DBFP: PLUG-AND-PLAY DUAL-BAND FUSION FIXED-POINT SAMPLING FOR SUBJECT-DRIVEN GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Subject-driven generation must preserve a reference subject's identity while executing a text edit—two objectives that compete under a single sampling condition. Even strong backbones leave this tension unresolved: following the prompt more closely drifts identity, while clinging to the reference weakens instruction following. Evaluation compounds the difficulty by scoring the two axes separately, without a compact criterion for their joint balance. To this end, we present Dual-Band Fusion Fixed-Point (DBFP), a training-free plug-and-play sampler. Given only a reference image I_c and an edit prompt p_{ed}, it builds a structure prompt that preserves subject attributes while absorbing salient edit keywords, pairs it with the original instruction, fuses the two guided velocities by early-stage frequency-band mixing in Fourier space, and refines the mixed step with a short Picard loop—without touching weights or embeddings. We further propose the Subject-Text Harmonic Score (STHS) as a compact measure of the coupled identity–edit balance. Experiments on multiple subject-driven generation backbones show that DBFP improves this balance and, under the same experimental protocol, yields stronger overall subject-driven generation performance as a drop-in sampling plug-in.

1 INTRODUCTION

Subject-driven generation asks a model to render the same subject as a reference image I_c under an edit prompt p_{ed} (Figure 1). The task is intrinsically dual-objective: the sampler must keep instance identity while carrying out the requested change. In practice this duality is hard to balance. Strengthening the text condition improves the edit but drifts the subject; strengthening the image condition preserves the subject but weakens the edit—so some pipelines gradually lose identity cues, while others over-preserve the reference and under-follow the instruction, limiting controllability and diversity. Most remedies retrain the generator or add learned adapters. In parallel, evaluation still largely reports identity (CLIP-I / DINO) and text compliance (CLIP-T) as separate columns (Ruiz et al., 2023). That practice is informative yet incomplete for ranking: a near-copy of I_c can dominate identity scores while failing the instruction, and no standard scalar jointly measures both axes or penalizes the imbalance. This paper addresses both the sampling trade-off and the ranking gap in three beats: (i) build two *compatible* texts from (I_c, p_{ed}) alone; (ii) fuse their guided velocities by frequency while layout is still plastic, then refine the mixed step velocity; (iii) rank with a score that penalizes identity–edit imbalance more strongly than an arithmetic mean of the same axes.

Sampling. We ask whether the identity–edit conflict can be controlled at sampling time by composing the model's own predictions, without weight updates. In a flow-matching generator with frozen parameters θ, the latent $z(t)$ obeys

$$\frac{dz(t)}{dt} = v_\theta\big(z(t), t; c\big), \tag{1}$$

where c is the conditioning. A spatial Fourier transform $\mathcal{F}$ (Section 3.4) and linearity give

$$\frac{d}{dt}\,\mathcal{F}\big(z(t)\big) = \mathcal{F}\big(v_\theta(z(t), t; c)\big), \tag{2}$$

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.

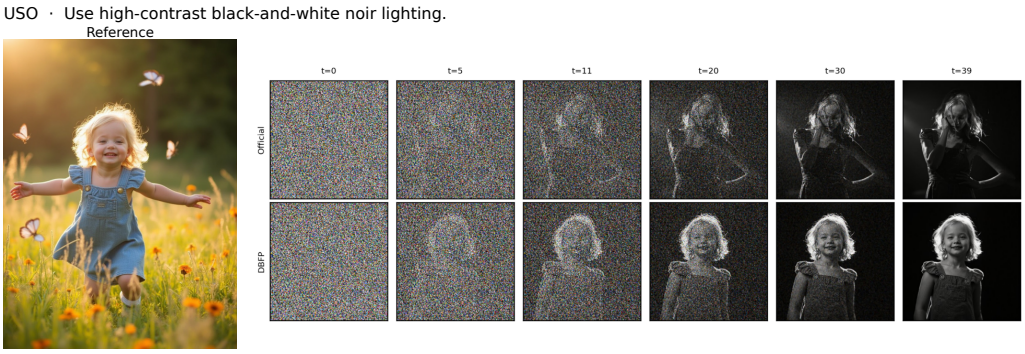

Figure 1: **Teaser (USO).** Left: reference image (height-aligned to the Official/DBFP traj strip). Right: Official (top) vs. Official+DBFP (bottom) mid-ODE decode thumbnails. Prompt above the panels. Mechanism: Figure 4; more tracks: Appendix H.

so band-wise control of the velocity is band-wise control of how the state evolves. The constructive implication is simple: if we can form two velocities that encode identity and edit respectively, we can assign their spectra to complementary bands rather than committing to a single scalar blend of conditions. We therefore build a structure branch from a target-aligned caption $p_{\mathrm{id}} = \mathcal{A}(I_c, p_{\mathrm{ed}})$ and an edit branch from p_{ed}, sharing the same reference I_c. Because noise-to-data trajectories form global layout early and refine texture later—an outline-first pattern for diffusion (Wang & Vastola, 2023) and a navigation-then-refinement split for flow matching (Liu et al., 2025a)—we apply $\mathrm{BandMix}$ only on that early window and revert to the edit-conditioned sampler afterward.

A one-shot mixture is queried at z_t yet consumed by the Euler step to $z_{t+\Delta t}$. We refine the *step velocity* itself: after each band fusion, form a probe $\tilde{z} = z_t + \Delta t\, v^\star$, re-evaluate both branches at $(\tilde{z}, t)$ with time held fixed, and update $v^\star$. The construction is related in spirit to ReNoise-style iterative consistency (Garibi et al., 2024), but iterates a forward velocity at time t rather than an inversion noise on a next latent. The outer schedule still advances exactly one Euler step per t_i.

Evaluation. Orthogonal to how we sample is how we *score*. Separate CLIP-I / DINO / CLIP-T columns cannot answer whether a method improved the trade-off or merely inflated identity by under-editing. We propose the Subject-Text Harmonic Score (STHS; Eqs. equation 18–equation 19): form a subject-fidelity score s_{id} by the geometric mean of CLIP-I and DINO, then take its harmonic mean with CLIP-T. The harmonic mean is strictly monotone in each positive argument and penalizes imbalance more strongly than an arithmetic mean (Section 3.8, Appendix E), in the spirit of ArtFID's coupling of two criteria (Wright & Ommer, 2022) but for similarity-type (higher-better) rather than distance-type inputs. We report STHS alongside the primitives so rankings remain interpretable.

Contributions.

- **Training-free dual-band fixed-point sampling (DBFP).** A plug-and-play chain—frozen prompt construction, Fourier BandMix of structure/edit velocities, and Picard refinement of the step velocity—that inserts into frozen subject-driven backbones from (I_c, p_{ed}) alone.
- **Subject-Text Harmonic Score (STHS).** A unified composite for ranking the identity–edit trade-off: geometric-mean identity s_{id} harmonic-meaned with CLIP-T.
- **Matched multi-backbone evidence.** Official vs. DBFP comparisons under matched seeds, steps, resolution, and guidance on SenseNova, USO, Qwen-Edit, BAGEL-7B-MoT, FLUX.2-klein-base-9B, HiDream-O1-Image, DreamOmni2, and OmniGen2, reporting primitives, s_{id}, STHS, and relative % deltas.

2 Related Work

We position DBFP against three nearby families: (i) learned subject adapters that change weights or tokens; (ii) training-free editors that intervene on attention or single-condition features; and (iii) scoreboards that list identity and text separately. Our contribution is a dual-field *composition* rule plus a joint ranking axis—not a new generator.

Diffusion and flow matching. Diffusion models learn a reverse process from noise to data (Ho et al., 2020); DDIM exposes deterministic trajectories useful for inversion and editing (Song et al., 2021). Flow matching learns a continuous velocity whose ODE transports noise to data (Lipman et al., 2023; Liu et al., 2023), and recent work scales rectified flow and diffusion transformers to high-resolution synthesis (Esser et al., 2024; Xie et al., 2025b; Xiao et al., 2025). Guidance for flow models is also being revisited to reduce off-manifold drift (Saini et al., 2025), while blockwise flow matching studies timestep-specialized velocity modules (Park et al., 2025). DBFP leaves these models frozen and changes only how their prediction fields are composed at sampling time.

Subject-driven generation. Reference-conditioned systems inject identity through learned tokens, adapters, or attention modules; IP-Adapter (Ye et al., 2023) and PhotoMaker (Li et al., 2024) are representative examples, with recent work extending personalization to frequency-aware reference features, serial standardization, full-body human customization, and multimodal personalization (Liu et al., 2025b; Xie et al., 2025a; Nam et al., 2025; Nguyen et al., 2025). Training-free editing methods instead manipulate attention or features at sampling time (Hertz et al., 2023; Cao et al., 2023; Tian et al., 2025). Spectral reweighting inside a frozen U-Net (FreeU; Si et al. (2024)) is related in using frequency structure, but acts on backbone feature maps of a single condition rather than fusing dual prediction fields. Scalar blends of conditions (or of velocities) likewise keep a single trade-off dial; BandMix instead assigns complementary bands. Our sampler is orthogonal to how the backend encodes I_c: it requires only that two prompts sharing the same reference can be evaluated at a shared state and time under a matched guidance recipe.

Iterative consistency. ReNoise refines a noise estimate during real-image inversion by repeatedly re-predicting at a trial next latent (Garibi et al., 2024). Fast inversion methods such as GNRI reduce the cost of real-image editing (Samuel et al., 2025). FlowEdit (Kulikov et al., 2025) instead performs inversion-free editing directly on flow-matching trajectories by modifying the ODE source distribution. We borrow only the idea of local consistency under iteration. In forward flow matching the iterated object is the *step velocity* at the current time t, queried at Euler probe states, not an inversion noise on a next latent.

Evaluation of subject-driven edits. DreamBooth-style protocols report CLIP-I / DINO against the reference and CLIP-T against the prompt (Ruiz et al., 2023). These primitives remain indispensable, yet treating them as an unordered scoreboard leaves methods under-specified whenever identity and text move in opposite directions. Composite aesthetic or style scores such as ArtFID (Wright & Ommer, 2022) couple two criteria multiplicatively, but for *distance*-type inputs (lower better). Subject-driven editing needs the complementary object: a higher-is-better coupling of *similarity*-type identity and text scores. STHS is designed for that setting (Section 3.8): it does not replace the primitives, but supplies the missing unified ranking axis.

