 8. We first pre-train the MMDiT backbone for 12,000 iterations, followed by 3,000 iterations to optimize the AS-Gate while freezing the backbone on 8 NVIDIA H20 GPUs.

Inference. During the inference stage, the number of inference steps is set to 50 to ensure a high-quality balance between generation fidelity and computational efficiency. We adopt Classifier-Free Guidance [20] with a scale of 4.0. All test images are generated with a resolution of 1376×768 for evaluation.

Table 1: Quantitative comparison with state-of-the-art methods.

	Conditions						Metrics					
Method	**L**	**T**	**H**	**C**	**G**	**I**	**Image-Align.↑**	**SSIM ↑**	**PSNR ↑**	**FID ↓**	**LPIPS ↓**	***ΔFC* ↓**
Text-based Lineart Colorization												
ControlNet [69]	✓	✓	×	×	×	×	0.6403	0.3079	7.7390	134.71	0.7920	10.13
Tag2pix [26]	✓	✓	×	×	×	×	0.6526	0.4525	7.4692	179.84	0.5938	3.02
Reference-based Lineart Colorization												
MagicColor [74]	✓	×	×	×	✓	×	0.7728	0.5631	10.69	118.33	0.5193	8.33
AniDoc [43]	✓	×	×	×	✓	×	0.8529	0.6638	17.08	60.47	0.3348	0.44
Manganinja [39]	✓	×	×	×	✓	×	0.9025	0.6914	16.63	46.86	0.3270	5.33
LVCD [23]	✓	×	×	×	✓	×	0.8967	0.8220	22.30	36.08	0.2699	4.65
AnimeColor [75]	✓	✓	×	×	✓	×	0.8376	0.6326	13.61	219.47	0.4286	8.07
Unified Multi-modal Generation Models												
DreamOmni2 [62]	✓	✓	×	×	×	×	0.7273	0.4046	9.01	218.95	0.7357	8.37
Flux-Kontext [30]	✓	✓	×	×	×	×	0.7448	0.4321	9.20	259.04	0.7112	2.92
Flux2-dev [29]	✓	✓	×	×	×	×	0.8061	0.4579	10.45	184.04	0.5634	3.61
Emu 3.5 [13]	✓	✓	×	×	×	×	0.8144	0.5055	11.71	187.85	0.5610	6.54
Seedream 4.5 [48]	✓	✓	×	×	×	×	0.8419	0.6073	11.13	179.87	0.5996	7.13
Nano Banana [12]	✓	✓	×	×	×	×	0.8415	0.5281	11.63	153.08	0.5466	8.28
Qwen-Image-Edit [58]	✓	✓	×	×	×	×	0.8473	0.5253	11.89	148.49	0.4682	8.34
Ours												
OmniColor	✓	✓	×	×	×	×	0.9242	0.7572	16.19	84.86	0.2817	5.44
OmniColor	✓	×	×	✓	×	×	0.9174	0.8189	19.85	86.69	0.2230	6.00
OmniColor	✓	×	×	✓	✓	×	0.9749	0.9149	29.72	29.37	0.0793	**0.09**
OmniColor	✓	✓	×	✓	×	×	0.9304	0.8121	19.47	79.64	0.2191	4.76
OmniColor	✓	✓	×	×	×	✓	0.9247	0.7591	16.31	81.78	0.2541	5.32
OmniColor	✓	✓	✓	×	×	×	0.9246	0.7611	16.38	91.49	0.2785	0.37
OmniColor	✓	✓	✓	✓	×	×	0.9310	0.8195	20.01	84.97	0.2112	0.11
OmniColor	✓	✓	✓	✓	×	✓	0.9312	0.8193	20.03	83.92	0.2115	0.64
OmniColor	✓	✓	✓	×	✓	×	**0.9769**	**0.9234**	**34.65**	**27.76**	**0.0770**	0.14

Note: L=lineart; T=text description; C=color hints; I=ID reference; H=sequential generation; G=using the first GT image of each shot as reference.

