BRAVER: ANIMATING RGB REFERENCES INTO REUSABLE RGBA VIDEOS

Anonymous authors
Paper under double-blind review

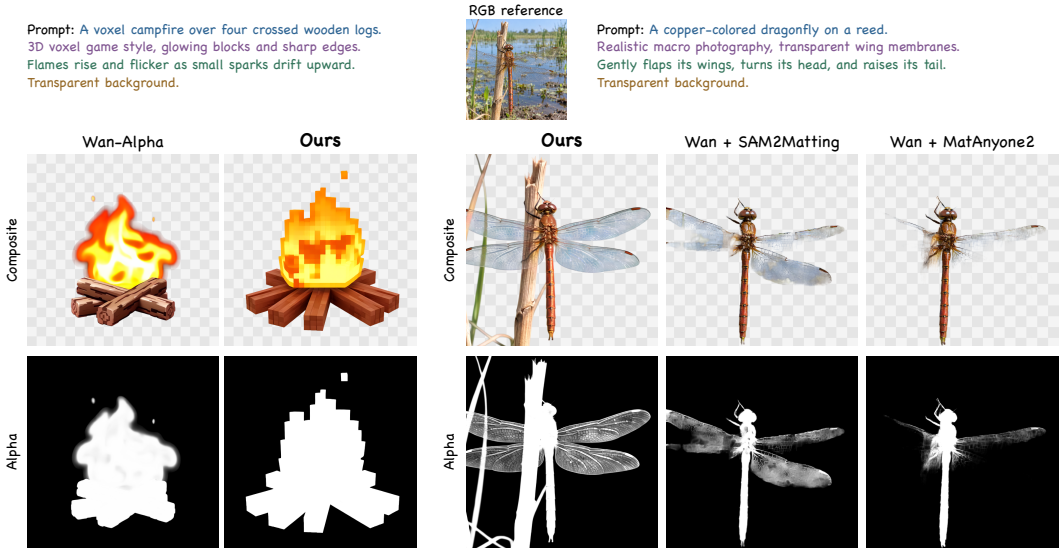


Figure 1: **Text- and RGB-guided RGBA video generation.** Left: a text-generated voxel campfire compared with Wan-Alpha. Right: the intended asset includes the dragonfly, reeds, and fine stems; the RGB-generation-plus-matting baselines produce inaccurate alpha mattes for the translucent wings. Selected frames show checkerboard composites above their corresponding alpha mattes. Translated, abbreviated prompts color-code subject, style, motion, and canvas. (§ 3.1)

ABSTRACT

Creators often begin with an image or video that captures the appearance or motion they want to preserve. Turning these references into reusable animated assets requires retaining the intended foreground while separating it from its original background, including through fine boundaries and translucent regions. This calls for both a faithful transparency representation and reference conditioning that distinguishes the subject from its surroundings. We introduce BRAVER (***B**ackground-**R**obust **A**nimation of **V**isual **E**lements for **R**euse*), a matte-free framework for text-conditioned generation, image-guided animation, and aligned foreground extraction from RGB videos. BRAVER combines a pretrained video prior with a compact RGBA autoencoder. Regional alpha supervision complements reconstruction over opaque interiors, boundaries, and translucent regions. Two-stage adaptation establishes an RGBA generation prior and incorporates visual inputs through shared adaptation and task modulation. Training references are composited over varied backgrounds while their original RGBA assets remain the targets, providing supervision for separating the foreground from its surroundings. The framework requires neither user-provided mattes nor a separate output-matting stage. We also curate ALPHAELEMENTS, an RGBA video collection spanning animated stickers, motion graphics, and visual effects, to support reconstruction and conditional generation. BRAVER outperforms the evaluated baselines in aesthetic quality and naturalness for text-conditioned generation, and reference consistency for image-guided animation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Advances in image and video generation have opened new possibilities for creating visual effects, animated stickers, and other digital assets (Esser et al., 2024; Wan Team et al., 2025; Yang et al., 2025c; Liu et al., 2025b). For these applications, however, a visually convincing RGB video is not the final product. Creators need foreground animations that can be placed into different scenes without carrying over their original backgrounds. This requires generating both foreground color and a temporally coherent alpha channel, including fine boundaries and partial transparency. The goal is therefore to produce **reusable RGBA video assets**.

Existing approaches pursue this goal through post-hoc decomposition or native RGBA generation. Generating RGB video and then applying matting (Yang et al., 2025b; 2026a; Shen et al., 2026) adds a separate inference stage; methods that require a user-supplied mask also introduce manual preparation. Automatic mask initialization reduces this burden, while recovering translucent structures and foreground colors from mixed RGB pixels remains challenging. Native RGBA generation (Wang et al., 2025a; Chen et al., 2026; Dong et al., 2026) offers a direct alternative, learning foreground appearance and transparency together without a separate matting stage. These advances establish a foundation for transparent-asset creation. For reusable assets, an important remaining goal is reliable local opacity: maintaining opaque interiors, preserving graded translucency, and keeping empty regions clear throughout the animation.

Reusable assets should also preserve the creator's visual intent. Creative work often begins with a reference image or an existing clip, which conveys appearance and motion without translating every detail into text. An RGB image serves as an **appearance anchor** for animation, whether drawn from existing material or produced by an image generator (Podell et al., 2024; Esser et al., 2024). An RGB video instead provides a foreground sequence to extract as RGBA while preserving its observed motion and temporal alignment. Text specifies content and guides generation, complementing these visual inputs. Visual-reference extensions have been explored in transparent-video generation (Wang et al., 2025a; Chen et al., 2026). Building on this progress, we seek a single RGBA generator that accepts text, RGB images, or RGB videos, without requiring users to prepare an alpha matte for visual inputs.

Reusing an asset across scenes requires preserving its foreground color and opacity structure when the background changes. This is particularly challenging at fine boundaries and in translucent regions, where visible RGB colors combine contributions from both the subject and its surroundings. Visual references thus contain both the appearance and motion to preserve and the background to leave behind. The RGBA representation must retain this opacity structure, while visual conditioning must preserve foreground evidence without copying the background. Together, these requirements support **preserving the intended asset independently of its reference background**.

We introduce BRAVER, a unified model for text-, image-, and video-conditioned RGBA generation, denoted T2RGBA, I2RGBA, and V2RGBA. BRAVER combines a pretrained video prior (Wan Team et al., 2025) with a compact RGBA autoencoder. Its reconstruction objective is supplemented by opaque-core supervision and Fine-grained Alpha Detail Enhancement (FADE), which supervises boundary and translucent regions. We first adapt the video prior to text-conditioned RGBA generation, then introduce image and video inputs through shared adaptation and task modulation. Images guide appearance during animation; videos guide aligned foreground extraction. For visual-conditioning training, we composite training assets over varied backgrounds while keeping the original RGBA videos as targets. We call this strategy Decoupling Background from Asset References (DEBAR). Supervision identifies the asset to recover from each composite. The resulting framework accepts ordinary RGB references without an input matte or a separate output-matting stage.

To support this learning, we curate ALPHAELEMENTS, a collection of Internet-sourced RGBA video assets spanning animated stickers, motion graphics, and visual effects. Its reconstruction training set contains approximately 236K videos, with a prompt-qualified generation training subset of approximately 59K video–text pairs. These data support learning both transparency reconstruction and conditional generation. Extensive experiments show that BRAVER improves alpha and premultiplied-RGB reconstruction over the baseline autoencoder (Dong et al., 2026). It outperforms the evaluated baselines in text alignment, aesthetic quality, naturalness, and temporal flickering for text-conditioned generation, and in aesthetic quality, motion smoothness, and subject consistency for image-conditioned generation.

