Perceptual quality assessment and enhancement of visual media using generative priors

Shreshth Saini

PhD student, The University of Texas at Austin

PhD Dissertation Defense · April 2026

Candidate and committee members



Shreshth Saini
Candidate
UT-Austin



Prof. Alan C. Bovik (Advisor)
Chair/Member
UT-Austin



Prof. Joydeep Ghosh
Member
UT-Austin



Prof. Diana Marculescu
Member
UT-Austin



Dr. Yan Ye
Member
Alibaba Group US



Dr. Balu Adsumilli
Member
Google Inc.

Dissertation roadmap

Part I. Perceptual quality assessment

- Beyond8Bits and BrightRate (reviewed)

  HDR-UGC data and quality model. CVPR 2026, WACV 2026

- LGDM (reviewed)

  Zero-shot image quality from diffusion priors. ICML 2025

- **HDR-Q**

  Multimodal LLM for HDR video. CVPR 2026

Part II. Generative enhancement

- **LumaFlux**

  SDR to HDR conversion. Preprint, 2026

- **Rectified-CFG++**

  Guidance for rectified flow models. NeurIPS 2025

- Future directions

Progress review recap

Beyond8Bits and LGDM

HDR-UGC benchmarks, and generative priors for perceptual quality

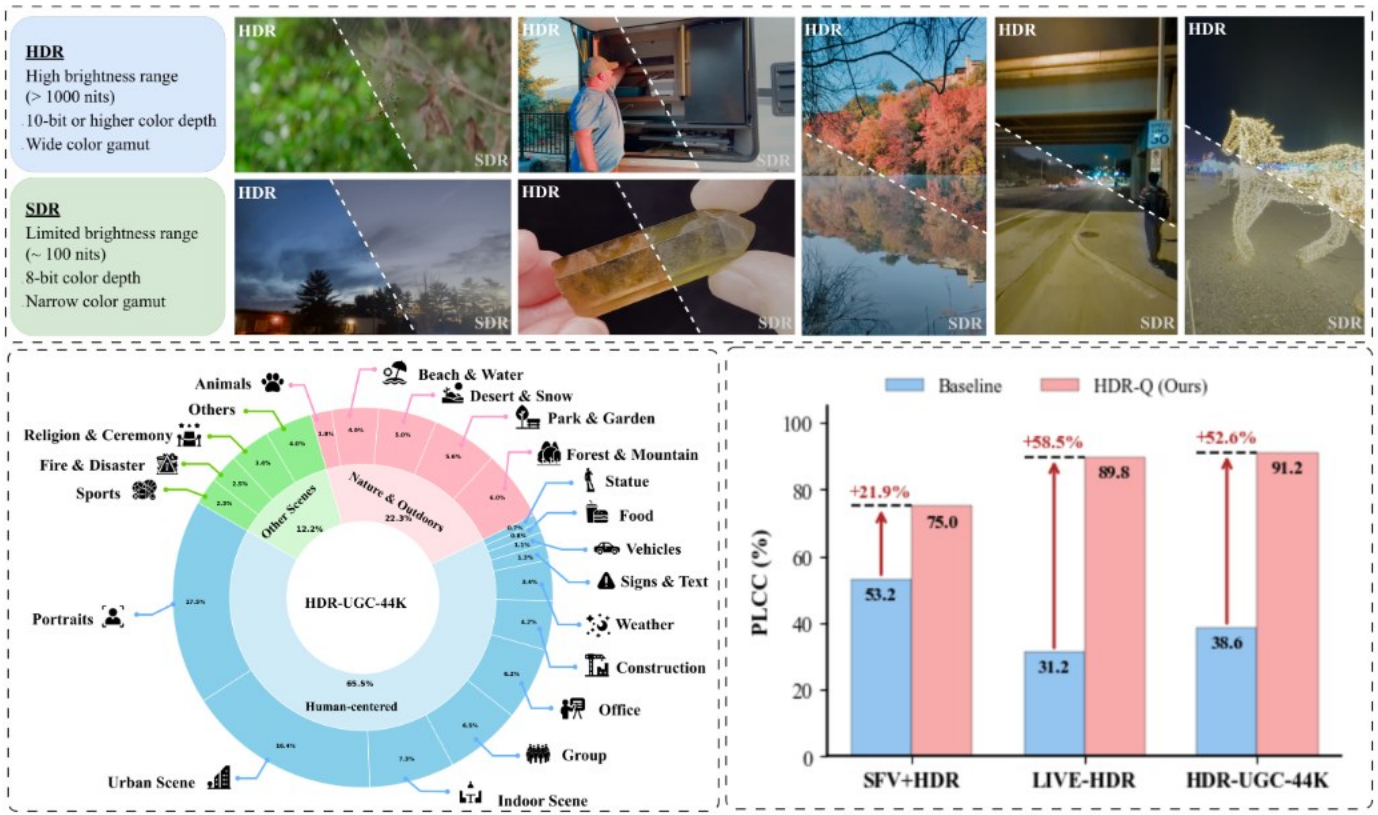
Beyond8Bits and BrightRate (review)

44,276
HDR-UGC clips, the largest such set

1.5M+
crowdsourced ratings

0.889
BrightRate SROCC, prior best 0.853

BrightRate is the first HDR-aware blind VQA model. It fuses HDR, CONTRIQUE and CLIP features with a temporal difference module.



HDR vs SDR gap, Beyond8Bits diversity, and HDR-Q improvements over baselines

Video sourcing
Vimeo and users
→
Bitrate ladder encoding
→
AMT study
35 ratings per video
→
SUREAL MOS aggregation

Beyond8Bits: Saini et al., CVPR 2026. BrightRate: Saini et al., WACV 2026 (Oral, Best Student Paper Award).
Earlier: HIDRO-VQA, Saini, Saha, Bovik, WACV 2024. CHUG, Saini et al., ICIP 2025.

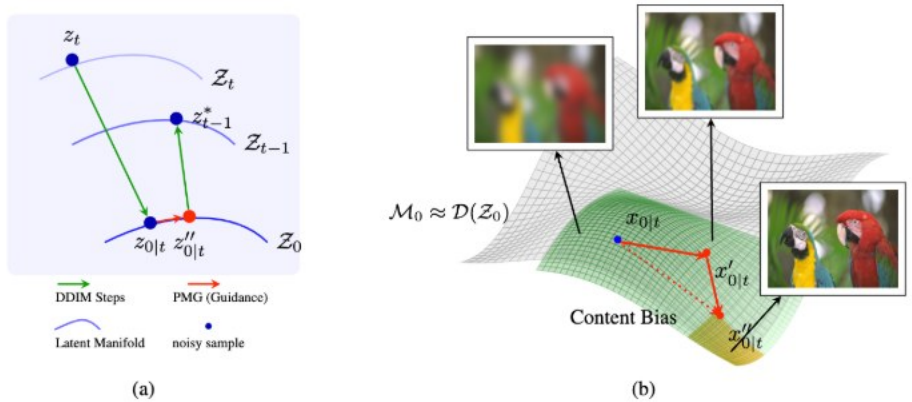
LGDM turns a diffusion model into a zero-shot quality judge (review)

