UNIRGBA: UNIFYING RGB AND ALPHA IN PRE-TRAINED VIDEO DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Native alpha is essential for compositing visual effects, yet video diffusion models remain RGB-only and matting pipelines underrepresent continuous transparency. We present **UniRGBA**, an interface-preserving extension of a pretrained video diffusion model for native RGBA video. UniRGBA-VAE expands the Wan2.1 3D causal VAE from RGB to RGBA by zero-initializing the new boundary channels and updating only symmetric $2{+}2$ layers (about 5% of VAE parameters), while retaining the original 16-channel latent interface. A KL-anchored two-stage objective learns alpha structure and sharp compositing details; standard DoRA then adapts the shared Wan2.1 DiT for RGBA T2V and I2V. On a held-out visual-effect benchmark, UniRGBA-VAE improves PSNR from 25.00 to 29.16 over a retrained Wan-Alpha baseline (+4.16 dB), including +5.33 dB RGB and +3.95 dB alpha gains. RGBA T2V improves text alignment (+0.28), naturalness (+0.33), and temporal flickering (+0.0050) over Wan-Alpha, while RGB transfer exposes a measurable -0.095 Dynamic Degree cost. These results show that native alpha can be added through a small VAE-side intervention while retaining a reusable latent interface.

Keywords. native RGBA video generation, alpha-aware video diffusion, transparent video synthesis, latent video autoencoders, parameter-efficient adaptation, text-to-video and image-to-video generation

1 INTRODUCTION

Video generation has advanced rapidly with large-scale diffusion models (Ho et al., 2020; Rombach et al., 2022; Ho et al., 2022; Blattmann et al., 2023; Singer et al., 2022; Guo et al., 2023; Wan et al., 2025; Zheng et al., 2024). Early video systems extended image diffusion models with temporal modules, while recent Diffusion Transformers (DiTs) (Peebles & Xie, 2023) learn long-range motion and appearance from large video–text corpora. Built on powerful visual-language representations (Radford et al., 2021), systems such as Wan2.1 (Wan et al., 2025), CogVideoX (Yang et al., 2024), HunyuanVideo (Kong et al., 2024), and Seedance (Gao et al., 2025) synthesize photorealistic videos with strong temporal coherence. They nevertheless operate in RGB space and produce fully opaque outputs, leaving transparency outside the generative interface.

Native alpha is essential for professional compositing. Particles, fire, smoke, glass reflections, animated stickers, and virtual gifts must be placed over arbitrary backgrounds without a second matting pass. Generating RGB first and estimating transparency afterward (Lin et al., 2021; 2022) can discard the soft, spatially diffuse opacity that defines visual effects: errors in a matte also change the color of the composited boundary. Training an RGBA model from scratch avoids this post-processing step but gives up the visual knowledge of a large RGB foundation model. Recent methods have explored transparent images and videos (Zhang & Agrawala, 2024; Yin et al., 2025; Wang et al., 2025a; Dong et al., 2025; Wang et al., 2025b; Ji et al., 2025), but two interface and data gaps remain.

First, matting-derived clips often obtain alpha by segmenting solid objects from RGB footage, so they encode the foreground as an opaque silhouette. Visual effects instead contain spatially non-local, continuously varying opacity; this mismatch produces color bleeding, chromatic fringes, or disconnected alpha blobs after compositing. The gap is especially visible for fireworks, particles,

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.

flames, smoke, glass-like boundaries, and soft halos, which are common in production assets but underrepresented by object-centric matting data.

Second, existing video extensions modify the interface consumed by a pretrained RGB generator. TransPixeler (Wang et al., 2025a) appends a separate alpha-token stream, increasing the sequence length processed by the Diffusion Transformer (DiT), whereas Wan-Alpha (Dong et al., 2025) shifts the RGB-A latent distribution and uses a bidirectional diffusion loss together with an object-centric Gaussian ellipse mask for noise guidance. These choices impose interface overhead or localized transparency priors that do not match full-frame effects. They also make direct reuse of the pretrained RGB latent-to-DiT pathway less immediate, precisely where a practical RGBA extension should preserve the strongest foundation-model priors.

We present **UniRGBA**, an interface-preserving extension that treats alpha as a fourth channel at the VAE boundary while retaining the shared 16-channel latent interface. This concentrates the modality change where it is needed: the DiT backbone keeps its original tokenization and conditioning interface, and the VAE learns the alpha boundary around a frozen RGB core. The result is a single representation that can be decoded as RGB or RGBA and consumed by the same text- and image-conditioned DiT. The design therefore separates representation learning from downstream generation: the VAE carries the new modality, while the pretrained DiT remains a reusable video prior.

Specifically, **UniRGBA-VAE** extends the Wan2.1 3D causal VAE to four-channel input and output. We zero-initialize the added alpha weights in the first and last convolutional layers, so an opaque RGB input starts from the pretrained RGB mapping. We then fine-tune only the outermost symmetric $2 + 2$ encoder/decoder layers (about 5% of VAE parameters), allowing the boundary layers to learn transparency while the intermediate temporal and semantic features remain fixed. This preserves the causal video compression pathway and makes the reconstruction transfer cost measurable on pure RGB clips. The symmetric update also gives the encoder and decoder matching capacity to introduce alpha features and reconstruct their high-frequency boundaries.

To stabilize the extension, **KL-Anchored Progressive Training** anchors the adapted posterior to the frozen RGB posterior with $\mathcal{L}_{\text{kl_ref}}$. A six-term reconstruction objective covers RGB, alpha, compositing, perceptual fidelity, the standard VAE prior, and the reference posterior, with a hard-alpha auxiliary for nearly opaque or transparent pixels. The first stage learns this reconstruction objective and stabilizes the shared RGB–alpha manifold; a second stage adds a lightweight compositing GAN after reconstruction has stabilized, sharpening particle edges, smoke wisps, glass-like boundaries, and other high-frequency effects. Because the latent interface is unchanged, standard DoRA modules adapt the pretrained Wan2.1 DiT for RGBA text-to-video (T2V) and image-to-video (I2V) generation without extra alpha tokens or a shifted noise schedule.

On our held-out visual-effect benchmark, UniRGBA-VAE raises RGBA PSNR from 25.00 to 29.16 over a retrained Wan-Alpha baseline (+4.16 dB), with +5.33 dB RGB and +3.95 dB alpha gains. Aligned RGBA T2V improves Text Alignment by +0.28, Naturalness by +0.33, and Temporal Flickering by +0.0050 over Wan-Alpha. The same checkpoint remains directly decodable for RGB, enabling direct evaluation of reconstruction and generation transfer. The shared interface also supports I2V: on the RGB transfer test, UniRGBA improves subject consistency, motion smoothness, temporal flickering, imaging quality, and naturalness, while the remaining Dynamic Degree decrease is reported explicitly. On the native RGBA I2V path, it reaches 0.9612 subject consistency, 0.9931 motion smoothness, 0.9913 temporal flickering, 4.18 text alignment, and 4.03 naturalness. These results show that native alpha can be added through a small VAE-side intervention while retaining a reusable video-generation interface.

Our contributions are:

- **A shared RGBA representation.** UniRGBA adds alpha at the VAE boundary and preserves the pretrained 16-channel latent interface, avoiding a second token stream or a shifted latent interface.

- **A mechanistic adaptation recipe.** Zero initialization, symmetric $2{+}2$ unfreezing, and a KL reference anchor adapt the VAE with about 5% trainable parameters; standard DoRA then transfers the shared DiT to RGBA T2V and I2V.

- **Evidence with explicit transfer costs.** Controlled reconstruction, generation, RGB, and I2V evaluations quantify gains in alpha fidelity and temporal quality together with the measurable motion cost of preserving the native alpha pathway.

2 Related Work

Large-Scale Video Generation. Diffusion models (Ho et al., 2020; Rombach et al., 2022; Ho et al., 2022) first brought high-fidelity synthesis to video by extending image generators with temporal modules (Singer et al., 2022; Wang et al., 2023; Blattmann et al., 2023; Guo et al., 2023). More recent Diffusion Transformers (DiTs) (Peebles & Xie, 2023) scale this paradigm with larger video–text corpora and longer-range spatiotemporal attention. Representative systems include Sora (Brooks et al., 2024), Open-Sora (Zheng et al., 2024), CogVideoX (Yang et al., 2024), HunyuanVideo (Kong et al., 2024), LTX-Video (HaCohen et al., 2024), Mochi (Team, 2024), Vidu (Bao et al., 2024), Seedance (Gao et al., 2025), and Wan2.1 (Wan et al., 2025). These foundation models provide strong priors for appearance, motion, and prompt alignment, but their latent interfaces are RGB-only. UniRGBA uses Wan2.1 as the pretrained video prior and adds alpha at the VAE boundary so that the DiT can be reused through its existing latent interface.