Our contributions can be summarized as follows:

1. **RGBA video reconstruction.** We adapt the Wan-Alpha autoencoder with reconstruction and regional alpha supervision. The complete VAE improves alpha and premultiplied-RGB reconstruction.
2. **Unified conditional RGBA generation.** We build a shared framework for text-conditioned generation, image-guided animation, and aligned video foreground extraction, using background-composited reference training without user-provided mattes.
3. **Dataset and results.** We curate ALPHAELEMENTS for RGBA reconstruction and conditional generation. BRAVER outperforms the evaluated baselines in aesthetic quality and naturalness for text-conditioned generation, and reference consistency for image-guided animation.

2 RELATED WORK

Native transparent and layered generation. Transparent image generation uses latent offsets (Zhang & Agrawala, 2024), alpha-aware autoencoding (Wang et al., 2025b), training-free sampling (Quattrini et al., 2025; Morita et al., 2025), and layered synthesis or decomposition supported by dedicated datasets (Zhang et al., 2023; Huang et al., 2024a; Tudosiu et al., 2024; Pu et al., 2025; Chen et al., 2025; Huang et al., 2025; 2026; Kang et al., 2025; Yang et al., 2025a; Yin et al., 2026). For video, TransPixeler pairs RGB and alpha tokens (Wang et al., 2025a), while TransAnimate extends LayerDiffuse with AnimateDiff motion modules (Chen et al., 2026; Guo et al., 2024). Other approaches learn native RGBA video representations (Dong et al., 2026; Li et al., 2025; Bai et al., 2025; Zhang et al., 2025; Ji et al., 2025) or generate multiple layers (Li et al., 2026b; Han et al., 2026). Visual-effect generators instead produce flattened RGB clips (Liu et al., 2025b; Mao et al., 2026; Li et al., 2026a). BRAVER uses one compact RGBA target sequence; Table 1 compares supported inputs.

Table 1: **Supported tasks and input requirements.** All methods output RGBA video without an input matte; rows follow publication year and include task-specific variants. ✓: reported; —: not reported; Ann.: announced but unreleased. †: code not publicly released. (§ 2)

		Condition		
Method		Text	RGB image	RGB video
TransPixeler	(2025a)	✓	Ann.	—
LayerFlow	(2025)	✓	—	✓
TransAnimate†	(2026)	✓	✓	—
Wan-Alpha	(2026)	✓	Ann.	—
BRAVER		✓	✓	✓

RGB-conditioned transparent video. TransPixeler describes an I2V extension with optional alpha input, but its official I2V weights remain unreleased (Wang et al., 2025a); TransAnimate uses RGB-only SparseCtrl signals (Chen et al., 2026). TransText animates glyph references with binary structural guidance (Zhang et al., 2026). Concurrent work targets game assets (Li et al., 2026c), while Wan-Alpha announces an unreleased I2V variant (Dong et al., 2026). RGB sticker animation lacks alpha output (Yan et al., 2024; Sinha et al., 2024; Yuan et al., 2025b;c). BRAVER addresses background ambiguity in ordinary RGB references by training on background-composited conditions paired with the original RGBA assets.

Reference-conditioned video, insertion, and matting. Video models condition on a key frame (Blattmann et al., 2023; Xing et al., 2024; Yang et al., 2025c; Wan Team et al., 2025), subject identity (Yuan et al., 2025a; Liu et al., 2025a; Jiang et al., 2025; Fei et al., 2025; Hu et al., 2025; Deng et al., 2026), or insert subjects into a target scene (Tu et al., 2025; Zhuang et al., 2025; Yang et al., 2026b; Kara et al., 2026). Their RGB outputs require decomposition for transparent reuse. Building on alpha compositing (Porter & Duff, 1984; Smith & Blinn, 1996), image matting uses trimaps or priors (Chuang et al., 2001; Levin et al., 2008; Xu et al., 2017), clean backgrounds (Sengupta et al., 2020), or trimap-free and promptable models (Ke et al., 2022; Yao et al., 2024; Li et al., 2024). Video matting estimates temporally coherent alpha (Lin et al., 2021; 2022; Yang et al., 2025b; 2026a; Ge et al., 2025; Lim et al., 2026; Shen et al., 2026); omnimatte decomposition also recovers object effects (Lu et al., 2021; Kasten et al., 2021; Lin et al., 2023; Lee et al., 2025; Samuel et al., 2025; Hu et al., 2026). Generation followed by matting must recover both foreground color and

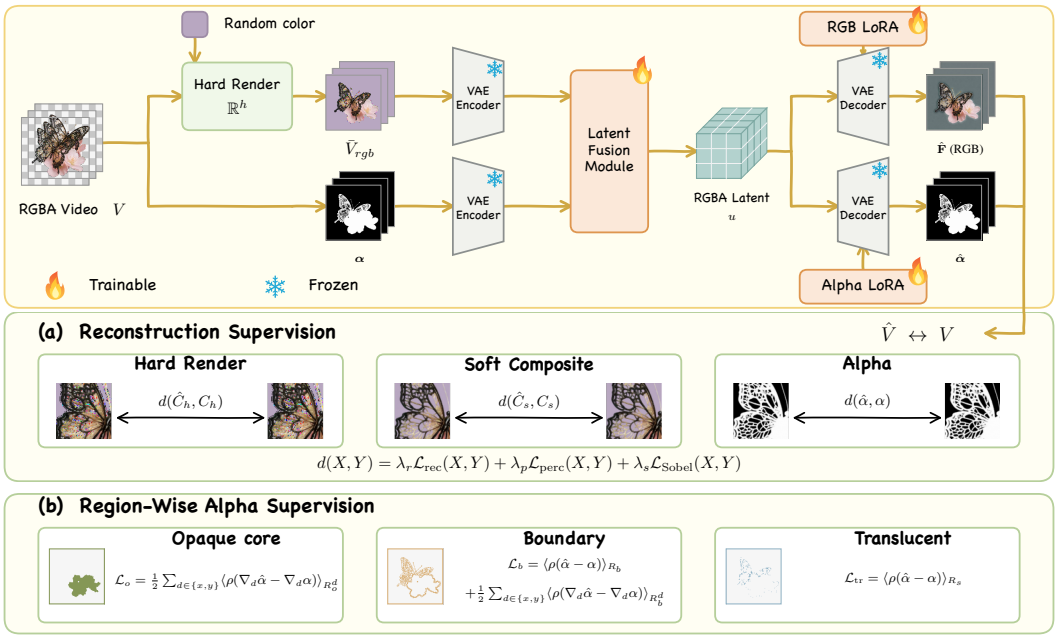


Figure 2: **RGBA VAE training.** Adapted from Wan-Alpha (Dong et al., 2026), two calls to a shared frozen encoder feed a trainable latent-fusion block; a shared latent is decoded into foreground RGB and alpha using trainable decoder adapters. Panel (a) illustrates paired reconstruction comparisons: C_h and C_s denote hard-rendered and soft-composited RGB; hats denote predictions. All three comparisons illustrate the same translucent-wing crop. The comparison d combines reconstruction, perceptual, and Sobel terms with symbolic weights. Regional supervision compares alpha values and spatial gradients on ground-truth masks. F denotes foreground RGB. (§ 3.2)

alpha from mixed RGB pixels. BRAVER predicts them jointly under RGBA supervision, including translucent regions and fine boundaries.