- Pretrained diffusion models already encode a perceptual manifold
- Perceptual manifold guidance (PMG) keeps sampling on it
- Features from many timesteps and U-Net layers of SD v1.5, no fine-tuning

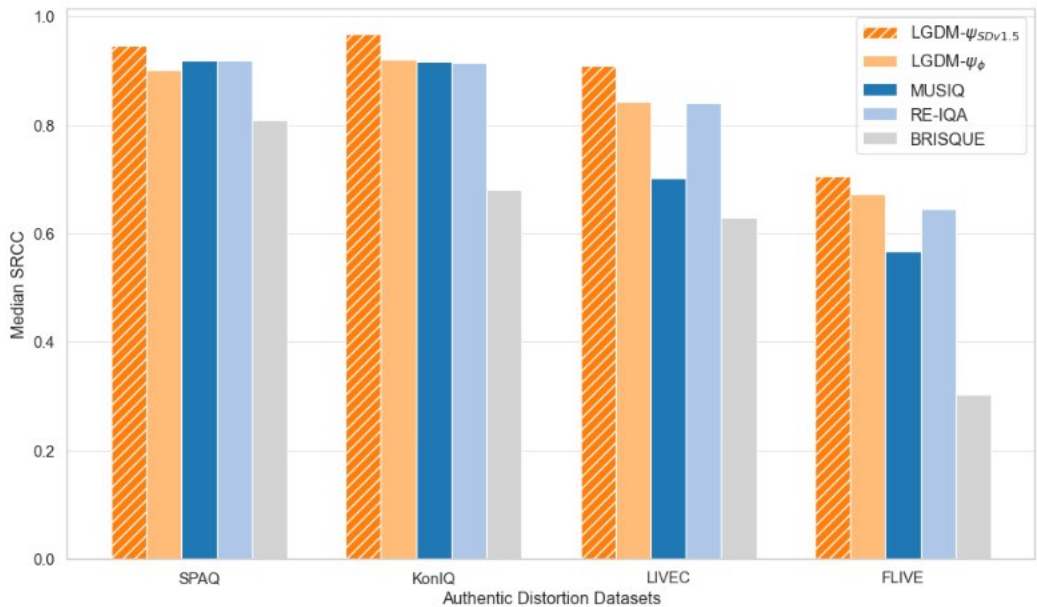
$$z^*_{t-1} = z_{t-1} + \zeta_2 \cdot \nabla_z \Phi(z_{0|t})$$

PMG guides denoising toward perceptually consistent manifold regions

Saini, Liao, Ye, Bovik. LGDM: Latent Guidance in Diffusion Models for Perceptual Evaluations. ICML 2025.



Perceptual manifold guidance, on-manifold quality-aware sampling



LGDM has the best median SRCC across authentic-distortion benchmarks

Five contributions

A pipeline from perception to generation for HDR visual media



Dashed boxes were covered at the progress review. Today's talk is the three on the right.

Both parts use generative models as evaluators and as enhancers of visual quality.

Part 1 of 3 · Perceptual quality assessment

HDR-Q, a multimodal LLM for HDR video quality

Can a model see HDR video and reason about its quality?

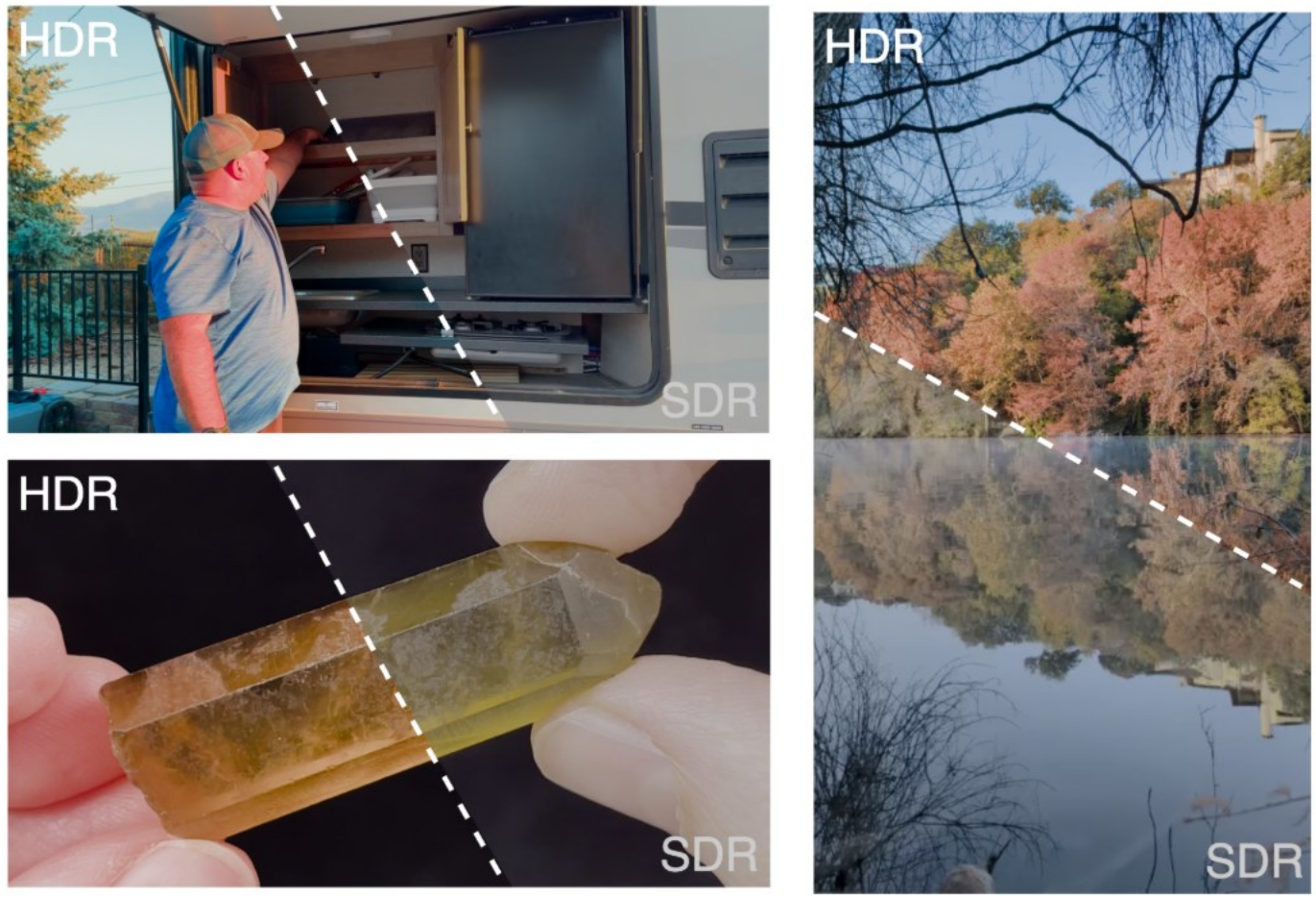
Saini, Chen, Birkbeck, Wang, Adsumilli, Bovik. Seeing Beyond 8bits: Subjective and Objective Quality Assessment of HDR-UGC Videos. CVPR 2026.