Transparent Image and Video Generation. Transparency has traditionally been added through matting or segmentation (Lin et al., 2021; 2022; Ravi et al., 2024). Image methods such as LayerDiffuse (Zhang & Agrawala, 2024), AlphaVAE (Wang et al., 2025b), and Qwen-Image-Layer (Yin et al., 2025) instead learn RGBA or layered representations directly; these methods establish useful image-level designs but do not model causal temporal compression. In video, LayerFlow (Ji et al., 2025), TransPixeler (Wang et al., 2025a), and Wan-Alpha (Dong et al., 2025) provide the closest native-RGBA comparisons. TransPixeler appends an alpha-token stream, increasing the sequence processed by the DiT. Wan-Alpha shifts the RGB–A latent distribution and uses bidirectional diffusion together with a Gaussian ellipse mask for noise guidance. These interfaces are effective reference points, but they either add token overhead or impose a localized transparency prior that does not cover full-frame effects. UniRGBA places the modality change at the VAE boundary and retains the pretrained 16-channel DiT interface, allowing the same downstream backbone to serve RGB and RGBA generation.

Variational Autoencoders for Video Latent Modeling. VAEs are the compression backbone of latent diffusion (Kingma & Welling, 2013; Rombach et al., 2022; Podell et al., 2023); video VAEs must additionally preserve causal temporal structure and high-frequency detail. 3D tokenizers such as MAGVIT (Yu et al., 2023a), MAGVIT-v2 (Yu et al., 2023b), CV-VAE (Zhao et al., 2024), and OmniTokenizer (Wang et al., 2024) improve spatiotemporal compression, while Wan-VAE (Wan et al., 2025) supplies the RGB backbone used here. UniRGBA extends that causal VAE with zero-initialized channel expansion, symmetric selective unfreezing, and a KL reference anchor. The adaptation updates about 5% of VAE parameters, leaving the intermediate RGB representation and the 16-channel interface available to the pretrained DiT; the transfer cost is then measured directly on RGB reconstruction and generation.

Positioning relative to image RGBA VAEs. AlphaVAE (Wang et al., 2025b) studies unified RGBA representation learning for images, whereas UniRGBA adapts a causal *video* VAE whose temporal compression interface is consumed by a pretrained video DiT. By preserving this interface, standard DoRA transfers the pretrained DiT to RGBA T2V/I2V with the same tokenization and latent width, without requiring alpha tokens, a shifted noise schedule, or a separate alpha branch. This positioning makes the contribution complementary to image-level RGBA VAEs and directly testable against native video baselines.

3 Method

Figures 1 and 2 provide an overview of our framework. We describe each component below.

Framework Overview. UniRGBA is a two-component framework that upgrades a pretrained RGB video generator into a native RGBA video generator through a focused, parameter-efficient

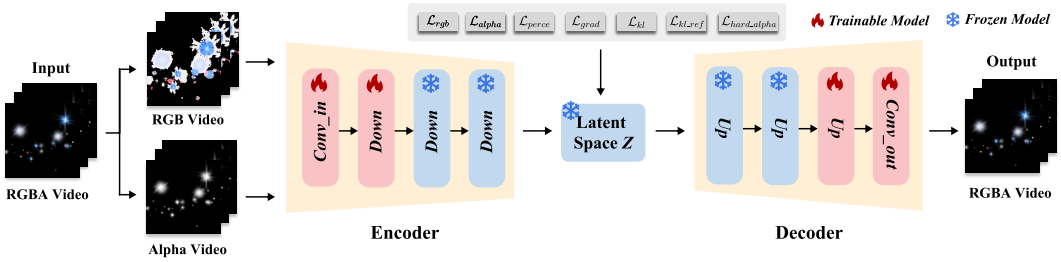

Figure 1: **Overview of UniRGBA-VAE.** UniRGBA-VAE extends the Wan2.1 VAE to 4-channel RGBA through zero-initialized channel expansion at Conv_in/Conv_out with selective layer fine-tuning (orange = trainable, blue = frozen). Training losses include pixel-level (L_{rgb}, L_{alpha}, L_{comp}, L_{hard_a}), perceptual (L_{perce}), and regularization (L_{kl}, L_{kl_ref}) terms.

extension of the RGB pathway. The design adds alpha at the VAE boundary, preserves latent dimensionality and tokenization, and keeps the pretrained DiT interface reusable. The separation is deliberate: the VAE learns the new channel and its compositing behavior, while the DiT continues to operate on the same-width latent sequence. At the *representation* level, **UniRGBA-VAE** extends the Wan2.1 3D causal VAE at its input and output boundaries. Zero initialization, symmetric $2 + 2$ unfreezing, and a KL reference anchor add alpha while keeping intermediate RGB features frozen. The six-term reconstruction objective is trained first, then augmented with a lightweight compositing GAN once the RGB–alpha representation is stable. At the *generation* level, standard DoRA adapts the pretrained Wan2.1 DiT directly on the shared 16-channel latent representation, with no alpha-token stream, shifted noise schedule, or separate alpha branch. The tokenization and latent width consumed by the pretrained DiT therefore remain unchanged; the experiments measure the resulting RGB transfer cost explicitly.

3.1 UniRGBA-VAE Architecture

Channel Expansion with Zero Initialization. We build upon the pretrained Wan2.1 3D causal VAE, which maps RGB video $\mathbf{x} \in \mathbb{R}^{T\times H\times W\times 3}$ to a shared latent representation with 16 channels and the model's causal temporal/spatial downsampling factors (for the valid frame counts used here, the temporal length is $1 + (T - 1)/4$). To support RGBA input $\mathbf{x} \in \mathbb{R}^{T\times H\times W\times 4}$, we expand the first convolutional layer (Conv_in) from 3 to 4 input channels and the final layer (Conv_out) from 3 to 4 output channels. Critically, the weights for the new alpha channel are *zero-initialized*, which provides an initialization-equivalent warm start: when the alpha channel is set to 1 (fully opaque), the expanded path is *identical* to the original Wan-VAE before any adaptation updates; this identity is not assumed after fine-tuning. This initialization strategy is in contrast to distribution-shifting approaches (Dong et al., 2025), which require additional bidirectional diffusion losses to restructure the latent space; our approach leaves the RGB latent interface unchanged and measures any post-training transfer cost directly.

Symmetric Selective Layer Unfreezing. Fine-tuning the entire VAE can degrade the strong RGB reconstruction capability acquired during large-scale pretraining. To adapt the pretrained VAE to 4-channel inputs without destroying this prior, we propose a symmetric unfreezing strategy. We train only the first k encoder layers and the last k decoder layers while keeping all intermediate layers frozen. Specifically, we unfreeze Conv_in + DownBlock1 (encoder side) and UpBlock3 + Conv_out (decoder side), corresponding to $k = 2$. By symmetrically unfreezing both the input encoder layers and the output decoder layers, the model can jointly optimize spatial-transparency feature extraction and high-frequency boundary reconstruction while keeping the intermediate RGB representation fixed. Under this symmetric $2 + 2$ configuration, only $\sim$5% of the total parameters are trainable.

3.2 KL-Anchored Progressive Training

Progressive Multi-Loss Objective. We design a comprehensive loss function with six complementary terms, each targeting a specific aspect of RGBA reconstruction quality:

$$\mathcal{L}_{\text{rgb}} = \|(\hat{\mathbf{x}}_{\text{rgb}} \odot \mathbf{a}) - (\mathbf{x}_{\text{rgb}} \odot \mathbf{a})\|_1 \tag{1}$$

$$\mathcal{L}_{\text{alpha}} = \|\hat{\mathbf{a}} - \mathbf{a}\|_1 \tag{2}$$

$$\mathcal{L}_{\text{comp}} = \|(\hat{\mathbf{x}}_{\text{rgb}} \odot \hat{\mathbf{a}} + \mathbf{b} \odot (1 - \hat{\mathbf{a}})) - (\mathbf{x}_{\text{rgb}} \odot \mathbf{a} + \mathbf{b} \odot (1 - \mathbf{a}))\|_1 \tag{3}$$

$$\mathcal{L}_{\text{perce}} = \text{LPIPS}(\hat{\mathbf{c}}_{\text{comp}}, \mathbf{c}_{\text{comp}}) \tag{4}$$

$$\mathcal{L}_{\text{kl}} = D_{\text{KL}}(q(\mathbf{z}|\mathbf{x}) \| \mathcal{N}(\mathbf{0}, \mathbf{I})) \tag{5}$$

$$\mathcal{L}_{\text{kl_ref}} = D_{\text{KL}}(q(\mathbf{z}|\mathbf{x}) \| q_{\text{ref}}(\mathbf{z}|\mathbf{x}_{\text{rgb}})) \tag{6}$$