3 Method

3.1 Task Formulation

Input and output. Let c be a text specification for the animation and τ the task. The visual condition x is absent for T2RGBA, an ordinary RGB image for I2RGBA, or an RGB video for V2RGBA. All tasks predict the same output:

$$y = (\mathbf{F}, \boldsymbol{\alpha}) \sim p_\theta(y \mid c, x, \tau), \tag{1}$$

where $\mathbf{F}$ is straight foreground RGB and $\boldsymbol{\alpha}$ is alpha. Visual inputs contain their original backgrounds and require no user-provided matte. Images constrain appearance; videos additionally provide temporal evidence. The video-conditioned training task is aligned RGB-to-RGBA conversion of the foreground sequence.

3.2 RGBA Autoencoder

The RGBA autoencoder defines the representation used by the video generator. Following Wan-Alpha (Dong et al., 2026), it encodes foreground RGB and alpha into one 16-channel latent at each spatiotemporal location (Figure 2). We retain this architecture and train it with reconstruction and region-wise alpha supervision.

The frozen encoder is shared across RGB and alpha calls; we train the latent-fusion module and decoder adapters. The base objective $\mathcal{L}_{\text{base}}$ combines reconstruction, perceptual, and Sobel losses on hard-rendered RGB, soft-composited RGB, and alpha. The supplementary material details the rendering, encoding, decoding, and base objective.

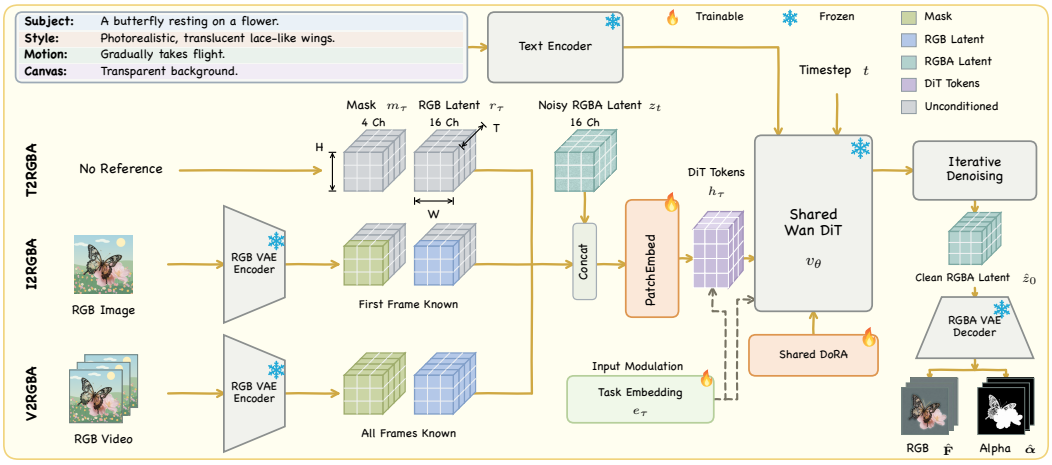


Figure 3: **Unified text-, image-, and video-conditioned RGBA generation.** Ordinary RGB references enter without prior matting. Temporal masks mark no reference, a known first frame, or a known clip. Shared PatchEmbed and task modulation feed the Wan DiT; iterative denoising and the RGBA VAE decoder produce foreground and alpha videos (Eqs. 4–6). PatchEmbed is partially trainable: its condition projection is trained, while its noisy-latent projection remains frozen. Reference-background construction is shown separately in Figure 4. (§ 3.3)

Region-wise alpha supervision. We distinguish opaque interiors R_o, boundary bands R_b, and intrinsically translucent regions R_s using ground-truth alpha. Let $e = \hat{\alpha} - \alpha$ and let $\langle \rho(e) \rangle_R$ denote the region-normalized Charbonnier error averaged over samples. In opaque interiors, we penalize spatial-gradient errors to suppress spurious alpha texture:

$$\mathcal{L}_o = \frac{1}{2} \sum_{d \in \{x,y\}} \langle \rho(\nabla_d e) \rangle_{R_o^d}, \tag{2}$$

where ∇_d is a spatial finite difference and R_o^d contains edges whose two endpoints lie in the opaque core. Alpha values remain supervised by $\mathcal{L}_{\text{base}}$.

Fine-grained alpha detail enhancement. FADE complements opaque-core supervision with boundary and translucent terms:

$$\mathcal{L}_{\text{FADE}} = \lambda_b \langle \rho(e) \rangle_{R_b} + \frac{\lambda_b}{2} \sum_{d \in \{x,y\}} \langle \rho(\nabla_d e) \rangle_{R_b^d} + \lambda_{\text{tr}} \langle \rho(e) \rangle_{R_s}. \tag{3}$$

Here R_b^d contains edges with at least one endpoint in the boundary band. The boundary terms supervise opacity and transition slopes, while the translucent term supervises intermediate alpha values. The complete VAE objective is $\mathcal{L}_{\text{base}} + \lambda_o \mathcal{L}_o + \mathcal{L}_{\text{FADE}}$. The coefficients λ_o, λ_b, and λ_{tr} weight opaque-core, boundary, and translucent supervision, respectively. Further details are provided in the supplementary material.

Latent representation for generation. The encoder and fusion module together define E_{RGBA} and produce the raw latent $u = E_{\text{RGBA}}(y)$. For DiT training, the target is $z_0 = N(u)$, where N applies the backbone's channel-wise latent normalization. This compact representation carries foreground and alpha jointly and is patchified as one video sequence. The autoencoder is trained before DiT adaptation and remains frozen within each DiT training stage.

3.3 Unified Conditional Diffusion Transformer

With the RGBA representation fixed, we adapt the pretrained Wan DiT to predict its denoising velocity under text, image, or video conditioning (Figure 3).

Noisy RGBA input. Flow matching (Lipman et al., 2023; Liu et al., 2023) interpolates between the clean target z_0 and a sampled noise tensor ϵ as $z_t = (1-t)z_0 + t\epsilon$. The DiT predicts the velocity $v^\star = \epsilon - z_0$ from z_t and the task conditions.

Condition construction. All tasks use the same 36-channel input projection. A frozen RGB VAE encoder E_{RGB} constructs a 16-channel reference latent r_τ; a four-channel temporal mask m_τ identifies known positions. T2RGBA sets both to zero. For I2RGBA, the RGB reference occupies the first frame of an otherwise zero-filled pixel-space clip, which is then encoded; the temporal mask marks only that frame as known. V2RGBA encodes the complete reference clip and marks all frames as known. The mask represents temporal availability, not foreground segmentation. The projection of the concatenated 36-channel input is implemented as

$$h_\tau = P_n(z_t) + P_c([m_\tau, r_\tau]), \tag{4}$$

where P_n is the frozen noisy-latent projection and P_c is the trainable condition projection.

Task modulation and shared adaptation. A learned task embedding e_τ modulates the resulting tokens:

$$h'_\tau = h_\tau + s_\tau \odot \mathrm{LN}(h_\tau) + b_\tau, \qquad (s_\tau, b_\tau) = W_{\mathrm{in}} e_\tau. \tag{5}$$

A second projection $W_{\mathrm{AdaLN}} e_\tau$ adds task-dependent residuals to the timestep modulation supplied to the DiT blocks. With the frozen text encoder E_{text}, the velocity is

$$v_\theta(z_t, t, c, x, \tau) = \mathrm{DiT}_{\theta_0,\Delta}(h'_\tau, E_{\mathrm{text}}(c), t; e_\tau). \tag{6}$$

The shared Wan DiT (Wan Team et al., 2025; Peebles & Xie, 2023) provides the video prior. The backbone weights θ_0 remain frozen during joint adaptation. All three tasks update the same DoRA (Liu et al., 2024) parameters Δ, condition projection, and task-modulation parameters, following parameter-efficient post-training (Hu et al., 2022).

