VISIBILITY INVARIANCE IS NOT ENOUGH FOR RGBA REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Variational autoencoders (VAEs) for transparent image generation must preserve both appearance and opacity. RGB values at zero opacity do not affect rendering, and premultiplication removes this invisible-color variation exactly. Does this make RGBA autoencoding better? We compare RGBA input representations by adapting two pretrained RGB VAEs, SDXL and Ostris. Three transformations with the same visibility invariance produce different reconstruction quality. After separate validation-based tuning, premultiplication increases alpha mean squared error (MSE) on held-out video frames from the training-source collection by 36.63% for SDXL and 25.68% for Ostris. On human photographs, the same SDXL models instead favor premultiplication, with 70.91% lower alpha MSE and no retraining; the Ostris difference remains uncertain. In SDXL generation with 40 Euler steps and guidance 5, the video-selected premultiplied VAEs improve image-distribution and text-alignment scores despite their worse video opacity reconstruction. These findings neither establish more accurate generated alpha maps nor extend the gains to other sampling settings. Removing invisible RGB guarantees input invariance, but its effect on reconstruction and generation must be evaluated for the intended model, data domain and use. Code is available at `https://anonymous.4open.science/r/RGBA-Invariance`.

1 INTRODUCTION

Transparent assets can be composited onto new backgrounds and edited as separate layers. Recent generative models support these uses by extending pretrained RGB latent models to transparency, as in LayerDiffuse and AlphaVAE (Zhang & Agrawala, 2024; Wang et al., 2025b), and by modeling multiple image or video layers (Yin et al., 2025; Li et al., 2026). An RGBA variational autoencoder (VAE) must preserve color and opacity while providing a latent space for generation. Choosing its input representation therefore requires evaluating both reconstruction and downstream use.

RGBA introduces an ambiguity absent from opaque RGB images. PNG stores color independently of alpha (World Wide Web Consortium, 2025), whereas compositing removes the contribution of color wherever alpha is zero (Porter & Duff, 1984). Different stored RGB values can thus produce identical renderings on every background. Changing these invisible target colors leaves a composite-based loss unchanged if the prediction stays fixed. Feeding the modified image to an encoder, however, can change its reconstruction. In a released AlphaVAE, changing only invisible RGB changes deterministic reconstructions (Figure 1). Reconstruction error can increase or decrease, so this sensitivity alone does not establish that invisible-color variation is harmful.

Premultiplication removes this ambiguity while retaining the information needed for rendering. Classical compositing uses color multiplied by opacity (Porter & Duff, 1984); LayerDiffuse also multiplies normalized RGB by alpha in its transparency-encoding path (Zhang & Agrawala, 2024). Equivalent renderings then give identical encoder inputs. Yet premultiplication also rescales partially visible colors and changes the inputs seen by pretrained RGB filters. Its exact invariance does not tell us how a VAE will learn from those inputs within a finite training budget.

We ask: *does removing invisible-color variation improve RGBA reconstruction and downstream generation?* Within each experiment, input variants share the decoder target, objective and training budget. We evaluate opacity, rendered appearance and generation separately: compositing combines color and opacity errors, while latent structure also affects diffusion (Skorokhodov et al., 2025).

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.

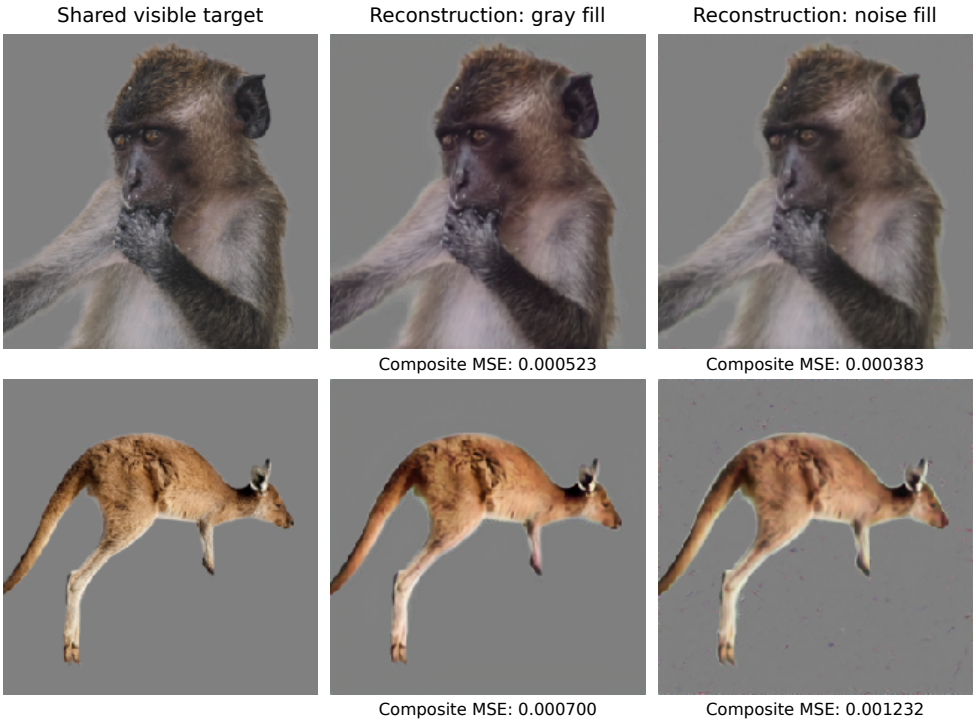

Figure 1: Identical target renderings can yield different reconstructions. In two predefined AlphaVAE examples, replacing gray RGB with uniform noise where alpha is exactly zero changes the output. Composite error decreases in the first example and increases in the second. The target composites remain identical.

Our study has three parts. **First**, we compare raw RGBA with three transformations that remove the same invisible-color variation. Their reconstruction quality differs despite their shared invariance. **Second**, we tune raw and premultiplied VAEs separately on video validation data. Opacity and rendered-image metrics can favor different inputs, and the opacity ranking can reverse when the same models are tested on photographs. **Third**, we use the video-selected VAEs in SDXL generation. Premultiplication improves generation scores with 40 Euler steps and guidance 5 despite worse video opacity reconstruction.

2 Related Work

Transparency, matting, and layered generation. Matting estimates opacity from composites (Levin et al., 2008; Xu et al., 2017), while layer decomposition recovers compositable layers that can include visual effects (Yang et al., 2025). Generative approaches produce objects with matting annotations (Hu et al., 2025), transparent images (Zhang & Agrawala, 2024; Wang et al., 2025b; Yu et al., 2026), or transparent and layered video (Wang et al., 2025a; Li et al., 2026). Qwen-Image-Layered supports editable layers through a unified RGB/RGBA VAE (Yin et al., 2025). These tasks motivate RGBA autoencoders that preserve both appearance and opacity. We compare their encoder input representations using known opacity targets and matched training conditions.

Invariance and reconstruction quality. Premultiplication removes invisible RGB while retaining renderings on every background (Porter & Duff, 1984). More broadly, Achille & Soatto (2018) study representations that discard nuisance variation while retaining task information. Work on domain invariance shows that invariant features need not yield low target-domain error (Zhao et al., 2019), though our equivalence concerns identical renderings rather than matched feature distributions. The perception–distortion tradeoff likewise distinguishes different notions of reconstruction quality (Blau & Michaeli, 2018). Here, the compositing equation provides a specific reason to

evaluate opacity and rendered-image error separately, even after invisible-color variation has been removed.

Visual autoencoders for generation. Visual autoencoders provide discrete representations in VQ-VAE and VQGAN (van den Oord et al., 2017; Esser et al., 2021) and continuous VAE latents in latent diffusion (Rombach et al., 2022). Representation design can target compactness, as in TiTok's one-dimensional tokens (Yu et al., 2024), or reuse pretrained representation encoders (Zheng et al., 2026). Latent spectral regularization also affects generation (Skorokhodov et al., 2025). We hold the architecture fixed within each input comparison and evaluate whether changing the RGBA input helps reconstruction and downstream generation.

3 Visibility Invariance and RGBA Fidelity

3.1 Rendering equivalence

Let $x = (C, a)$ be an RGBA image of height H and width W, with straight RGB color $C \in [0,1]^{3\times H\times W}$, opacity $a \in [0,1]^{1\times H\times W}$ and premultiplied color $p = aC$. Rendering onto RGB background $b \in [0,1]^3$, broadcast over pixels, gives

$$R_b(C, a) = p + (1 - a)b. \tag{1}$$

All calculations use the stored RGB values, without color-profile conversion or transfer-function inversion. Given another image $x' = (C', a')$ with premultiplied color $p' = a'C'$, we say that x and x' are rendering-equivalent, written $x \sim x'$, if $R_b(x) = R_b(x')$ for every background b. Rendering on black gives p; rendering on white then determines a. Thus

$$x \sim x' \quad \Longleftrightarrow \quad (p, a) = (p', a'). \tag{2}$$

We call an input representation *visibility-invariant* if rendering-equivalent images give identical encoder inputs.

3.2 Input canonicalization

We compare four encoder inputs. The transformations preserve alpha and affect only the encoder input; decoder targets and outputs remain straight RGBA. Let u denote a pixel location and $c \in [0,1]^3$ a fixed RGB color shared across pixels. Brackets below concatenate RGB and alpha channels.

Raw RGBA. The input is $[C, a]$, including the stored RGB values where $a = 0$. Rendering-equivalent images can therefore give different encoder inputs.

Hard filling. We replace color only at completely transparent pixels:

$$C_{\mathrm{hard}}(u) = \begin{cases} C(u), & a(u) > 0, \\ c, & a(u) = 0. \end{cases} \tag{3}$$

The encoder receives $[C_{\mathrm{hard}}, a]$. Hard gray uses $c = 1/2$ in each channel and leaves all partially visible colors unchanged.

Visible padding. Visible padding copies the nearest nonzero-alpha color into each zero-alpha pixel, using Euclidean image-grid distance and a fixed rule for ties. An entirely transparent image uses c, set to gray in the four-input comparison. It leaves visible colors and alpha unchanged and does not enlarge the image.

Centered premultiplication. We transform the input as

$$T_c(C, a) = [aC + (1 - a)c, a]. \tag{4}$$

Unlike hard filling and visible padding, this also changes partially visible colors: they approach c as opacity decreases. Our primary comparison uses gray, $c = 1/2$ in each channel. This convention follows directly from normalizing RGB to $[-1, 1]$:

$$a(2C - 1) = 2[aC + (1 - a)/2] - 1. \tag{5}$$

Multiplying normalized RGB by alpha is therefore gray-centered premultiplication.