HDR is a different signal space

Every modern phone captures 10-bit HDR by default

- HDR10 is 10-bit, the BT.2020 wide gamut and the perceptual quantizer (PQ)
- Peaks of 1000 nits or more, against roughly 100 for SDR
- New artifacts: highlight clipping, near-black banding, color blooming, exposure flicker

SigLIP and CLIP take 8-bit sRGB, so they are structurally blind to these.



HDR keeps luminance and color detail that SDR collapses

Where current methods break

Method family	HDR input	Perceptual grounding	Reasoning	Continuous MOS	Interpretable	SROCC
Classical (BRISQUE, VMAF)	no	no	no	yes	no	0.41
Deep VQA (FastVQA, DOVER)	no	partial	no	yes	no	0.51
HDR-specific (HIDRO-VQA)	yes	yes	no	yes	no	0.85
MLLM-VQA (Q-Insight, DeQA)	no	no	yes	partial	yes	0.52
HDR-Q (ours)	yes	yes	yes	yes	yes	**0.92**

MOS is the mean opinion score. HIDRO-VQA: Saini, Saha, Bovik, WACV 2024.

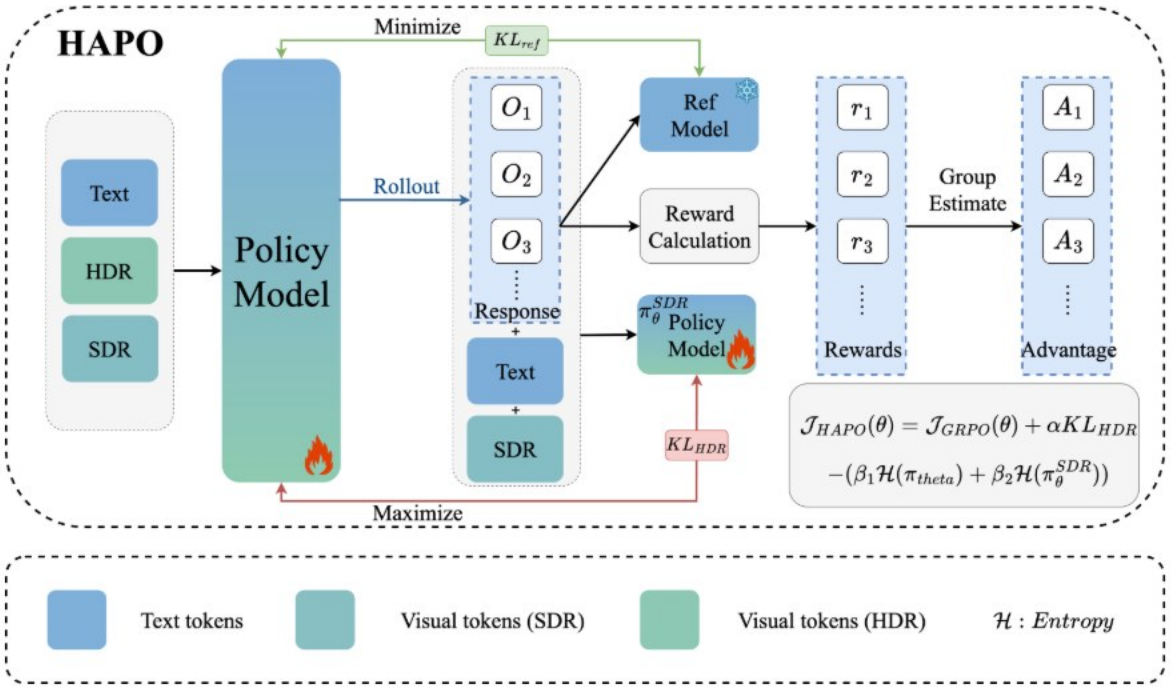
Only HDR-Q has HDR-aware perception, a continuous score and readable reasoning.

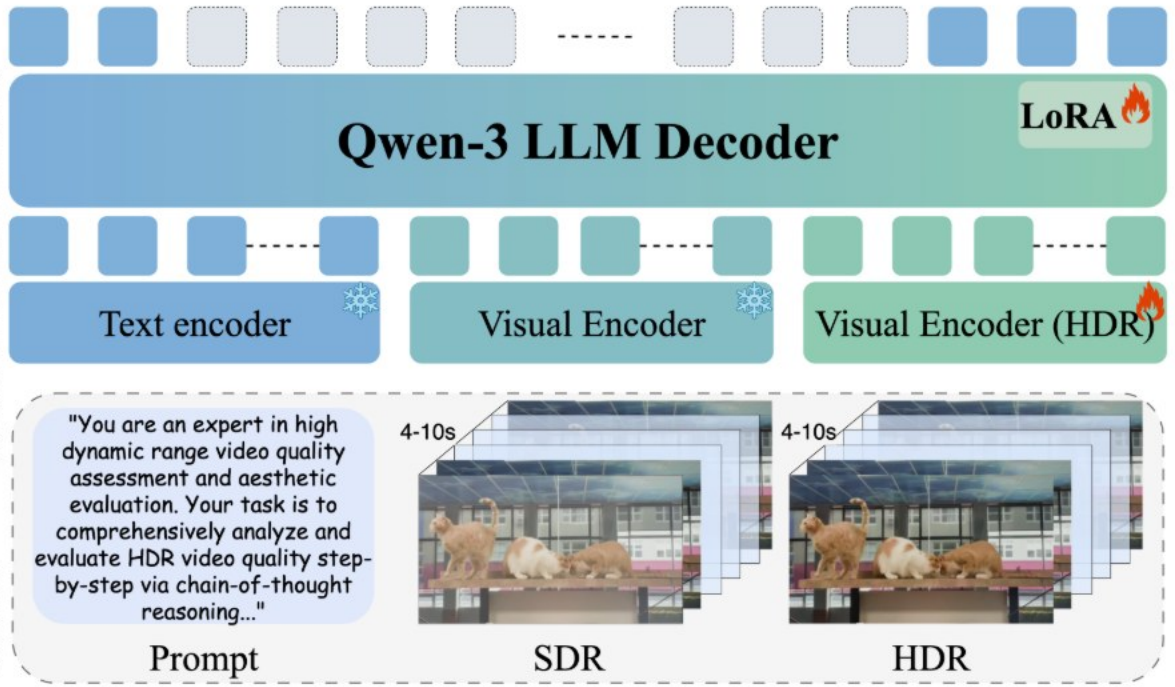
Three obstacles

- **O1. SDR-pretrained vision encoders.** They cannot represent 10-bit PQ.
- **O2. A continuous score from tokens.** The LLM writes text, and we need a MOS.
- **O3. Modality neglect.** Group relative policy optimization (GRPO) ignores the HDR tokens.

With GRPO, adding the HDR input lowers SROCC from 0.891 to 0.875.

HDR-Q architecture





HDR-aware encoder (O1)

SigLIP-2 fine-tuned contrastively on native 10-bit PQ, with separate HDR and SDR pathways.

Structured output (O2)

Ovis2.5 (9B), rank-4 LoRA. <think> reasoning, then an <answer> score. Gaussian reward, σ = 3.