where $\hat{\mathbf{x}}_{\text{rgb}}$ and $\hat{\mathbf{a}}$ denote predicted RGB and alpha channels, $\mathbf{b}$ is a randomly sampled background color used for compositing supervision, and $\mathcal{L}_{\text{perce}}$ uses LPIPS (Zhang et al., 2018) as the perceptual distance. The reference posterior q_{ref} is computed by applying the frozen RGB VAE to the RGB projection of the same clip and is compared in the shared 16-channel latent coordinates. For hard-alpha regions (where $\mathbf{a} \in \{0, 1\}$), we additionally apply:

$$\mathcal{L}_{\text{hard_alpha}} = \|(\hat{\mathbf{a}} \odot \mathbf{m}_{\text{hard}}) - (\mathbf{a} \odot \mathbf{m}_{\text{hard}})\|_1 \tag{7}$$

where $\mathbf{m}_{\text{hard}}$ is a binary mask identifying regions with alpha values close to 0 or 1. The progressive training objective combines six primary pixel-, perceptual-, and regularization-level terms with a separate hard-alpha auxiliary:

$$\mathcal{L}_{\text{recon}} = \mathcal{L}_{\text{rgb}} + \mathcal{L}_{\text{alpha}} + 0.5 \cdot \mathcal{L}_{\text{comp}} + 0.5 \cdot \mathcal{L}_{\text{perce}} + 10^{-6} \cdot \mathcal{L}_{\text{kl}} + 10^{-10} \cdot \mathcal{L}_{\text{kl_ref}} + 0.5 \cdot \mathcal{L}_{\text{hard_alpha}} \tag{8}$$

Latent Distribution Anchoring via $\mathcal{L}_{\text{kl_ref}}$. A critical challenge in fine-tuning large-scale VAEs is latent manifold drift: as the model adapts to 4-channel input, the learned posterior $q(\mathbf{z}|\mathbf{x})$ may diverge from the pretrained RGB posterior, causing the DiT (which was trained on the original latent space) to receive out-of-distribution inputs at inference time. We introduce $\mathcal{L}_{\text{kl_ref}}$, a KL divergence from the fine-tuned posterior to the frozen reference VAE posterior q_{ref}. Although its configured weight is small (10^{-10}), this term acts as a *soft anchor* to the original RGB latent distribution: the learned alpha representation is a residual extension anchored to the RGB posterior, keeping the latent statistics consumable by the pretrained DiT.

Two-Stage GAN-Augmented Training. To further refine fine-grained reconstruction quality (fine edges of particles, smoke wisps, and semi-transparent textures), we train the final reconstruction model in two stages under a single cosine learning-rate schedule. We introduce a lightweight patch-based adversarial objective $\mathcal{L}_{\text{GAN}}$ on the alpha-composited output only in the second stage. *Stage-1* trains the VAE from the zero-initialized RGBA warm start using only the reconstruction loss in Eq. 8:

$$\mathcal{L}_{\text{stage-1}} = \mathcal{L}_{\text{recon}}. \tag{9}$$

Stage-2 continues training from the Stage-1 checkpoint along the same cosine schedule, augmenting the objective with a small-weight adversarial term to refine fine-grained details after the reconstruction loss has stabilized:

$$\mathcal{L}_{\text{stage-2}} = \mathcal{L}_{\text{recon}} + 0.02 \cdot \mathcal{L}_{\text{GAN}}. \tag{10}$$

Adding the GAN objective after the reconstruction-only Stage-1 checkpoint raises PSNR by 2.91 dB while preserving LPIPS, sharpening the fine structures that matter for visual-effect compositing; the small mask-region trade-off is reported in Table 1.

3.3 Alpha-Aware DiT Adaptation for T2V and I2V

To generate RGBA videos, we adapt the pretrained Wan2.1 diffusion transformer on top of our expanded UniRGBA-VAE. Lightweight parameter-efficient tuning tests whether the unified RGBA latent space supports downstream generation while keeping the large text-video backbone largely frozen. We retain the original tokenization and latent width, with no alpha-token stream, extra conditioning pathway, or dedicated alpha attention.

T2V adaptation. During T2V training, the text prompt is encoded by the frozen umT5 encoder, UniRGBA-VAE provides RGBA latent targets, and the DiT denoises latent videos conditioned on text embeddings and timestep embeddings. We insert lightweight DoRA-style adapters (Liu et al.,

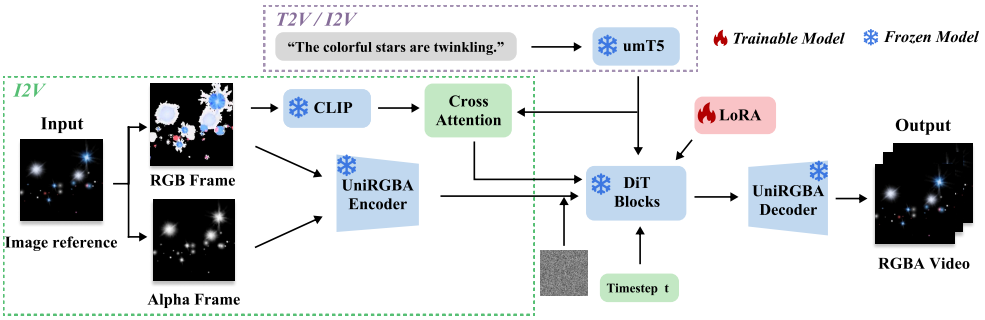

Figure 2: **Adapter adaptation in the shared RGBA latent interface.** For T2V, frozen umT5 text features and timestep embeddings condition adapter-tuned DiT blocks, whose denoised latents are decoded by UniRGBA-VAE. For I2V, a reference RGBA frame is additionally encoded into visual conditions in the same latent space and injected into the shared DiT backbone; our experiments use standard DoRA adapters.

2024) into the attention projections and feed-forward layers of all DiT blocks, while freezing the remaining backbone weights. Concretely, we adapt the standard Wan2.1 DiT modules `q`, `k`, `v`, `o`, `ffn.0`, and `ffn.2`, following the standard DoRA configuration used by Wan-Alpha (Dong et al., 2025). This isolates the contribution of the unified RGBA representation, keeps the adapter footprint small, and enables RGBA-specific generation with the pretrained conditioning pathway.

I2V extension. The same shared latent interface also supports image-to-video generation without introducing a separate alpha-specific branch. Given a reference RGBA image, we encode it with UniRGBA-VAE to obtain reference latent tokens. Following the conditioning design of Wan2.1 I2V, the same reference image also provides frozen visual features from the image encoder, which are concatenated with the text context. The DiT then predicts a sequence of RGBA latents conditioned on text, timestep, and reference-image cues, after which UniRGBA-VAE decodes the result into an RGBA video. T2V and I2V therefore share the same downstream backbone: T2V uses text-only conditioning, whereas I2V augments this backbone with reference-image conditions in the same unified RGBA latent space.

4 Experiments

4.1 Experimental Setup

We curate an RGBA video dataset from historical special-effect asset packages used in live-streaming production. The data cover particles, flames, smoke, fireworks, stickers, virtual gifts, and object-centric cases including humans, animals, solid objects, glass-like media, and fine hair or halo boundaries. We deduplicate by asset identity before splitting and keep held-out clips disjoint from all training captions. The final VAE uses 17K training clips and 100 held-out RGBA clips; DiT adaptation uses 2,357 captioned clips and a separate 100-clip test set, while RGB transfer is evaluated on 5K WebVid reconstruction clips and 100 RGB T2V clips.

4.2 Evaluation Metrics and Baselines

For VAE reconstruction, we report PSNR, SSIM, and LPIPS, with PSNR decomposed into RGB, alpha, and hard-mask regions. For generation, we report public VBench/LAION metrics together with fixed-rubric Text Alignment and Naturalness scores; RGB-only generation additionally reports Dynamic Degree and Temporal Spatial Consistency. The native-RGBA video comparison is restricted to Wan-Alpha and TransPixeler-Wan, the reproducible video baselines with RGBA generation configurations; image-only RGBA methods are discussed as related work rather than treated as video baselines. All RGBA comparisons use the same 832×480 evaluation protocol used by the baselines, while RGB results use the original Wan2.1 model. Metric definitions and uncertainty estimates are provided in Appendix B.

Table 1: **Reconstruction on the RGBA benchmark and RGB-only WebVid.** Bold marks the best point estimate for each metric within each dataset; arrows indicate the preferred direction.