Hard filling, visible padding and premultiplication all give identical inputs for rendering-equivalent images. At nonzero alpha, $C = p/a$ is fixed by the rendering; at zero alpha, each transformation replaces the stored color without using it. We call these transformations canonicalizations; each preserves all information needed for rendering. For binary alpha, hard filling and premultiplication coincide when they use the same c; their difference comes from partially transparent pixels.

3.3 What invariance guarantees

If preprocessing T gives identical inputs for rendering-equivalent images $x \sim x'$, a deterministic encoder gives identical posterior parameters. Latent samples also match when they share the same random draw. Let $\hat{x} = (\hat{C}, \hat{a})$ be a reconstruction, with premultiplied color $\hat{p} = \hat{a}\hat{C}$. For fixed $\hat{x}$, a loss ℓ that depends on the target only through its composites satisfies $\ell(\hat{x}, x) = \ell(\hat{x}, x')$. This equality alone does not require identical encoder responses to the two inputs.

The guarantee is therefore specific: canonicalization makes the encoder insensitive to RGB changes at zero alpha. Its effect on reconstruction quality depends on how the transformed colors interact with pretrained filters and subsequent training. The three canonicalizations can differ in this respect even though they remove exactly the same variation.

3.4 Rendered and opacity fidelity

Mean squared error (MSE) on a composite combines color and opacity errors, so it need not track opacity error alone. To make this relationship explicit, for a pixel define errors $\Delta p = \hat{p} - p$ and $\Delta a = \hat{a} - a$. Here $\mathbb{E}_b$ averages over random background b, $\|\cdot\|_2$ is the Euclidean norm, and Cov and tr denote covariance and trace. Expanding around the mean gives

$$\mathbb{E}_b\|\Delta p - b\Delta a\|_2^2 = \|\Delta p - \mathbb{E}[b]\Delta a\|_2^2 + \mathrm{tr}(\mathrm{Cov}\, b)(\Delta a)^2. \tag{6}$$

For background channels with mean $1/2$ and common variance s_b, let $\mathcal{R}_{s_b}$ be this expected error averaged over RGB channels and pixels, D_{gray} the corresponding MSE on gray background $b = (1/2, 1/2, 1/2)$, and A the alpha MSE. Then

$$\mathcal{R}_{s_b} = D_{\mathrm{gray}} + s_b A, \qquad A = \mathrm{MSE}_\alpha\,. \tag{7}$$

For uniformly distributed background colors, $s_b = 1/12$. We call this expected error *composite MSE* throughout the experiments. Equation (7) is a centered form of the established background-moment objective (Wang et al., 2025b). Its two terms can move in opposite directions: an improvement in the gray-background term can outweigh a rise in alpha MSE. We therefore evaluate rendered appearance and opacity separately.

4 Reconstruction Experiments

4.1 Experimental setup

Models and comparisons. We adapt the RGB VAEs from SDXL (Podell et al., 2024; madebyollin, 2023) and Ostris (Ostris, 2024) to output straight RGBA, retaining four- and sixteen-channel latents, respectively. All VAE parameters are finetuned. Inputs are compared within each initializer.

The initial four-input experiment compares raw RGBA, hard gray, visible padding and gray premultiplication using three paired training seeds and 10k updates on a local RGBA image collection at 256^2. The main experiment then compares raw and premultiplied inputs after separate validation-based tuning on VideoMatte at 256^2. It uses nine training settings and seven paired seeds for each initializer and input: 252 models trained for 20k updates each. Data and budgets differ between experiments but remain matched within each comparison.

Training objective and hyperparameter selection. All input variants use the objective

$$\mathcal{L} = D_{\mathrm{gray}} + (1/12 + \lambda_\alpha)A + \lambda_L L_{\mathrm{train}} + 10^{-6}\overline{\mathrm{KL}}. \tag{8}$$

Here λ_α and λ_L weight extra alpha supervision and perceptual loss; $D_{\mathrm{gray}} + A/12$ is composite MSE by Equation (7). L_{train} is AlexNet learned perceptual image patch similarity (LPIPS), averaged

over black, gray and white composites. $\overline{\mathrm{KL}}$ is the posterior-to-standard-normal Kullback–Leibler divergence, averaged over the batch and latent elements. The four-input experiment uses $\lambda_\alpha = \lambda_L = 0$ and learning rate 10^{-5}. The main grid fixes $\lambda_L = 0.05$ and tests the same learning rates and alpha weights for both inputs.

For each initializer, input and seed, we select the learning rate and λ_α that minimize validation $A + \kappa L$ at the final training step, where L is evaluation LPIPS. The validation weight κ is separate from the training weights. The main comparison sets $\kappa = 0$ to select by opacity error alone; larger values include rendered-image quality in selection. Training duration is fixed throughout this selection. Appendix B reports all nine candidates.

Data and test splits. We use *internal* for a held-out split of the training-source collection and *external* for a separately supplied test collection. Neither is used for finetuning. VideoMatte has 6,080 training and 576 validation frames. Its internal test contains 768 frames in 45 source groups; its external test contains 80 frames from five official-test clips. Source groups collect detected duplicate or related content and are kept separate across splits. The same video-selected models are tested on all 85 PhotoMatte85 photographs, without retraining or photograph-based selection.

The local RGBA collection supports the initial comparison and downstream generation. It has 6,718 training images, 372 validation images and 358 internal-test images; the external AlphaTest collection contains 902 images. We reserve PhotoMatte transfer for video-trained models; Appendix A explains overlap with the local external collection and details data preparation.

Metrics and uncertainty. We report composite MSE, alpha MSE or mean absolute error (MAE), and rendered LPIPS (Zhang et al., 2018); lower values are better. Evaluation LPIPS uses AlexNet on a wider set of backgrounds than the training loss, as specified in Appendix A. Means and sample standard deviations (SDs) summarize scores on the full test set across training seeds. Differences are premultiplied minus raw unless another reference is stated.

Crossed bootstrap intervals resample paired training seeds and source groups. For tuned models, *fixed-choice* intervals retain the selected hyperparameters, while *reselected* intervals also resample validation groups and repeat selection. Our main tuned results use reselected intervals. We additionally report *seed-only* Student's t intervals, which vary training seeds while holding the observed test collection and selection fixed. All confidence intervals (CIs) are pointwise 95% intervals; an interval that includes zero does not establish equivalence.

4.2 Exact invariance does not determine reconstruction quality

Figure 2 shows that hard gray raises external composite MSE for SDXL but lowers it for Ostris. Visible padding does not consistently lower composite MSE and LPIPS across the two initializers. Gray premultiplication lowers both metrics for both initializers under these fixed settings. Thus inputs that all remove invisible RGB variation still differ in reconstruction quality.

The transformations treat partially visible color and pretrained inputs differently. Most hidden RGB in the local data is constant black, so replacing it with gray changes a regular input pattern rather than removing random noise. Appendix D tests black and gray centers and compatibility with the first pretrained layer. We next test whether premultiplication also improves opacity reconstruction after separate tuning for each input.

4.3 Premultiplication can worsen opacity reconstruction after tuning

For every initializer, input and seed, the main validation rule selects learning rate 10^{-4} and added alpha-MSE weight $\lambda_\alpha = 3$; the paired runs therefore use the same selected hyperparameters. Despite separate input-specific selection, premultiplication raises internal-video alpha MSE by 36.63% for SDXL and 25.68% for Ostris (Table 1); the paired intervals for relative change are positive (Figure 3).

External-video mean alpha MSE also increases, by 48.27% for SDXL and 28.66% for Ostris, but evidence for this ordering weakens when uncertainty across the five source groups is included. Both seed-only difference intervals are positive. With source-group resampling, SDXL's reselected inter-

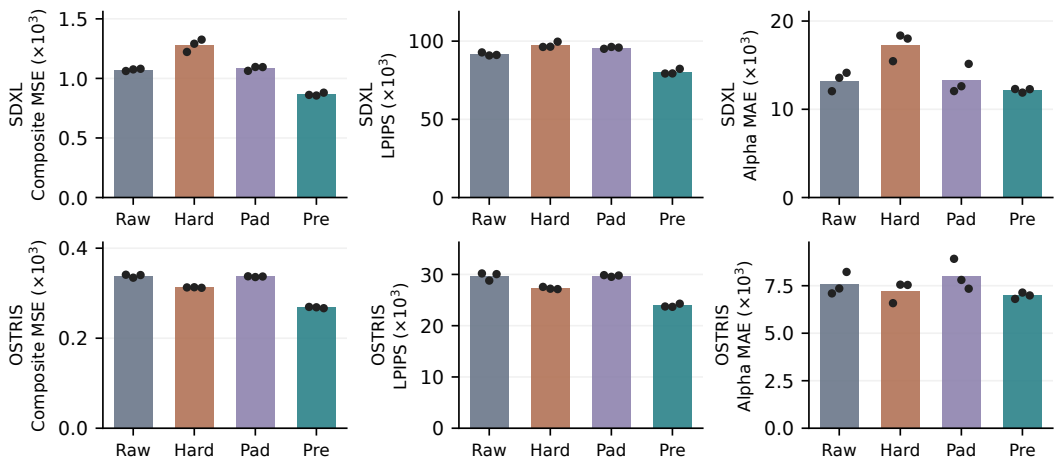

Figure 2: Four-input comparison on the external test. Bars show means over three training seeds; dots show individual seeds. Raw/Hard/Pad/Pre denote raw RGBA, hard gray, visible padding and gray premultiplication. The latter three remove the same invisible RGB variation. Alpha error is MAE; panel scales and initializer capacities differ. Appendix D reports both test sets and paired intervals.

Table 1: VideoMatte test results after selection by alpha MSE, with seven paired seeds. Raw/Premul. cells show means and ± sample SD; bold marks the lower mean within each model, test and metric. Δ is premultiplied minus raw, with reselected 95% CIs below. Alpha and composite MSE use 10^{-6} units; LPIPS uses 10^{-3} units. Figure 3 shows relative alpha-MSE changes.