HAPO training (O3)

HDR-aware policy optimization, built from a contrastive KL, dual entropy and an entropy-weighted advantage.

Adding the HDR input made GRPO worse

We call it modality neglect

0.8914
HDR-Q, SDR input only

0.8753
standard GRPO, HDR + SDR input

0.9206
HAPO, HDR + SDR input

- GRPO weights all tokens equally
- The model can shortcut through the text context
- Familiar SDR features win over unfamiliar HDR ones

HAPO rewards different outputs for HDR and SDR inputs.

The contrastive KL at the core of HAPO

If the model ignores HDR, its HDR and SDR outputs match and $D_{KL} \approx 0$. So we **maximize** it.

$$K_{HDR}(\theta) = D_{KL}(\pi^{HDR}_{\theta} \,\|\, \pi^{SDR}_{\theta}) \tag{1}$$

Maximized with coefficient $\gamma = 0.5$ in the HAPO objective

Formal guarantee

$$I_{\theta*}(\text{output}; \text{HDR} \mid \text{SDR}) \geq \gamma \cdot K_{HDR}(\theta^*) - \kappa \tag{2}$$

Mutual information between the output and the HDR content is lower-bounded, so the policy must use HDR.

The HAPO objective

GRPO plus three mechanisms that keep the policy on the HDR input

$$J_{\mathrm{HAPO}}(\theta) = E[\Sigma \min(\rho \cdot \tilde{A}, \mathrm{clip}(\ldots) \cdot \tilde{A})] - \beta\, D_{\mathrm{KL}}(\pi^{\mathrm{HDR}} \,\|\, \pi_{\mathrm{ref}}) + \gamma\, K_{\mathrm{HDR}} - H_{\mathrm{dual}}$$

Term	What it does	Setting
$\tilde{A}$	Entropy-weighted advantage (HEW)	$\lambda_{\mathrm{HEW}} = 0.3$
$\beta\, D_{KL}$	Reference KL, for stability	$\beta = 0.02$
$+\gamma\, K_{HDR}$	Contrastive KL, for HDR grounding	$\gamma = 0.5$
$-\, H_{dual}$	Dual entropy, against collapse	$\eta_1 = 0.01$, $\eta_2 = 0.05$

HDR-Q reaches 0.921 SROCC on Beyond8Bits

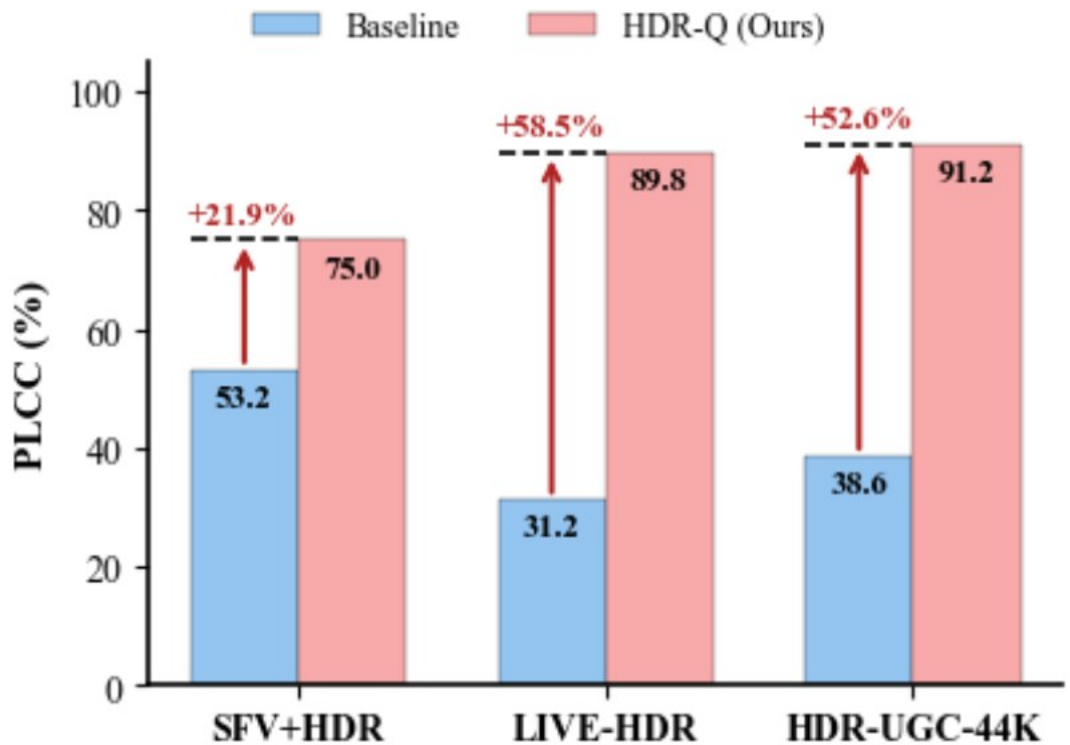
Method	SROCC ↑	PLCC ↑	RMSE ↓
BRISQUE	0.410	0.469	11.70
CONTRIQUE	0.625	0.605	15.02
CONVIQT	0.799	0.810	8.48
HIDRO-VQA	<u>0.851</u>	<u>0.878</u>	<u>6.09</u>
Q-Insight (best MLLM)	0.517	0.562	20.78
HDR-Q (SDR only)	0.891	0.890	7.42
HDR-Q (full)	**0.921**	**0.912**	**5.16**

Beyond8Bits (Saini et al., CVPR 2026). Bold is best, underline second best. HIDRO-VQA: Saini, Saha, Bovik, WACV 2024.

0.921
Beyond8Bits

0.908
LIVE-HDR, zero-shot

0.725
SFV+HDR, zero-shot



Zero-shot generalization, with no fine-tuning on the target datasets

HDR-Q names real artifacts where the baseline invents them

Question



You are an expert HDR video quality assessor. You analyze user-generated HDR videos with 10-bit PQ transfer, wide color gamut (e.g., BT.2020). You must provide step-by-step HDR-aware reasoning followed by a structured answer. Rate the overall quality of the video in 0-100 range, where 0 represent extremely poor quality and 100 represent excellent quality. First output the thinking process in <think></think> tags and then output the final answer with only score in <answer></answer>.

Ovis 2.5

<think> The image shows an indoor scene ... Overall, the details are well preserved, the colors look natural. ... However, further checking is necessary. ... The brightness of the image appears uneven: the trees on the left seem clearly underexposed, while the right side appears overexposed due to strong light. The plant colors lean toward a grayish green, suggesting poor color reproduction. The purple flowers also show color blooming, making the hues look unnatural.Therefore, the final score is 85.</think> <answer>85</answer>

HAPO (ours)

<think>The video shows a lush indoor botanical space... highlights in the mist roll off smoothly without harsh clipping, ... the hues remain stable without notable chroma shifts. Some fine detail is softened in the mistier regions, but this aligns with the natural ambience. Motion through the foliage likely feels gentle, with no obvious signs of jitter or smear. Overall, the scene benefits from balanced luminance, strong color reproduction, and clean gradients, giving the video a polished HDR appearance.</think> <answer>{"82"}</answer>

- **Ovis2.5 baseline** hallucinates underexposure, color blooming and unnatural hues.
- **HDR-Q** reports what is there: highlights that roll off without clipping, stable hues, no jitter or smear.