3 Method

Overview. Subject-driven sampling receives a reference image I_c and an edit prompt p_{ed}. DBFP keeps the generator frozen and intervenes only in how conditions and prediction fields are composed. The pipeline has three stages (Figure 2): (i) a frozen constructor builds a target-aligned structural prompt $p_{\mathrm{id}} = \mathcal{A}(I_c, p_{\mathrm{ed}})$ once per sample; (ii) at selected times, two guided velocities under $c_{\mathrm{id}} = (I_c, p_{\mathrm{id}})$ and $c_{\mathrm{ed}} = (I_c, p_{\mathrm{ed}})$—each produced by the backbone's official guided predictor CFG_B—are fused by a complementary Fourier operator $\mathrm{BandMix}$; (iii) the fused *step velocity* is refined by short Picard iteration at fixed time t, then used for a single Euler update. Stages (i)–(iii) form one training-free chain: $\mathcal{A}$ supplies compatible dual texts; BandMix fuses the resulting fields in the Fourier domain; Picard makes the mixed velocity locally consistent with the state it induces.

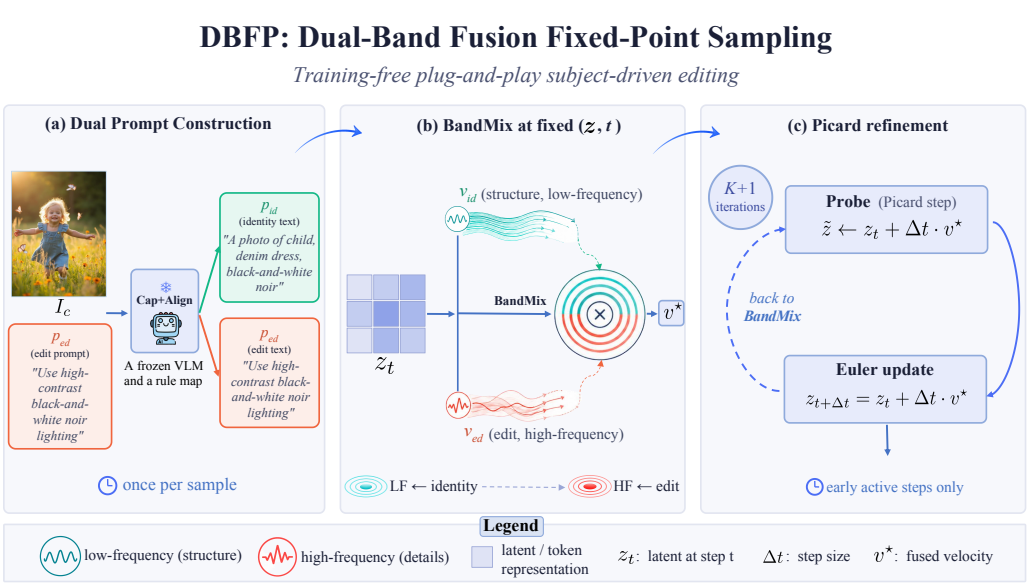

Figure 2: **DBFP pipeline.** (a) Frozen dual-prompt construction $\mathcal{A}$ yields p_{id} alongside p_{ed}. (b) BandMix fuses identity (LF) and edit (HF) guided velocities at fixed (z, t). (c) Picard refinement of the step velocity on early active steps, then one Euler update.

3.1 Setup

Let I_c be the reference image and p_{ed} the edit prompt. Parameters θ are frozen. The latent state is $z_t \in \mathbb{R}^{B\times L\times D}$ with $L = HW$ tokens arranged on an $H \times W$ grid (Rombach et al., 2022); we treat this grid as a spatial field for frequency analysis (an inductive bias, not an architectural assumption). A flow-matching backend (Lipman et al., 2023) predicts $v_\theta(z_t, t; c)$. Sampling uses a discrete schedule $0 = t_0 < \cdots < t_N = 1$ with step sizes $\Delta t_i = t_{i+1} - t_i > 0$. At a generic step we write $(t, \Delta t)$ for $(t_i, \Delta t_i)$ and use the Euler update

$$z_{t+\Delta t} = z_t + \Delta t\, v_\theta(z_t, t; c). \tag{3}$$

3.2 Frozen prompt constructor

The user-facing interface is (I_c, p_{ed}) only. The structure-branch text is produced once by a *frozen deterministic constructor*

$$p_{\mathrm{id}} = \mathcal{A}(I_c, p_{\mathrm{ed}}) = \mathrm{Align}\big(\mathrm{Cap}(I_c),\, p_{\mathrm{ed}}\big), \tag{4}$$

where Cap and Align have no trainable parameters. Cap is a one-shot structural caption of I_c from a frozen VLM (Qwen2.5-VL-7B-Instruct): category and coarse appearance attributes, not fine texture and not a new scene. Captions are cacheable offline; the sampler never calls the VLM during the trajectory. Align is a training-free rule map (Appendix F): it drops caption attributes that conflict with p_{ed} (e.g., old pose words under a pose edit; old colors under a recolor) and appends short target-state phrases extracted from p_{ed} (pose, color, style, or lighting), yielding a compact string of the form "`A photo of ...`". Thus p_{id} describes the *same* subject in the *requested* edit state. This is deliberate upstream design for BandMix: the structure branch should reinforce identity-consistent geometry *without* fighting the edit branch in text space.

3.3 Dual conditioned fields

A single text condition cannot resolve the identity–edit conflict: increasing its weight either under-edits or over-edits. We therefore query the *same* frozen network twice, under two conditions that share the reference image and differ only in text,

$$c_{\mathrm{id}} = (I_c, p_{\mathrm{id}}), \qquad c_{\mathrm{ed}} = (I_c, p_{\mathrm{ed}}), \tag{5}$$

and reuse them for the whole trajectory (across all schedule times and Picard probes, only the state and the time embedding change). At a shared state z and time t,

$$v^{\mathrm{id}}(z,t) = \mathrm{CFG}_B(z,t;c_{\mathrm{id}}), \tag{6}$$

$$v^{\mathrm{ed}}(z,t) = \mathrm{CFG}_B(z,t;c_{\mathrm{ed}}), \tag{7}$$

where B denotes the subject-driven backbone under study (e.g., SenseNova, USO, Qwen-Edit, BAGEL), and $\mathrm{CFG}_B(z,t;c)$ is that backbone's *official* guided one-step prediction at (z,t) under condition c (Ho & Salimans, 2022)—possibly involving several internal forwards and backend-specific renorm, such as single-scale guidance or image/text multi-scale variants in the style of InstructPix2Pix (Brooks et al., 2023). We do not invent a universal expansion of CFG_B; the method only consumes its output fields.

Matched-CFG rule. Both branches use the identical CFG_B recipe and hyperparameters; they differ solely in the text that enters c_{id} versus c_{ed}. Under this rule the residual $v^{\mathrm{id}} - v^{\mathrm{ed}}$ isolates the textual roles of p_{id} (target-aligned structure) versus p_{ed} (full edit instruction), rather than mismatched guidance bookkeeping. We label the first branch v^{id} because its role under BandMix is to supply identity-consistent structure via its low-frequency content; the fact that p_{id} absorbs edit-state words is an upstream design choice that prevents the two branches from conflicting in text space.

3.4 BANDMIX

Given two guided fields whose residual already encodes the identity–edit conflict, we fuse them by *frequency assignment* rather than by a scalar blend of conditions. Reshape each field from $L = HW$ tokens to an $H \times W$ layout and apply an orthonormal 2-D DFT $\mathcal{F}$ independently per batch element and latent channel (Appendix D.1). With unshifted frequencies $f_y(m), f_x(n) \in [-1/2, 1/2)$, radial frequency

$$r(m,n) = \sqrt{f_y(m)^2 + f_x(n)^2}, \tag{8}$$

and a soft radial gate

$$M_\rho(m,n) = \mathrm{sigmoid}\big(\kappa\,[\rho - r(m,n)]\big), \qquad \overline{M}_\rho(m,n) = 1 - M_\rho(m,n), \tag{9}$$

where $\mathrm{sigmoid}(u) = 1/(1+e^{-u})$, cutoff $\rho > 0$, and sharpness $\kappa > 0$ (we fix κ=8), define

$$\mathrm{BandMix}(v^{\mathrm{id}}, v^{\mathrm{ed}}; \rho, \kappa) := \mathcal{F}^{-1}\Big(M_\rho \odot \mathcal{F}(v^{\mathrm{id}}) + \overline{M}_\rho \odot \mathcal{F}(v^{\mathrm{ed}})\Big). \tag{10}$$

Because $M_\rho + \overline{M}_\rho = 1$, every frequency receives a soft convex assignment and there are no spectral holes. The operating step velocity is simply

$$v^\star = \mathrm{BandMix}(v^{\mathrm{id}}, v^{\mathrm{ed}}; \rho, \kappa); \tag{11}$$

ρ is the primary mixing hyperparameter.

Why mix velocities. By Eq. equation 2, band control of velocity is band control of state evolution. For one Euler step with Eq. equation 11, linearity of $\mathcal{F}$ gives

$$\mathcal{F}(z_t + \Delta t\, v^\star) = \mathcal{F}(z_t) + \Delta t\big(M_\rho \odot \mathcal{F}(v^{\mathrm{id}}) + \overline{M}_\rho \odot \mathcal{F}(v^{\mathrm{ed}})\big), \tag{12}$$

i.e., the Fourier update of the state is exactly the BandMix spectrum of the two branches (Appendix D.2). Filtering z *after* an ordinary edit-only step does not share this guarantee: a single condition never jointly determines the increment under both texts.