Model	Statistic	Internal			External		
		Alpha MSE ↓	Comp. MSE ↓	LPIPS ↓	Alpha MSE ↓	Comp. MSE ↓	LPIPS ↓
SDXL	Raw	**93.23** ± 1.75	397.5 ± 5.5	10.096 ± 0.146	**73.22** ± 1.72	192.6 ± 5.6	5.166 ± 0.164
	Premul.	127.38 ± 7.39	**381.3** ± 2	**9.469** ± 0.092	108.57 ± 7.09	**180.3** ± 2.1	**4.654** ± 0.085
	Δ	34.15	−16.23	−0.627	35.35	−12.32	−0.513
	95% CI	[19.39, 52.4]	[−24.16, −10.37]	[−0.912, −0.406]	[−2.917, 90.72]	[−33.61, 7.224]	[−1.067, −0.094]
Ostris	Raw	**16.16** ± 2.10	82.5 ± 5.5	**1.791** ± 0.122	**12.88** ± 2.24	**43.2** ± 2.9	**0.977** ± 0.066
	Premul.	20.32 ± 1.07	**82.1** ± 1.4	1.812 ± 0.061	16.57 ± 1.54	43.8 ± 1.5	1.013 ± 0.049
	Δ	4.152	−0.412	0.021	3.692	0.579	0.036
	95% CI	[1.062, 8.095]	[−4.312, 3.044]	[−0.076, 0.102]	[0.022, 8.007]	[−2.083, 2.856]	[−0.025, 0.095]

val includes zero; Ostris's fixed-choice interval includes zero, while its reselected lower bound is just above zero.

Dependence on the validation target. The selected models are unchanged at $\kappa = 0, 0.001, 0.01$. At $\kappa = 0.1$, some seeds select weaker alpha supervision ($\lambda_\alpha = 1$ or $1/3$). Both initializers have smaller mean internal alpha-MSE differences, with reselected intervals that include zero, giving weaker evidence of an opacity penalty under the combined validation objective. The selected learning rate remains unchanged.

Rendered and opacity fidelity disagree. For SDXL, premultiplication lowers composite MSE on internal video and LPIPS on both video tests, with negative reselected difference intervals. Its external composite-MSE interval includes zero. For Ostris, the composite-MSE and LPIPS intervals include zero on both video tests. As Equation (7) allows, rendered and alpha errors can move differently. Even the opacity metric matters: Ostris's mean alpha MAE decreases while MSE increases, but the video MAE intervals include zero (Appendix B).

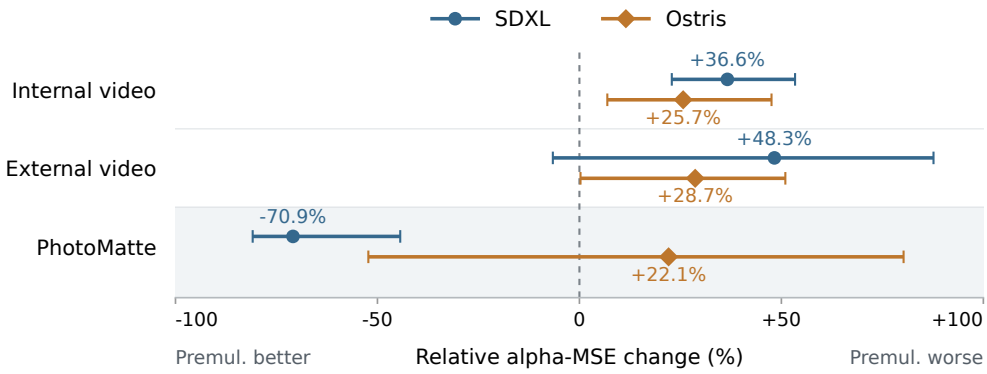

Figure 3: Relative alpha-MSE change across domains for the same video-selected models (premultiplied/raw minus one, in percent). Markers distinguish SDXL and Ostris; bars show reselected paired 95% CIs. Negative values favor premultiplication. SDXL reverses ordering on photographs; Ostris's photographic interval includes zero. Table 1 gives absolute video scores; Appendix B gives results for all validation settings.

4.4 Opacity rankings can reverse across domains

On PhotoMatte, the same SDXL models reduce alpha MSE by 70.91%, with negative difference intervals and lower error in all seven paired runs (Figure 3). This reverses the video ranking without retraining or photograph-based selection.

For Ostris, premultiplication instead raises mean alpha MSE by 22.08%, with a difference interval that includes zero. It lowers mean composite MSE for both initializers, but Ostris's reselected LPIPS difference interval includes zero. The clear opacity reversal therefore applies to SDXL; the photographic opacity ordering remains uncertain for Ostris.

Large errors on a few images can dominate the mean. One image accounts for 66.82% of SDXL's net alpha-MSE decrease; removing it still leaves a negative mean difference. For Ostris, premultiplication lowers error on 71/85 images but raises the overall mean, with 61.20% of that net increase coming from the same image. All primary estimates use the full test set; Appendix B.4 gives the source and image influence analysis.

5 Downstream Generation

5.1 Experimental setup

Models and matched adaptation. We adapt SDXL with seven frozen raw/premultiplied VAE pairs selected on video validation data. These VAEs were trained for 20k updates with learning rate 10^{-4}, $\lambda_\alpha = 3$ and $\lambda_L = 0.05$. The *Joint* study pairs each VAE training seed with one generator training seed, so variation comes from both stages. The *Fixed* study uses seven generator training seeds with one VAE pair, measuring generator variation conditional on those VAEs. Each study has seven raw/premultiplied pairs of generator runs. They share one pair, giving 26 distinct adaptations, and are analyzed separately.

We adapt the SDXL UNet with low-rank adaptation (LoRA), keeping the VAEs, base weights and text encoders frozen (Podell et al., 2024; Hu et al., 2022). Both inputs use matched random draws and the same training budget: 20k updates on the 6,718 local image–caption pairs at 512^2, with learning rate 10^{-4}. The diffusion objective is

$$\mathcal{L}_{\text{diff}} = \frac{1}{d}\mathbb{E}\left[\|\epsilon_\theta(z_t, t, y) - \epsilon\|_2^2\right]. \tag{9}$$

Here d is the number of latent elements, z_t is a scaled VAE posterior sample with noise added at diffusion timestep t, and y is SDXL conditioning. The predictor ϵ_θ uses trainable LoRA parameters

θ to predict standard Gaussian noise ϵ. Expectation covers data and training randomness. This objective has no direct alpha, composite or perceptual loss.

Evaluation and uncertainty. Evaluation uses 358 internal and 902 external prompts, with four matched noise draws per prompt, 40 Euler steps and classifier-free guidance (CFG) 5. We measure generated image-distribution quality with kernel inception distance (KID) (Bińkowski et al., 2018) and text alignment with contrastive language–image pretraining (CLIP) cosine similarity using ViT-B/32 (Radford et al., 2021). Both scores are averaged over black, gray and white composites. Lower KID and higher CLIP are better.

Crossed intervals independently resample paired training runs, prompt source groups and reference-image source groups. All four noise draws for a prompt remain together. Seed-only Student's t intervals hold the prompt and reference collections fixed. Appendix C gives the adaptation and evaluation protocols, including how source groups are linked.

5.2 GENERATION QUALITY

Table 2: Generation with 40 Euler steps and CFG 5, using seven paired runs per study. Raw/Premul. cells show means and ± sample SD; bold marks lower KID or higher CLIP within each study and test. Δ is premultiplied minus raw, with crossed 95% CIs below. All values use 10^{-3} units. Joint varies both VAE and generator seeds; Fixed varies generator seeds for one VAE pair. The studies share one pair. Appendix C gives seed-only intervals.

Study	Statistic	Internal		External	
		KID ↓	CLIP ↑	KID ↓	CLIP ↑
Joint	Raw	2.138 ± 0.412	315.69 ± 1.80	3.899 ± 0.694	313.89 ± 1.36
	Premul.	**1.416** ± 0.267	**319.85** ± 0.91	**2.989** ± 0.435	**317.66** ± 1.22
	Δ	–0.722	4.165	–0.911	3.77
	95% CI	[−1.376, −0.171]	[2.774, 5.642]	[−1.59, −0.262]	[2.596, 4.971]
Fixed	Raw	2.104 ± 0.393	316.71 ± 0.93	3.706 ± 0.616	314.71 ± 0.88
	Premul.	**1.539** ± 0.211	**319.51** ± 1.22	**3.063** ± 0.743	**317.61** ± 2.01
	Δ	–0.565	2.796	–0.644	2.903
	95% CI	[−1.141, −0.048]	[1.568, 3.989]	[−1.279, −0.053]	[1.492, 4.162]

With 40 Euler steps and CFG 5, premultiplication lowers mean KID and raises mean CLIP in both studies and on both test sets (Table 2). The crossed difference intervals support all four KID improvements, while the Fixed study's external seed-only interval includes zero. All four CLIP difference intervals are positive with both methods. Both studies favor premultiplication on average, despite their different sources of training variation. These video-selected VAEs give better generation scores despite worse video opacity reconstruction. Because the tasks and test domains differ, this comparison does not show that poorer opacity reconstruction causes better generation.

Additional checks. Internal prompts and reference images can share a source. Resampling their linked groups together still gives negative KID difference intervals: $[-1.239, -0.247]$ for Joint and $[-1.033, -0.115]$ for Fixed, in 10^{-3} units. The gain is less certain on individual backgrounds: gray and white KID difference intervals remain negative, while those on black include zero. Appendix C.3 examines other sampling settings with a smaller prompt subset, leaving their KID gains unresolved; Appendix C.4 reports reconstruction at the generation resolution to compare errors on known targets.

Qualitative analysis. Figure 4 shows that models trained with premultiplied inputs can still produce alpha artifacts, including uneven opacity within the generated leopard. These examples highlight a limitation of the aggregate scores: improvements in KID and CLIP do not ensure artifact-free alpha maps.

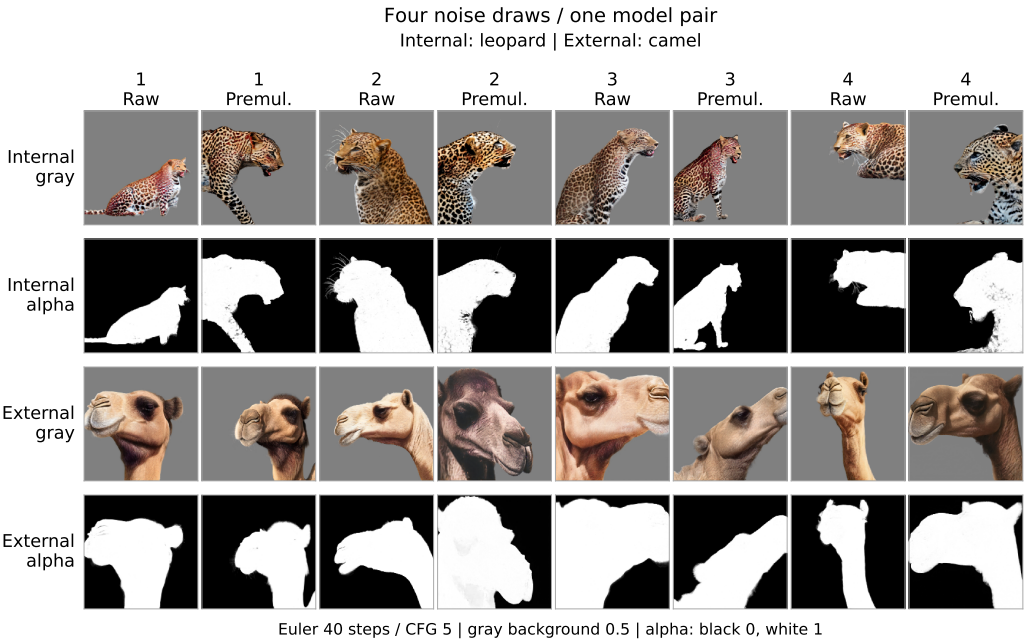

Figure 4: Generated composites on gray backgrounds and alpha maps for an internal leopard prompt and an external camel prompt. Raw/Premul. columns show one paired training run and four matched noise draws, with 40 Euler steps and CFG 5. Prompts and the paired run are chosen without inspecting output quality.