Same prompt to both models. HDR-Q identifies true HDR artifacts (highlight preservation, banding, color fidelity).

The contrastive KL matters most

Variant	SROCC	RMSE	Token entropy
GRPO baseline	0.81	10.73	0.20
+ HDR encoder	0.83	8.96	0.24
HAPO without contrastive KL	0.86	7.10	0.29
HAPO without HEW	0.88	6.11	0.27
HAPO without dual entropy	0.91	5.82	0.26
HDR-Q (full)	**0.92**	**5.15**	**0.33**

Ablation. Rows 3 to 5 each remove one component from the full model.

1. **Contrastive KL.** 0.92 falls to 0.86 without it, the largest drop.
2. **HEW.** 0.92 falls to 0.88. It sends gradient to quality-relevant tokens.
3. **Dual entropy.** 0.92 falls to 0.91. It prevents collapse.
4. **HDR encoder.** Lifts the GRPO baseline from 0.81 to 0.83.

Token entropy rises from 0.20 to 0.33, so the model keeps exploring.

Takeaways

1. **The first MLLM for HDR video quality.** 0.92 SROCC against 0.52 for prior MLLMs.

2. **HAPO fixes modality neglect.** Three ablated mechanisms, one mutual-information guarantee.

3. **It generalizes zero-shot.** 0.908 on LIVE-HDR and 0.725 on SFV+HDR with no fine-tuning.

Saini, Chen, Birkbeck, Wang, Adsumilli, Bovik. Seeing Beyond 8bits: Subjective and Objective Quality Assessment of HDR-UGC Videos. CVPR 2026.

Part 2 of 3 · Generative enhancement

LumaFlux: Structure-Aware SDR-to-HDR Conversion with Monotone Luminance Flows

Can a frozen image generator turn 8-bit SDR into faithful 10-bit HDR?

Saini, Gedik, Birkbeck, Wang, Adsumilli and Bovik, arXiv:2604.02787

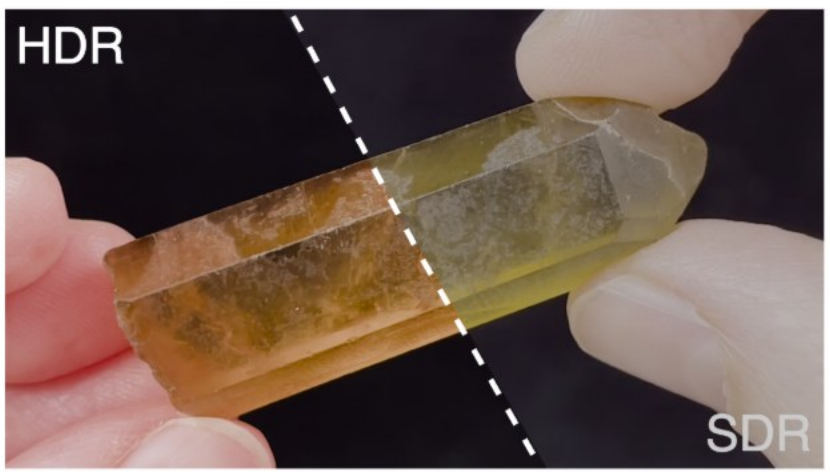
The question

Can a frozen generative prior be steered by the physics of the SDR frame?

- Inverse tone mapping (ITM) turns 8-bit SDR into 10-bit HDR.
- Our HDR is PQ (Perceptual Quantizer) in BT.2020, mastered at 1,000 nits.
- Image models know scenes, but nothing about nits or gamuts.
- So LumaFlux freezes the model and conditions it on measurements.

What gets lost, on three frames





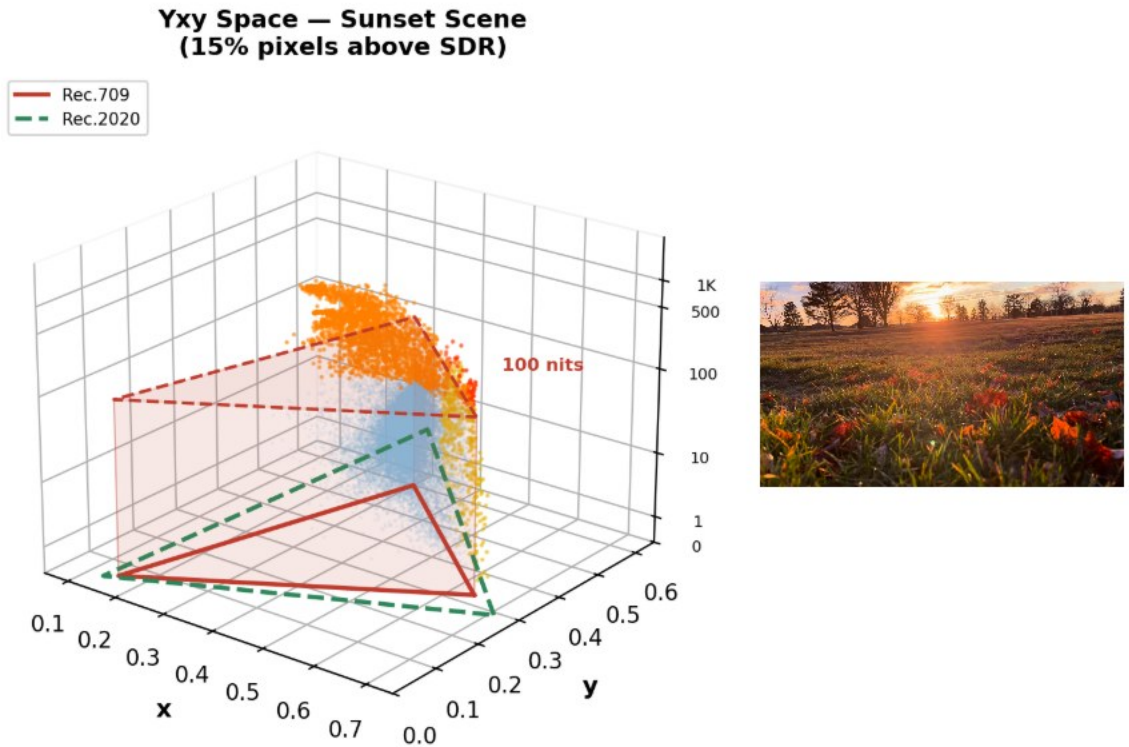


Each frame is split: HDR in the upper-left triangle, SDR in the lower-right.