3.4 Training

The VAE supplies fixed RGBA targets for DiT adaptation. We construct RGB references from these targets for visual conditioning, then train the generator in two stages.

Background-augmented references. DEBAR (**DE**coupling **B**ackground from **A**sset **R**eferences) constructs RGB conditions by compositing training assets over diverse backgrounds, following compositing-based augmentation (Dwibedi et al., 2017; Ghiasi et al., 2021). For foreground $\mathbf{F}_r$, alpha $\boldsymbol{\alpha}_r$, and background B, the linear-light composite is

$$x_B = \boldsymbol{\alpha}_r \odot \mathbf{F}_r + (1 - \boldsymbol{\alpha}_r) \odot B. \tag{7}$$

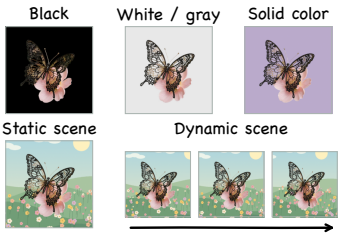


Figure 4: **DEBAR backgrounds.** One asset on varied backgrounds; a pan illustrates motion. (§ 3.4)

I2RGBA uses the first frame of the selected window; V2RGBA uses the aligned clip. Backgrounds include solid colors, static natural scenes, and temporally coherent natural videos (Figure 4). In every case the target remains the original RGBA asset: background variation affects the condition, not the desired output. The resulting RGB reference supplies x to the condition encoder, while the original RGBA clip supplies z_0 for denoising supervision.

Two-stage adaptation. We first adapt the pretrained Wan T2V model to RGBA targets using DoRA and text-only flow matching. Merging these updates into the backbone establishes a frozen RGBA video prior for the second stage. We then introduce fresh DoRA, the visual condition projection, and task modulation for joint T2RGBA, I2RGBA, and V2RGBA training. The condition projection is initialized from Wan's I2V model; the noisy-latent projection retains the merged first-stage weights.

Training objective. During joint adaptation, training alternates among the three tasks, which share the flow-matching objective

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}\left[w(t)\|v_\theta(z_t, t, c, x, \tau) - (\epsilon - z_0)\|_2^2\right]. \tag{8}$$

Text-only steps additionally regularize the prediction toward the frozen first-stage model. The complete objective and initialization details appear in the supplementary material.

Inference. Given text and an optional RGB reference, the sampler integrates the predicted velocity from standard Gaussian noise to $\hat{z}_0$. After inverse latent normalization, the frozen decoder produces $(\hat{\mathbf{F}}, \hat{\boldsymbol{\alpha}}) = D_{\mathrm{RGBA}}(N^{-1}(\hat{z}_0))$. An external T2I model can supply the image reference without joint training.

Table 2: **VAE reconstruction evaluation.** Including a FADE ablation. PSNR in dB; best scores in bold. (§ 4.2)

Method	PSNR ↑	SSIM ↑	LPIPS ↓
	Alpha		
Wan-Alpha	39.043	0.99585	0.00885
BRAVER w/o FADE	39.545	0.99599	0.00792
BRAVER	**39.948**	**0.99668**	**0.00585**
	Premultiplied RGB		
Wan-Alpha	35.557	0.97787	0.01655
BRAVER w/o FADE	35.692	0.97923	0.01542
BRAVER	**36.943**	**0.98132**	**0.01252**

Table 3: **Text-to-RGBA evaluation.** Comparison on 100 text prompts using white-background composites. Alignment and naturalness are rated from 1 to 5 by Seed2.1 Pro; the other scores use VBench. Higher is better for all metrics. (§ 4.2)

Method	Align. ↑	Aesth. ↑	Natur. ↑	Smooth. ↑	Flicker ↑
Wan2.1-T2V + MatAnyone	4.20	0.5701	4.82	0.9900	0.9874
Wan-Alpha	3.90	0.5817	4.85	**0.9930**	0.9914
LayerFlow	2.73	0.4558	2.97	0.9666	0.9647
TransPixeler	2.77	0.4039	3.66	0.9830	0.9809
BRAVER	**4.25**	**0.5969**	**4.87**	0.9928	**0.9919**

4 EXPERIMENTS

4.1 SETUP

Data and experimental settings. Training draws on VideoMatting108 (Zhang et al., 2021), Deep Video Matting (DVM) (Sun et al., 2021), and VideoMatte240K (Lin et al., 2021), together with web-collected sticker, visual-effect, and game assets curated as ALPHAELEMENTS, with asset-level deduplication and identity-level splits. The VAE trains its latent-fusion module and rank-128 decoder adapters; DiT adaptation uses rank-32 DoRA for T2RGBA, followed by fresh shared rank-64 DoRA for joint training with clips of at most 49 frames. Matched DiT ablations fix the RGBA VAE, while external generators retain their native codecs and checkpoints. Our evaluation suite covers VAE reconstruction and text-, image-, and video-to-RGBA generation. It comprises approximately 3K RGBA clips for VAE reconstruction, 100 text prompts, 100 generated RGB references with animation prompts, and 100 background-composited videos containing 50 sticker and 50 effect clips. Image-to-RGBA evaluation supplies no ground-truth input mattes; baseline pipelines estimate their own masks when needed. Training and evaluation details appear in the supplementary material.

Metrics. VAE reconstruction uses PSNR, SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018) separately on alpha and premultiplied RGB. Text-to-RGBA evaluation follows the five dimensions of Wan-Alpha (Dong et al., 2026): text alignment, aesthetic quality, naturalness, motion smoothness, and temporal flickering. VBench (Huang et al., 2024b) provides the aesthetic, smoothness, and flickering scores; Seed2.1 Pro rates subject/style alignment and naturalness from 1 to 5 under a shared rubric. All scores use white-background composites. I2RGBA additionally reports VBench-I2V subject consistency. Blinded human judgments assess overall visual quality and alpha cleanliness. Reconstruction scores apply only where spatially and temporally aligned ground truth exists, not to unconstrained future frames. The supplementary material specifies rendering, aggregation, and sampling settings.

4.2 QUANTITATIVE RESULTS

VAE reconstruction evaluation. On approximately 3K validation videos, the complete BRAVER VAE improves all six metrics over Wan-Alpha (Dong et al., 2026), gaining 0.905 dB alpha PSNR and 1.386 dB premultiplied-RGB PSNR (Table 2). Removing FADE worsens all six metrics, reducing alpha and premultiplied-RGB PSNR by 0.403 and 1.251 dB, respectively. All variants use identical evaluation clips and preprocessing with native encode/decode conventions (see supplementary material).

Text-to-RGBA evaluation. Table 3 compares Wan-Alpha (Dong et al., 2026), LayerFlow (Ji et al., 2025), the released CogVideoX-5B (Yang et al., 2025c) TransPixeler (Wang et al., 2025a), and Wan2.1-T2V (Wan Team et al., 2025) + MatAnyone (Yang et al., 2025b) with automatic mask initialization. Under method-specific settings (see supplementary material), BRAVER leads on the 100 prompts in alignment, aesthetics, naturalness, and anti-flickering. These scores measure composites, not alpha quality.