Evaluation Metrics. We evaluate the quality across two aspects: *1) Frame Similarity:* To assess the quality and fidelity of individual generated frames relative to the ground truth, we employ several widely-used metrics. We use **FID** [19] to measure visual realism and **PSNR**, **LPIPS**, and **SSIM** [57] to quantify pixel-level and perceptual reconstruction accuracy. Additionally, we calculate **Image-Align.** as the cosine similarity between the CLIP [46] embeddings of the generated images and their corresponding GT images to evaluate semantic alignment. *2) Frame Consistency:* To evaluate the temporal coherence across a sequence, we measure consistency following the methodology established in VBench [25]. Specifically, for a sequence of N frames, we extract features f_i using a pre-trained vision backbone and calculate the average cosine similarity between consecutive frames. The frame consistency FC is defined as:

$$FC = \frac{1}{N-1} \sum_{i=1}^{N-1} \frac{f_i \cdot f_{i+1}}{|f_i||f_{i+1}|} \tag{6}$$

To quantify how accurately the generated sequence mimics the temporal dynamics of the ground truth, we define the final **Consistency Error** ΔFC as the absolute difference between the consistency of GT images FC_{GT} and generated images FC_{Gen}:

$$\Delta FC = |FC_{GT} - FC_{Gen}| \tag{7}$$

A smaller ΔFC indicates that the generated sequence better preserves the semantic and temporal flow inherent in the ground truth.

4.3 Quantitative Evaluation

Comparison with the state-of-the-art. We evaluate our framework against representative lineart colorization methods published in the past two years, covering reference-guided, text-driven, and multi-modal approaches. As reported in Tab. 1, our method consistently outperforms these methods under identical input configurations. Specifically, in the L+T setting, our model surpasses all unified generation competitors by a significant margin in all metrics. This performance gain demonstrates the effectiveness of our proposed framework in bridging the gap between sparse structural inputs and high-fidelity synthesis, ensuring superior structural integrity and semantic alignment.

Multi-condition synergy. Our framework supports arbitrary combinations of heterogeneous control signals, and we show several representative configurations in Tab. 1. A prominent observation is that combining multiple modalities consistently elevates the output quality across all metrics. For instance, the introduction of sequential generation (Condition H) substantially enhances temporal stability over basic L+T inputs, reducing the ΔFC from 5.44 to 0.37. And the further integration of sparse color hints on L+T+H conditions provides additional localized refinement, leading to enhanced rendering precision. This demonstrates the scalability of our approach for complex, production-level workflows. Furthermore, utilizing the first frame of a shot as a spatial reference (Condition G) yields the highest fidelity, as this adjacent frame provides dense color information that effectively guides the diffusion process.

4.4 User Study

We conducted a user study with 10 participants (all holding a bachelor's degree or higher) to evaluate the perceptual quality of OmniColor. Each participant evaluated the full test set under the L+T conditions. For each case, results from four models were presented in a randomized order. Participants selected the preferred result separately across three dimensions: Structural Fidelity, Instruction Following, and Visual Quality. As shown in Table 2, OmniColor significantly outperforms all competitors. Notably, it achieves a dominant 81.0% preference rate in Structural Fidelity, far exceeding the second-best method (9.3%). Our model also leads in Instruction Following (42.1%) and Visual Quality (59.9%),

resulting in a 61.0% overall preference. These results validate that our specialized gating and architectural design effectively bridge the gap between rigid structural constraints and high-fidelity semantic generation.

4.5 Qualitative Evaluation

Prompt-driven Synthesis. We compare our method against Tag2pix [26] and multi-modal generation methods in Fig. 3. We observe that these models struggle to achieve perfect alignment with the lineart, particularly in multi-person scenes. In contrast, our approach maintains strict adherence to the input structures. Furthermore, our model demonstrates superior alignment with textual prompts, accurately rendering the colors for hair and garments and faithfully capturing the atmospheric lighting and background details described in the instructions.

Reference-based Colorization. Fig. 4 showcases the performance of our framework in reference-based colorization.Our method ensures high consistency between the reference and the target frames across diverse and challenging scenarios, including character rotations, dramatic motion, zoom-in, and zoom-out. Notably, our model exhibits remarkable stability even when the reference provides incomplete information. For instance, in cases where the reference only displays the character's back (e.g., the 4th column), our framework successfully infers and maintains the correct facial features and hair colors.