Metric	RGBA benchmark			WebVid	
	Wan-Alpha	UniRGBA-VAE (Stage-1)	UniRGBA-VAE (Stage-2)	UniRGBA-VAE	Wan2.1-VAE
PSNR↑	25.00	26.25	**29.16**	33.41	**36.73**
SSIM↑	0.891	**0.933**	0.932	0.945	**0.964**
LPIPS↓	0.0936	0.0808	**0.0801**	0.0401	**0.0213**
PSNR_{rgb}↑	24.42	26.16	**29.75**	33.41	**36.73**
$\text{PSNR}_{\text{rgb_mask}}$↑	20.83	**22.76**	22.49	–	–
$\text{PSNR}_{\text{alpha}}$↑	30.50	32.56	**34.45**	–	–
$\text{PSNR}_{\text{alpha_mask}}$↑	**22.54**	21.91	22.02	–	–

Table 2: **Progressive loss ablation.** E002 adds the KL reference term to E001; E003 jointly adds the compositing and hard-alpha terms. Bold marks the best point estimate in each metric column.

ID	Configuration	PSNR↑	PSNR_{rgb}↑	$\text{PSNR}_{\text{alpha}}$↑	SSIM↑	LPIPS↓
E001	$\mathcal{L}_{\text{rgb}} + \mathcal{L}_{\text{alpha}} + \mathcal{L}_{\text{perce}} + \mathcal{L}_{\text{kl}}$ (Baseline)	19.93	26.56	15.96	0.841	0.210
E002	+ $\mathcal{L}_{\text{kl_ref}}$	24.46	**33.50**	20.46	**0.886**	**0.159**
E003	Full loss (Eq. 8): + $\mathcal{L}_{\text{comp}}$ + $\mathcal{L}_{\text{hard_alpha}}$	**25.53**	32.80	**22.14**	0.880	0.165

4.3 Main Results: VAE Reconstruction

Table 1 summarizes the final two-stage reconstruction model trained on the 17K high-quality RGBA subset; ablations use the 1K RGBA set. Stage-1 already exceeds Wan-Alpha on the aggregate PSNR, SSIM, LPIPS, RGB, and alpha metrics, while the hard alpha-mask score remains slightly lower; Stage-2 adds a lightweight GAN objective for sharper composited details. The same VAE is also evaluated on RGB-only WebVid to measure transfer cost.

Superior RGB and Alpha Fidelity. Compared to Wan-Alpha, UniRGBA-VAE raises RGBA PSNR from 25.00 to 29.16 (+4.16 dB), with +5.33 dB RGB and +3.95 dB alpha gains. SSIM increases from 0.891 to 0.932 and LPIPS decreases from 0.0936 to 0.0801. The gains across both channels show that the boundary update learns transparency while retaining high-fidelity color reconstruction.

Effect of the Stage-2 GAN-augmented schedule. Stage-2 raises PSNR from 26.25 to 29.16 (+2.91 dB) and preserves perceptual fidelity (LPIPS 0.0808 $\rightarrow$ 0.0801) by adding adversarial supervision on the alpha-composited output. The small decrease in $\text{PSNR}_{\text{rgb_mask}}$ (22.76 $\rightarrow$ 22.49) is reported alongside the gain rather than folded into the aggregate score.

Backward compatibility on pure RGB data. On RGB-only WebVid, the same checkpoint remains directly decodable and reaches PSNR 33.41 and LPIPS 0.0401. The corresponding Wan2.1-VAE scores (36.73 and 0.0213) quantify the reconstruction cost of adding native alpha.

Together, the table separates the two effects: boundary-aware reconstruction drives the gain over Wan-Alpha, while the second-stage GAN adds 2.91 dB PSNR with nearly unchanged LPIPS. The same checkpoint remains directly decodable for RGB, making the transfer cost measurable rather than implicit.

4.4 Loss Ablation

Table 2 presents our progressive loss ablation, starting from a minimal baseline and incrementally adding loss terms. All experiments use 1K training videos at 832×480 resolution with 5 frames.

Key findings: Adding $\mathcal{L}_{\text{kl_ref}}$ (E002) yields the largest single improvement (+4.53 dB PSNR), while the compositing and hard-alpha terms jointly add +1.68 dB $\text{PSNR}_{\text{alpha}}$ (E003). The configured coefficient is reported in Appendix B. Additional ablation studies regarding the layer unfreezing strategy, training resolution, and frame count are provided in Appendix C.

Table 3: **T2V generation on RGBA and RGB evaluation sets.** Bold marks the best point estimate within each evaluation set; paired uncertainty tests are reported in Appendix B.

Metric	RGBA, T2V			RGB, T2V	
	Wan-Alpha	TransPixeler-Wan	Ours	Ours	Wan2.1
Aesthetic Quality↑	**0.5232**	0.4490	0.4694	**0.6116**	0.5471
Text Alignment↑	3.9794	3.3540	**4.2600**	3.8889	**4.0566**
Naturalness↑	3.6939	3.1460	**4.0200**	**3.9821**	3.8438
Motion Smoothness↑	0.9927	0.9935	**0.9948**	0.9896	**0.9898**
Temporal Flickering↑	0.9878	0.9903	**0.9928**	**0.9823**	0.9800
Dynamic Degree (RGB)↑	–	–	–	0.3846	**0.4800**
Temporal Spatial Consistency (RGB)↑	–	–	–	**0.2388**	0.2283

4.5 DiT Fine-tuning with Standard DoRA

To evaluate downstream generation after adapting Wan2.1 for RGBA, we fine-tune the DiT with standard DoRA adapters (see Section 3.3) on 2,357 RGBA video-caption pairs and hold out 100 RGBA videos for testing. Table 3 summarizes aligned T2V comparisons using public VBench/LAION metrics and fixed-rubric Text Alignment/Naturalness scores; the full protocol is in Appendix B.

On aligned RGBA T2V evaluation, UniRGBA has the highest point estimates for Text Alignment (4.2600), Naturalness (4.0200), Motion Smoothness (0.9948), and Temporal Flickering (0.9928). The paired gains over Wan-Alpha are +0.28 [0.13,0.43] in Text Alignment, +0.33 [0.16,0.49] in Naturalness, and +0.0050 [0.0028,0.0072] in Temporal Flickering; the motion-smoothness interval includes zero and Aesthetic Quality remains higher for Wan-Alpha. For RGB T2V, we replace only the VAE without additional DiT fine-tuning; this zero-shot swap raises Aesthetic Quality and Naturalness while reducing Dynamic Degree by 0.095 $[-0.15, -0.04]$, quantifying the transfer cost of native alpha.

The shared interface also supports I2V without a second alpha branch. On the RGB transfer test, UniRGBA improves subject consistency (0.9731 vs. 0.9708, $p = .021$), motion smoothness (0.9917 vs. 0.9901, $p = .012$), temporal flickering (0.9846 vs. 0.9822, $p = .014$), imaging quality (0.6903 vs. 0.6841, $p = .019$), and naturalness (3.98 vs. 3.85, $p = .026$) relative to Wan2.1-I2V. The transfer cost is concentrated in motion extent: Dynamic Degree decreases from 0.4796 to 0.4388 ($\Delta = -0.0408$, 95% CI $[-0.0779, -0.0037]$, $p = .038$), while aesthetic quality, background consistency, and text alignment remain statistically comparable. On the native RGBA path, UniRGBA reaches 0.9612 subject consistency, 0.9931 motion smoothness, 0.9913 temporal flickering, 4.18 text alignment, and 4.03 naturalness. Full paired intervals and the complete I2V table are reported in Appendix B.

Human preference and transfer behavior. A forced-choice study on fixed visual-effect prompts prefers UniRGBA in 83.1% of reconstruction, 66.4% of T2V, and 93.5% of I2V comparisons. The T2V study covers native-RGBA baselines and Wan2.1+BiRefNet, while the I2V study uses the executable RGB-to-alpha reference; confidence intervals and the complete transfer table are reported in Appendix B.

Cross-baseline and stress-case analysis. The gap to TransPixeler-Wan is concentrated in the capabilities most affected by the representation interface: UniRGBA improves Text Alignment by $+0.91$ and Naturalness by $+0.87$, with additional gains in Aesthetic Quality ($+0.020$, $p = .0325$) and Temporal Flickering ($+0.0025$, $p = .0365$); the Motion Smoothness difference is small and not significant ($+0.0013$, $p = .3600$). The advantage is concentrated in semantic and alpha-aware temporal fidelity. Across the held-out object-centric cases, the qualitative improvements are most visible around hair strands, glass-like boundaries, soft halos, and fine particles, where a matting pass tends to create colored fringes or disconnected alpha blobs. The same failure modes appear in the original S1–S3 panels, providing a visual stress test for the compositing and hard-alpha terms.