3.5 PICARD REFINEMENT OF THE STEP VELOCITY

A one-shot mixture $\mathrm{BandMix}(v^{\mathrm{id}}(z_t,t), v^{\mathrm{ed}}(z_t,t); \rho, \kappa)$ is queried at z_t but consumed by the Euler map to $z_{t+\Delta t}$. We therefore refine the *step velocity* at fixed schedule time t. Write

$$\Phi(z,t) := \mathrm{BandMix}\big(v^{\mathrm{id}}(z,t), v^{\mathrm{ed}}(z,t); \rho, \kappa\big) \tag{13}$$

for the fused velocity map at a given state and time. Ideal local consistency asks for a velocity that reproduces itself when re-evaluated at its own Euler image,

$$v^\dagger = \Phi\big(z_t + \Delta t\, v^\dagger,\, t\big). \tag{14}$$

We approximate $v^\dagger$ by $K{+}1$ Picard evaluations with probe $\tilde{z}^{(0)} = z_t$:

$$v_k^\star = \Phi\big(\tilde{z}^{(k)}, t\big), \qquad \tilde{z}^{(k+1)} = z_t + \Delta t\, v_k^\star, \qquad k = 0, \dots, K, \tag{15}$$

then take the single scheduled step

$$z_{t+\Delta t} = z_t + \Delta t\, v_K^\star \quad \big(= \tilde{z}^{(K+1)}\big). \tag{16}$$

Here K counts *extra* refinements after the initial evaluation: $K{=}0$ is one-shot BandMix at z_t (1 call to Φ); $K{=}1$ uses 2 calls. Two design choices keep the outer schedule intact. (i) *Time stays at t.* Every call to Φ uses the current schedule time, so we refine $v(\,\cdot\,, t)$ rather than advance the time embedding. (ii) *Probes are anchored at z_t.* The rule $\tilde{z}^{(k+1)} = \tilde{z}^{(k)} + \Delta t\, v_k^\star$ would silently take $K{+}1$ global Euler steps inside one schedule index; anchoring at z_t prevents that. The construction is related in spirit to ReNoise-style iterative consistency (Garibi et al., 2024), but the iterated object is a forward step velocity at time t, not an inversion noise on a next latent. We do not claim contraction of Φ. Empirically, successive probe updates shrink sharply after the first refinement (Section 4.4), so main tables use $K{=}1$ (two Φ calls) as a compute–consistency tradeoff; $K{=}0$ vs. $K{=}1$ decode scores appear in Table 2.

3.6 Active schedule

Because layout is largely committed early (Wang & Vastola, 2023; Liu et al., 2025a), BandMix/Picard run only on

$$\mathcal{S} = \{i \in \{0, \dots, N-1\} : i < \lfloor \gamma N \rfloor\}, \tag{17}$$

with early fraction γ; later steps use ordinary edit-conditioned Euler under c_{ed}.

3.7 Scope

We implement the plug-in for backends that expose a CFG-aggregated velocity field at a shared (z, t) under matched guidance—covering both controllable-CFG pipelines (separate cond/uncond forwards) and embedded-CFG pipelines (joint forwards with image and text conditions) (DFT details in Appendix D.1). One trajectory is summarized by Algorithm 1 (Appendix A): construct p_{id} once, run BandMix inside a short Picard probe on active steps, take one Euler step from z_{t_i} with the final $v^\star$, and otherwise follow $\mathrm{CFG}_B(\cdot; c_{\mathrm{ed}})$.

3.8 Subject-Text Harmonic Score

Subject-driven editing still lacks a standard scalar for ranking the identity–edit trade-off (Ruiz et al., 2023): separately listed CLIP-I / DINO / CLIP-T cannot tell whether a method improved both goals or merely inflated identity by under-editing. We define the Subject-Text Harmonic Score (STHS) to fill that gap.

Primitives. CLIP-I is the cosine between CLIP image embeddings (Radford et al., 2021) of the output and the reference (OpenAI CLIP ViT-B/32); DINO is the corresponding cosine between DINOv2 (Oquab et al., 2024) ViT-S/14 features (`dinov2_vits14`, embed dimension 384—not ViT-B/14 / dinov2-base); CLIP-T is the cosine between the CLIP embedding of the output and that of the full edit prompt. All three are L2-normalized cosines in $[-1, 1]$ (no ReLU truncation); empirically they fall in $[0, 1]$.

Composite. For each test case we first form a subject-fidelity score by the geometric mean of the two identity cosines,

$$s_{\mathrm{id}} = \sqrt{\text{CLIP-I} \cdot \text{DINO}}, \tag{18}$$

so that identity is credited only when *both* CLIP-I and DINO agree. We then take the harmonic mean with CLIP-T,

$$\mathrm{STHS} = \frac{2\, s_{\mathrm{id}}\, \text{CLIP-T}}{s_{\mathrm{id}} + \text{CLIP-T}}, \tag{19}$$

with a shared defensive floor $\max(x, 0)$ on cosines. Dataset-level STHS is the mean of per-case scores (not the HM of means). Monotonicity / imbalance properties and the ArtFID analogy are deferred to Appendix E; tables always also list CLIP-I / DINO / CLIP-T so STHS remains interpretable.

Table 1: Cross-backbone summary on the shared n=800 protocol (Eqs. equation 18–equation 19; Appendix E). Higher is better for CLIP-I / DINO / CLIP-T / s_{id} / STHS. Δ is the relative change $100\times(DBFP-\mathrm{Official})/\mathrm{Official}$ (%).

Backbone	Sampler	CLIP-I ↑	DINO ↑	CLIP-T ↑	s_{id} ↑	STHS ↑	ΔI (%)	ΔD (%)	ΔT (%)	Δs_{id} (%)	ΔS (%)
SenseNova	Official	0.841	0.763	0.239	0.799	0.365	—	—	—	—	—
SenseNova	DBFP	0.870	0.811	0.241	0.837	0.372	+3.4	+6.3	+0.8	+4.8	+1.9
USO	Official	0.884	0.808	0.254	0.840	0.386	—	—	—	—	—
USO	DBFP	0.897	0.839	0.253	0.863	0.388	+1.5	+3.8	−0.4	+2.7	+0.5
Qwen-Edit	Official	0.853	0.764	0.265	0.803	0.394	—	—	—	—	—
Qwen-Edit	DBFP	0.875	0.793	0.267	0.828	0.399	+2.6	+3.8	+0.8	+3.1	+1.3
BAGEL	Official	0.782	0.665	0.265	0.719	0.384	—	—	—	—	—
BAGEL	DBFP	0.796	0.697	0.269	0.741	0.391	+1.8	+4.8	+1.5	+3.1	+1.8
FLUX.2	Official	0.827	0.735	0.270	0.776	0.395	—	—	—	—	—
FLUX.2	DBFP	0.842	0.761	0.270	0.797	0.398	+1.8	+3.5	+0.0	+2.7	+0.8
HiDream-O1	Official	0.818	0.776	0.262	0.793	0.389	—	—	—	—	—
HiDream-O1	DBFP	0.838	0.812	0.265	0.821	0.395	+2.4	+4.6	+1.1	+3.5	+1.5
DreamOmni2	Official	0.905	0.851	0.232	0.874	0.363	—	—	—	—	—
DreamOmni2	DBFP	0.917	0.876	0.235	0.893	0.368	+1.3	+2.9	+1.3	+2.2	+1.4
OmniGen2	Official	0.876	0.804	0.249	0.837	0.379	—	—	—	—	—
OmniGen2	DBFP	0.880	0.811	0.254	0.843	0.386	+0.5	+0.9	+2.0	+0.7	+1.8

4 EXPERIMENTS

4.1 SETUP, METRICS, AND CONTROLS

We evaluate eight subject-driven backbones—SenseNova-U1 (Diao et al., 2026), USO (Wu et al., 2026b), Qwen-Image-Edit-2511 (Wu et al., 2025a; Qwen Team, 2025) (Qwen-Edit), BAGEL-7B-MoT (Deng et al., 2025), FLUX.2-klein-base-9B (Black Forest Labs, 2025; 2026), HiDream-O1-Image (Cai et al., 2026), DreamOmni2 (Xia et al., 2026), and OmniGen2 (Wu et al., 2026a)—under one shared protocol: the same 40×20 (n=800; 40 subjects $\times$ 20 prompts) UNO-1M-filtered split (Wu et al., 2025b), recipe $(\rho,\gamma,K)=(0.08,0.30,1)$, and $p_{\mathrm{id}}=\mathcal{A}(I_c,p_{\mathrm{ed}})$ (Appendix F). For each backbone the baseline is its official subject-driven sampler; the only controlled change is Official sampling versus the same frozen weights with DBFP (identical content image, edit prompt, seed, and backbone knobs). We report CLIP-I / DINO / CLIP-T, s_{id}, and STHS (Section 3.8) as means and Official→DBFP relative Δ (%).

4.2 QUANTITATIVE COMPARISON

Table 1 reports Official vs. DBFP under the matched n=800 protocol. Across all eight backbones, DBFP raises s_{id} and STHS, while CLIP-T stays flat to slightly higher on seven backends and drops only 0.4% on USO—identity gains without a systematic text-following penalty, the imbalance STHS is designed to penalize. A single shared (ρ,γ,K) transfers from weaker baselines (largest Δs_{id} on SenseNova) to strong ones such as DreamOmni2 / OmniGen2 without per-backbone retuning. Relative Δs_{id} spans $+0.7\%$–$+4.8\%$, and ΔSTHS remains positive throughout.

4.3 QUALITATIVE COMPARISON

Figure 3 makes the Table 1 gains concrete under the same matched protocol. Official+DBFP typically restores subject cues that Official alone dilutes—logos and facial geometry on SenseNova/USO, fine jewelry on FLUX.2, small marks on OmniGen2—while still executing the requested style or lighting. Shared content image, instruction, and seed isolate the sampler, so residual differences are attributable to BandMix and Picard rather than to prompt or initialization effects. The gallery therefore shows *what* the quantitative deltas recover; extended panels are in Appendix I.

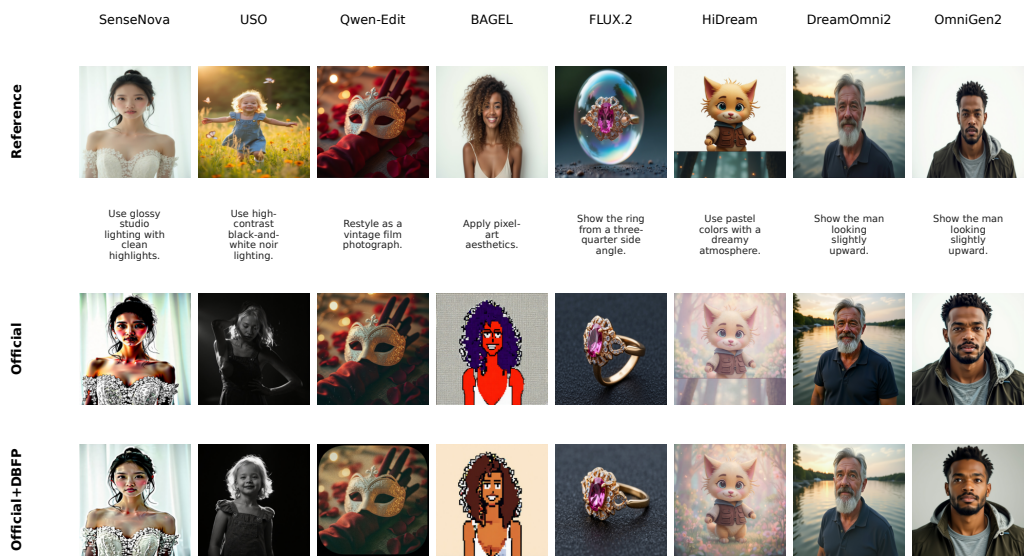

Figure 3: **Cross-backbone qualitative gallery** (shared protocol). Three rows (Reference / Official / Official+DBFP) and eight backbone columns; the edit prompt for each column sits in the middle blank band. Additional six-row galleries per backbone are in Appendix I.