Multi-modal Control. Fig. 5 shows the results by integrating ID maps and sparse color hints. Compared to the L+T configuration, the inclusion of an ID reference significantly enhances the precision of character-specific attributes, such as intricate makeup and hair highlights. Moreover, our framework demonstrates exceptional sensitivity to user-provided color hints. It supports a wide range of simultaneous color constraints (up to 10 distinct hints) and effectively propagates color from sparse pixel blocks to the corresponding semantic regions.

Sequential Generation. We demonstrate the potential of OmniColor for sequential generation. By utilizing previously generated frames as historical conditions (L+T+H), our model maintains impressive temporal continuity in character outfits, makeup, and background elements, as shown in Fig. 6. When guided by the first GT frame of a shot (L+T+H+G), the synthesized sequence achieves a visual fidelity nearly indistinguishable from the ground truth. This underscores the robustness of our iterative generation strategy in mitigating error accumulation and ensuring stable, high-quality colorization for extended sequences.

Real-world Lineart Colorization. To further evaluate the practical generalization ability of OmniColor, we conduct experiments on real-world linearts collected from the Pinterest[3]. We use both the input lineart and text prompt as guidance for generation. As shown in Fig. 7, OmniColor consistently produces visually plausible and structurally faithful colorization results. The generated images preserve the original hand-drawn boundaries while assigning coherent colors to characters and background regions, demonstrating the robustness of our framework in real-world creative scenarios.

[3] `https://www.pinterest.com/search/pins/?q=lineart`

Table 2: Results of the User Study. We report the average preference rates (%) of human evaluators comparing our OmniColor against three SoTA methods across three dimensions. The best results are highlighted in **bold**.

Method	Structural Fidelity	Instruction Following	Visual Quality	Overall Pref.
Seedream 4.5 [48]	2.1	17.6	11.8	10.5
Nano Banana [12]	7.6	27.8	15.2	16.9
Qwen-Image-Edit [58]	9.3	12.5	13.1	11.6
OmniColor (Ours)	**81.0**	**42.1**	**59.9**	**61.0**