The adaptation footprint also clarifies why these gains transfer across tasks. Only the VAE boundary layers (about 5% of VAE parameters) and lightweight DoRA adapters are updated; the 16-channel latent/token interface and the pretrained DiT backbone remain unchanged. Consequently,

the UniRGBA-VAE checkpoint is reusable for native RGBA decoding and RGB reconstruction, while the corresponding DoRA adapters support T2V and I2V; the reported RGB Dynamic Degree decrease makes the remaining transfer cost explicit. The preference results are consistent with this modular behavior, but are reported as supporting evidence alongside the per-clip metrics and paired uncertainty tests.

All generation comparisons use the same prompts, seeds, sampler, step count, guidance scale, output resolution, and five-background compositing. We therefore interpret the paired intervals as differences under a shared evaluation protocol rather than as changes caused by sampling or post-processing. The uncertainty pattern is also informative: the largest gains are stable for alignment, naturalness, and flicker, whereas the RGB Dynamic Degree reduction is the main measurable cost of preserving the native alpha pathway. This separation identifies motion diversity as the next engineering target while preserving the representation-level finding.

5 Qualitative Results

Figure 3 compares UniRGBA with Wan-Alpha and TransPixeler-Wan on three visual-effect T2V prompts. All outputs use four uniformly sampled frames composited on a checkerboard so that color and semi-transparent structure can be inspected together; additional reconstruction and I2V examples appear in the Appendix.

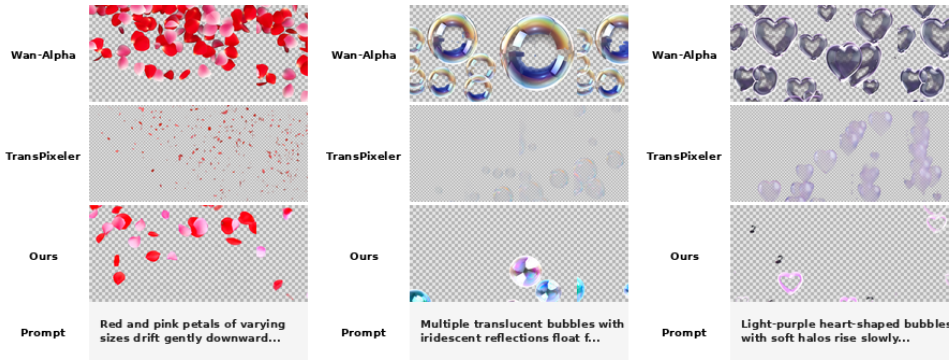

Figure 3: **Three-way qualitative comparison on visual-effect T2V prompts.** Wan-Alpha (top), TransPixeler-Wan (middle), and Ours (bottom, DoRA on Wan2.1-T2V-14B) are stacked, all rendered on a checkerboard background to visualize transparency. Our method produces transparency and color content that is visually competitive with Wan-Alpha while clearly outperforming TransPixeler on challenging non-elliptical visual-effect scenes.

6 Conclusion

We presented **UniRGBA**, a native RGBA extension of a pretrained RGB video generator. Zero-initialized boundary expansion, symmetric 2+2 unfreezing, and KL-anchored progressive training adapt the VAE with about 5% trainable parameters while retaining the 16-channel latent interface; standard **DoRA** then transfers the shared DiT to RGBA T2V and I2V. On visual-effect data, UniRGBA improves RGBA reconstruction by 4.16 dB PSNR over a retrained Wan-Alpha baseline and obtains the strongest aligned T2V point estimates for text alignment, naturalness, and temporal flickering, with uncertainty reported for each comparison. RGB and I2V evaluations quantify the transfer cost, including the 0.095 Dynamic Degree decrease in RGB T2V. Broader public RGBA benchmarks, higher-resolution evaluation, and released evaluation artifacts are directions for future work.

AI USE STATEMENT

Qwen-VL assisted with initial captions for internal video assets; human annotators checked the captions, and the authors verified the manuscript, data, experiments, and claims.

ETHICS STATEMENT

The work uses licensed or internally owned visual-effect assets and does not collect personal or sensitive data. We will document the provenance and usage restrictions of any released artifacts and will not release proprietary source videos.

REPRODUCIBILITY STATEMENT

The manuscript specifies data splits, preprocessing, optimization schedules, model modules, sampling settings, evaluation implementations, and statistical procedures. The primary RGBA assets are proprietary, so the submission documents a reproducible RGB evaluation path and the curation protocol; subject to licensing and anonymization, we will release configuration files, evaluation code, model weights, and a synthetic or curated benchmark subset with the final artifact.

REFERENCES

Fan Bao, Chendong Xiang, Gang Yue, Guande He, Hongzhou Zhu, Kaiwen Zheng, Min Zhao, Shilong Liu, Yaole Wang, and Jun Zhu. Vidu: a highly consistent, dynamic and skilled text-to-video generator with diffusion models, 2024.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Leo Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1(8):1, 2024.

Haotian Dong, Wenjing Wang, Chen Li, and Di Lin. Wan-alpha: High-quality text-to-video generation with alpha channel. *arXiv e-prints*, pp. arXiv–2509, 2025.

Yu Gao, Haoyuan Guo, Tuyen Hoang, Weilin Huang, Lu Jiang, Fangyuan Kong, Huixia Li, Jiashi Li, Liang Li, Xiaojie Li, et al. Seedance 1.0: Exploring the boundaries of video generation models. *arXiv preprint arXiv:2506.09113*, 2025.

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. *arXiv preprint arXiv:2307.04725*, 2023.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022.

Sihui Ji, Hao Luo, Xi Chen, Yuanpeng Tu, Yiyang Wang, and Hengshuang Zhao. Layerflow: A unified model for layer-aware video generation. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–10, 2025.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian L Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Real-time high-resolution background matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8762–8771, 2021.

Shanchuan Lin, Linjie Yang, Imran Saleemi, and Soumyadip Sengupta. Robust high-resolution video matting with temporal guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 238–247, 2022.

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 *Forty-first International Conference on Machine Learning*, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 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. *arXiv preprint arXiv:2307.01952*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. URL https://arxiv.org/abs/2408.00714.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

Genmo Team. Mochi 1. https://github.com/genmoai/models, 2024.

Qwen Team. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Junke Wang, Yi Jiang, Zehuan Yuan, Binyue Peng, Zuxuan Wu, and Yu-Gang Jiang. Omnitokenizer: A joint image-video tokenizer for visual generation. *Advances in Neural Information Processing Systems*, 37:28281–28295, 2024.

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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18229–18239, 2025a.

Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *Advances in Neural Information Processing Systems*, 36:7594–7611, 2023.

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.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Shengming Yin, Zekai Zhang, Zecheng Tang, Kaiyuan Gao, Xiao Xu, Kun Yan, Jiahao Li, Yilei Chen, Yuxiang Chen, Heung-Yeung Shum, Lionel M. Ni, Jingren Zhou, Junyang Lin, and Chenfei Wu. Qwen-image-layered: Towards inherent editability via layer decomposition, 2025. URL https://arxiv.org/abs/2512.15603.

Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, et al. Magvit: Masked generative video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10459–10469, 2023a.

Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, et al. Language model beats diffusion–tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023b.

Lvmin Zhang and Maneesh Agrawala. Transparent image layer diffusion using latent transparency. *arXiv preprint arXiv:2402.17113*, 2024.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Sijie Zhao, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Muyao Niu, Xiaoyu Li, Wenbo Hu, and Ying Shan. Cv-vae: A compatible video vae for latent generative video models. *Advances in Neural Information Processing Systems*, 37:12847–12871, 2024.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024.

APPENDIX

This Appendix is organized into the following sections: Section A provides the detailed experimental setup and baseline protocol; Section B gives metric definitions, uncertainty estimates, and generation-level ablations; Section C presents layer, resolution, and frame-count ablations; Section D shows qualitative reconstruction demos; Section E discusses limitations and future work; and Section F reports I2V results and visualizations.

A DETAILED EXPERIMENTAL SETUP

A.1 DATASETS

We curate an internal RGBA video dataset from historical special-effect asset packages used in live-streaming production, covering a wide range of *visual-effect* transparent content (e.g., particles, flame, smoke, fireworks, animated stickers, virtual gifts, and object-centric assets with hair, glass-like surfaces, and soft halos). We construct RGBA samples by extracting and standardizing alpha-bearing assets, then perform a two-stage labeling pipeline: (1) automatic captioning with Qwen-VL (Team, 2023), followed by (2) manual verification and filtering to remove low-quality, ambiguous, or duplicated samples.