4.4 Mechanism Analysis

A gallery cannot separate late decoding luck from intervention along the ODE. Figure 4 closes that gap on USO (FLUX-based), Qwen-Edit, SenseNova, and HiDream (the remaining Table 1 backbones build on FLUX and/or Qwen foundations). Panel (a): mid-ODE decode DINO vs. the reference (Official dashed/open; DBFP solid). USO/Qwen-Edit separate earlier inside/near the BandMix window; SenseNova and HiDream often diverge later yet still pull apart by late probes (Appendix H). Panel (b): Picard residuals $\|\Delta^{(k)}\|/\|\Delta^{(1)}\|$ drop by $k{=}1$ and flatten for $k{>}1$—empirical diminishing returns that motivate $K{=}1$ (Table 2). Panels (c)–(f): early LF id–edit gap (mean$\pm$s.e.m.; shaded window) is nontrivial on every lineage, so BandMix mixes a real residual timed to $\mathcal{S}$. Panels (g)–(j): later $L_{k,t}$ iterates add little beyond $k{=}1$. Taken together, early identity separation and a collapsing Picard residual support the shared $K{=}1$ recipe: Fig. 3 shows *what*, while this figure localizes *when/how*.

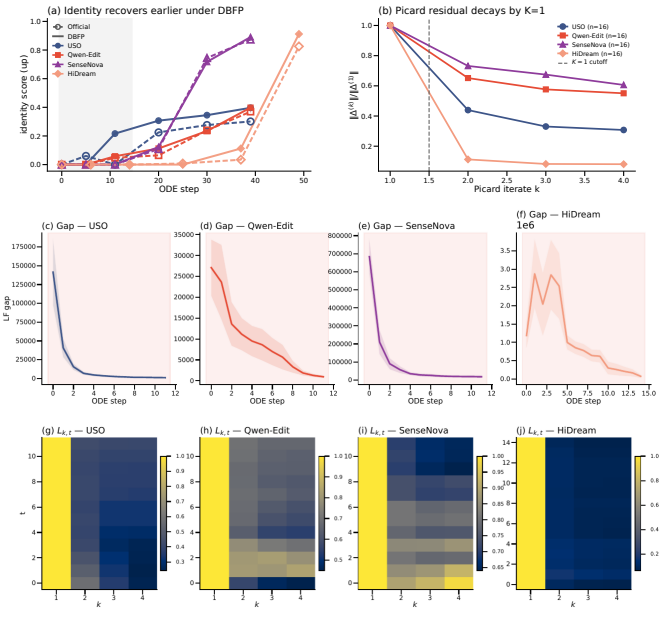

Figure 4: **Mechanism diagnostics** on four lineage representatives (USO [FLUX-based], Qwen-Edit, SenseNova, HiDream; the remaining Table 1 backbones build on FLUX and/or Qwen foundations). Row 1: (a) ODE identity; (b) Picard residual vs. k. Row 2: (c)–(f) LF id–edit gap. Row 3: (g)–(j) residual maps, $k = 1, \ldots, 4$.

4.5 Ablation study

We ablate Picard depth and (ρ, γ) on a shared n=16 ID subset (same controls as Table 1; main transfer n=800), on the same four *lineage representatives* as Figure 4 (USO [FLUX-based], Qwen-Edit, SenseNova, HiDream; the remaining Table 1 backbones build on FLUX and/or Qwen foundations). **Picard depth.** Official vs. one-shot BandMix (K=0) vs. full DBFP (K=1): K=0 mixes once per active step; K=1 adds one Picard probe. Table 2: K=1 matches or improves STHS over K=0 on all four lineages (largest Δ on SenseNova / HiDream; USO nearly tied), so we keep K=1. (ρ, γ) **sensitivity.** Figure 5: mean±s.e.m. on the same n=16 cases, sweeping ρ (at γ=0.30) and γ (at ρ=0.08) with K=1. Absolute STHS and ΔSTHS place $(0.08, 0.30)$ near-best or on a broad plateau, motivating the shared recipe in Table 1. Too small a ρ under-uses identity LF mass; too large a ρ or γ can trade editability for identity without a matching STHS gain.

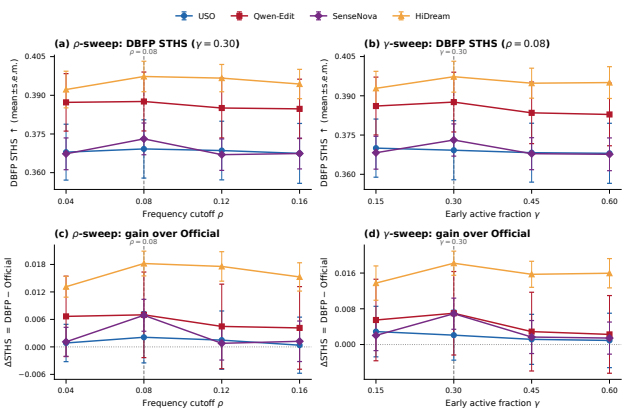

Figure 5: ρ/γ sensitivity on four lineage representatives (USO [FLUX-based], Qwen-Edit, SenseNova, HiDream; mean±s.e.m., n=16 matched cases each, K=1). (a–b) DBFP STHS; (c–d) paired ΔSTHS against fixed Official. Dashed: default $(\rho, \gamma) = (0.08, 0.30)$.

4.6 User study

Automatic scores do not settle perceptual preference, so we ran a forced 2AFC study with N=87 volunteers (45 computing, 42 non-computing). Each participant evaluated the same 16 Official vs. DBFP pairs per backbone (128 comparisons/person), with randomized left/right placement. Figure 6 shows that DBFP is preferred on every backbone.

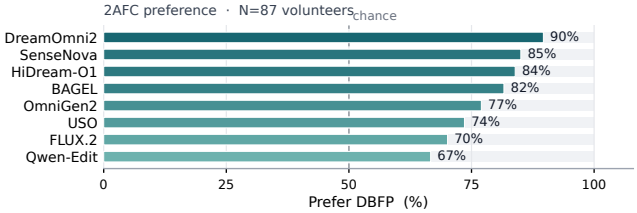

Figure 6: **User study** (N=87; 16 pairs/backbone/person). Bars show preference for DBFP over Official with randomized sides; the dashed line marks chance (50%).

5 Conclusion and Discussion

We presented DBFP, a training-free sampler that constructs dual text conditions from (I_c, p_{ed}), fuses their guided velocities through early Fourier BandMix, and applies a short Picard refinement, all without modifying model weights or embeddings. Across eight frozen backbones, the same recipe consistently shifts generation toward a better identity–edit balance: the automatic metrics, qualitative comparisons, mechanism probes, and human preferences all point in the same direction.

The gain is not free, but the overhead is limited: during active $\mathcal{S}$ steps, DBFP evaluates one additional Φ call at K=1, and the residual plateau makes larger Picard budgets unnecessary in our setting. The main assumptions are that the two velocity fields admit spatially aligned latent representations, and that Align remains an English, rule-based prompt procedure; relaxing these constraints is left for future work.

AI USE STATEMENT

Generative AI assisted minor language polishing, figure/table layout polishing, and auxiliary code. Authors manually verified all AI-assisted artifacts and take full responsibility for the research ideas, experimental design, data analysis, and final content.

ETHICS STATEMENT

Evaluation uses public research assets; benchmark references are synthetic UNO-1M pairs (Wu et al., 2025b). The preference study recruited consenting adult volunteers who compared model outputs only (no collection of biometric identity data). DBFP is training-free and releases no private identity data; we discourage deceptive misuse of subject-driven editing.

REPRODUCIBILITY STATEMENT

Section 3 / Algorithm 1 specify the sampler; Section 4 the fairness protocol and shared $(\rho, \gamma, K) = (0.08, 0.30, 1)$; Appendices A–I give metrics and implementation notes. We will release code, configs, and evaluation scripts upon acceptance.

REFERENCES

Black Forest Labs. FLUX.2: Frontier visual intelligence, 2025. URL https://bfl.ai/blog/flux-2. Technical blog post; no formal arXiv or conference paper available.

Black Forest Labs. FLUX.2 [klein] 9B Base. Hugging Face model card, 2026. URL https://huggingface.co/black-forest-labs/FLUX.2-klein-base-9B.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. InstructPix2Pix: Learning to follow image editing instructions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Qi Cai, Jingwen Chen, Chengmin Gao, Zijian Gong, Yehao Li, Yingwei Pan, Yi Peng, Zhaofan Qiu, Kai Yu, Yiheng Zhang, Hao Ai, Siying Bai, Yang Chen, Zhihui Chen, Fengbin Gao, Ying Guo, Dong Li, Zhen Shen, Leilei Shi, Jing Wang, Siyu Wang, Yimeng Wang, Rui Zheng, Ting Yao, and Tao Mei. HiDream-O1-Image: A natively unified image generative foundation model with pixel-level unified transformer. *arXiv preprint arXiv:2605.11061*, 2026.

Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. MasaCtrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *IEEE/CVF International Conference on Computer Vision*, 2023.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Haiwen Diao, Penghao Wu, Hanming Deng, Jiahao Wang, Shihao Bai, Silei Wu, Weichen Fan, Wenjie Ye, Wenwen Tong, Xiangyu Fan, Yan Li, Yubo Wang, Zhijie Cao, Zhiqian Lin, Zhitao Yang, Zhongang Cai, Yuwei Niu, Yue Zhu, Bo Liu, Chengguang Lv, Haojia Yu, Haozhe Xie, Hongli Wang, Jianan Fan, Jiaqi Li, Jiefan Lu, Jingcheng Ni, Junxiang Xu, Kaihuan Liang, Lianqiang Shi, Linjun Dai, Linyan Wang, Oscar Qian, Peng Gao, Pengfei Liu, Qingping Sun, Rui Shen, Ruisi Wang, Shengnan Ma, Shuang Yang, Siyi Xie, Siying Li, Tianbo Zhong, Xiangli Kong, Xuanke Shi, Yang Gao, Yongqiang Yao, Yves Wang, Zhengqi Bai, Zhengyu Lin, Zixin Yin, Wenxiu Sun, Ruihao Gong, Quan Wang, Lewei Lu, Lei Yang, Ziwei Liu, and Dahua Lin. SenseNova-U1: Unifying multimodal understanding and generation with NEO-unify architecture. *arXiv preprint arXiv:2605.12500*, 2026. Technical report.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning*, 2024.

Daniel Garibi, Or Patashnik, Andrey Voynov, Hadar Averbuch-Elor, and Daniel Cohen-Or. ReNoise: Real image inversion through iterative noising. In *European Conference on Computer Vision*, 2024.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *International Conference on Learning Representations*, 2023.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. FlowEdit: Inversion-free text-based editing using pre-trained flow models. In *IEEE/CVF International Conference on Computer Vision*, pp. 19721–19730, 2025.

Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. PhotoMaker: Customizing realistic human photos via stacked ID embedding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

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

Haoming Liu, Jinnuo Liu, Yanhao Li, Liuyang Bai, Yunkai Ji, Yuanhe Guo, Shenji Wan, and Hongyi Wen. From navigation to refinement: Revealing the two-stage nature of flow-based diffusion models through oracle velocity. *arXiv preprint arXiv:2512.02826*, 2025a.

Mushui Liu, Dong She, Jingxuan Pang, Qihan Huang, Jiacheng Ying, Wanggui He, Yuanlei Hou, and Siming Fu. TFCustom: Customized image generation with time-aware frequency feature guidance. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2714–2723, 2025b.

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

Jisu Nam, Soowon Son, Zhan Xu, Jing Shi, Difan Liu, Feng Liu, Seungryong Kim, and Yang Zhou. Visual persona: Foundation model for full-body human customization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18630–18641, 2025.

Thao Nguyen, Krishna Kumar Singh, Jing Shi, Trung Bui, Yong Jae Lee, and Yuheng Li. Yo'Chameleon: Personalized vision and language generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14438–14448, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Dogyun Park, Taehoon Lee, Minseok Joo, and Hyunwoo J. Kim. Blockwise flow matching: Improving flow matching models for efficient high-quality generation. In *Advances in Neural Information Processing Systems*, 2025.

Qwen Team. Qwen-Image-Edit-2511. Official model release, 2025. URL `https://github.com/QwenLM/Qwen-Image`. Released December 23, 2025.

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 *International Conference on Machine Learning*, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Shreshth Saini, Shashank Gupta, and Alan C. Bovik. Rectified-CFG++ for flow based models. In *Advances in Neural Information Processing Systems*, 2025.

Dvir Samuel, Barak Meiri, Haggai Maron, Yoad Tewel, Nir Darshan, Shai Avidan, Gal Chechik, and Rami Ben-Ari. Lightning-fast image inversion and editing for text-to-image diffusion models. In *International Conference on Learning Representations*, 2025.

Chenyang Si, Ziye Huang, Yuming Jiang, and Ziwei Liu. FreeU: Free lunch in diffusion U-Net. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Feng Tian, Yixuan Li, Yichao Yan, Shanyan Guan, Yanhao Ge, and Xiaokang Yang. PostEdit: Posterior sampling for efficient zero-shot image editing. In *International Conference on Learning Representations*, 2025.

Binxu Wang and John J. Vastola. Diffusion models generate images like painters: An analytical theory of outline first, details later. *arXiv preprint arXiv:2303.02490*, 2023.

Matthias Wright and Björn Ommer. ArtFID: Quantitative evaluation of neural style transfer. In *Pattern Recognition (DAGM GCPR)*, 2022.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-Image technical report. *arXiv preprint arXiv:2508.02324*, 2025a. Technical report.

Chenyuan Wu, Jiahao Wang, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, Ziyi Xia, Ze Liu, Chaofan Li, Haoge Deng, Kun Luo, Bo Zhang, Jiajun Zhang, Dong Liu, Defu Lian, Xinlong Wang, Zhongyuan Wang, Tiejun Huang, and Zheng Liu. OmniGen2: Towards instruction-aligned multimodal generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21964–21975, 2026a.

Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-more generalization: Unlocking more controllability by in-context generation. In *IEEE/CVF International Conference on Computer Vision*, 2025b.

Shaojin Wu, Mengqi Huang, Yufeng Cheng, Wenxu Wu, Jiahe Tian, Yiming Luo, Fei Ding, and Qian He. Unified customized generation by disentangled reward modeling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 34418–34427, 2026b.

Bin Xia, Bohao Peng, Yuechen Zhang, Junjia Huang, Jiyang Liu, Jingyao Li, Haoru Tan, Sitong Wu, Chengyao Wang, Yitong Wang, Xinglong Wu, Bei Yu, and Jiaya Jia. DreamOmni2: Multimodal instruction-based generation and editing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29275–29284, 2026.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. OmniGen: Unified image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13294–13304, 2025.

Cong Xie, Han Zou, Ruiqi Yu, Yan Zhang, and Zhenpeng Zhan. SerialGen: Personalized image generation by first standardization then personalization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2847–2856, 2025a.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. SANA: Efficient high-resolution text-to-image synthesis with linear diffusion transformers. In *International Conference on Learning Representations*, 2025b.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Appendix for DBFP:

Plug-and-Play Dual-Band Fusion Fixed-Point Sampling for Subject-Driven Generation

A PSEUDOCODE

Algorithm 1 is the full sampling procedure referenced from Section 3.7.

Algorithm 1 DBFP (one trajectory). Φ is Eq. equation 13; active indices $\mathcal{S}$ are Eq. equation 17.

Require: frozen v_θ; I_c, p_{ed}; frozen constructor $\mathcal{A}$; $\{t_i\}_{i=0}^{N}$; ρ, κ; γ; K

$p_{\mathrm{id}} \leftarrow \mathcal{A}(I_c, p_{\mathrm{ed}})$ ▷ Eq. equation 4; once
$c_{\mathrm{id}} \leftarrow (I_c, p_{\mathrm{id}}),\ c_{\mathrm{ed}} \leftarrow (I_c, p_{\mathrm{ed}})$
initialize z_{t_0}
for $i = 0, \ldots, N-1$ **do**
  $\Delta t \leftarrow t_{i+1} - t_i$
  if $i \in \mathcal{S}$ **then**
    $\tilde{z} \leftarrow z_{t_i}$
    for $k = 0, \ldots, K$ **do**
      $v^\star \leftarrow \Phi(\tilde{z}, t_i)$ ▷ BandMix of dual guided fields
      $\tilde{z} \leftarrow z_{t_i} + \Delta t\, v^\star$ ▷ probe; not a schedule advance
    end for
    $z_{t_{i+1}} \leftarrow z_{t_i} + \Delta t\, v^\star$
  else
    $z_{t_{i+1}} \leftarrow z_{t_i} + \Delta t\, \mathrm{CFG}_B(z_{t_i}, t_i; c_{\mathrm{ed}})$
  end if
end for
return decode z_{t_N}

B PICARD DEPTH ABLATION (K=0 VS. K=1)

Table 2 reports Official vs. one-shot BandMix (K=0) vs. DBFP (K=1) on the shared n=16 cases used for Figure 5, with $(\rho, \gamma) = (0.08, 0.30)$, on the same four lineage representatives as Figure 4. K=0 applies BandMix once per active step; K=1 adds one Picard probe (two Φ calls). Mean STHS: K=1 is best or tied on every lineage.

C ρ/γ SENSITIVITY (EXTRA NOTES)

Figure 5 reports the sweep on four representative backbones—USO (FLUX-based), Qwen-Edit, SenseNova, and HiDream—chosen to cover distinct model designs in the broader evaluation. All sweeps use a shared set of n=16 matched cases and one fixed Official baseline per backbone ($\rho \in \{0.04, 0.08, 0.12, 0.16\}$ at γ=0.30; $\gamma \in \{0.15, 0.30, 0.45, 0.60\}$ at ρ=0.08; K=1). The default $(0.08, 0.30)$ remains near-best or on a broad plateau across these representatives, so we retain the shared recipe in Table 1.

D DETAILED DERIVATIONS

D.1 DISCRETE FOURIER TRANSFORM OF A TOKEN VELOCITY FIELD

For one channel of a field, written $v[y, x]$ with spatial indices $y \in \{0, \ldots, H-1\}$ and $x \in \{0, \ldots, W-1\}$ on the $H \times W$ token grid, the orthonormal DFT evaluated at frequency indices (m, n) is

$$\big(\mathcal{F}v\big)[m, n] = \frac{1}{\sqrt{HW}} \sum_{y=0}^{H-1} \sum_{x=0}^{W-1} v[y, x] \exp\Big[-2\pi \mathrm{i} \Big(\frac{my}{H} + \frac{nx}{W}\Big)\Big]. \tag{20}$$

Table 2: Picard depth ablation on four lineage representatives (n=16, $(\rho, \gamma) = (0.08, 0.30)$). Mean STHS. Δ is the relative change $100\times(K{=}1-K{=}0)/(K{=}0)$ (%).

Backbone	Official	K=0	K=1	$\Delta(K{=}1-K{=}0)$ (%)
USO	0.3671	0.3686	0.3692	+0.2
Qwen-Edit	0.3806	0.3843	0.3876	+0.9
SenseNova	0.3662	0.3672	0.3731	+1.6
HiDream	0.3791	0.3927	0.3973	+1.2

The inverse transform uses the opposite sign and the same normalization. Because the input is real, the spectrum has conjugate symmetry. The gates in Eq. equation 9 are real and centrosymmetric in the unshifted frequency coordinates, so the inverse mixture is real up to numerical precision.