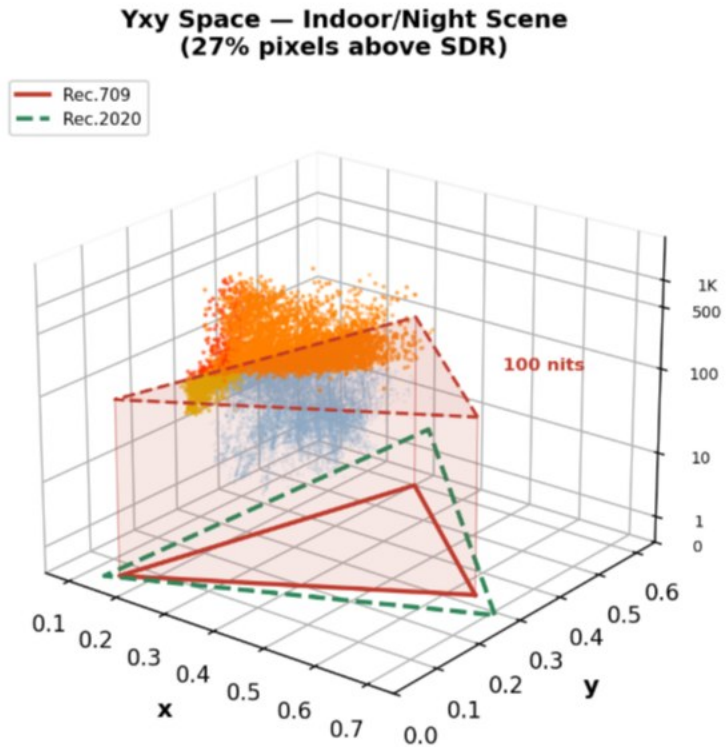
- **Shadows.** Detail inside the fridge is gone.
- **Color.** Foliage and sky lose saturation.
- **Highlights.** The crystal's speculars merge into one white patch.

The map is many-to-one, so a converter has to invent what is missing.

From the SDR input space to the HDR output space



A sunset clip. The shaded box is the SDR volume; orange points are HDR pixels above the 100-nit plane.





An indoor night clip. It has a larger share of pixels above the SDR ceiling than the sunset.

The input lives inside the box. The output has to fill the volume around it, without moving what is already right.

Prior work and where it fails

Analytic operators

Reinhard, BT.2446

One fixed curve per scene.
Clipped highlights stay clipped.

Learned converters

HDRTVNet++, FMNet, ITM-LUT, KUNet

Excellent in domain, tied to the degradation they trained on.

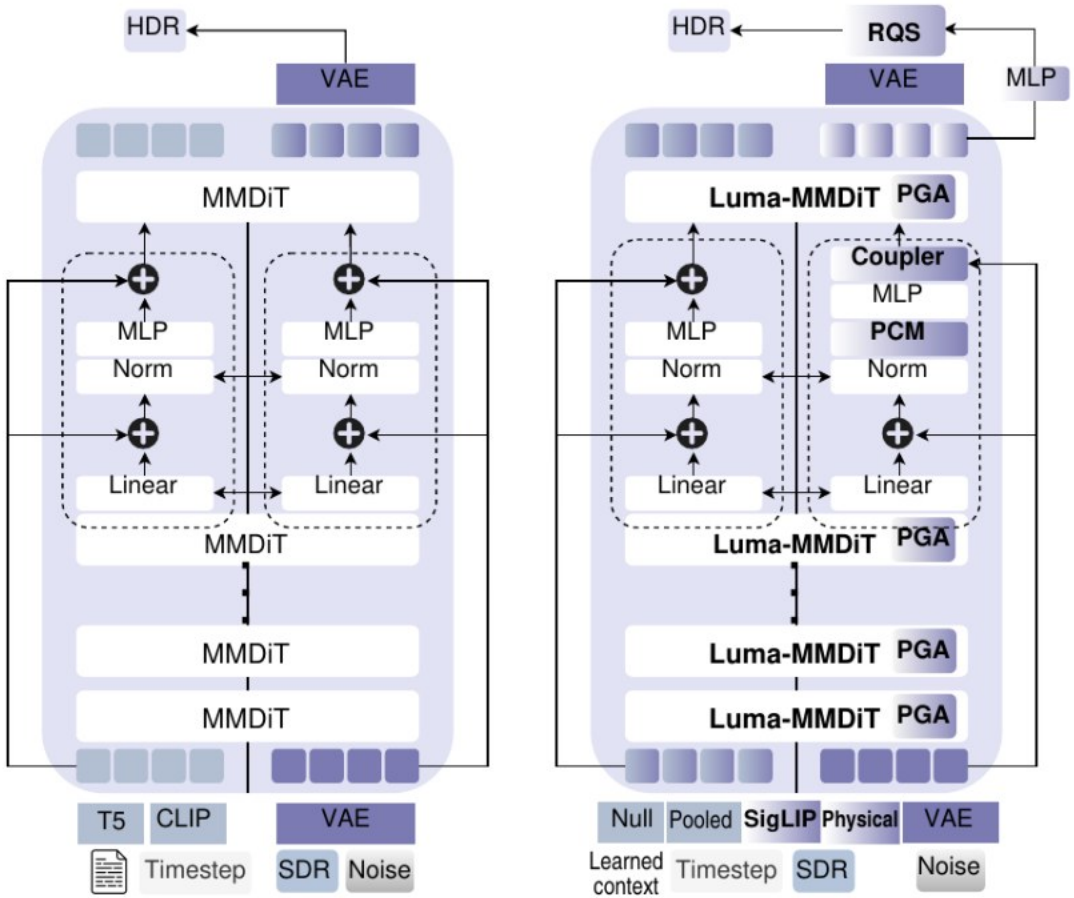
Generative priors

LEDiff, X2HDR

A real structural prior, but color and tone drift.

LumaFlux keeps the prior and pins down the physics.

A frozen FLUX transformer, steered by what the frame measures



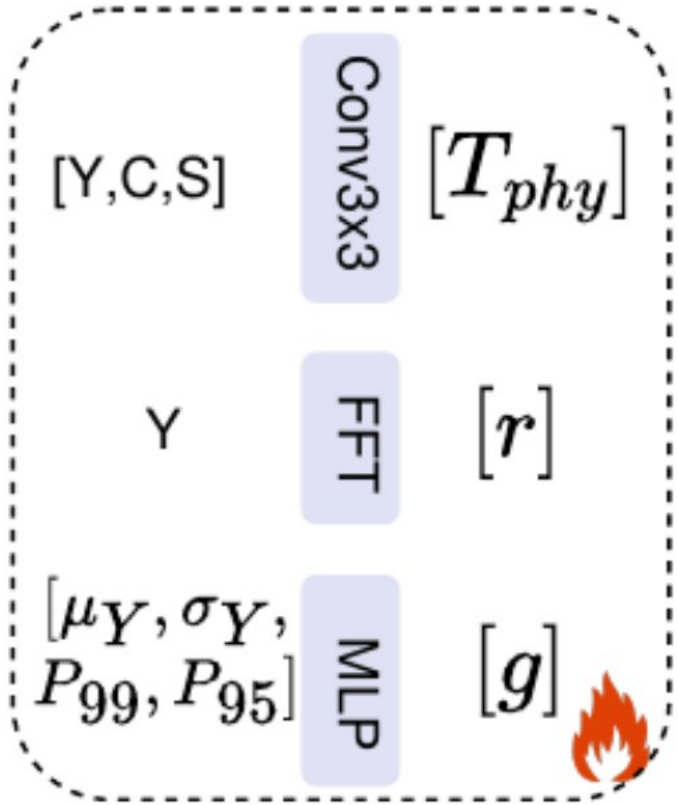
Left: direct adaptation, driven by a text prompt. Right: frozen blocks plus PGA, PCM, a coupler and the RQS decoder.