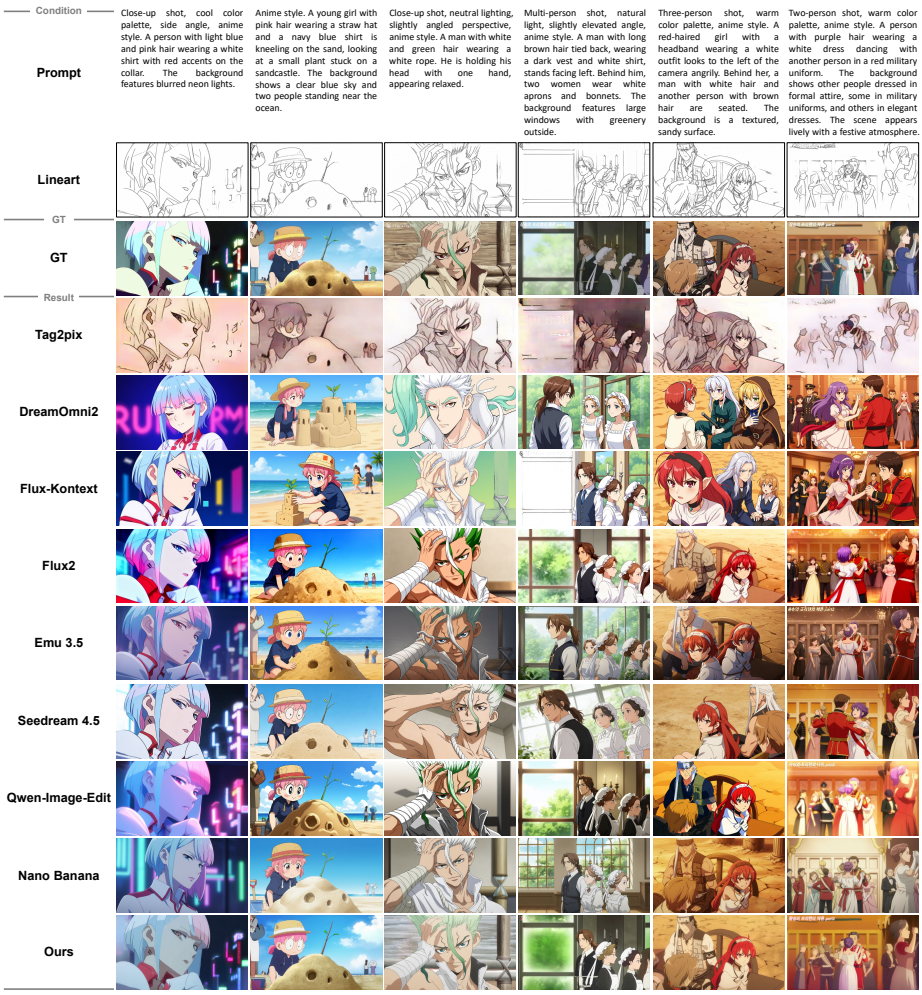

Fig. 3: Qualitative comparison of prompt-based lineart colorization. We compare our method with Tag2pix [26] and SoTA multi-modal generation models. Our framework achieves superior structural fidelity and prompt alignment.

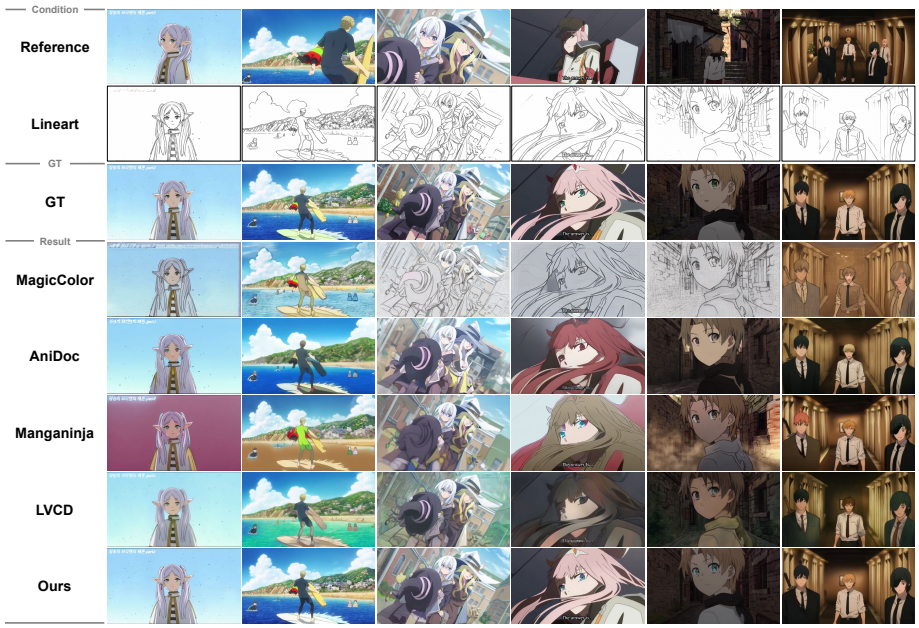

Fig. 4: Qualitative results of reference-based lineart colorization. Our method maintains robust identity and stylistic consistency even under various motions (rotation, zoom) and significant viewpoint changes.

Fig. 5: Synergistic control with multiple modalities. We show the effects of combining lineart (L) and text (T) with ID references (I) and color hints (C). The results demonstrate that our model effectively integrates multi-modal signals.