6 Discussion and Limitations

For compositing, evaluate alpha accuracy and appearance on the intended backgrounds: average composite error can hide opacity errors or errors on specific backgrounds. For generation, inspect alpha maps alongside KID and CLIP, which assess rendered images rather than opacity directly.

These recommendations come from finite-budget adaptation of two VAEs with different capacities and pretraining. Selected alpha weights reach the upper limit of the tested grid, so the selected models need not be optimal for either input. Longer training or a wider search could change the comparison.

VideoMatte has only five external source groups, and PhotoMatte contains 85 images from one studio without verified subject independence. Content, compression, framing and alpha statistics differ across domains, leaving the cause of the reversal unresolved. Generation covers four-channel SDXL, one latent scale and one adaptation setup, without measuring alpha-map accuracy or human preference. With few seeds and uncertain source independence, nominal CI coverage is not guaranteed.

7 Conclusion

Removing invisible RGB guarantees input invariance but does not guarantee better RGBA reconstruction. Equally invariant inputs yield different reconstruction quality, and premultiplication can improve rendered appearance while worsening opacity reconstruction even after separate tuning. For SDXL, the opacity ranking reverses between video and photographs without retraining. The video-selected premultiplied VAEs also improve SDXL generation scores with 40 Euler steps and guidance 5, although alpha artifacts remain. Together, these results show that the value of an input convention depends on the model, metric and test domain. Visibility invariance is a useful property to establish, but selecting an RGBA representation also requires evidence from reconstruction and downstream use.

AI USE STATEMENT

We used generative AI tools to create and edit software code, assist with translation, and revise manuscript wording and organization. We also used these tools to discuss experimental comparisons and limitations and to help interpret reported results, confidence intervals and visual examples. We did not use generative AI tools to generate synthetic datasets, clean or reformat datasets, or perform thematic coding of qualitative data. Developing theoretical models, formulating mathematical claims, writing proofs and proposing hypotheses were not part of this use. The authors take responsibility for the final methods, analyses, text, code and artifacts.

REPRODUCIBILITY STATEMENT

The complete code is available in the anonymous repository linked in the abstract. Appendix A describes data preparation, VAE training and statistical evaluation; Appendix C describes diffusion adaptation and sampling. The appendices and code together specify the implementation details, hyperparameters and random seeds used in the experiments.

REFERENCES

Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research*, 19(50):1–34, 2018. URL https://jmlr.org/papers/v19/17-646.html.

Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1801.01401.

Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6228–6237, June 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Blau_The_Perception-Distortion_Tradeoff_CVPR_2018_paper.html.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12873–12883, June 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Esser_Taming_Transformers_for_High-Resolution_Image_Synthesis_CVPR_2021_paper.html.

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*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Xiaobin Hu, Xu Peng, Donghao Luo, Xiaozhong Ji, Jinlong Peng, Zhengkai Jiang, Jiangning Zhang, Taisong Jin, Chengjie Wang, and Rongrong Ji. DiffuMatting: Synthesizing arbitrary objects with matting-level annotation. In *Computer Vision – ECCV 2024*, volume 15126 of *Lecture Notes in Computer Science*, pp. 396–413. Springer Nature Switzerland, 2025. doi: 10.1007/978-3-031-73113-6_23. URL https://doi.org/10.1007/978-3-031-73113-6_23.

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. URL https://doi.org/10.1109/TPAMI.2007.1177.

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. *arXiv preprint arXiv:2508.04228v2*, 2026. URL https://arxiv.org/abs/2508.04228v2.

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. URL https://openaccess.thecvf.com/content/CVPR2021/html/Lin_Real-Time_High-Resolution_Background_Matting_CVPR_2021_paper.html.

madebyollin. SDXL-VAE-FP16-Fix. Hugging Face model card, 2023. URL https://huggingface.co/madebyollin/sdxl-vae-fp16-fix. Accessed September 25, 2026.

Ostris. Ostris VAE – KL-f8-d16. Hugging Face model card, 2024. URL https://huggingface.co/ostris/vae-kl-f8-d16. Accessed September 25, 2026.

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*, pp. 1862–1874, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/081b08068e4733ae3e7ad019fe8d172f-Paper-Conference.pdf.

Thomas Porter and Tom Duff. Compositing digital images. *ACM SIGGRAPH Computer Graphics*, 18(3):253–259, 1984. doi: 10.1145/964965.808606. URL https://doi.org/10.1145/964965.808606.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

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 (CVPR)*, pp. 10684–10695, June 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.html.

Ivan Skorokhodov, Sharath Girish, Benran Hu, Willi Menapace, Yanyu Li, Rameen Abdal, Sergey Tulyakov, and Aliaksandr Siarohin. Improving the diffusability of autoencoders. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 55876–55905. PMLR, 2025. URL https://proceedings.mlr.press/v267/skorokhodov25a.html.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/7a98af17e63a0ac09ce2e96d03992fbc-Paper.pdf.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18229–18239, June 2025a. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_TransPixeler_Advancing_Text-to-Video_Generation_with_Transparency_CVPR_2025_paper.html.

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:2507.09308v1*, 2025b. URL https://arxiv.org/abs/2507.09308v1.

World Wide Web Consortium. Portable network graphics (PNG) specification (third edition). W3C Recommendation, 24 June 2025, 2025. URL https://www.w3.org/TR/2025/REC-png-3-20250624/.

Ning Xu, Brian Price, Scott Cohen, and Thomas Huang. Deep image matting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2970–2979, July 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Xu_Deep_Image_Matting_CVPR_2017_paper.html.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7643–7653, June 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yang_Generative_Image_Layer_Decomposition_with_Visual_Effects_CVPR_2025_paper.html.

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. *arXiv preprint arXiv:2512.15603v1*, 2025. URL https://arxiv.org/abs/2512.15603v1.

Hao Yu, Jinglin Wang, Jiabo Zhan, Rui Chen, Zile Wang, Huaisong Zhang, Hongyu Li, Xinrui Chen, Yongxian Wei, and Chun Yuan. OmniAlpha: Aligning transparency-aware generation via multi-task unified reinforcement learning. *arXiv preprint arXiv:2511.20211v2*, 2026. URL https://arxiv.org/abs/2511.20211v2.

Qihang Yu, Mark Weber, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. An image is worth 32 tokens for reconstruction and generation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 128940–128966. Curran Associates, Inc., 2024. doi: 10.52202/079017-4096. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/e91bf7dfba0477554994c6d64833e9d8-Paper-Conference.pdf.

Lvmin Zhang and Maneesh Agrawala. Transparent image layer diffusion using latent transparency. *ACM Transactions on Graphics*, 43(4):1–15, 2024. doi: 10.1145/3658150. URL https://doi.org/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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_The_Unreasonable_Effectiveness_CVPR_2018_paper.html.

Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7523–7532. PMLR, 2019. URL https://proceedings.mlr.press/v97/zhao19a.html.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *International Conference on Learning Representations*, pp. 35791–35820, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/3c4141c12660ad3625eb4ae845e0a6f9-Paper-Conference.pdf.

A Data, Training, and Statistical Protocols

A.1 Data preparation and separation

Local RGBA collection. The initial reconstruction comparison and downstream adaptation use a collection of 8,645 candidate RGBA images. We group exact duplicates, images with identical visible content, and images whose rendered-image and alpha perceptual hashes both have Hamming distance at most 16. Connected components define the groups. We remove 1,197 candidate images linked to the external collection, then assign whole groups to training, validation and internal testing, yielding 6,718, 372 and 358 images, respectively. The separate external AlphaTest collection contains 902 images from its supplied AIM and validation subsets, with 500 and 402 images, respectively. Both subsets are used only for testing here. There are 336 internal-test groups and 858 external-test groups. These checks screen for duplicates but do not establish subject independence. Source images retain their original access terms and are not redistributed as a new dataset.

We resize and crop before applying any input transformation. Float32 antialiased bilinear interpolation resizes premultiplied RGB and alpha, after which straight RGB is recovered wherever alpha is nonzero. At zero alpha, we retain the interpolated original RGB. Images are resized by their shorter side and center-cropped to 256^2 for the initial comparison or 512^2 for generation, without color-profile conversion.

Among the original candidates, 8,099 have constant black hidden RGB, 537 have varying hidden RGB, and nine have no zero-alpha pixels. In the external collection, 901 have constant black hidden RGB and one has varying hidden RGB. The constant-color criterion uses channel standard deviation below 10^{-6}. Hidden-color noise is therefore an intervention for measuring sensitivity, not a description of typical stored colors. Every prepared local test image contains partially transparent pixels; the binary-alpha statement in Section 3.2 is a mathematical observation rather than a separate test condition.

VideoMatte240K. The standard-definition video collection contains 479 training-source clips and five official-test clips (Lin et al., 2021). We sample 16 evenly spaced frames per clip, including the endpoints and using nearest-even rounding, giving 7,744 JPEG foreground/alpha pairs. Images are resized to fit 256^2 while preserving aspect ratio, using premultiplied RGB and alpha, and centered on a zero-RGBA canvas. A 768×432 frame becomes 256×144, with 56 padded rows on each side. Interpolation overshoots are clamped. Fifteen fully transparent pairs remain in the source counts.

We group clips linked by exact matches or by rendered-image and alpha perceptual hashes both within Hamming distance four. The 382 resulting groups capture detected duplicate content, not verified subject identities. Training-source clips linked to the official test set are excluded; remaining groups are assigned to disjoint splits (Table 3). The largest original group contains 57 clips. JPEG alpha, repeated frames and uncertain subject identities limit the independence of these groups.

Table 3: VideoMatte split after group separation. The excluded row is not used in training or evaluation.