We use straight (unpremultiplied) alpha in $[0, 1]$ and preserve RGB values in transparent pixels; all compositing is performed in linearized RGB before conversion to the display space. We use different subsets for reconstruction and generation experiments:

- **VAE ablation set**: 1K RGBA videos for rapid loss ablations and other small-scale experimental iterations.
- **VAE main training set**: a curated high-quality RGBA subset with 17K videos, used for the final reconstruction model reported in Table 1.
- **T2V generation set**: approximately 2K RGBA videos (2,357 samples) with Qwen-VL captions verified by human annotators for DoRA training, plus a held-out test set of 100 RGBA videos for generation evaluation.
- **WebVid evaluation sets**: for RGB-only VAE reconstruction, we use a 5K-sample WebVid test set; for RGB T2V generation, we use a manually selected 100-sample WebVid test subset. These RGB sets are disjoint from all RGBA assets.

A.2 Implementation Details

VAE Training. All VAE experiments use AdamW on 32 GPUs with a per-GPU batch size of 1. The 2+2 update trains about 5% of VAE parameters; the remaining VAE core and the RGB reference VAE are frozen. Because our dataset contains mixed resolutions, we normalize each sample by resizing while preserving aspect ratio and applying center-crop/pad to the target training resolution. Unless otherwise specified, we use 5 frames at 832×480 for the loss ablation, 128×128 for the resolution and layer-unfreezing ablations on the 1K ablation set, and scale to 832×480 for the final reconstruction model trained on the 17K high-quality subset.

Two-Stage Schedule. For the final reconstruction model, we train for 100 epochs on the 17K high-quality subset under a single cosine learning-rate schedule with 640 warmup steps, peaking at 5×10^{-5} after warmup and decaying to 1.5×10^{-5} at epoch 100. Training is split into two stages based on the loss configuration: *Stage-1 (epochs 0–30)* uses only the six primary reconstruction terms plus a hard-alpha auxiliary (Eq. 8); *Stage-2 (epochs 30–100)* continues along the same cosine schedule (lr $\approx 4.3 \times 10^{-5}$ at the epoch-30 transition, decaying to 1.5×10^{-5} at epoch 100) and additionally introduces a 0.02-weighted GAN objective to refine fine-grained details. Unless otherwise specified, reported Stage-1/Stage-2 numbers correspond to the end-of-stage checkpoints of this schedule. The GAN is a 3D patch discriminator on composited outputs, updated with the same optimizer family; all baselines use the same spatial/temporal resolution, sampler, number of steps, guidance scale, prompts, and five-background compositing. Wan-Alpha's VAE is retrained on our RGBA data under the same budget. Wan-Alpha's generator uses its official rank-32 DoRA recipe, while TransPixeler-Wan uses the public Wan branch with the aligned budget and protocol.

B Evaluation Protocol and Statistical Analysis

Metric definitions. PSNR and SSIM are computed per video after decoding; LPIPS uses the public AlexNet implementation. Hard-mask PSNR is restricted to pixels with alpha below 0.05 or above 0.95. For generation, Motion Smoothness, Temporal Flickering, Dynamic Degree, and Temporal Spatial Consistency use the released VBench implementation; Aesthetic Quality uses the public LAION-Aesthetic predictor. Text Alignment measures prompt-video semantic faithfulness and Naturalness measures physical/temporal plausibility using a fixed VLM rubric on a 1–5 scale. The evaluation rubric and aggregation procedure are specified here; release artifacts will follow the licensing and anonymization constraints described in the reproducibility statement.

Uncertainty. We evaluate 100 held-out clips with fixed prompts, seeds, sampler, steps, guidance, and five-background compositing. We report per-clip means, 10,000-bootstrap 95% confidence intervals for paired differences, and paired Wilcoxon tests with Holm correction. Bold in the main table marks the best point estimate; uncertainty tests are reported separately.

Table S1: **RGBA T2V paired differences.** Δ is UniRGBA minus the comparison method; all metrics favor positive differences. Wan-Alpha comparisons include 95% bootstrap CIs. † marks differences with Holm-corrected paired $p < .05$, including negative differences.

	Wan-Alpha Δ		TransPixeler-Wan Δ	
Metric	Δ [95% CI]	p	Δ	p
Text Alignment	$+0.28^{\dagger}$ $[+0.13, +0.43]$	$< .001$	$+0.91^{\dagger}$	$< .001$
Naturalness	$+0.33^{\dagger}$ $[+0.16, +0.49]$	$< .001$	$+0.87^{\dagger}$	$< .001$
Aesthetic Quality	$-0.054^{\dagger}$ $[-0.073, -0.035]$	$< .001$	$+0.020^{\dagger}$	.0325
Motion Smoothness	$+0.0021$ $[-0.0007, +0.0049]$	.140	$+0.0013$	.3600
Temporal Flickering	$+0.0050^{\dagger}$ $[+0.0028, +0.0072]$	$< .001$	$+0.0025^{\dagger}$	.0365

Relative to Wan-Alpha, the significant gains concentrate on Text Alignment, Naturalness, and Temporal Flickering, whereas Aesthetic Quality decreases and Motion Smoothness is inconclusive. Relative to TransPixeler-Wan, UniRGBA improves every significant metric listed here, with the largest differences in Text Alignment and Naturalness. This pattern isolates the strongest benefits of the shared RGBA interface while making the aesthetic trade-off explicit.

For RGB T2V, the VAE swap improves Aesthetic Quality ($+0.065$ $[+0.047, +0.082]$) but reduces Dynamic Degree (-0.095 $[-0.15, -0.04]$) and Text Alignment (-0.168 $[-0.33, -0.01]$); the remaining metrics are within the reported uncertainty. The KL-reference coefficient is fixed at 10^{-10} in the released configuration and is reported as an implementation detail rather than a universal optimum.

Table S2: **Generation with Stage-1 and Stage-2 decoder checkpoints.** Both decode the same DiT outputs. Stage-2 includes additional training with the GAN objective. Bold marks the best point estimate within each evaluation set; all metrics are higher-is-better.

	RGBA T2V		RGB T2V	
Metric	Stage-2	Stage-1	Stage-2	Stage-1
Aesthetic Quality	**0.4694**	0.4097	**0.6116**	0.5287
Text Alignment	**4.2600**	4.1422	**3.8889**	3.7987
Naturalness	**4.0200**	3.8514	**3.9821**	3.8015
Motion Smoothness	0.9948	**0.9949**	0.9896	**0.9898**
Temporal Flickering	0.9928	**0.9935**	0.9823	**0.9833**
Dynamic Degree	–	–	**0.3846**	0.3797
Temporal Spatial Consistency	–	–	0.2388	**0.2396**

Forced-choice preference study. The study compares outputs on fixed visual-effect prompts and reports the percentage of selections for each method with 95% confidence intervals. The corrected comparisons include Wan-Alpha, TransPixeler-Wan where an executable video baseline is available, and Wan2.1+BiRefNet as the RGB-to-alpha reference.

Table S3: **Forced-choice preference shares.** Values are percentages with 95% CIs in brackets; bold marks the highest share in each task. Each row sums to 100% over the available comparison methods.

Task	UniRGBA	Wan-Alpha	TransPixeler-Wan	Wan2.1+BiRefNet
VAE reconstruction	**83.1** [79.3,86.4]	13.9 [10.9,17.4]	–	3.0 [1.6,5.0]
T2V generation	**66.4** [62.1,70.4]	23.2 [19.7,27.1]	7.4 [5.4,10.0]	3.0 [1.8,4.9]
I2V generation	**93.5** [90.6,95.7]	–	–	6.5 [4.3,9.4]

Table S4: **RGB and RGBA I2V evaluation.** Bold marks the best RGB point estimate; all metrics are higher-is-better. Δ and p compare RGB UniRGBA with Wan2.1-I2V. RGBA results are reported separately because no comparable public baseline is available.

Metric	RGB Ours	RGB Wan2.1-I2V	Δ [95% CI]	p	RGBA Ours
I2V Subject Consistency	**0.9731**	0.9708	$+0.0023\,[+0.0005, +0.0041]$	.021	0.9612
I2V Background Consistency	0.9687	**0.9695**	$-0.0008\,[-0.0031, +0.0015]$	.402	0.9540
Motion Smoothness	**0.9917**	0.9901	$+0.0016\,[+0.0004, +0.0028]$	.012	0.9931
Temporal Flickering	**0.9846**	0.9822	$+0.0024\,[+0.0006, +0.0042]$	.014	0.9913
Aesthetic Quality	**0.5842**	0.5798	$+0.0044\,[-0.0037, +0.0125]$	.271	0.5021
Imaging Quality	**0.6903**	0.6841	$+0.0062\,[+0.0011, +0.0113]$	.019	0.6434
Dynamic Degree	0.4388	**0.4796**	$-0.0408\,[-0.0779, -0.0037]$	.038	–
Text Alignment (0–5)	**3.92**	3.88	$+0.04\,[-0.11, +0.19]$	.583	4.18
Naturalness (0–5)	**3.98**	3.85	$+0.13\,[+0.02, +0.24]$	.026	4.03

C Additional Ablation Studies

C.1 Layer Unfreezing Strategy

Table S5 compares different unfreezing strategies on an early data-scaling run. While the 2+1 configuration achieves the highest pixel-level PSNR, the symmetric 2+2 configuration obtains substantially better perceptual quality (LPIPS 0.082 vs. 0.0995) and structural similarity (SSIM 0.936 vs. 0.898). This indicates that unfreezing both the encoder and the decoder symmetrically allows the model to learn more coherent RGBA representations. We therefore select 2+2 as our default configuration.