D.2 BandMix state increment

By Eq. equation 10, under $v^\star = \Phi(z, t)$ (Eq. equation 13) and $M_\rho + \overline{M}_\rho = 1$, the Euler increment is

$$z_{t+\Delta t} - z_t = \Delta t\, \mathcal{F}^{-1}\big(M_\rho \odot \mathcal{F}(v^{\mathrm{id}})\big) + \Delta t\, \mathcal{F}^{-1}\big(\overline{M}_\rho \odot \mathcal{F}(v^{\mathrm{ed}})\big). \tag{21}$$

Each frequency receives one soft convex assignment; the sigmoid gate reduces ringing relative to a hard cutoff.

E Subject-Text Harmonic Score (STHS)

A contribution of this paper is an evaluation axis, not only a sampler. Section 3.8 argues that subject-driven editing lacks a standard unified ranking signal and defines Eqs. equation 18–equation 19. Here we record why STHS is higher-is-better, why the harmonic mean is the right aggregator, and how it should be read next to the cross-backbone summary (Table 1).

Subject-driven evaluation usually lists CLIP-I, DINO, and CLIP-T separately. A model that nearly copies the reference can look strong on identity cosines while ignoring the edit. STHS compresses the two goals into one score by (i) taking the geometric mean of the identity cosines (s_{id}) and (ii) taking their harmonic mean with CLIP-T. Cosines are used as-is when nonnegative; we apply a shared defensive floor $\max(x, 0)$ only if a cosine is negative, so Official and DBFP remain comparable.

Monotonicity. For $a, b > 0$, the harmonic mean $\mathrm{HM}(a, b) = 2ab/(a + b)$ is strictly increasing in each argument. Fixing b and differentiating yields

$$\frac{\partial \mathrm{HM}}{\partial a} = \frac{2b^2}{(a+b)^2} > 0, \tag{22}$$

and symmetry gives $\partial \mathrm{HM}/\partial b > 0$. Consequently, if $a_1 < a_2$ and $b_1 < b_2$, then $\mathrm{HM}(a_1, b_1) < \mathrm{HM}(a_2, b_1) < \mathrm{HM}(a_2, b_2)$. Because both s_{id} and CLIP-T are similarity scores (larger is better), any joint improvement raises STHS. The direction "larger is better" is therefore inherited from the underlying cosines; the harmonic mean only decides *how* imbalance is penalized.

Why harmonic, not arithmetic. For $a, b > 0$ one has $\mathrm{HM} \le \mathrm{GM} \le \mathrm{AM}$, with equality iff $a{=}b$. When the two arguments diverge, HM falls farthest below AM. Example: a balanced case $(s_{\mathrm{id}}, \text{CLIP-T}) = (0.714, 0.82)$ gives AM 0.767 vs. HM 0.763; a copy-reference case $(0.95, 0.30)$ gives AM 0.625 vs. HM 0.454. We adopt HM so that "high identity, failed edit" is scored as a failure rather than a mild trade-off, in the same spirit as ArtFID's multiplicative coupling of two criteria (Wright & Ommer, 2022), but with similarity (higher-better) inputs instead of distance (lower-better) inputs.

Aggregation. We compute Eq. equation 19 on every test case, then average over cases. Averaging cosines first and then taking HM is *not* equivalent and is not used.

F RULE-BASED ALIGNER Align

The map $\mathrm{Align}(\mathrm{cap}, p_{\mathrm{ed}})$ in Eq. equation 4 is deterministic and training-free. Let the VLM caption be a comma-separated list of short attributes. We first detect conflict axes from p_{ed} (and an optional edit-type tag when available): pose/view, color, style medium, lighting, and color temperature. Attributes on an active axis are dropped (pose/view words are always dropped, since target pose is re-injected). Remaining attributes are truncated to a small budget (category plus a few appearance tokens). We then append short target-state phrases parsed from p_{ed} when present—e.g. "seated", "front-facing standing", "pastel colors", "soft watercolor"—and wrap the result as "`A photo of ...`". Examples: a standing fox caption under a seated edit becomes "`A photo of fox, orange, white, black, puffy jacket, seated.`"; under a pastel edit, old color tokens are removed and "`pastel colors`" is appended. Full keyword lists match the released implementation.

G IMPLEMENTATION CHECKLIST

Practical implementation checks used for all matched Official vs. DBFP pairs:

- Match the initial noise/latent, random seed, resolution, time schedule, solver steps, and all backend-specific guidance scales.
- Build both branches from the same reference image and use the same CFG_B recipe and hyperparameters.
- Evaluate both branches at the same current or probe state and the same schedule time.
- Reshape the spatial velocity tokens to an $H \times W$ grid and apply the 2-D FFT over the two spatial axes independently for each batch element and latent channel.
- Construct the vertical and horizontal frequency bins with `torch.fft.fftfreq(H)` and `torch.fft.fftfreq(W)`, respectively, and pair them with the native, unshifted ordering of `torch.fft.fft2`; no `fftshift` is applied.
- Use `norm="ortho"` for both `fft2` and `ifft2`, and apply the complementary gates M_ρ and $1 - M_\rho$ constructed from the same radial frequency grid.
- Re-query the step velocity at the anchored probe $\tilde{z} = z_t + \Delta t\, v^\star$, and take only one scheduled Euler step from z_t using the final $v^\star$.
- For compute accounting, count backend-specific conditional, image-only (when applicable), and unconditional forward passes separately.

H DECODE TRAJECTORY PANELS

Figure 7 presents mid-ODE decode trajectories for USO, Qwen-Edit, SenseNova, and HiDream-O1, complementing the mechanism diagnostics in Figure 4. With the reference, edit prompt, and sampling recipe fixed, the panels show when the Official and DBFP trajectories begin to separate.

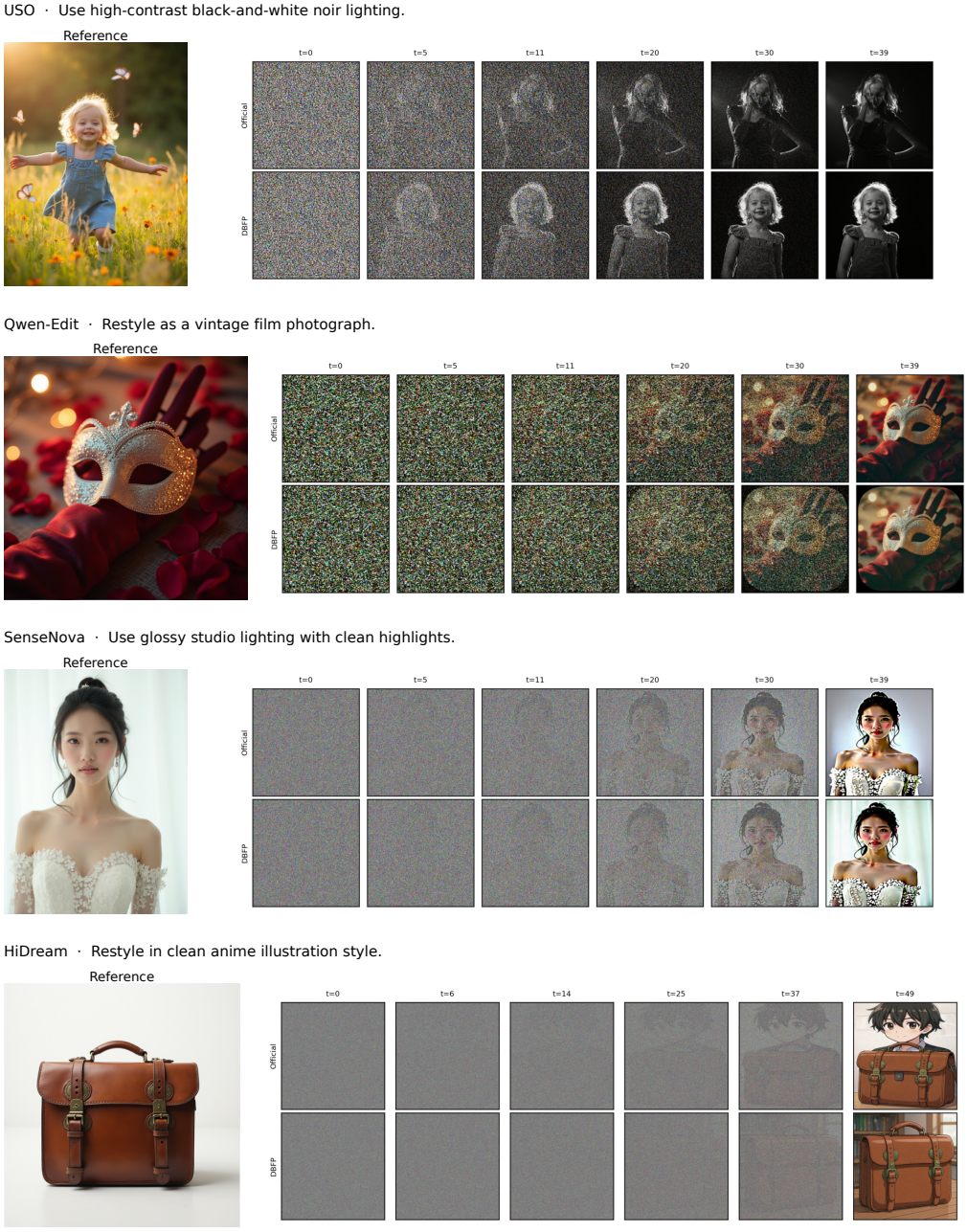

Figure 7: Reference (left; height-aligned) and Official (top) vs. DBFP (bottom) mid-ODE decode thumbnails on USO, Qwen-Edit, SenseNova, and HiDream-O1 under the shared recipe, with the edit prompt above each strip. Complements the teaser and main mechanism figures.

I Appendix qualitative set

Per-backbone six-row galleries under the same protocol follow (each row includes the edit prompt as vector text).



Figure 8: Additional SenseNova qualitative comparisons. Columns are Reference, SenseNova, and SenseNova+DBFP.