Matte-free image-to-RGBA. On 100 shared RGB references and motion prompts, BRAVER outperforms Wan2.1-I2V (Wan Team et al., 2025) followed by SAM2Matting (Shen et al., 2026), MatAny-

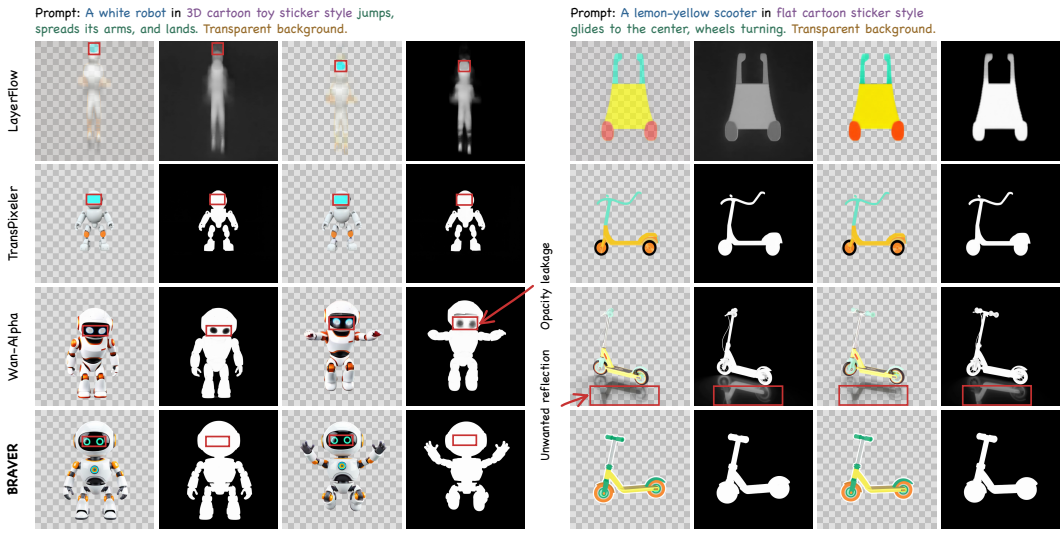


Figure 5: **Qualitative T2RGBA comparison.** LayerFlow, TransPixeler, Wan-Alpha, and **BRAVER** generate a jumping cartoon toy robot and a gliding scooter in cartoon style from corresponding method-adapted prompts. Red boxes mark the robot's eye regions across all methods and unwanted floor reflections in Wan-Alpha's scooter results, where extraneous scene content is encoded in the generated alpha. Wan-Alpha exhibits opacity leakage in the eye regions, illustrating the local-opacity issue discussed in Section 1. (§ 4.3)

Table 4: **Matte-free image-to-RGBA.** VBench scores; Subject: I2V subject consistency. Higher is better. (§ 4.2)

Method		Aesth. ↑	Smooth. ↑	Subject ↑
RGB generation + matting				
Wan2.1-I2V + SAM2Matting	(2025; 2026)	0.6004	0.9887	0.8312
Wan2.1-I2V + MatAnyone	(2025; 2025b)	0.6025	0.9885	0.8398
Wan2.1-I2V + MatAnyone 2	(2025; 2026a)	0.6060	0.9887	0.8385
Native RGBA generation				
BRAVER		**0.6266**	**0.9948**	**0.8560**

Table 5: **DiT design ablation.** Simple backgrounds use solid colors. Aesth.: aesthetic quality; Subject: I2V subject consistency; α MAD: mean absolute alpha error ($\times 10^3$). (§ 4.2)

Variant	T2V Aesth. ↑	I2V Subject ↑	V2V α MAD ↓
Simple backgrounds	0.5790	0.8528	322.251
w/o task modulation	0.5811	0.8550	41.524
BRAVER	**0.5969**	**0.8560**	**31.531**

one (Yang et al., 2025b), or MatAnyone 2 (Yang et al., 2026a) across all three metrics (Table 4). Baselines use automatic BiRefNet first-frame masks and generated RGB as foreground color. Subject consistency includes the reference background. See supplementary material for sampling details.

Video-to-RGBA evaluation. Results and settings appear in the supplementary material.

Ablation studies. Removing task modulation disables both Input FiLM and task-dependent AdaLN (Table 5). Solid-background-only training yields poor foreground separation on varied backgrounds, with V2V alpha MAD of 322.251 versus 31.531 for BRAVER.

Human preference. We recruited 43 participants to compare anonymized, left/right-randomized video pairs for overall visual quality and alpha cleanliness. Figure 6 summarizes 215 comparisons per baseline across five baselines; details appear in the supplementary material.

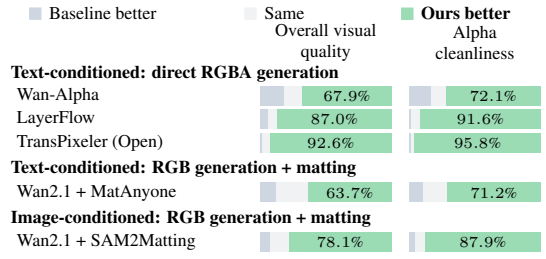


Figure 6: **GSB user preference.** Good/same/bad judgments from 43 participants; 215 comparisons per baseline. Labels show the percentage preferring ours. (§ 4.2)

4.3 Qualitative Results

Qualitative comparisons. Figures 5 and 7 compare checkerboard composites and alpha across sampled frames for text-, image-, and video-conditioned generation.

Generation from T2I references. The 100 generated RGB references in Table 4 cover characters, products, decorative elements, fine structures, and translucent effects. Animating these GPT-image-2.5 (OpenAI, 2026) outputs with BRAVER evaluates the text-to-image-to-RGBA workflow without input mattes.

Video-to-RGBA comparison. We compare BRAVER with MatAnyone 2 (Yang et al., 2026a), MatAnyone (Yang et al., 2025b), SAM2Matting (Shen et al., 2026), and VideoMaMa (Lim et al., 2026) on 50 sticker and 50 effect clips. All methods receive the same composited RGB videos; BRAVER additionally uses text, while the matting baselines use a ground-truth first-frame mask (see supplementary material). Figure 7(b) shows a pencil sticker and a progressively drawn heart arrow, comparing the recovered foreground structure across frames.

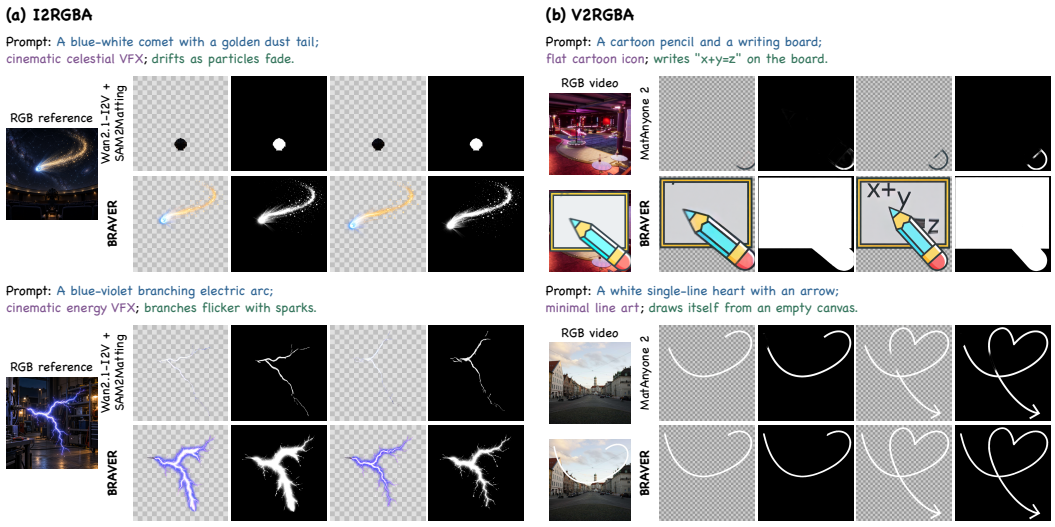


Figure 7: **Qualitative I2RGBA and V2RGBA comparisons.** (a) RGB-reference animation with Wan2.1-I2V + SAM2Matting and **BRAVER**. (b) Video-conditioned extraction of a pencil sticker and a drawn heart arrow. The shown MatAnyone 2 results omit much of the pencil asset and segments of the arrow. Initialization settings appear in the supplementary material. (§ 4.3)