Table S5: **Layer unfreezing ablation.** $k_e{+}k_d$ denotes the number of unfrozen encoder and decoder layers. Bold marks the best point estimate in each metric column.

ID	Strategy (k_e+k_d)	PSNR↑	SSIM↑	LPIPS↓	MSE↓
E004	1+1	28.12	0.896	0.112	0.0076
E005	2+1	**29.04**	0.898	0.0995	0.0023
E006	1+2	27.50	0.910	0.111	0.0085
E007	2+2	26.63	**0.936**	**0.082**	**0.0017**

C.2 Training Resolution and Frame Count

Table S6 isolates the effects of training resolution and frame count. The main comparison uses 832×480 because this is the common 480p-class setting used by the native-RGBA baselines; it keeps the comparison fair while leaving 720p and 1080p scaling as future engineering work. Fixing the temporal window to 5 frames reveals an inverse relationship between spatial resolution and reconstruction metrics: lower resolutions generally yield higher PSNR and lower LPIPS. We hypothesize this occurs because a fixed-capacity VAE can focus more effectively on learning the underlying alpha channel structure when relieved of the burden of modeling high-frequency spatial details. Fixing the resolution to 256×256 shows a strict monotonic decay in reconstruction quality as the sequence length grows from 5 to 21 frames, indicating that shorter temporal windows allow the model to achieve better per-frame compression fidelity.

Table S6: **Resolution and frame count ablations.** Left: spatial resolution with 5 frames. Right: temporal length at 256×256. Bold marks the best point estimate in each metric column within each panel.

Resolution	PSNR↑	SSIM↑	LPIPS↓	ID
832×480	25.53	0.880	0.165	E003
256×256	**27.81**	0.897	0.141	E008
128×128	27.74	**0.899**	**0.111**	E009

Frames	PSNR↑	SSIM↑	LPIPS↓	ID
5	**27.81**	**0.897**	**0.141**	E008
9	26.36	0.890	0.157	E010
13	25.91	0.886	0.162	E011
21	24.65	0.883	0.167	E012

D Additional Qualitative Reconstruction Demos

As a complement to the quantitative reconstruction results in Section 4.3, Figures S1–S3 present selected held-out RGBA examples. For each example, we show a triptych consisting of RGB composited on a black background, RGB composited on a checkerboard background, and the alpha matte. Within each panel, the first row shows the reconstruction produced by **UniRGBA**, the second row shows the corresponding reconstruction from **Wan-Alpha**, and the third row shows the **ground-truth input video**. This layout makes it easy to jointly inspect foreground appearance, semi-transparent boundaries, and the consistency between RGB appearance and alpha structure across methods.

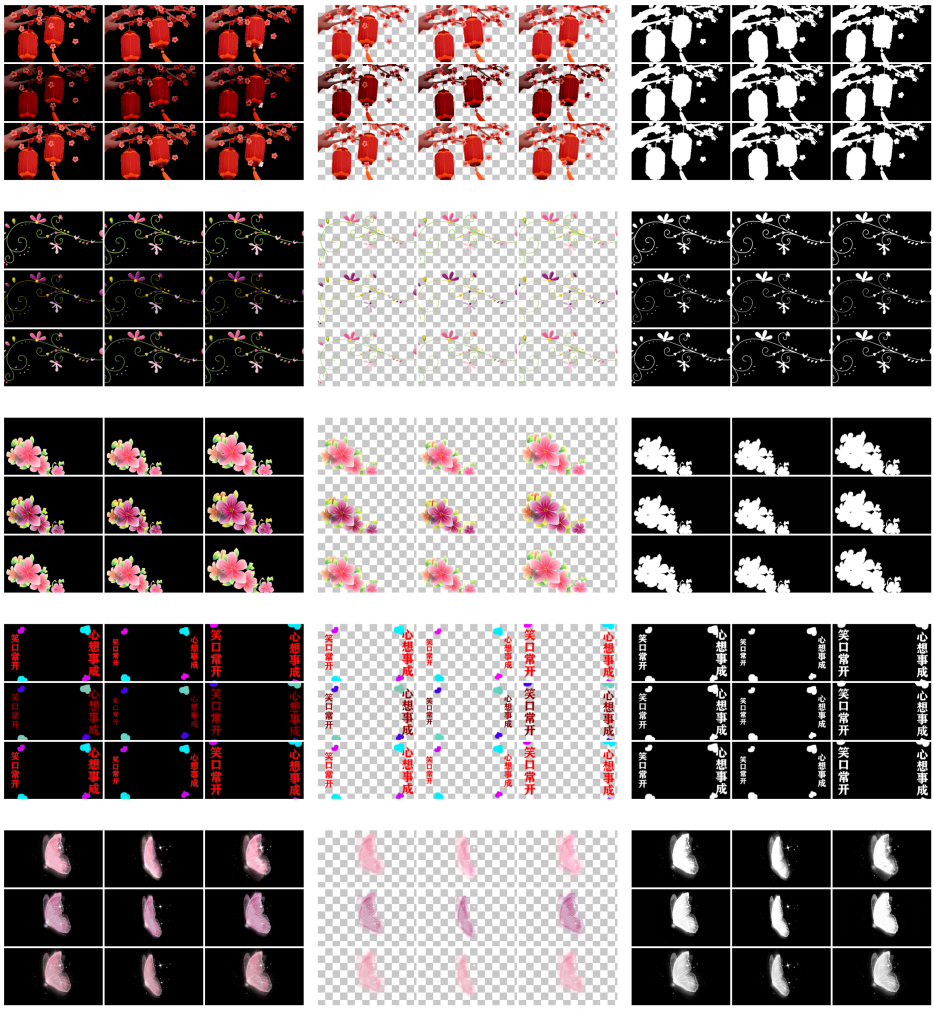

Figure S1: **Additional qualitative reconstruction demos (page 1/3).** Five selected held-out RGBA videos. Within each triptych, the same example is shown as RGB composited on a black background (left), RGB composited on a checkerboard background (center), and the alpha matte (right). Within each panel, rows correspond to UniRGBA reconstruction (top), Wan-Alpha reconstruction (middle), and the ground-truth input video (bottom). The examples highlight strong color preservation together with clean and spatially coherent transparency on diverse visual-effect assets.