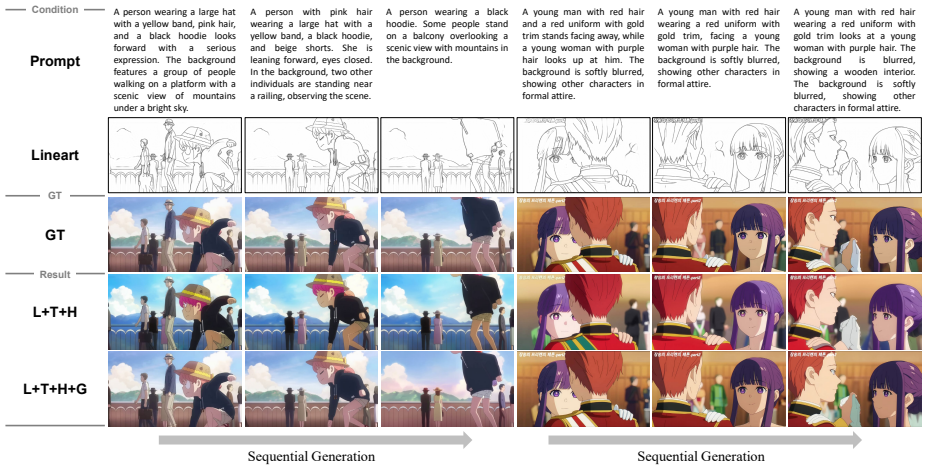

Fig. 6: Visual demonstration of sequential generation. The sequence is generated using historical frames (H) and the 1st GT frame of the target shot (G). OmniColor ensures high temporal stability and visual coherence across the entire sequence.

Fig. 7: Visual results on real-world linearts. Our method generalizes well to diverse hand-drawn styles and produces high-quality colorization results.

4.6 Ablation Study

In this section, we evaluate the contribution of each key component in the L+T+H setting. As shown in Tab. 3, adopting **VLM-only encoding** as the backbone significantly optimizes inference latency, reducing the inference time from 452s to 203s while maintaining comparable performance. Building upon this baseline, the **DFA loss** significantly enhances structural fidelity, with SSIM increasing from 0.7287 to 0.7500 and PSNR rising by 0.79. These improvements confirm that dense feature-level supervision provides stronger constraints for preserving fine-grained lineart structures and reducing local distortions during colorization. Further incorporating the **TRE** mechanism enhances perceptual quality (FID 105.90 $\rightarrow$ 97.31) by filtering redundant history, while simultaneously boosting efficiency. It reduces history tokens from 7521 to 5203, achieving a 25% speedup without compromising generation performance. This demonstrates

Table 3: Ablation study of the proposed components. #Tokens denotes the average token overhead of long-term history frames per sample.

Components				Metrics							
VLM-only	DFA	TRE	AS-Gate	Image-Align.↑	SSIM ↑	PSNR ↑	FID ↓	LPIPS ↓	ΔFC ↓	#Tokens ↓	Time (s) ↓
×	×	×	×	0.9141	0.7371	15.23	110.32	0.3052	0.74	7521	452
✓	×	×	×	0.9143	0.7287	14.93	112.29	0.3143	0.61	7521	203
✓	✓	×	×	0.9222	0.7500	15.72	105.90	0.2943	**0.17**	7521	203
✓	✓	✓	×	0.9240	0.7588	16.23	97.31	0.2805	0.56	**5203**	**150**
✓	✓	✓	✓	**0.9246**	**0.7611**	**16.38**	**91.49**	**0.2785**	0.37	**5203**	152

that TRE not only improves efficiency but also helps the model focus on informative semantic changes across frames. Finally, our full model with **AS-Gate** achieves the best performance across all quality metrics (e.g., 91.49 FID and 0.9246 Image-Align.), demonstrating that adaptive gating effectively balances spatial hints and historical semantic references.

5 Conclusion

In this paper, we present OmniColor, a unified framework designed to meet the demands of multi-modal lineart colorization. By systematically categorizing control signals into spatially-aligned and semantic-reference conditions, we provide a structured approach to managing heterogeneous inputs. Our dual-path encoding strategy, reinforced by the DFA loss, ensures high-fidelity structural integrity and precise color mapping for spatial constraints. For semantic-reference inputs, the VLM-only encoding scheme captures compact yet informative guidance from text, ID references, and historical frames. Moreover, the TRE mechanism further removes redundant temporal content, enabling efficient long-range conditioning for sequential generation. Extensive evaluations demonstrate that OmniColor outperforms existing methods in both visual quality and temporal consistency. We believe our framework offers a practical and scalable solution for the animation industry, paving the way for more intuitive AI-assisted creative workflows.

Future work can focus on further accelerating inference to better satisfy the low-latency requirements of interactive lineart colorization. Another promising direction is to extend techniques beyond strictly aligned lineart inputs toward more abstract semantic sketches, thereby broadening its applicability to early-stage creative design.

Acknowledgements

This work has been supported by the Hong Kong Research Grants Council under the Theme-based Research Scheme (project no. T41-517/25-N), and the National Natural Science Foundation of China (Grant No.: 62372314) The experimental part of this work was supported by The Centre for Large AI Models (CLAIM) of The Hong Kong Polytechnic University.