5 Conclusion

We introduced BRAVER, a unified framework for creating reusable RGBA animations from text, RGB images, and RGB videos without user-provided mattes or a separate output-matting stage. Its compact RGBA representation couples foreground appearance with opacity, while shared adaptation and task modulation extend a pretrained video prior to all three tasks. Regional alpha supervision improves transparency reconstruction, and background-augmented reference training teaches the model to preserve the intended asset while separating it from its surroundings. Together with AlphaElements, these components support both reference-guided creation and video foreground extraction. Experiments show improved VAE reconstruction, higher aesthetic quality and naturalness for text-conditioned generation, and gains across all three image-conditioned metrics over the evaluated baselines. Ablations support the contributions of FADE, background diversity, and task modulation, while human judgments favor BRAVER over the evaluated baselines in visual quality and alpha cleanliness. Foreground ambiguity in cluttered references remains a source of alpha errors, and severe occlusion and refraction are difficult to resolve from image evidence alone. Future work will explore more precise foreground control, longer animations, and multiple interacting layers.

AI USE STATEMENT

We used generative AI assistants for research ideation, mathematical formulation, experimental design and analysis, coding and data preparation, literature review, writing, and figure preparation. Uses of image and video generators are described in the relevant data and experimental sections.

ETHICS STATEMENT

Transparent-video generation supports animation and compositing but may also enable misleading or unauthorized derivative content. Responsible use requires respect for privacy, consent, and intellectual-property rights. Human evaluation involved 43 externally recruited volunteers. We confirmed permissions for the research use of source data, generation models, and generated content.

REPRODUCIBILITY STATEMENT

The supplementary material provides notation and conventions, background-composited reference construction, dataset and evaluation details, metric computation, sampling settings, and training configurations.

REFERENCES

Jingqi Bai, Jingkai Zhou, Benzhi Wang, Weihua Chen, Yang Yang, Zhen Lei, and Fan Wang. LayerAnimate for transparent video generation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10887999.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. doi: 10.48550/arXiv.2311.15127.

Junwen Chen, Heyang Jiang, Yanbin Wang, Keming Wu, Ji Li, Chao Zhang, Keiji Yanai, Dong Chen, and Yuhui Yuan. PrismLayers: Open data for high-quality multi-layer transparent image generative models. *arXiv preprint arXiv:2505.22523*, 2025.

Xuewei Chen, Zhimin Chen, and Yiren Song. TransAnimate: Taming layer diffusion to generate RGBA video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 6083–6092, 2026.

Yung-Yu Chuang, Brian Curless, David H. Salesin, and Richard Szeliski. A Bayesian approach to digital matting. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pp. II–264–II–271, 2001. doi: 10.1109/CVPR.2001.990970.

Yufan Deng, Yuanyang Yin, Xun Guo, Yizhi Wang, Zhiyuan Fang, Shenghai Yuan, Yiding Yang, Angtian Wang, Bo Liu, Haibin Huang, and Chongyang Ma. MAGREF: Masked guidance for any-reference video generation with subject disentanglement. In *International Conference on Learning Representations (ICLR)*, pp. 1–32, 2026.

Haotian Dong, Wenjing Wang, Chen Li, Jing Lyu, and Di Lin. Video generation with stable transparency via shiftable RGB-A distribution learner. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1885–1894, 2026.

Debidatta Dwibedi, Ishan Misra, and Martial Hebert. Cut, paste and learn: Surprisingly easy synthesis for instance detection. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. doi: 10.1109/ICCV.2017.146.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12606–12633. PMLR, 2024.

Zhengcong Fei, Debang Li, Di Qiu, Jiahua Wang, Yikun Dou, Rui Wang, Jingtao Xu, Mingyuan Fan, Guibin Chen, Yang Li, and Yahui Zhou. SkyReels-A2: Compose anything in video diffusion transformers. *arXiv preprint arXiv:2504.02436*, 2025.

Yongtao Ge, Kangyang Xie, Guangkai Xu, Li Ke, Mingyu Liu, Longtao Huang, Hui Xue, Hao Chen, and Chunhua Shen. Generative video matting. In *ACM SIGGRAPH Conference Papers*, pp. 1–10, 2025. doi: 10.1145/3721238.3730642.

Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D. Cubuk, Quoc V. Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. doi: 10.1109/CVPR46437.2021.00294.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning. In *International Conference on Learning Representations (ICLR)*, 2024.

Kyujin Han, Seungjoo Shin, and Sunghyun Cho. Explicit layer modeling for video object insertion and layer decomposition. *arXiv preprint arXiv:2607.25802*, 2026.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Teng Hu, Zhentao Yu, Zhengguang Zhou, Sen Liang, Yuan Zhou, Qin Lin, and Qinglin Lu. HunyuanCustom: A multimodal-driven architecture for customized video generation. *arXiv preprint arXiv:2505.04512*, 2025.

Yihan Hu, Xuelin Chen, and Xiaodong Cun. EasyOmnimatte: Taming pretrained inpainting diffusion models for end-to-end video layered decomposition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 43341–43351, 2026.

Dingbang Huang, Wenbo Li, Yifei Zhao, Xinyu Pan, Yanhong Zeng, and Bo Dai. PSDiffusion: Harmonized multi-layer image generation via layout and appearance alignment. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3233–3242, 2026. doi: 10.1109/WACV61042.2026.00316.

Junjia Huang, Pengxiang Yan, Jinhang Cai, Jiyang Liu, Zhao Wang, Yitong Wang, Xinglong Wu, and Guanbin Li. DreamLayer: Simultaneous multi-layer generation via diffusion model. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3357–3366, 2025. doi: 10.1109/ICCV51701.2025.00321.

Runhui Huang, Kaixin Cai, Jianhua Han, Xiaodan Liang, Renjing Pei, Guansong Lu, Songcen Xu, Wei Zhang, and Hang Xu. LayerDiff: Exploring text-guided multi-layered composable image synthesis via layer-collaborative diffusion model. In *European Conference on Computer Vision (ECCV)*, volume 15134 of *Lecture Notes in Computer Science*, pp. 144–160, 2024a. doi: 10.1007/978-3-031-73116-7_9.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818, 2024b. doi: 10.1109/CVPR52733.2024.02060.

Sihui Ji, Hao Luo, Xi Chen, Yuanpeng Tu, Yiyang Wang, and Hengshuang Zhao. LayerFlow: A unified model for layer-aware video generation. In *ACM SIGGRAPH Conference Papers*, pp. 1–10, 2025. doi: 10.1145/3721238.3730662.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. VACE: All-in-one video creation and editing. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17191–17202, 2025. doi: 10.1109/ICCV51701.2025.01597.

Kyoungkook Kang, Gyujin Sim, Geonung Kim, Donguk Kim, Seungho Nam, and Sunghyun Cho. LayeringDiff: Layered image synthesis via generation, then disassembly with generative knowledge. *arXiv preprint arXiv:2501.01197*, 2025.