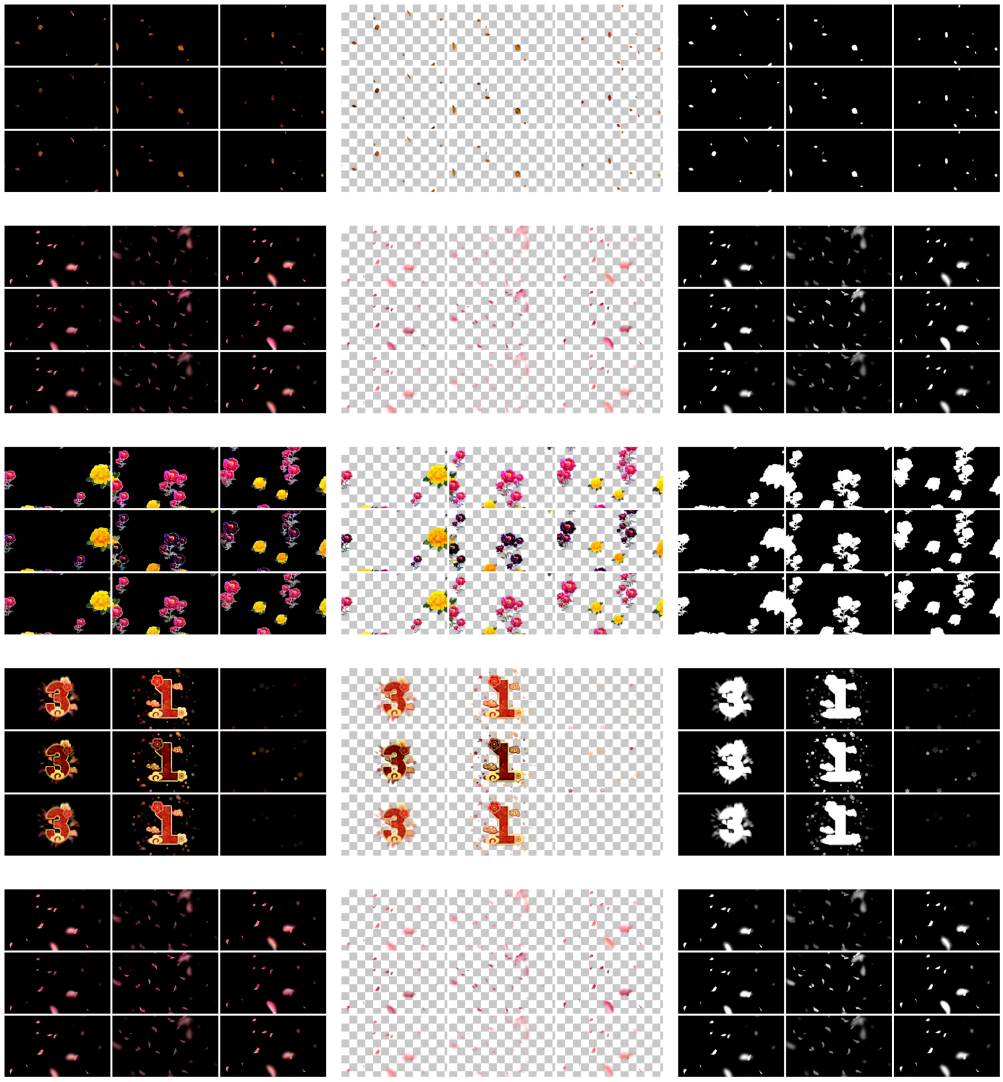

Figure S2: **Additional qualitative reconstruction demos (page 2/3).** Five more selected held-out RGBA videos, shown in the same triptych format as Figure S1. Rows correspond to UniRGBA reconstruction, Wan-Alpha reconstruction, and ground truth. Across these cases, UniRGBA preserves both high-frequency color structure and soft alpha transitions that are critical for live-streaming visual-effect assets.

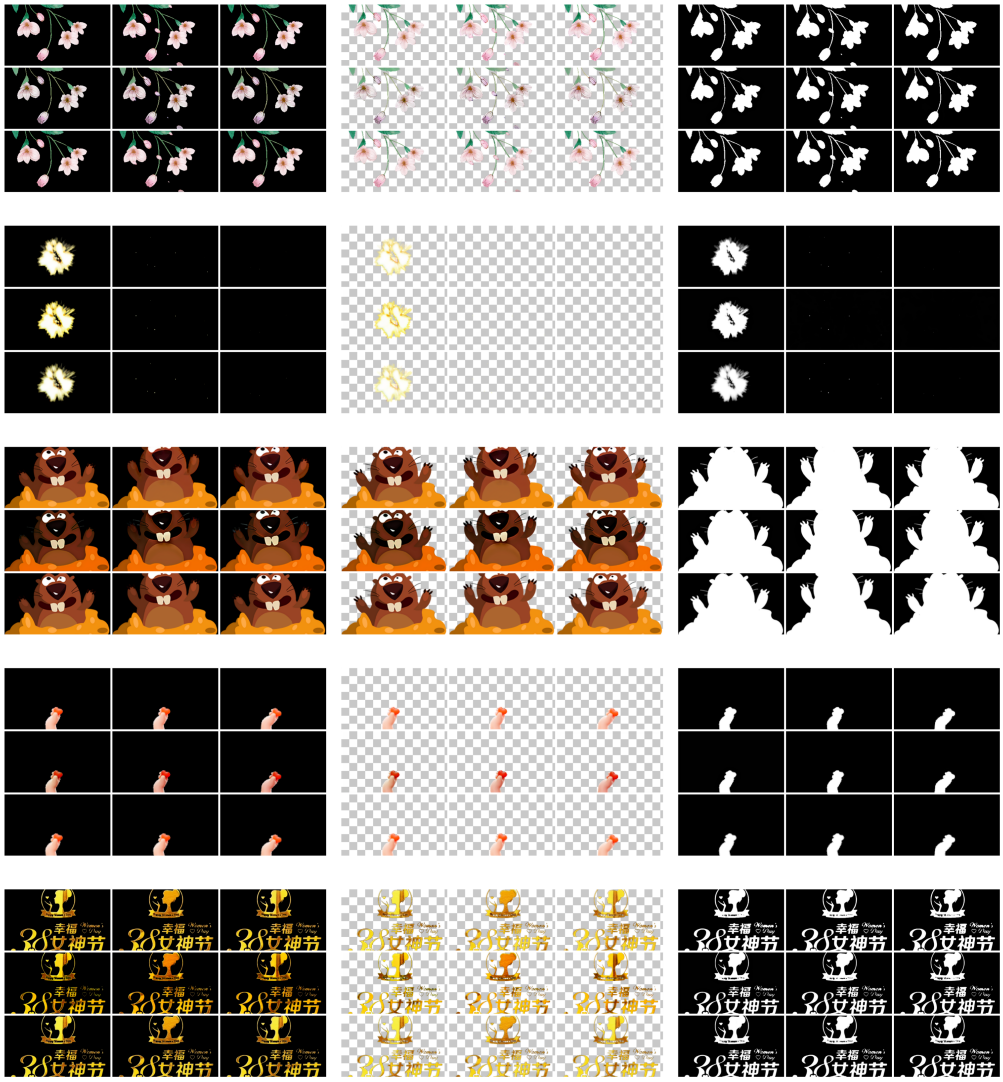

Figure S3: **Additional qualitative reconstruction demos (page 3/3).** Five further selected held-out RGBA videos, again shown as black-composited RGB, checkerboard-composited RGB, and alpha, with rows corresponding to UniRGBA reconstruction, Wan-Alpha reconstruction, and ground truth. These additional examples reinforce the robustness of UniRGBA on challenging semi-transparent structures, thin boundaries, and complex visual-effect textures.

E LIMITATIONS AND FUTURE WORK

UniRGBA has three limitations. First, the primary RGBA data are internal and the current DoRA training set is limited to about 2K visual-effect videos; a public benchmark with broader object- and scene-centric coverage would make cross-work comparison more direct. Second, our current pipeline is evaluated at 832×480; scaling the unified VAE and DiT adapter to higher spatial resolutions (e.g., 720p/1080p) and longer temporal windows remains an open engineering direction. Finally, we restrict adaptation to lightweight DoRA for controlled comparison with Wan-Alpha; exploring full-parameter fine-tuning and larger adapter capacity may further improve high-level aesthetic attributes. The release plan is therefore conditional on asset licensing and focuses on code, weights, evaluation scripts, and a synthetic or curated subset that can be redistributed.

F SUPPLEMENTARY I2V EVALUATION AND VISUALIZATION

Table S4 reports the quantitative I2V comparison. Figure S4 provides an additional qualitative RGBA I2V example with six uniformly sampled frames to illustrate temporal coherence, alpha-aware motion evolution, and preservation of delicate semi-transparent structures over time.

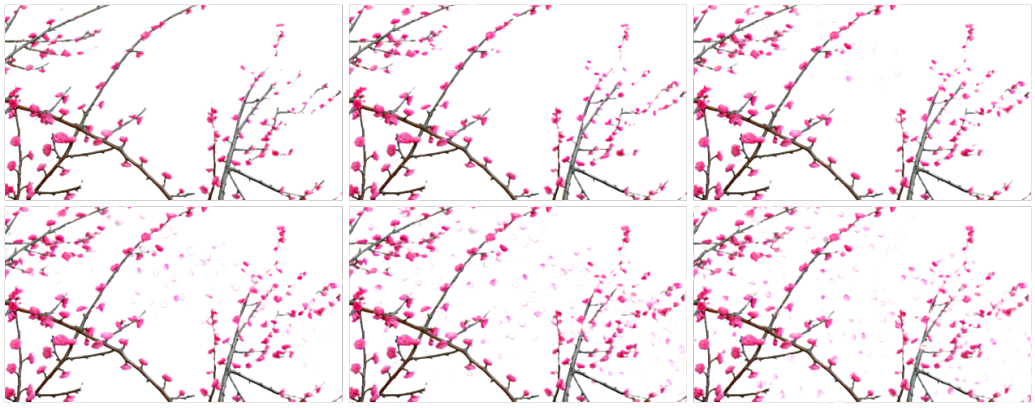

Figure S4: **Supplementary qualitative RGBA I2V result.** Six uniformly sampled frames from an RGBA image-to-video generation example. The prompt is: "A tree branch is covered with pink blossoms, and petals slowly drift through the air. The branch is dark brown, the flowers are densely distributed, and some petals have already fallen, forming a trajectory extending from the lower-left to the upper-right. The petals are pink and white, swaying gently in the wind." The generated sequence maintains coherent blossom appearance while producing temporally smooth petal motion and soft semi-transparent details.