Split	Frames	Clips	Groups
Training	6,080	380	297
Validation	576	36	35
Internal test	768	48	45
Official external test	80	5	5
Excluded external-linked candidates	240	15	1

PhotoMatte85. We retain all 85 original RGBA PNGs from Background Matting V2 (Lin et al., 2021). Most are 2304×3456; resizing, visible-color recovery and centered padding follow the video procedure. Exact duplicates, identical visible content, and rendered-image and alpha perceptual hashes both within Hamming distance four define 82 visual groups. The photographs come from one studio, and subject identities are unverified. These checks find no matches to the prepared VideoMatte frames but cannot exclude shared subjects.

PhotoMatte overlaps with local AlphaTest, so it is not an independent addition to that external image collection. We restrict the photographic transfer comparison to video-trained models selected only on video validation data. Locally finetuned models are excluded from this comparison. The dataset's research-use terms are distinct from the accompanying software license.

A.2 Inputs, initialization and training objectives

Input and output conventions. Section 3.2 defines the four encoder inputs. Visible padding uses a distance transform with deterministic tie handling. After the input transformation, all four channels are normalized to $[-1, 1]$. Decoder outputs are rescaled to straight RGBA and are never unpremultiplied.

Initialization. Compatible RGB weights are copied from the SDXL and Ostris VAEs; added alpha-input and alpha-output weights are zero. A normalized alpha-output bias of zero gives initial decoded opacity 0.5, while a bias of one gives opacity one. At 256^2 input, SDXL has four latent channels and Ostris has sixteen, both at spatial size 32^2. The repository records exact pretrained versions and software environments.

Training. All variants minimize Equation (8) with sampled diagonal-Gaussian posteriors and unclamped predictions. The initial four-input experiment and color-center controls use $\lambda_\alpha = \lambda_L = 0$, initial opacity 0.5, learning rate 10^{-5}, batch size eight and 10k updates, with three paired seeds per input. The video grid uses initial opacity one, $\lambda_L = 0.05$, batch size eight and 20k updates, with seven paired seeds for each of the nine settings in Table 4.

All VAE parameters are trained with float32 AdamW, first- and second-moment decay rates of 0.9 and 0.999, a constant learning rate, zero weight decay and global gradient-norm clipping at one. The video grid uses fixed preprocessed images without online random cropping or background blending. Validation is performed every 500 updates in the initial experiment and every 1,000 in the video grid. All comparisons use the final checkpoint, including when an earlier checkpoint has a better validation score.

Table 4: Complete training grid. Each row is tested with both initializers, both inputs and all seven seeds. The added alpha coefficient is in addition to the $1/12$ term in Equation (8). All rows share $\lambda_L = 0.05$ and 20k updates.

Learning rate	**Added λ_α**
10^{-5}	0
10^{-5}	$1/3$
3×10^{-5}	0
3×10^{-5}	$1/3$
10^{-4}	0
10^{-4}	$1/3$
10^{-4}	1
10^{-4}	3
3×10^{-4}	$1/3$

A.3 Evaluation and uncertainty

Reconstruction metrics. We decode the posterior mean and clamp RGBA to $[0, 1]$. Composite MSE is the analytical expectation over uniformly distributed background colors, $D_{\text{gray}} + A/12$. Evaluation LPIPS uses AlexNet and averages nine backgrounds: black, gray, white, red, green, blue, two checkerboards with cell widths eight and 32 and levels 0.2 and 0.8, and fixed RGB noise. Alpha MSE and MAE directly compare predicted and target opacity. For hidden-RGB interventions, *rendered drift* is composite MSE between reconstructions before and after the intervention, rather than error against the target. Bold table entries mark the better means within the groups defined by each caption; difference estimates and CI bounds use regular type.

Crossed bootstrap. Absolute differences are premultiplied minus raw unless another reference is stated. Let s index seeds and i index images, d_{si} be their paired metric difference, and $g(i)$ be the source group containing image i. In a crossed bootstrap draw, w_s and v_g count the sampled occurrences of seed s and group g. The mean paired difference is estimated by

$$\widehat{\Delta}^* = \frac{\sum_{s,i} w_s v_{g(i)} d_{si}}{(\sum_s w_s)(\sum_i v_{g(i)})}. \tag{10}$$

This preserves the image-weighted target when groups differ in size. Backgrounds are not separate replicates. We use percentile intervals with 2,000 draws for the initial experiment and canonicalization controls. Selected-model comparisons also use 2,000 draws, except for the main external-video alpha estimates, which use 500,000.

Selection and interval definitions. Hyperparameters are selected separately for each initializer, input and seed by final-step validation $A + \kappa L$, with $\kappa \in \{0, 0.001, 0.01, 0.1\}$. Ties favor the lower learning rate, then the lower added alpha weight. Fixed-choice intervals resample test groups and paired seeds while retaining these selections. Reselected intervals additionally resample validation groups independently, repeat selection for each input and seed, and evaluate the selected models on the resampled test groups. In the seven-seed experiments, seed-only intervals use Student's t distribution with six degrees of freedom over paired full-test means, keeping test images and validation choices fixed.

All intervals are pointwise, without multiple-comparison correction. Repeated evaluation, reused seeds and additional bootstrap draws do not provide independent training replications. With few seeds, only five external video groups and uncertain subject dependence in photographs, nominal coverage is not guaranteed. An interval that includes zero does not demonstrate equivalence; no equivalence margin is specified.

Relative changes. For Figure 3, let A_{raw} and A_{pre} be aggregate alpha MSE for raw and premultiplied inputs. Relative change is $100(A_{\mathrm{pre}}/A_{\mathrm{raw}} - 1)\%$. Each bootstrap draw repeats input-specific validation selection and recomputes both means using the same resampled seeds and test groups, then takes their ratio. Relative CIs use the percentiles of these ratios. They use the same paired draws as the absolute-difference intervals: 2,000 for internal video and photographs, and 500,000 for external video. Exact random seeds and implementation details are recorded in the repository.

B Training-Grid Results and Cross-Domain Transfer

B.1 The complete nine-candidate comparison

Tables 5 and 6 report validation means over seven seeds for all nine training settings in Table 4. These means summarize the candidates; selection uses each input and seed separately. The grid was expanded from six to nine settings after inspecting earlier results. The three added settings use the same seven seeds and contribute 84 runs to the total of 252. Every run completes 20k updates and is evaluated at its final checkpoint on validation and both full test sets.

At $\kappa = 0, 0.001, 0.01$, all 28 initializer/input/seed selections choose learning rate 10^{-4} and $\lambda_\alpha = 3$. At $\kappa = 0.1$, the learning rate is unchanged, but some selections use less alpha supervision. SDXL raw selects $\lambda_\alpha = 3$ in three seeds and 1 in four; SDXL premultiplication selects 3 in all seven. Ostris raw selects 3, 1 and $1/3$ in two, four and one seeds, respectively; Ostris premultiplication selects 3 in four seeds and 1 in three. Across the four validation objectives, 40 distinct trained models are selected.

The alpha-only selections reach the largest tested alpha weight. The highest tested learning rate, 3×10^{-4}, is never selected and has larger training-loss spikes and greater variation across seeds, especially for Ostris. Recorded losses and diagnostics remain finite; gradient clipping occurs in zero to 8.54% of updates across runs. These finite-budget results do not establish convergence or a globally optimal choice of hyperparameters. Full training curves and per-seed results are available in the repository.

Table 5: SDXL: validation means over seven seeds for all nine training settings, in 10^{-3} units. For each metric and input (Raw or Premul.), bold marks the lowest mean across the nine settings.

Learning rate	λ_α	Input	Alpha MSE ↓	LPIPS ↓	Comp. MSE ↓
10^{-5}	0	Raw	0.4487	8.333	0.336
		Premul.	0.5137	7.653	0.2977
10^{-5}	$1/3$	Raw	0.3754	8.228	0.3315
		Premul.	0.4111	7.416	**0.293**
3×10^{-5}	0	Raw	0.3204	8.089	0.3316
		Premul.	0.428	7.88	0.3103
3×10^{-5}	$1/3$	Raw	0.2432	7.995	0.3259
		Premul.	0.2924	7.474	0.3036
10^{-4}	0	Raw	0.2319	7.634	0.315
		Premul.	0.3328	7.474	0.3042
10^{-4}	$1/3$	Raw	0.1725	**7.519**	**0.3112**
		Premul.	0.2359	7.242	0.298
10^{-4}	1	Raw	0.1324	7.72	0.3128
		Premul.	0.1692	**7.213**	0.2984
10^{-4}	3	Raw	**0.0914**	8.096	0.3184
		Premul.	**0.103**	7.445	0.304
3×10^{-4}	$1/3$	Raw	0.1427	8.187	0.3152
		Premul.	0.2114	7.385	0.3001

Table 6: Ostris: validation means over seven seeds for all nine training settings, in 10^{-3} units. For each metric and input (Raw or Premul.), bold marks the lowest mean across the nine settings.

Learning rate	λ_α	Input	Alpha MSE ↓	LPIPS ↓	Comp. MSE ↓
10^{-5}	0	Raw	0.0717	1.745	0.0773
		Premul.	0.0729	1.624	0.0724
10^{-5}	$1/3$	Raw	0.0501	1.707	0.0770
		Premul.	0.0554	1.634	0.0732
3×10^{-5}	0	Raw	0.0502	1.49	0.0702
		Premul.	0.0563	**1.438**	**0.0670**
3×10^{-5}	$1/3$	Raw	0.0362	1.472	**0.0698**
		Premul.	0.0422	1.447	0.0678
10^{-4}	0	Raw	0.0512	1.53	0.0704
		Premul.	0.0589	1.49	0.0695
10^{-4}	$1/3$	Raw	0.0334	**1.466**	0.0712
		Premul.	0.0397	1.457	0.0672
10^{-4}	1	Raw	0.0224	1.478	0.0702
		Premul.	0.0284	1.451	0.0684
10^{-4}	3	Raw	**0.0166**	1.556	0.0737
		Premul.	**0.0187**	1.57	0.0720
3×10^{-4}	$1/3$	Raw	0.1769	10.2	0.2388
		Premul.	0.0664	2.92	0.1045

B.2 Selected results and sensitivity to the validation target

Table 7 gives absolute alpha-MSE differences for the models in Figure 3. Its reselected intervals use the same paired observations, validation objective and resampling draws as the figure's relative intervals.

Table 7: Alpha-MSE differences for the video-selected models in Figure 3, in 10^{-6} units. Cells show premultiplied minus raw, with reselected 95% CIs below. Resampling includes paired seeds, test groups and video-validation groups; photographs are used only for testing.