Ozgur Kara, Yujia Chen, Ming-Hsuan Yang, James M. Rehg, Wen-Sheng Chu, and Du Tran. Layer-aware video composition via split-then-merge. In *European Conference on Computer Vision (ECCV)*, volume 17028 of *Lecture Notes in Computer Science*, pp. 110–130, 2026. doi: 10.1007/978-3-032-37624-4_7.

Yoni Kasten, Dolev Ofri, Oliver Wang, and Tali Dekel. Layered neural atlases for consistent video editing. *ACM Transactions on Graphics*, 40(6), 2021. doi: 10.1145/3478513.3480546.

Zhanghan Ke, Jiayu Sun, Kaican Li, Qiong Yan, and Rynson W. H. Lau. MODNet: Real-time trimap-free portrait matting via objective decomposition. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 36, pp. 1140–1147, 2022. doi: 10.1609/aaai.v36i1.19999.

Yao-Chih Lee, Erika Lu, Sarah Rumbley, Michal Geyer, Jia-Bin Huang, Tali Dekel, and Forrester Cole. Generative omnimatte: Learning to decompose video into layers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12522–12532, 2025. doi: 10.1109/CVPR52734.2025.01168.

Anat Levin, Dani Lischinski, and Yair Weiss. A closed-form solution to natural image matting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):228–242, 2008. doi: 10.1109/TPAMI.2007.1177.

Baolu Li, Yiming Zhang, Qinghe Wang, Liqian Ma, Xiaoyu Shi, Xintao Wang, Pengfei Wan, Zhenfei Yin, Yunzhi Zhuge, Huchuan Lu, and Xu Jia. VFXMaster: Unlocking dynamic visual effect generation via in-context learning. In *ACM SIGGRAPH Conference Papers*, pp. 1–9. ACM, 2026a. doi: 10.1145/3799902.3811208.

Guangzhao Li, Kangrui Cen, Baixuan Zhao, Yi Xin, Siqi Luo, Guangtao Zhai, Lei Zhang, and Xiaohong Liu. LayerT2V: A unified multi-layer video generation framework. In *International Conference on Machine Learning (ICML)*, 2026b. URL `https://layert2v.github.io/`.

Jiachen Li, Jitesh Jain, and Humphrey Shi. Matting anything. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 1775–1785, 2024. doi: 10.1109/CVPRW63382.2024.00184.

Menghao Li, Zhenghao Zhang, Junchao Liao, Long Qin, and Weizhi Wang. TransVDM: Motion-constrained video diffusion model for transparent video synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10889873.

Zhe Li, Honghao Qiao, Zhixin Xu, Qijie Wang, Bo Peng, and Dawei Li. Alpha as an efficiency signal: Visibility-routed RGBA image-to-video generation. *arXiv preprint arXiv:2608.09355*, 2026c.

Sangbeom Lim, Seoung Wug Oh, Jiahui Huang, Heeji Yoon, Seungryong Kim, and Joon-Young Lee. VideoMaMa: Mask-guided video matting via generative prior. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3845–3855, 2026.

Geng Lin, Chen Gao, Jia-Bin Huang, Changil Kim, Yipeng Wang, Matthias Zwicker, and Ayush Saraf. OmnimatteRF: Robust omnimatte with 3D background modeling. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23471–23480, 2023.

Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian L. Curless, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. Real-time high-resolution background matting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. doi: 10.1109/CVPR46437.2021.00865.

Shanchuan Lin, Linjie Yang, Imran Saleemi, and Soumyadip Sengupta. Robust high-resolution video matting with temporal guidance. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022. doi: 10.1109/WACV51458.2022.00319.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Lijie Liu, Tianxiang Ma, Bingchuan Li, Zhuowei Chen, Jiawei Liu, Gen Li, Siyu Zhou, Qian He, and Xinglong Wu. Phantom: Subject-consistent video generation via cross-modal alignment. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14951–14961, 2025a. doi: 10.1109/ICCV51701.2025.01387.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. DoRA: Weight-decomposed low-rank adaptation. In *International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32100–32121. PMLR, 2024.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023.

Xinyu Liu, Ailing Zeng, Wei Xue, Harry Yang, Wenhan Luo, Qifeng Liu, and Yike Guo. VFX creator: Animated visual effect generation with controllable diffusion transformer. *arXiv preprint arXiv:2502.05979*, 2025b.

Erika Lu, Forrester Cole, Tali Dekel, Andrew Zisserman, William T. Freeman, and Michael Rubinstein. Omnimatte: Associating objects and their effects in video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. doi: 10.1109/CVPR46437.2021.00448.

Fangyuan Mao, Aiming Hao, Jintao Chen, Dongxia Liu, Xiaokun Feng, Jiashu Zhu, Meiqi Wu, Chubin Chen, Jiahong Wu, and Xiangxiang Chu. Omni-Effects: Unified and spatially-controllable visual effects generation. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 40, pp. 7927–7935, 2026. doi: 10.1609/aaai.v40i10.37737.

Ryugo Morita, Stanislav Frolov, Brian Bernhard Moser, Takahiro Shirakawa, Ko Watanabe, Andreas Dengel, and Jinjia Zhou. TKG-DM: Training-free chroma key content generation diffusion model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13031–13040, 2025. doi: 10.1109/CVPR52734.2025.01216.

OpenAI. Introducing ChatGPT Images 2.5, September 2026. URL `https://openai.com/index/introducing-chatgpt-images-2-5/`.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations (ICLR)*, 2024.

Thomas Porter and Tom Duff. Compositing digital images. In *Proceedings of ACM SIGGRAPH*, pp. 253–259, 1984. doi: 10.1145/800031.808606.

Yifan Pu, Yiming Zhao, Zhicong Tang, Ruihong Yin, Haoxing Ye, Yuhui Yuan, Dong Chen, Jianmin Bao, Sirui Zhang, Yanbin Wang, Lin Liang, Lijuan Wang, Ji Li, Xiu Li, Zhouhui Lian, Gao Huang, and Baining Guo. ART: Anonymous region transformer for variable multi-layer transparent image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7952–7962, 2025. doi: 10.1109/CVPR52734.2025.00745.

Fabio Quattrini, Vittorio Pippi, Silvia Cascianelli, and Rita Cucchiara. Alfie: Democratising RGBA image generation with no $$$. In *European Conference on Computer Vision (ECCV) Workshops*, volume 15627 of *Lecture Notes in Computer Science*, pp. 38–55. Springer, 2025. doi: 10.1007/978-3-031-92808-6_3.

Dvir Samuel, Matan Levy, Nir Darshan, Gal Chechik, and Rami Ben-Ari. OmnimatteZero: Fast training-free omnimatte with pre-trained video diffusion models. In *ACM SIGGRAPH Asia Conference Papers*, pp. 1–11, 2025. doi: 10.1145/3757377.3763917.

Soumyadip Sengupta, Vivek Jayaram, Brian Curless, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. Background matting: The world is your green screen. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. doi: 10.1109/CVPR42600.2020.00236.

Ruiqi Shen, Guangquan Jie, Chang Liu, and Henghui Ding. SAM2Matting: Generalized image and video matting. In *European Conference on Computer Vision (ECCV)*, pp. 295–313, 2026. doi: 10.1007/978-3-032-37132-4_17.