References

1. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
2. Cao, R., Mo, H., Gao, C.: Line art colorization based on explicit region segmentation. In: Computer Graphics Forum. vol. 40, pp. 1–10. Wiley Online Library (2021)
3. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025)
4. Carrillo, H., Clément, M., Bugeau, A., Simo-Serra, E.: Diffusart: Enhancing line art colorization with conditional diffusion models. In: CVPR. pp. 3486–3490 (2023)
5. Casey, E., Pérez, V., Li, Z.: The animation transformer: Visual correspondence via segment matching. In: ICCV. pp. 11323–11332 (2021)
6. Castellano, B.: PySceneDetect, `https://github.com/Breakthrough/PySceneDetect`
7. Chan, C., Durand, F., Isola, P.: Learning to generate line drawings that convey geometry and semantics. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7915–7925 (2022)
8. Chen, J., Xu, Z., Pan, X., Hu, Y., Qin, C., Goldstein, T., Huang, L., Zhou, T., Xie, S., Savarese, S., et al.: Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. arXiv preprint arXiv:2505.09568 (2025)
9. Chen, L., Bai, S., Chai, W., Xie, W., Zhao, H., Vinci, L., Lin, J., Chang, B.: Multimodal representation alignment for image generation: Text-image interleaved control is easier than you think. arXiv preprint arXiv:2502.20172 (2025)
10. Chen, X., Zhang, Z., Zhang, H., Zhou, Y., Kim, S.Y., Liu, Q., Li, Y., Zhang, J., Zhao, N., Wang, Y., et al.: Unireal: Universal image generation and editing via learning real-world dynamics. In: CVPR. pp. 12501–12511 (2025)
11. Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C.: Janus-pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811 (2025)
12. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
13. Cui, Y., Chen, H., Deng, H., Huang, X., Li, X., Liu, J., Liu, Y., Luo, Z., Wang, J., Wang, W., et al.: Emu3.5: Native multimodal models are world learners. arXiv preprint arXiv:2510.26583 (2025)
14. Dai, Y., Zhou, S., Li, Q., Li, C., Loy, C.C.: Learning inclusion matching for animation paint bucket colorization. In: CVPR. pp. 25544–25553 (2024)
15. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., et al.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025)
16. Dou, Z., Wang, N., Li, B., Wang, Z., Li, H., Liu, B.: Dual color space guided sketch colorization. TIP **30**, 7292–7304 (2021)
17. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: ICML (2024)

18. Eva, Y., Jamarun, N., Leo, A., Malik, C., et al.: Visual game character design to engage generation z in an effort to develop anti-corruption behavior in indonesian society. Journal of Urban Culture Research **25**, 64–82 (2025)
19. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. NeurIPS **30** (2017)
20. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
21. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
22. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR. pp. 7132–7141 (2018)
23. Huang, Z., Zhang, M., Liao, J.: Lvcd: reference-based lineart video colorization with diffusion models. TOG **43**(6), 1–11 (2024)
24. Huang, Z., Zhao, N., Liao, J.: Unicolor: A unified framework for multi-modal colorization with transformer. TOG **41**(6), 1–16 (2022)
25. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: CVPR. pp. 21807–21818 (2024)
26. Kim, H., Jhoo, H.Y., Park, E., Yoo, S.: Tag2pix: Line art colorization using text tag with secat and changing loss. In: ICCV. pp. 9056–9065 (2019)
27. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
28. Labs, B.F.: Flux. `https://github.com/black-forest-labs/flux` (2024)
29. Labs, B.F.: FLUX.2: Frontier Visual Intelligence. `https://bfl.ai/blog/flux-2` (2025)
30. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., et al.: Flux.1 kontext: Flow matching for in-context image generation and editing in latent space. arXiv preprint arXiv:2506.15742 (2025)
31. Li, Y.k., Lien, Y.H., Wang, Y.S.: Style-structure disentangled features and normalizing flows for diverse icon colorization. In: CVPR. pp. 11244–11253 (2022)
32. Li, Z., Geng, Z., Kang, Z., Chen, W., Yang, Y.: Eliminating gradient conflict in reference-based line-art colorization. In: ECCV. pp. 579–596. Springer (2022)
33. Liang, Z., Li, Z., Zhou, S., Li, C., Loy, C.C.: Control color: Multimodal diffusion-based interactive image colorization: Z. liang et al. IJCV **133**(11), 7897–7923 (2025)
34. Ling, P., Chen, L., Zhang, P., Chen, H., Jin, Y.: Freedrag: Point tracking is not you need for interactive point-based image editing. arXiv preprint arXiv:2307.04684 **3**(6) (2023)
35. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
36. Liu, F.L., Fu, H., Wang, X., Ye, W., Wan, P., Zhang, D., Gao, L.: Sketchvideo: Sketch-based video generation and editing. In: CVPR. pp. 23379–23390 (2025)
37. Liu, S., Han, Y., Xing, P., Yin, F., Wang, R., Cheng, W., Liao, J., Wang, Y., Fu, H., Han, C., et al.: Step1x-edit: A practical framework for general image editing. arXiv preprint arXiv:2504.17761 (2025)
38. Liu, Y., Qin, Z., Wan, T., Luo, Z.: Auto-painter: Cartoon image generation from sketch by using conditional wasserstein generative adversarial networks. Neurocomputing **311**, 78–87 (2018)