- **Physically-Guided Adaptation (PGA)** gates a low-rank update with luminance and gradients.
- **Perceptual Cross-Modulation (PCM)** adds scene context from SigLIP.
- **Rational-Quadratic Spline (RQS)** decoder calibrates luma and keeps its ordering.
- **Bridge.** SDR latent to HDR latent in 8 Euler steps.

0.57%

of 12.48B weights trained (71.3M)

PGA is a LoRA told what it is looking at



T_{phys} per token, from luminance, gradient, saturation; g in $\mathbb{R}^{16}$ and r (FFT energy, 8 bands) per frame.

$$R_{v,i} = (\alpha\, A_v B_v + \beta\, I)\, G_{phys,i}\, (I + n_{spec}\, \mathrm{Diag}(g_{FFT}))$$

- Rank 8, image-token value projection only.
- A physical gate sets adaptation per token and head.
- Constant gains instead of Ψ cost 0.81 dB and 6.09 ΔE_{ITP}.

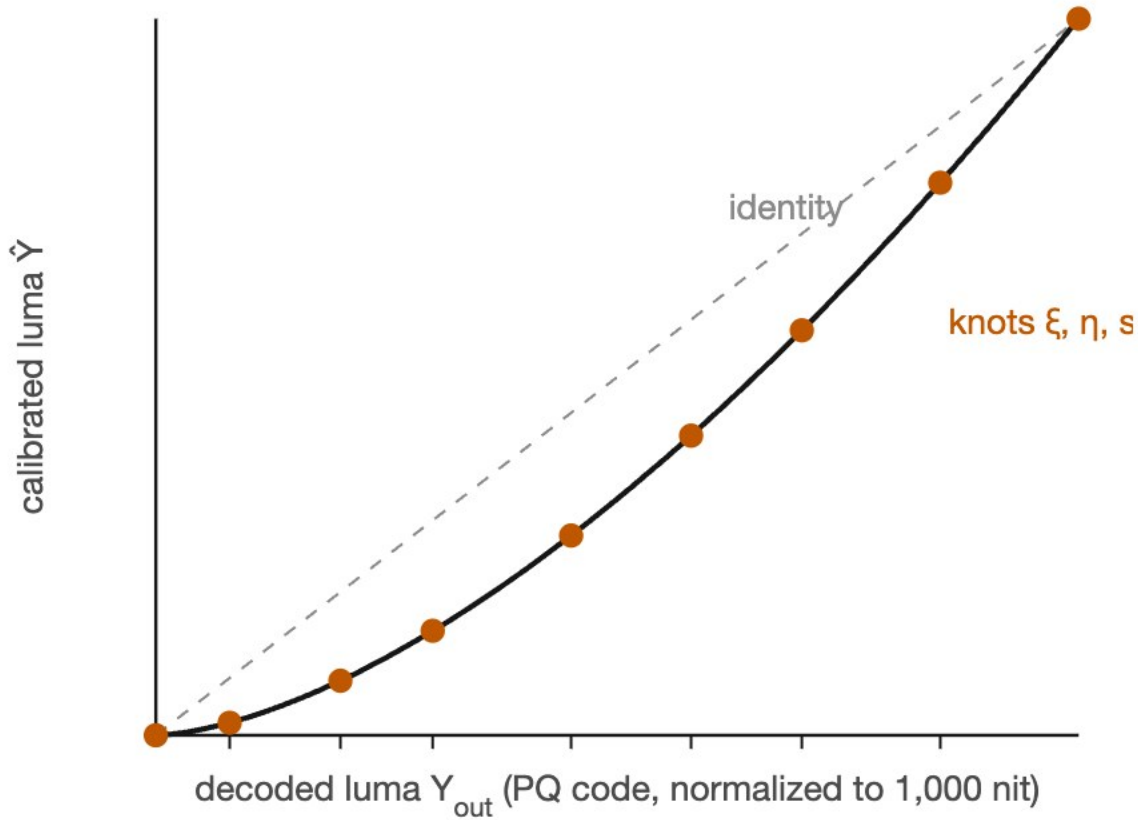
Scene context from SigLIP, and a monotone tone decoder

PCM sets one scale and shift per frame from pooled SigLIP tokens.

$$\hat{Y} = \mathrm{RQS}(Y_{out}; \xi, \eta, s) \tag{4}$$

$$\hat{x}_{hdr} = M_{RGB}([\hat{Y}, \hat{U}, \hat{V}]) \text{ in PQ BT.2020}$$

- 25 spline parameters per frame, all positive.
- So decoded luma keeps its ordering.
- Removing it costs 9.69 dB.



8 bins: 8 widths ξ, 8 heights η, 9 knot slopes s, from a 21.9K-parameter head that starts at identity.

A 314,396-pair corpus of user and professional video

314,396
SDR-HDR pairs in PQ BT.2020

2,092
source videos

144
GPU-hours on four GH200 GPUs, about 36 hours

Source	What it adds
CHUG (Saini et al., ICIP 2025)	856 user-generated references, 846 train and 10 reserved
LIVE-TMHDR	40 references with 120 expert-graded SDR clips
HDRTV1K	1,235 native training pairs
Sol Levante	one HDR10 master
Synthetic SDR	307,752 rows: eight tone-mapping operators, three CRF settings
Mix	171,384 user-generated (UGC), 143,012 professional (PGC)

All HDR is 16-bit PQ BT.2020. A manifest and code regenerate the corpus in about four CPU hours.

Training	Setting
Loss	flow matching, two L1 terms in linear light, spline smoothness
Crops	512^2, 1,024 tokens
Schedule	100,000 steps, global batch 16, learning rate 2e-4
Trained	71.3M parameters; the backbone stays frozen
In-domain checkpoint	HDRTV1K only, 50,000 steps, adapters only

The in-domain checkpoint is labeled wherever it appears. Every other result uses the main, mixed-corpus model.

LumaFlux is best in every column of the primary track

LumaEval synthetic track, 100 held-out frames under eight tone-mapping degradations, one scoring stack

Method	PU21-PSNR ↑	PU21-SSIM ↑	ΔE_{ITP} ↓	HDR-LPIPS ↓	HDR-VDP-3 ↑	FR-HIDRO ↓
BT.2446c	<u>23.74</u>	<u>0.8291</u>	<u>61.60</u>	0.307	5.611	0.753
Reinhard operator	23.62	0.8286	61.76	<u>0.295</u>	<u>5.726</u>	0.753
HDRTVNet++	23.03	0.8192	66.67	0.328	5.417	<u>0.703</u>
FMNet	22.86	0.8176	67.92	0.330	5.412	0.724
ITM-LUT	22.92	0.8142	67.41	0.315	5.496	0.761
VAE + RQS (control)	18.54	0.7647	111.67	0.381	3.673	0.924
LumaFlux (ours)	**24.23**	**0.8294**	**56.54**	**0.269**	**5.812**	**0.631**

Macro-averages over the same 100 IDs and 78 video-variant groups. Bold is best, underline second best. PU21-PSNR-Y is in the full table (backup).