Animesh Sinha, Bo Sun, Anmol Kalia, Arantxa Casanova, Elliot Blanchard, David Yan, Winnie Zhang, Tony Nelli, Jiahui Chen, Hardik Shah, Licheng Yu, Mitesh Kumar Singh, Ankit Ramchandani, Maziar Sanjabi, Sonal Gupta, Amy Bearman, and Dhruv Mahajan. Text-to-sticker: Style tailoring latent diffusion models for human expression. In *European Conference on Computer Vision (ECCV)*, volume 15128 of *Lecture Notes in Computer Science*, pp. 36–52. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72897-6_3.

Alvy Ray Smith and James F. Blinn. Blue screen matting. In *Proceedings of ACM SIGGRAPH*, pp. 259–268, 1996. doi: 10.1145/237170.237263.

Yanan Sun, Guanzhi Wang, Qiao Gu, Chi-Keung Tang, and Yu-Wing Tai. Deep video matting via spatio-temporal alignment and aggregation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6975–6984, 2021. doi: 10.1109/CVPR46437.2021.00690.

Yuanpeng Tu, Hao Luo, Xi Chen, Sihui Ji, Xiang Bai, and Hengshuang Zhao. VideoAnydoor: High-fidelity video object insertion with precise motion control. In *ACM SIGGRAPH Conference Papers*, pp. 1–11. ACM, 2025. doi: 10.1145/3721238.3730647.

Petru-Daniel Tudosiu, Yongxin Yang, Shifeng Zhang, Fei Chen, Steven McDonagh, Gerasimos Lampouras, Ignacio Iacobacci, and Sarah Parisot. MULAN: A multi layer annotated dataset for controllable text-to-image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22413–22422, 2024. doi: 10.1109/CVPR52733.2024.02115.

Wan Team, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. doi: 10.48550/arXiv.2503.20314.

Luozhou Wang, Yijun Li, Zhifei Chen, Jui-Hsien Wang, Zhifei Zhang, He Zhang, Zhe Lin, and Ying-Cong Chen. TransPixeler: Advancing text-to-video generation with transparency. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18229–18239, 2025a. doi: 10.1109/CVPR52734.2025.01699.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

Zile Wang, Hao Yu, Jiabo Zhan, and Chun Yuan. AlphaVAE: Unified end-to-end RGBA image reconstruction and generation with alpha-aware representation learning. *arXiv preprint arXiv:2507.09308*, 2025b.

Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. DynamiCrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision (ECCV)*, volume 15104 of *Lecture Notes in Computer Science*, pp. 399–417, 2024. doi: 10.1007/978-3-031-72952-2_23.

Ning Xu, Brian Price, Scott Cohen, and Thomas Huang. Deep image matting. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. doi: 10.1109/CVPR.2017.41.

David Yan, Winnie Zhang, Luxin Zhang, Anmol Kalia, Dingkang Wang, Ankit Ramchandani, Miao Liu, Albert Pumarola, Edgar Schoenfeld, Elliot Blanchard, Krishna Narni, Yaqiao Luo, Lawrence Chen, Guan Pang, Ali Thabet, Peter Vajda, Amy Bearman, and Licheng Yu. Animated stickers: Bringing stickers to life with video diffusion. *arXiv preprint arXiv:2402.06088*, 2024.

Jinrui Yang, Qing Liu, Yijun Li, Soo Ye Kim, Daniil Pakhomov, Mengwei Ren, Jianming Zhang, Zhe Lin, Cihang Xie, and Yuyin Zhou. Generative image layer decomposition with visual effects. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7643–7653, 2025a. doi: 10.1109/CVPR52734.2025.00716.

Peiqing Yang, Shangchen Zhou, Jixin Zhao, Qingyi Tao, and Chen Change Loy. MatAnyone: Stable video matting with consistent memory propagation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7299–7308, 2025b. doi: 10.1109/CVPR52734.2025.00684.

Peiqing Yang, Shangchen Zhou, Kai Hao, and Qingyi Tao. MatAnyone 2: Scaling video matting via a learned quality evaluator. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 37476–37485, 2026a.

Shuzhou Yang, Xiaoyu Li, Xiaodong Cun, Guangzhi Wang, Lingen Li, Ying Shan, and Jian Zhang. GenCompositor: Generative video compositing with diffusion transformer. In *International Conference on Learning Representations (ICLR)*, pp. 65477–65501, 2026b.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations (ICLR)*, 2025c.

Jingfeng Yao, Xinggang Wang, Shusheng Yang, and Baoyuan Wang. ViTMatte: Boosting image matting with pre-trained plain vision transformers. *Information Fusion*, 103:102091, 2024. doi: 10.1016/j.inffus.2023.102091.

Shengming Yin, Zekai Zhang, Zecheng Tang, Kaiyuan Gao, Xiao Xu, Kun Yan, Jiahao Li, Yilei Chen, Yuxiang Chen, Heung-Yeung Shum, Lionel M. Ni, Junyang Lin, and Chenfei Wu. Qwen-Image-Layered: Towards inherent editability via layer decomposition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16196–16205, 2026.

Shenghai Yuan, Jinfa Huang, Xianyi He, Yunyang Ge, Yujun Shi, Liuhan Chen, Jiebo Luo, and Li Yuan. Identity-preserving text-to-video generation by frequency decomposition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12978–12988, 2025a. doi: 10.1109/CVPR52734.2025.01211.

Zhiqiang Yuan, Jiapei Zhang, Ying Deng, Yeshuang Zhu, Jie Zhou, and Jinchao Zhang. VSD2M: Large-scale vision-language sticker dataset for multi-frame animated sticker generation. In *IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2025b. doi: 10.1109/ICME59968.2025.11209368.

Zhiqiang Yuan, Ting Zhang, Peixiang Luo, Ying Deng, Jiapei Zhang, Zexi Jia, Jinchao Zhang, and Jie Zhou. RDTF: Resource-efficient dual-mask training framework for multi-frame animated sticker generation. *arXiv preprint arXiv:2503.17735*, 2025c.

Fei Zhang, Zijian Zhou, Bohao Tang, Sen He, Hang Li, Zhe Wang, Soubhik Sanyal, Pengfei Liu, Viktar Atliha, Tao Xiang, Mengmeng Xu, and Semih Günel. TransText: Alpha-as-RGB representation for transparent text animation. In *ACM International Conference on Multimedia (MM)*, 2026.

Lvmin Zhang and Maneesh Agrawala. Transparent image layer diffusion using latent transparency. *ACM Transactions on Graphics*, 43(4):100:1–100:15, 2024. doi: 10.1145/3658150.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018. doi: 10.1109/CVPR.2018.00068.

Ting Zhang, Zhiqiang Yuan, Yeshuang Zhu, Jie Zhou, and Jinchao Zhang. ILDiff: Generate transparent animated stickers by implicit layout distillation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10888420.

Xinyang Zhang, Wentian Zhao, Xin Lu, and Jeff Chien. Text2Layer: Layered image generation using latent diffusion model. *arXiv preprint arXiv:2307.09781*, 2023.

Yunke Zhang, Chi Wang, Miaomiao Cui, Peiran Ren, Xuansong Xie, Xian-sheng Hua, Hujun Bao, Qixing Huang, and Weiwei Xu. Attention-guided temporally coherent video object matting. In *ACM International Conference on Multimedia (MM)*, pp. 5128–5137, 2021. doi: 10.1145/3474085.3475623.

Shaobin Zhuang, Zhipeng Huang, Binxin Yang, Ying Zhang, Fangyikang Wang, Canmiao Fu, Chong Sun, Zheng-Jun Zha, Chen Li, and Yali Wang. Get in video: Add anything you want to the video. *arXiv preprint arXiv:2503.06268*, 2025.