Model	Video internal	Video external	PhotoMatte
SDXL	34.15 [19.39, 52.4]	35.35 [−2.917, 90.72]	−572.1 [−1580, −122]
Ostris	4.152 [1.062, 8.095]	3.692 [0.022, 8.007]	49.19 [−290, 184]

Tables 8 and 9 compare intervals that include different sources of uncertainty: training seeds, test groups and validation selection. They report all three methods at $\kappa = 0$ and $\kappa = 0.1$. The intermediate values $\kappa = 0.001, 0.01$ select the same models as $\kappa = 0$ and therefore have the same point estimates; their full results are in the repository. At $\kappa = 0.1$, internal alpha-MSE differences remain positive on average but are smaller, and their reselected intervals include zero. The selected models for each validation objective are kept unchanged across all test domains.

Table 8: SDXL: video test results under two validation selection objectives, over seven paired seeds. Δ is premultiplied minus raw, with reselected 95% CIs below; additional rows give fixed-choice and seed-only 95% CIs. Composite and alpha MSE use 10^{-6} units; LPIPS uses 10^{-3} units. At $\kappa = 0$, external alpha-MSE bootstrap intervals use 500,000 draws.

κ	Statistic	Internal			External		
		Comp. MSE	**LPIPS**	**Alpha MSE**	**Comp. MSE**	**LPIPS**	**Alpha MSE**
0	Δ	−16.23	−0.627	34.15	−12.32	−0.513	35.35
	95% CI	[−24.16, −10.37]	[−0.912, −0.406]	[19.39, 52.4]	[−33.61, 7.224]	[−1.067, −0.094]	[−2.917, 90.72]
	Fixed-choice	[−24.16, −10.37]	[−0.912, −0.406]	[19.39, 52.4]	[−33.61, 7.224]	[−1.067, −0.094]	[−2.917, 90.72]
	Seed-only CI	[−21.82, −10.64]	[−0.823, −0.431]	[26.68, 41.62]	[−17.59, −7.042]	[−0.715, −0.311]	[27.95, 42.74]
0.1	Δ	−8.893	−0.211	12.06	−5.548	−0.203	16.94
	95% CI	[−17.46, −2.251]	[−0.651, 0.014]	[−12.67, 38.81]	[−25, 12.12]	[−0.775, 0.24]	[−19.71, 70.47]
	Fixed-choice	[−16.06, −1.849]	[−0.430, −0.002]	[−8.477, 35.34]	[−22.62, 12.11]	[−0.727, 0.220]	[−18.66, 67.69]
	Seed-only CI	[−16.72, −1.069]	[−0.444, 0.023]	[−11.99, 36.11]	[−12.93, 1.83]	[−0.353, −0.053]	[−3.788, 37.67]

Table 9: Ostris: video test results under two validation selection objectives, over seven paired seeds. Δ is premultiplied minus raw, with reselected 95% CIs below; additional rows give fixed-choice and seed-only 95% CIs. Composite and alpha MSE use 10^{-6} units; LPIPS uses 10^{-3} units. At $\kappa = 0$, external alpha-MSE bootstrap intervals use 500,000 draws.

κ	Statistic	Internal			External		
		Comp. MSE	**LPIPS**	**Alpha MSE**	**Comp. MSE**	**LPIPS**	**Alpha MSE**
0	Δ	−0.412	0.021	4.152	0.579	0.036	3.692
	95% CI	[−4.312, 3.044]	[−0.076, 0.102]	[1.062, 8.095]	[−2.083, 2.856]	[−0.025, 0.095]	[0.022, 8.007]
	Fixed-choice	[−4.353, 3.02]	[−0.076, 0.101]	[1.083, 8.122]	[−2.016, 2.856]	[−0.025, 0.092]	[−0.050, 8.047]
	Seed-only CI	[−5.182, 4.357]	[−0.090, 0.132]	[1.976, 6.327]	[−2.191, 3.348]	[−0.031, 0.103]	[1.965, 5.419]
0.1	Δ	1.78	0.057	3.646	1.252	0.054	3.904
	95% CI	[−1.329, 3.613]	[−0.038, 0.117]	[−3.25, 16.26]	[−2.792, 3.037]	[−0.022, 0.094]	[−1.698, 15.91]
	Fixed-choice	[0.108, 3.734]	[−0.009, 0.125]	[−3.949, 12.01]	[−0.882, 3.49]	[0.008, 0.102]	[−1.888, 12.09]
	Seed-only CI	[0.083, 3.476]	[−0.005, 0.120]	[−6.137, 13.43]	[−0.054, 2.559]	[0.008, 0.099]	[−3.24, 11.05]

Alpha MAE gives a different view of the errors (Table 10). Under alpha-MSE selection, Ostris's mean alpha MAE decreases by 27.12% on internal video and 29.27% on external video, while alpha MSE increases. The video MAE fixed-choice and reselected intervals include zero. On photographs, MAE decreases by 29.57%; fixed-choice and reselected difference intervals are negative, but the seed-only interval includes zero. Squared error gives greater weight to large errors than absolute error does, so the two means can move in opposite directions.

Table 10: Alpha MAE after alpha-MSE selection, over seven paired seeds, in 10^{-3} units. Bold marks the lower Raw/Premul. mean within each model and test. Δ is premultiplied minus raw, with reselected 95% CIs below; additional rows give fixed-choice and seed-only 95% CIs.

Model	Statistic	Video internal	Video external	PhotoMatte
SDXL	Raw	**1.048**	**0.830**	3.054
	Premul.	1.144	1.047	**1.781**
	Δ	0.096	0.217	−1.273
	95% CI	[−0.013, 0.199]	[0.012, 0.488]	[−2.695, −0.563]
	Fixed-choice	[−0.013, 0.199]	[0.012, 0.488]	[−2.695, −0.563]
	Seed-only CI	[−0.028, 0.219]	[−0.005, 0.439]	[−1.767, −0.779]
Ostris	Raw	1.181	1.041	2.953
	Premul.	**0.861**	**0.736**	**2.08**
	Δ	−0.320	−0.305	−0.873
	95% CI	[−0.829, 0.176]	[−0.725, 0.174]	[−1.803, −0.134]
	Fixed-choice	[−0.853, 0.080]	[−0.732, 0.065]	[−1.767, −0.126]
	Seed-only CI	[−0.949, 0.309]	[−0.817, 0.208]	[−1.919, 0.173]

B.3 PHOTOMATTE TRANSFER WITH THE SAME SELECTED MODELS

Tables 11 and 12 extend the opacity comparison in Section 4.4 to composite MSE, LPIPS and alpha MAE. They report models selected at $\kappa = 0$ and $\kappa = 0.1$, with all three interval methods. All comparisons use video-only validation selection, seven paired seeds and all 85 photographs. The intermediate validation weights reuse the alpha-only selections.

The interval methods agree on SDXL's main alpha-MSE improvement, while Ostris's alpha-MSE ordering remains unresolved despite a positive mean difference. Its reselected LPIPS difference interval also includes zero. Figure 5 complements these aggregate results with a predefined image and training seed: premultiplication lowers alpha MSE for both models, while the composites look similar at this scale. For Ostris, this example differs from the overall mean, motivating the image-level analysis below. All primary scores retain every image.

Table 11: SDXL: PhotoMatte results for video-selected models, using all 85 photographs and seven paired seeds. Values use 10^{-3} units. Raw/Premul. rows show means; bold marks the lower mean within each metric and validation objective. Δ is premultiplied minus raw, with reselected 95% CIs below; additional rows give fixed-choice and seed-only 95% CIs.

κ	Statistic	Comp. MSE ↓	LPIPS ↓	Alpha MSE ↓	Alpha MAE ↓
0.0	Raw	0.8887	18.77	0.8068	3.054
	Premul.	**0.625**	**13.47**	**0.2347**	**1.781**
	Δ	−0.2637	−5.302	−0.5721	−1.273
	95% CI	[−0.542, −0.119]	[−6.89, −4.03]	[−1.58, −0.122]	[−2.69, −0.563]
	Fixed-choice	[−0.542, −0.119]	[−6.89, −4.03]	[−1.58, −0.122]	[−2.69, −0.563]
	Seed-only CI	[−0.336, −0.191]	[−6.25, −4.36]	[−0.842, −0.302]	[−1.77, −0.779]
0.1	Raw	0.9396	18.48	1.211	3.619
	Premul.	**0.625**	**13.47**	**0.2347**	**1.781**
	Δ	−0.3146	−5.009	−0.9759	−1.838
	95% CI	[−0.743, −0.121]	[−7, −3.7]	[−2.97, −0.148]	[−4.06, −0.757]
	Fixed-choice	[−0.739, −0.123]	[−6.87, −3.67]	[−2.98, −0.153]	[−4.07, −0.752]
	Seed-only CI	[−0.415, −0.214]	[−5.63, −4.39]	[−1.56, −0.392]	[−2.7, −0.973]

Table 12: Ostris: PhotoMatte results for video-selected models, using all 85 photographs and seven paired seeds. Values use 10^{-3} units. Raw/Premul. rows show means; bold marks the lower mean within each metric and validation objective. Δ is premultiplied minus raw, with reselected 95% CIs below; additional rows give fixed-choice and seed-only 95% CIs.

κ	Statistic	Comp. MSE ↓	LPIPS ↓	Alpha MSE ↓	Alpha MAE ↓
0.0	Raw	0.2383	5.15	**0.2228**	2.953
	Premul.	**0.2077**	**4.472**	0.272	**2.08**
	Δ	–0.0306	–0.679	0.0492	–0.873
	95% CI	[–0.109, –0.0119]	[–1.81, 0.184]	[–0.29, 0.184]	[–1.8, –0.134]
	Fixed-choice	[–0.0546, –0.0111]	[–1.47, 0.236]	[–0.0385, 0.191]	[–1.77, –0.126]
	Seed-only CI	[–0.0486, –0.0126]	[–1.13, –0.228]	[–0.0234, 0.122]	[–1.92, 0.173]
0.1	Raw	0.4235	7.408	1.117	4.693
	Premul.	**0.2141**	**4.58**	**0.3652**	**2.39**
	Δ	–0.2094	–2.827	–0.7517	–2.303
	95% CI	[–0.453, –0.039]	[–4.23, –0.767]	[–1.93, 0.0502]	[–3.59, –0.429]
	Fixed-choice	[–0.516, –0.0604]	[–4.67, –1.47]	[–2.31, –0.0456]	[–4.28, –0.986]
	Seed-only CI	[–0.348, –0.0712]	[–4.5, –1.15]	[–1.46, –0.0412]	[–3.64, –0.969]

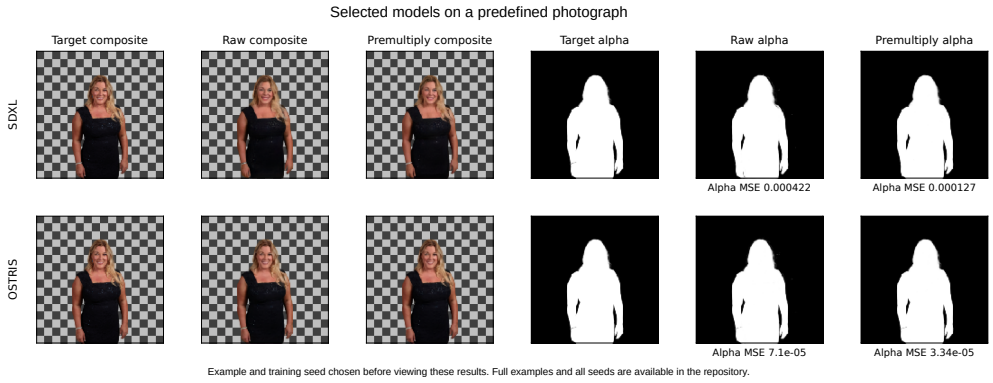

Figure 5: PhotoMatte reconstruction with both initializers. The photograph and training seed were chosen before inspecting these reconstructions. Checkerboard composites and alpha maps use common display scales.