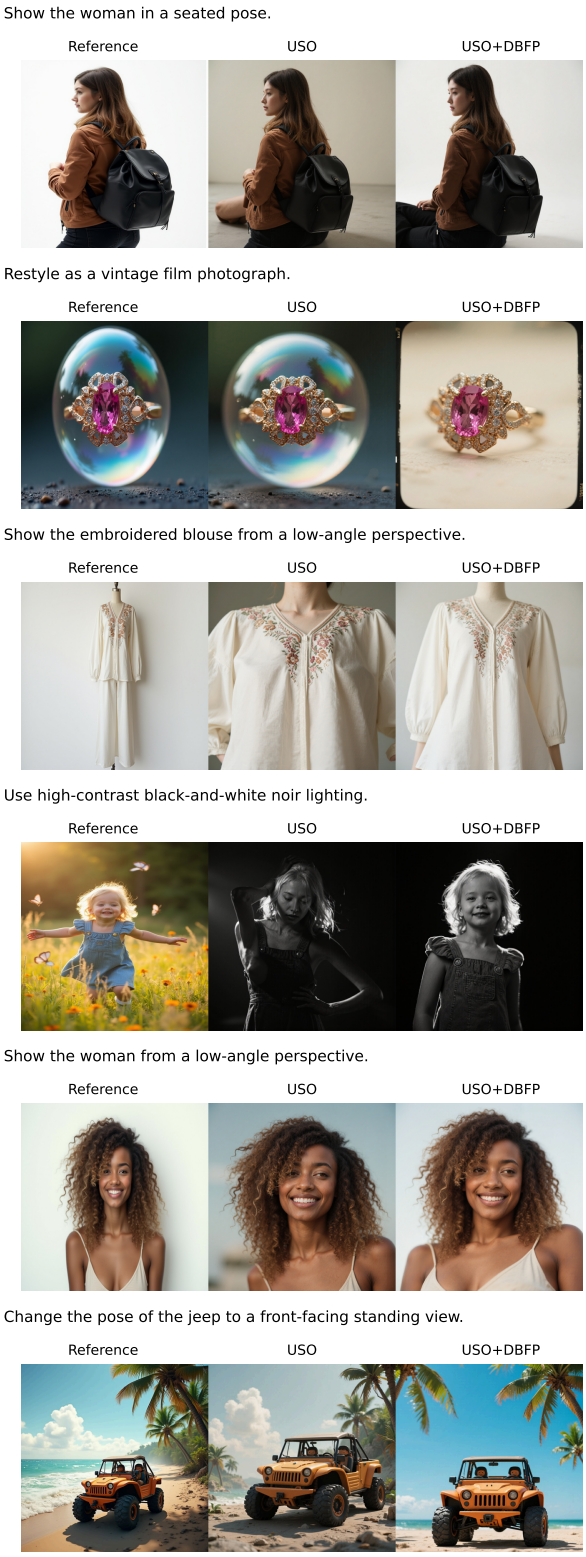

Figure 9: Additional USO qualitative comparisons. Columns are Reference, USO, and USO+DBFP.

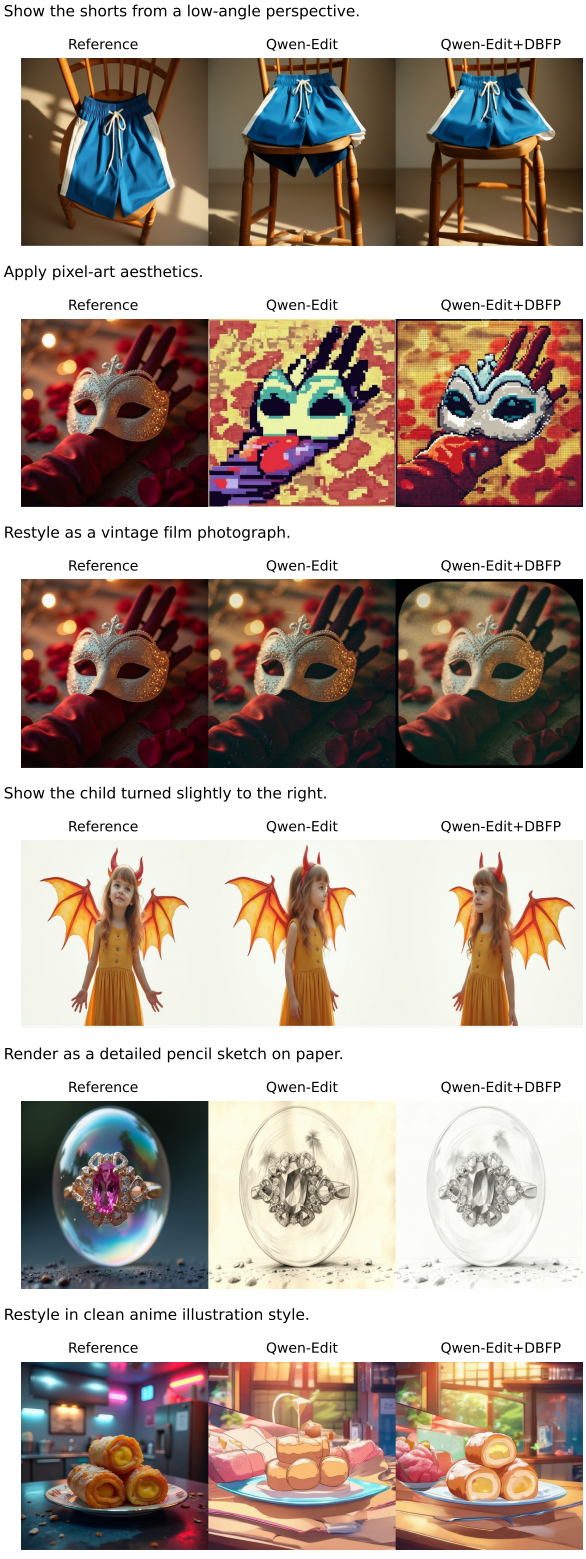

Figure 10: Additional Qwen-Image-Edit-2511 qualitative comparisons. Columns are Reference, Qwen-Edit, and Qwen-Edit+DBFP.

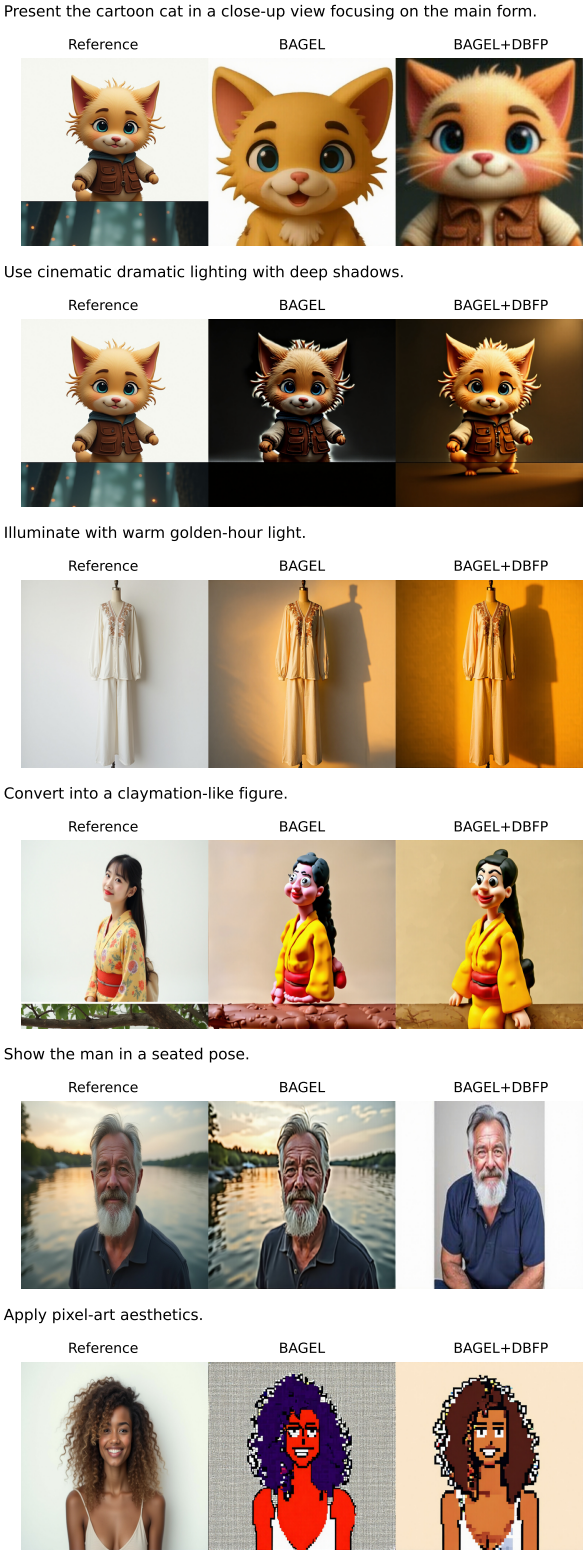

Figure 11: Additional BAGEL-7B-MoT qualitative comparisons. Columns are Reference, BAGEL, and BAGEL+DBFP.

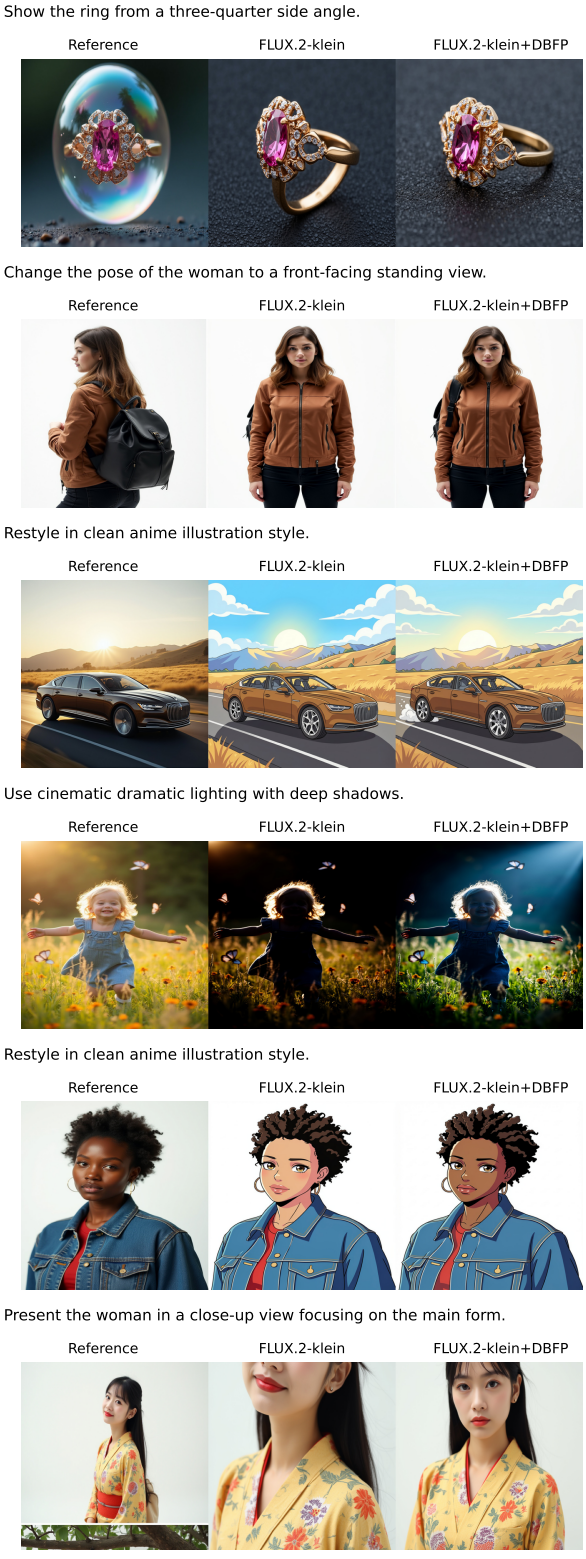

Figure 12: Additional FLUX.2-klein-base-9B qualitative comparisons. Columns are Reference, FLUX.2-klein, and FLUX.2-klein+DBFP.