39. Liu, Z., Cheng, K.L., Chen, X., Xiao, J., Ouyang, H., Zhu, K., Liu, Y., Shen, Y., Chen, Q., Luo, P.: Manganinja: Line art colorization with precise reference following. In: CVPR. pp. 5666–5677 (2025)
40. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
41. Ma, Y., Liu, X., Chen, X., Liu, W., Wu, C., Wu, Z., Pan, Z., Xie, Z., Zhang, H., Yu, X., et al.: Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In: CVPR. pp. 7739–7751 (2025)
42. Maejima, A., Kubo, H., Funatomi, T., Yotsukura, T., Nakamura, S., Mukaigawa, Y.: Graph matching based anime colorization with multiple references. In: SIGGRAPH. pp. 1–2 (2019)
43. Meng, Y., Ouyang, H., Wang, H., Wang, Q., Wang, W., Cheng, K.L., Liu, Z., Shen, Y., Qu, H.: Anidoc: Animation creation made easier. In: CVPR. pp. 18187–18197 (2025)
44. Pan, X., Shukla, S.N., Singh, A., Zhao, Z., Mishra, S.K., Wang, J., Xu, Z., Chen, J., Li, K., Juefei-Xu, F., et al.: Transfer between modalities with metaqueries. arXiv preprint arXiv:2504.06256 (2025)
45. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: ICCV. pp. 4195–4205 (2023)
46. 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: ICML. pp. 8748–8763. PmLR (2021)
47. Sangkloy, P., Lu, J., Fang, C., Yu, F., Hays, J.: Scribbler: Controlling deep image synthesis with sketch and color. In: CVPR. pp. 5400–5409 (2017)
48. Seedream, T., Chen, Y., Gao, Y., Gong, L., Guo, M., Guo, Q., Guo, Z., Hou, X., Huang, W., Huang, Y., et al.: Seedream 4.0: Toward next-generation multimodal image generation. arXiv preprint arXiv:2509.20427 (2025)
49. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
50. Song, Y., Dong, W., Wang, S., Zhang, Q., Xue, S., Yuan, T., Yang, H., Feng, H., Zhou, H., Xiao, X., et al.: Query-kontext: An unified multimodal model for image generation and editing. arXiv preprint arXiv:2509.26641 (2025)
51. Swain, M.J., Ballard, D.H.: Indexing via color histograms. In: Active perception and robot vision, pp. 261–273. Springer (1992)
52. Team, C.: Chameleon: Mixed-modal early-fusion foundation models. arXiv preprint arXiv:2405.09818 (2024)
53. Tong, S., Fan, D., Li, J., Xiong, Y., Chen, X., Sinha, K., Rabbat, M., LeCun, Y., Xie, S., Liu, Z.: Metamorph: Multimodal understanding and generation via instruction tuning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17001–17012 (2025)
54. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
55. Wang, P., Shi, Y., Lian, X., Zhai, Z., Xia, X., Xiao, X., Huang, W., Yang, J.: Seededit 3.0: Fast and high-quality generative image editing. arXiv preprint arXiv:2506.05083 (2025)
56. Wang, X., Zhang, X., Luo, Z., Sun, Q., Cui, Y., Wang, J., Zhang, F., Wang, Y., Li, Z., Yu, Q., et al.: Emu3: Next-token prediction is all you need. arXiv preprint arXiv:2409.18869 (2024)

57. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. TIP **13**(4), 600–612 (2004)
58. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
59. Wu, C., Chen, X., Wu, Z., Ma, Y., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C., et al.: Janus: Decoupling visual encoding for unified multimodal understanding and generation. In: CVPR. pp. 12966–12977 (2025)
60. Wu, S., Yan, X., Liu, W., Xu, S., Zhang, S.: Self-driven dual-path learning for reference-based line art colorization under limited data. IEEE Transactions on Circuits and Systems for Video Technology **34**(3), 1388–1402 (2023)
61. Wu, S., Yang, Y., Xu, S., Liu, W., Yan, X., Zhang, S.: Flexicon: Flexible icon colorization via guided images and palettes. In: MM. pp. 8662–8673 (2023)
62. Xia, B., Peng, B., Zhang, Y., Huang, J., Liu, J., Li, J., Tan, H., Wu, S., Wang, C., Wang, Y., et al.: Dreamomni2: Multimodal instruction-based editing and generation. arXiv preprint arXiv:2510.06679 (2025)
63. Xie, J., Mao, W., Bai, Z., Zhang, D.J., Wang, W., Lin, K.Q., Gu, Y., Chen, Z., Yang, Z., Shou, M.Z.: Show-o: One single transformer to unify multimodal understanding and generation. arXiv preprint arXiv:2408.12528 (2024)
64. Xie, J., Yang, Z., Shou, M.Z.: Show-o2: Improved native unified multimodal models. arXiv preprint arXiv:2506.15564 (2025)
65. Yan, D., Yuan, L., Wu, E., Nishioka, Y., Fujishiro, I., Saito, S.: Colorizediffusion: Adjustable sketch colorization with reference image and text. arXiv preprint arXiv:2401.01456 (2024)
66. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
67. Zhang, L., Li, C., Simo-Serra, E., Ji, Y., Wong, T.T., Liu, C.: User-guided line art flat filling with split filling mechanism. In: CVPR. pp. 9889–9898 (2021)
68. Zhang, L., Li, C., Wong, T.T., Ji, Y., Liu, C.: Two-stage sketch colorization. TOG **37**(6), 1–14 (2018)
69. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV. pp. 3836–3847 (2023)
70. Zhang, Q., Wang, B., Wen, W., Li, H., Liu, J.: Line art correlation matching feature transfer network for automatic animation colorization. In: WACV. pp. 3872–3881 (2021)
71. Zhang, X., Wei, X.Y., Wu, J., Zhang, T., Zhang, Z., Lei, Z., Li, Q.: Compositional inversion for stable diffusion models. In: AAAI. vol. 38, pp. 7350–7358 (2024)
72. Zhang, X., Wei, X., Hu, W., Wu, J., Wu, J., Zhang, W., Zhang, Z., Lei, Z., Li, Q.: A survey on personalized content synthesis with diffusion models. Machine Intelligence Research **22**(5), 817–848 (2025)
73. Zhang, X., Zhang, W., Wei, X., Wu, J., Zhang, Z., Lei, Z., Li, Q.: Generative active learning for image synthesis personalization. In: ACM MM. pp. 10669–10677 (2024)
74. Zhang, Y., Ma, Y., Wang, B., Chen, Q., Wang, Z.: Magiccolor: Multi-instance sketch colorization. In: ICCV. pp. 15205–15217 (2025)
75. Zhang, Y., Wang, L., Wang, H., Wu, D., Lin, Z., Wang, F., Song, L.: Animecolor: Reference-based animation colorization with diffusion transformers. In: ACM MM. pp. 6682–6690 (2025)
76. Zhou, C., Yu, L., Babu, A., Tirumala, K., Yasunaga, M., Shamis, L., Kahn, J., Ma, X., Zettlemoyer, L., Levy, O.: Transfusion: Predict the next token and diffuse images with one multi-modal model. arXiv preprint arXiv:2408.11039 (2024)