B.4 Differences across sources and influential images

For each image, we average the premultiplied-minus-raw alpha-MSE difference across the seven paired seeds. For models selected by alpha MSE, removing any one source group leaves the sign of the overall mean difference unchanged in all six initializer/test combinations. On internal video, SDXL has 579 images with positive differences and 189 with negative differences; Ostris has 531 positive and 237 negative. On external video, the corresponding counts are 52 positive and 28 negative for SDXL, and 56 positive and 24 negative for Ostris.

On PhotoMatte, SDXL has 81 images with negative differences and four with positive differences. Ostris has 71 negative and fourteen positive, despite its positive overall mean. SDXL's median difference is -2.14×10^{-4} and its trimmed mean is -2.03×10^{-4}. For Ostris, these values are -4.77×10^{-5} and -4.34×10^{-5}. The trimmed means discard the lowest and highest 10% of image-level differences, rounded down to eight images at each end. These are descriptive sensitivity checks; primary estimates and intervals retain all 85 images.

One image examined after the main analysis contributes 66.82% of SDXL's net alpha-MSE decrease and 61.20% of Ostris's net increase. Removing it leaves mean differences of approximately -1.92×10^{-4} for SDXL and $+1.93 \times 10^{-5}$ for Ostris. These contributions are signed: improvements and deteriorations can cancel, so a single image's share of the net change can exceed 100%.

Reconstructions of this image for all seven seeds are available in the repository. Both raw and premultiplied inputs can produce a partly transparent reconstruction of an opaque shirt, revealing the checkerboard background. Premultiplication gives higher error in one SDXL run, while raw error is substantially lower in two Ostris runs. This example explains a large contribution to the mean; it does not estimate the frequency of this defect.

C Downstream Protocol and Sensitivity

C.1 Selected VAEs and generator training

Paired studies. Section 5 defines the Joint and Fixed studies and their fourteen frozen SDXL VAEs. Each VAE was trained for 20k updates at the selected learning rate 10^{-4}, $\lambda_\alpha = 3$ and $\lambda_L = 0.05$. Selection uses video validation data and precedes generator training. Joint includes variation from both training stages but does not estimate their contributions separately. Fixed uses one predefined VAE pair and the same seven generator seeds as Joint. The studies share one pair of runs, giving 26 distinct LoRA training runs, and are analyzed separately.

Optimization. Adaptation minimizes Equation (9) on all 6,718 local training images and their supplied captions at 512^2. The 372 local validation images are available for evaluation. The pretrained SDXL UNet weights and both text encoders remain frozen (Podell et al., 2024); rank-64 LoRA (Hu et al., 2022) uses scaling parameter 64, no adapter dropout and float32 trainable parameters. Four devices share a global batch of eight. Training uses 20k AdamW updates with learning rate 10^{-4}, a 500-step linear warmup followed by a constant rate, zero weight decay and gradient clipping at one. We retain checkpoints at 5k, 10k and 20k updates; all reported main results use 20k.

During training, fresh posterior samples are drawn from stored means and standard deviations. Both inputs use latent scale 0.13025, without separate fitted normalization. Diffusion timesteps are uniform over $\{0, \dots, 999\}$, with the pretrained SDXL forward-noising schedule and the unweighted noise-prediction objective. Conditioning combines text embeddings with fixed image-size and crop metadata. Paired runs share adapter initialization, training-image draws, posterior and diffusion noise, timesteps, and caption dropout with probability 0.1. The repository records precision and random-number details.

Sampling and examples. The main generation test uses the original Euler scheduler with 40 steps, guidance 5 and zero negative conditioning. It covers 358 internal and 902 external prompts, with four fixed initial-noise draws per prompt shared across models. A separate set of 32 validation prompts is evaluated at all three checkpoints without using those scores to select a checkpoint. All 26 training runs and planned generations finish without nonfinite values.

Figure 4 uses the first internal leopard and external camel prompts in a predefined ordering, one predefined paired run and all four noise draws. No outputs are filtered by quality. Columns 1–4 correspond to matched noise draws. Full galleries, training curves and diagnostics are available in the repository.

C.2 Metrics and confidence intervals

Generation metrics. Reference and generated RGBA images are quantized to eight bits, composited on black, gray and white, and quantized to RGB8 in the same way. KID (Bińkowski et al., 2018) averages the full polynomial U statistic across these backgrounds. It uses 2,048-dimensional Inception features and kernel $k(u, v) = (u^\top v/2048 + 1)^3$, where $u, v \in \mathbb{R}^{2048}$ are image feature vectors. Self-pairs are excluded from within-set terms, and each term is normalized by its number of pairs. This unbiased estimator can be negative. Reference sets contain all 358 internal or 902 external images; generated and reference images on the external test are unpaired.

CLIP (Radford et al., 2021) uses cosine similarity between normalized ViT-B/32 image and text features, computed in float32 without learned scaling or clamping. Only the fixed generation directive is removed from the supplied content text. We do not compute paired reconstruction metrics for generated outputs without paired targets.

Intervals and source dependence. Each study has seven paired training runs. Seed-only intervals use Student's t distribution with six degrees of freedom; crossed intervals use 10,000 bootstrap draws over paired training runs and source groups. Prompt groups are defined by the associated source-image groups. All four outputs for a prompt stay together when its group is resampled. The primary calculation resamples reference-image groups independently. The internal sensitivity check instead resamples linked prompt and reference groups together. These choices change the resampling distribution, not the observed mean differences. Backgrounds and noise draws are not independent training runs.

Table 13 supplements Table 2 with seed-only intervals for the same generation setting. All intervals are pointwise. Section 5 reports the linked-group check and the KID comparisons on individual backgrounds.

Table 13: Generation with 40 Euler steps and CFG 5. Paired differences are premultiplied minus raw, with crossed 95% CIs below; additional rows give seed-only 95% CIs conditional on the test collections. KID and CLIP use 10^{-3} units. Crossed intervals resample paired training runs and source groups.

Study	Statistic	Internal		External	
		KID	CLIP	KID	CLIP
Joint	Δ	–0.722	4.165	–0.911	3.77
	95% CI	[−1.376, −0.171]	[2.774, 5.642]	[−1.59, −0.262]	[2.596, 4.971]
	Seed-only CI	[−1.254, −0.191]	[2.859, 5.471]	[−1.691, −0.13]	[2.51, 5.029]
Fixed	Δ	–0.565	2.796	–0.644	2.903
	95% CI	[−1.141, −0.048]	[1.568, 3.989]	[−1.279, −0.053]	[1.492, 4.162]
	Seed-only CI	[−0.996, −0.134]	[1.974, 3.618]	[−1.378, 0.091]	[1.355, 4.451]

C.3 SAMPLER SENSITIVITY

We evaluate 32 fixed prompts per test at all combinations of 20 or 40 Euler steps and guidance 3, 5 or 7.5, retaining the same trained models and four noise draws per prompt. Outputs at 40 steps and guidance 5 are reused. Tables 14 and 15 report all six settings for both studies and tests.

Across these 24 KID comparisons, sixteen mean differences are positive and eight are negative; all crossed intervals include zero. All Joint CLIP difference intervals are positive with both methods. Fixed CLIP mean differences are positive throughout, but three crossed intervals and one seed-only interval include zero. With this smaller prompt subset, the KID ordering remains unresolved across the tested settings; differences in point estimates alone do not establish sampler dependence.

Table 14: Joint study: six sampling settings, each using 32 prompts per test and four noise draws per prompt. Cells show premultiplied-minus-raw differences with crossed 95% CIs below, in 10^{-3} units for both KID and CLIP. All KID difference intervals include zero.

Steps	CFG	Internal		External	
		KID	CLIP	KID	CLIP
20	3	–0.276 [−2.098, 2.389]	7.74 [2.638, 13.08]	0.095 [−2.915, 3.911]	6.489 [1.83, 11.52]
20	5	–0.217 [−1.858, 2.244]	7.378 [3.254, 11.64]	0.828 [−1.739, 4.435]	5.146 [0.789, 9.543]
20	7.5	0.305 [−1.42, 2.418]	6.264 [2.408, 10.17]	1.405 [−1.898, 5.581]	4.981 [1.241, 8.882]
40	3	–0.303 [−2.021, 2.195]	7.922 [3.704, 12.62]	0.197 [−2.184, 3.51]	6.377 [1.915, 11.08]
40	5	–0.149 [−1.547, 1.911]	6.933 [2.884, 11.31]	0.579 [−2.32, 4.578]	4.855 [1.419, 8.53]
40	7.5	0.208 [−1.88, 2.188]	5.69 [2.271, 9.046]	1.532 [−1.601, 6.351]	3.937 [0.432, 7.422]

Table 15: Fixed study: six sampling settings, each using 32 prompts per test and four noise draws per prompt. Cells show premultiplied-minus-raw differences with crossed 95% CIs below, in 10^{-3} units for both KID and CLIP. All KID difference intervals include zero.