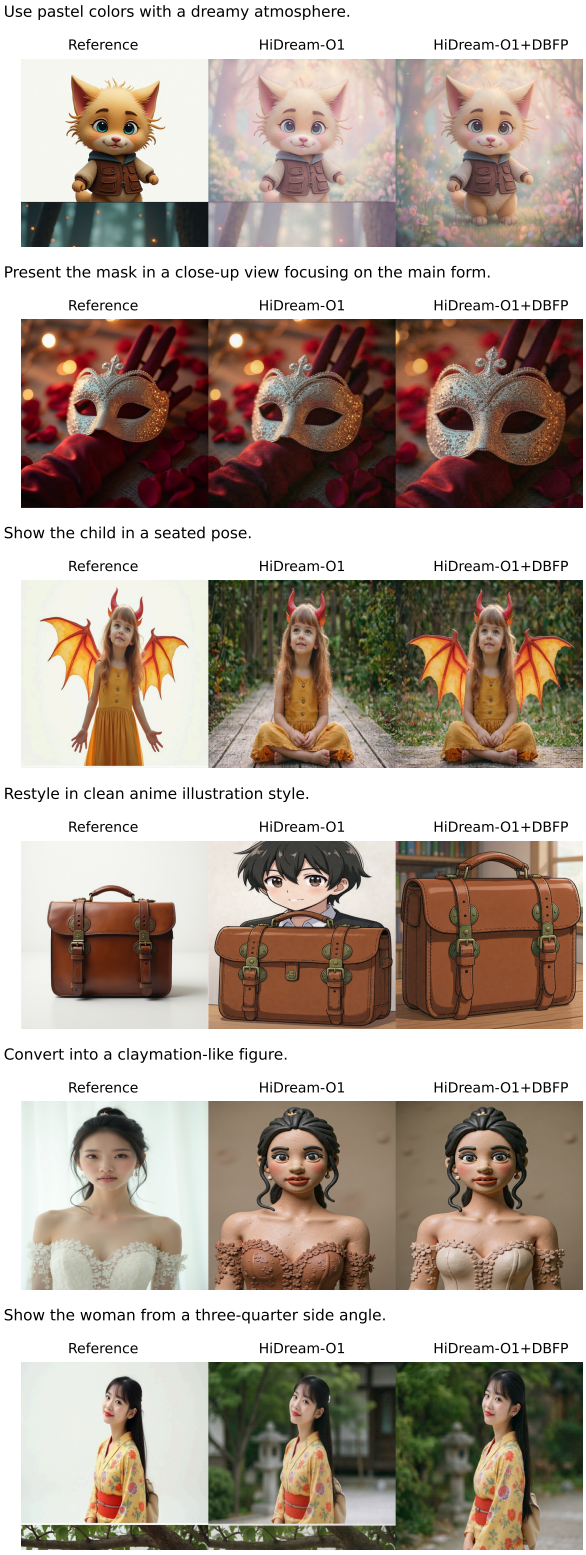

Figure 13: Additional HiDream-O1-Image qualitative comparisons. Columns are Reference, HiDream-O1, and HiDream-O1+DBFP.

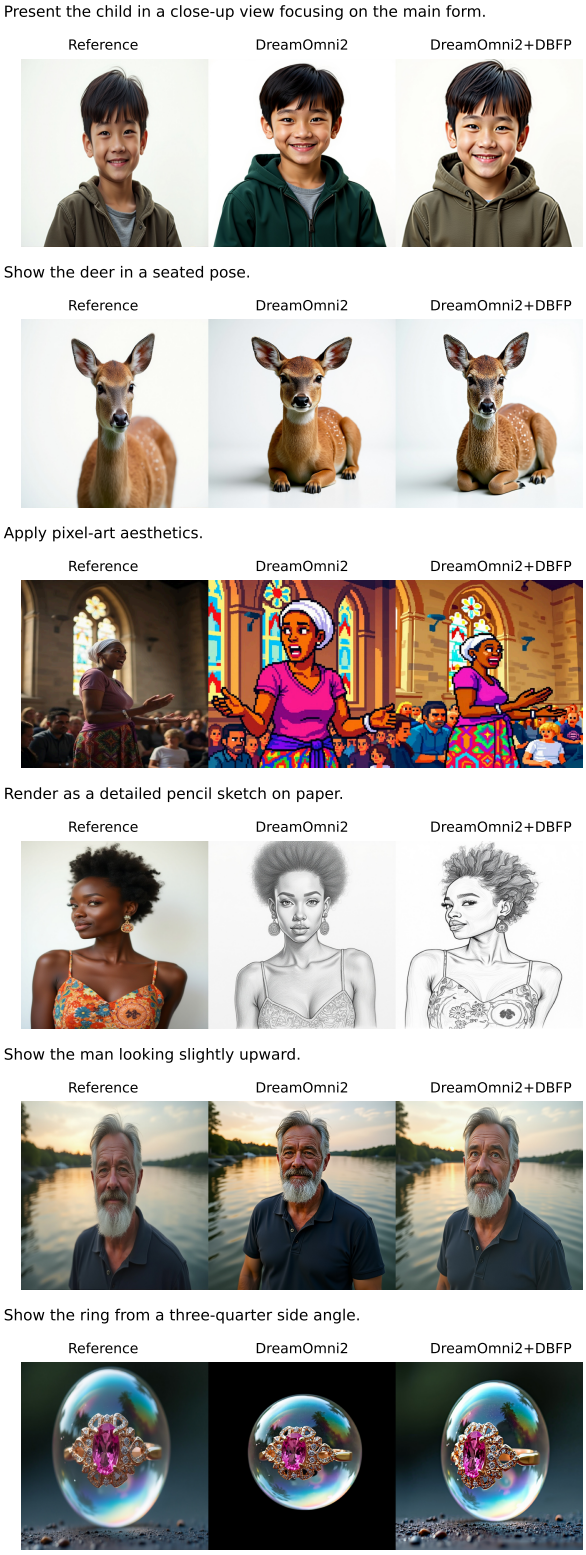

Figure 14: Additional DreamOmni2 qualitative comparisons. Columns are Reference, DreamOmni2, and DreamOmni2+DBFP.

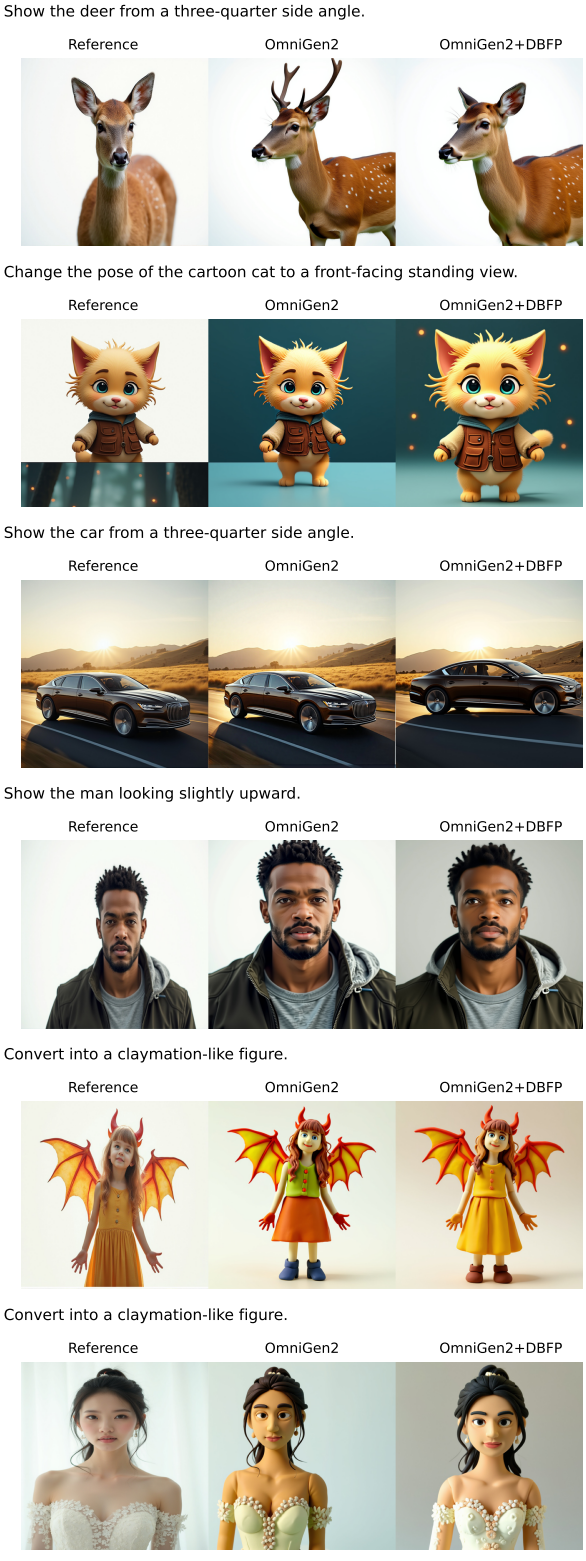

Figure 15: Additional OmniGen2 qualitative comparisons. Columns are Reference, OmniGen2, and OmniGen2+DBFP.