Steps	CFG	Internal		External	
		KID	CLIP	KID	CLIP
20	3	−0.823 [−3.022, 1.345]	4.454 [0.013, 9.31]	0.898 [−1.346, 4.125]	6.334 [1.625, 10.83]
20	5	−0.525 [−2.612, 1.748]	4.175 [0.205, 8.19]	0.868 [−1.045, 3.947]	5.271 [1.322, 9.297]
20	7.5	0.181 [−1.936, 2.424]	2.755 [−1.15, 6.46]	0.618 [−2.208, 4.346]	5.661 [1.715, 9.644]
40	3	−0.597 [−2.772, 1.483]	4.476 [0.759, 8.564]	0.760 [−1.52, 3.701]	5.478 [1.284, 9.584]
40	5	−0.558 [−2.778, 1.457]	2.862 [−1.007, 6.655]	0.477 [−1.71, 4.211]	5.957 [2.197, 9.73]
40	7.5	0.507 [−1.674, 2.521]	2.487 [−1.467, 6.474]	1.364 [−1.867, 6.068]	6.216 [2.256, 10.43]

C.4 Reconstruction at the generation resolution

To evaluate reconstruction on the image domain and resolution used for generation, we apply the fourteen frozen VAEs to all local validation and test images at 512^2, without further training. Reconstructions use posterior means and are compared with their source images. Table 16 reports both full test sets, extending the 256^2 VideoMatte and PhotoMatte evaluations.

Composite MSE, averaged analytically over uniform background colors, decreases by 13.60% internally and 22.34% externally. LPIPS decreases by 9.02% and 11.00%, respectively, with negative difference intervals. Alpha MSE changes by $+1.70\%$ internally and -16.73% externally; crossed and seed-only intervals include zero for both alpha MSE and MAE.

Average composite MSE improves, but error on black increases. Black-background MSE rises by 77.09% internally and 73.21% externally, with higher error in every seed and positive crossed and seed-only difference intervals. Gray- and white-background MSE decrease. These reconstruction results concern images with known targets and do not measure the accuracy of generated alpha maps.

Table 16: Reconstruction at 512^2 with fourteen frozen video-selected VAEs. Raw/Premul. rows show means; bold marks the lower mean within each metric and test. Δ is premultiplied minus raw, with crossed 95% CIs below; additional rows give seed-only 95% CIs and relative changes. Values use 10^{-3} units, except relative changes in percent. Composite MSE averages over uniform background colors, as in Equation (7).

Statistic	Internal		External	
	Comp. MSE	Black MSE	Comp. MSE	Black MSE
Raw	2.208	**2.261**	1.916	**1.542**
Premul.	**1.908**	4.003	**1.488**	2.671
Δ	−0.300	1.743	−0.428	1.129
95% CI	[−0.576, −0.028]	[1.103, 2.528]	[−0.632, −0.245]	[0.869, 1.425]
Seed-only CI	[−0.429, −0.172]	[1.492, 1.994]	[−0.570, −0.286]	[0.974, 1.284]
Change (%)	−13.6	77.09	−22.34	73.21

Statistic	Internal		External	
	Gray MSE	White MSE	Gray MSE	White MSE
Raw	1.767	3.919	1.555	3.732
Premul.	**1.459**	**1.606**	**1.188**	**1.506**
Δ	−0.308	−2.314	−0.368	−2.226
95% CI	[−0.513, −0.117]	[−3.286, −1.514]	[−0.497, −0.256]	[−2.962, −1.585]
Seed-only CI	[−0.393, −0.223]	[−2.725, −1.902]	[−0.442, −0.293]	[−2.722, −1.73]
Change (%)	−17.42	−59.03	−23.64	−59.65

Statistic	Internal			External		
	Alpha MSE	Alpha MAE	LPIPS	Alpha MSE	Alpha MAE	LPIPS
Raw	**5.292**	**17.648**	64.907	4.326	14.156	59.91
Premul.	5.382	17.87	**59.052**	**3.602**	**13.548**	**53.319**
Δ	0.090	0.221	−5.854	−0.724	−0.608	−6.591
95% CI	[−1.05, 1.284]	[−1.805, 2.344]	[−8.416, −3.256]	[−1.77, 0.264]	[−2.159, 1.081]	[−8.755, −4.493]
Seed-only CI	[−0.595, 0.775]	[−1.506, 1.948]	[−8.372, −3.337]	[−1.69, 0.242]	[−2.458, 1.241]	[−9.053, −4.129]
Change (%)	1.698	1.254	−9.02	−16.73	−4.298	−11

D Canonicalization Controls

D.1 Input sensitivity and the four-input comparison

Released AlphaVAE. The intervention in Figure 1 replaces gray RGB with uniform noise only where alpha is exactly zero in each of the 902 external images. Deterministic reconstructions change for 900 images. Mean rendered drift, defined as composite MSE between the two reconstructions, is 8.169, with a 95% CI of $[7.878, 8.466]$, in 10^{-4} units. Repeated inference without the intervention gives identical outputs, and all target renderings remain unchanged. The released AlphaVAE was trained with a direct RGBA adversarial loss as well as reconstruction losses; this intervention therefore measures input sensitivity without isolating the effect of one training loss. Reconstruction error relative to the target can either increase or decrease.

Matched input comparison. Tables 17 and 18 report both local test sets and paired intervals for the initial four-input experiment. Each initializer uses three paired seeds per input and 10k updates. Raw, hard gray, visible padding and gray premultiplication share targets and training budgets, with no separate hyperparameter tuning. Among the three visibility-invariant transformations, reconstruction quality differs.

Table 17: SDXL: four-input comparison over three paired seeds. Mean rows show full-test scores; Δ rows show differences from raw, with crossed seed/source-group 95% CIs below. Composite MSE uses 10^{-6} units; LPIPS and alpha MAE use 10^{-3} units. Bold marks the lowest mean across the four inputs for each metric and test.

Input	Statistic	Internal			External		
		Comp. MSE	LPIPS	Alpha MAE	Comp. MSE	LPIPS	Alpha MAE
Raw	Mean	1069.8	86.682	14.793	1072.8	91.595	13.251
Hard	Mean	1371.5	93.236	19.678	1279.5	97.454	17.274
	Δ	302	6.55	4.89	207	5.86	4.02
	95% CI	[183, 464]	[3.4, 9.94]	[3.49, 6.61]	[160, 263]	[3.59, 8.37]	[3.28, 4.86]
Pad	Mean	1087.1	90.567	14.619	1084.8	95.742	13.274
	Δ	17.3	3.88	−0.173	12	4.15	0.023
	95% CI	[−12.2, 47.4]	[1.67, 5.77]	[−1.41, 0.951]	[−4.01, 30.3]	[2.36, 5.59]	[−0.887, 0.895]
Premul.	Mean	**856.97**	**74.232**	**14.15**	**866**	**80.303**	**12.155**
	Δ	−213	−12.5	−0.642	−207	−11.3	−1.1
	95% CI	[−261, −166]	[−15, −9.77]	[−2.04, 0.821]	[−237, −182]	[−13.4, −8.97]	[−2.09, 0.162]

Table 18: Ostris: four-input comparison over three paired seeds. Mean rows show full-test scores; Δ rows show differences from raw, with crossed seed/source-group 95% CIs below. Composite MSE uses 10^{-6} units; LPIPS and alpha MAE use 10^{-3} units. Bold marks the lowest mean across the four inputs for each metric and test.

Input	Statistic	Internal			External		
		Comp. MSE	LPIPS	Alpha MAE	Comp. MSE	LPIPS	Alpha MAE
Raw	Mean	365.26	30.519	8.410	338.61	29.681	7.559
Hard	Mean	337.97	28.054	8.013	312.68	27.305	7.226
	Δ	−27.3	−2.46	−0.397	−25.9	−2.38	−0.333
	95% CI	[−36.6, −18.1]	[−3.21, −1.68]	[−0.862, 0.086]	[−34, −19.4]	[−2.98, −1.63]	[−0.713, 0.121]
Pad	Mean	365.27	30.622	8.841	336.83	29.727	8.021
	Δ	0.0114	0.104	0.431	−1.79	0.046	0.462
	95% CI	[−7.25, 6.49]	[−0.513, 0.775]	[−0.831, 1.68]	[−8.21, 3.59]	[−0.47, 0.677]	[−0.834, 1.8]
Premul.	Mean	**280.09**	**23.1**	**7.527**	**268.29**	**23.904**	**6.978**
	Δ	−85.2	−7.42	−0.883	−70.3	−5.78	−0.581
	95% CI	[−109, −66.5]	[−9.07, −5.91]	[−1.47, −0.405]	[−82.7, −60.3]	[−6.57, −4.99]	[−1.22, −0.14]

D.2 COLOR CENTERS

Table 19 compares black ($c = 0$) and gray ($c = 1/2$ in each channel) for SDXL hard filling and premultiplication. It uses the same three paired seeds and training budget as the initial comparison; paired intervals are available in the repository. Premultiplication follows Equation (4). Black hard filling leaves unchanged the hidden RGB of most images in this collection.

Changing hard filling from gray to black lowers mean composite MSE, LPIPS and alpha MAE on both tests. For premultiplication, the same center change raises mean composite MSE and LPIPS while lowering alpha MAE. The centers remove exactly the same invisible-color variation but lead to different reconstruction scores.

Table 19: SDXL color-center comparison, with means over three seeds and the same budget as the four-input experiment. Values use 10^{-3} units. Bold marks the lower Gray/Black mean for each metric and test, separately within Hard fill and Premul.

Input	Center	Internal			External		
		Comp. MSE ↓	LPIPS ↓	Alpha MAE ↓	Comp. MSE ↓	LPIPS ↓	Alpha MAE ↓
Hard fill	Gray	1.372	93.24	19.68	1.28	97.45	17.27
	Black	**1.074**	**87.21**	**14.52**	**1.082**	**91.74**	**13.46**
Premul.	Gray	**0.857**	**74.23**	14.15	**0.866**	**80.3**	12.16
	Black	1.166	78.34	**12.77**	1.206	84.91	**11.62**

D.3 Pretrained compatibility control

This SDXL control asks whether the effect of changing the color center partly reflects compatibility with pretrained weights. Black- and gray-centered inputs differ by an affine transformation involving RGB and alpha. We compensate for this change by adjusting the first layer's alpha-channel weights and bias. Zero padding leaves a residual near image edges, which a separate fixed boundary correction removes when both adjustments are applied. We test all four combinations with three paired seeds and the initial experiment's budget. Together, the adjustments approximately match the gray-centered encoder's output at initialization; subsequent training can make the outputs diverge.

The affine adjustment lowers composite MSE and LPIPS with or without the boundary correction, but raises internal alpha MSE, with positive paired difference intervals in both cases. Intervals for the additional boundary correction and its interaction with the affine adjustment include zero. Pretrained compatibility can affect reconstruction, though these results do not isolate initial output matching as the cause. The repository provides the adjustment, checks and full results.