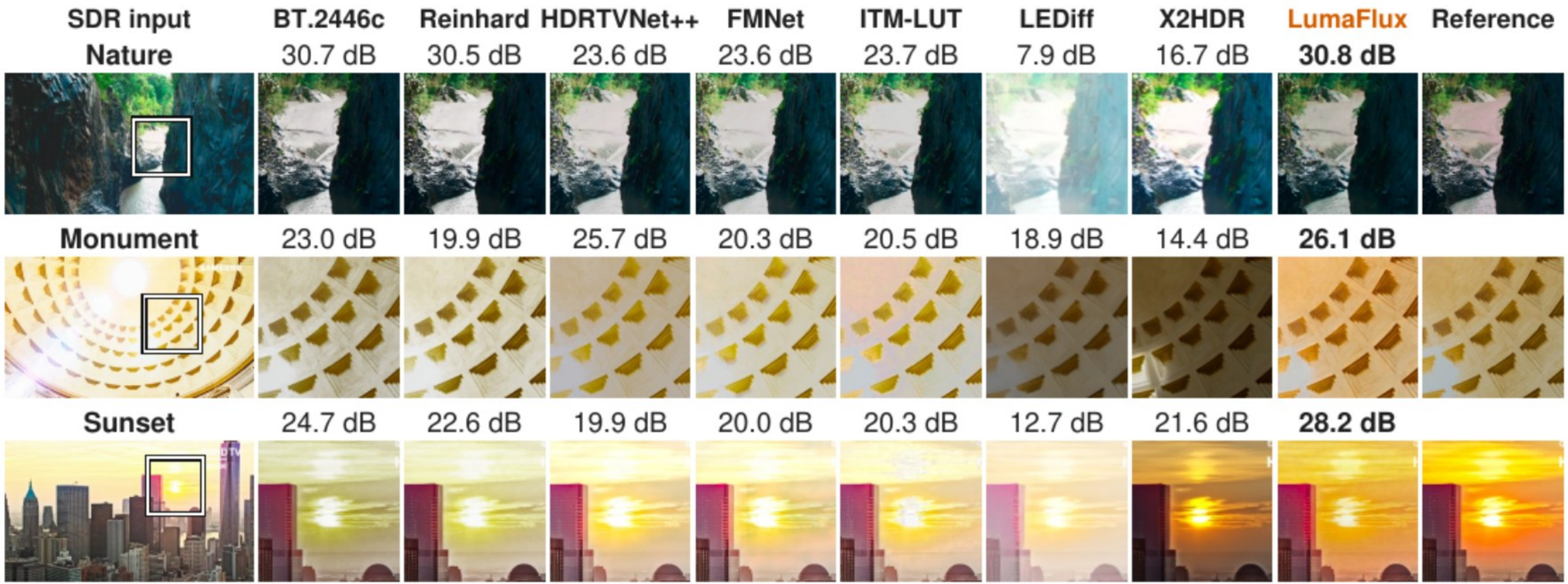
24.23 dB
PU21-PSNR, 0.49 dB above the best baseline

56.54
ΔE_{ITP}, 5.06 lower

0.631
FR-HIDROVQA, from 0.703 for HDRTVNet++

FR-HIDROVQA is a full-reference distance on HIDRO-VQA features (Saini et al., WACV 2024).

Three scenes, every method



SDR input with the patch box, then each method's crop. Full-frame PU21-PSNR is above each crop, best in bold.

On Sunset LumaFlux reaches 28.22 dB; the next best, BT.2446c, reaches 24.67.

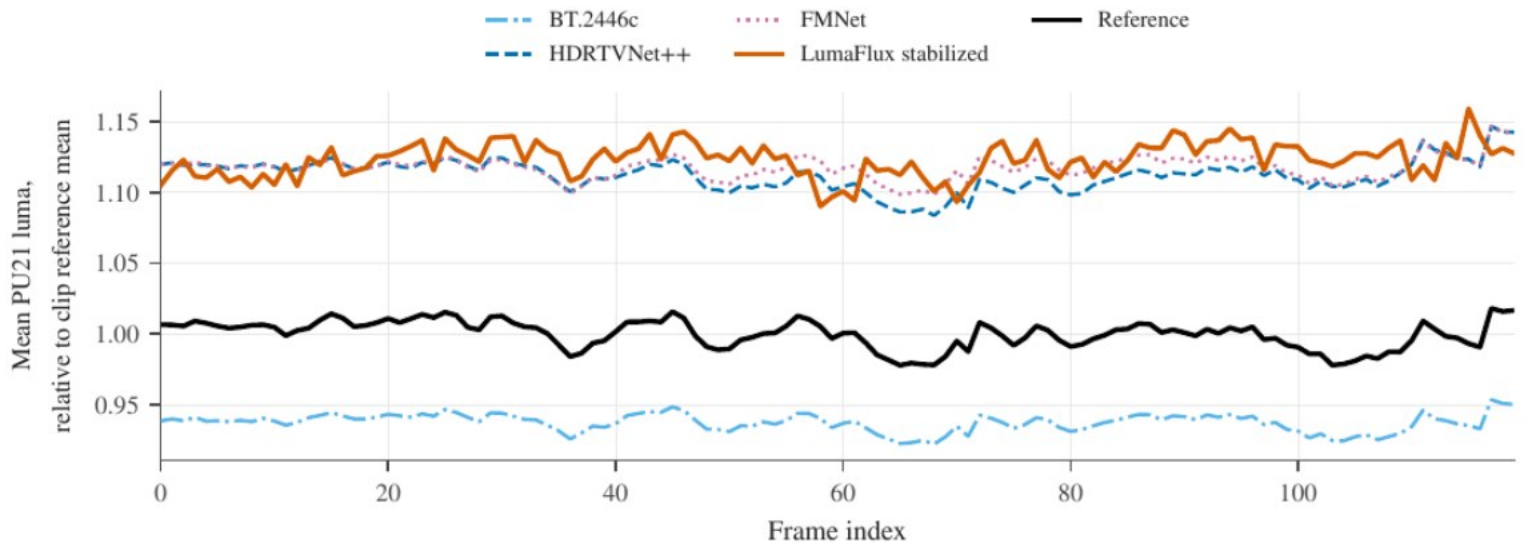
In the ablation, the tone decoder matters most

Row	Configuration	PSNR (Δ vs g)	ΔE_{ITP} (Δ vs g)
(c)	PGA, no spectral gate	28.57 (–0.62)	33.48 (+5.66)
(d)	PGA + spectral gate	**29.56 (+0.37)**	28.38 (+0.57)
(f)	full design, seed-0 run 1	28.67 (–0.52)	35.00 (+7.18)
(g)	full design, seed-0 run 2 (ours)	29.19 (ref.)	**27.81 (ref.)**
(h)	full, linear spline instead of RQS	28.74 (–0.45)	32.16 (+4.35)
(i)	full, no RQS decoder	19.50 (–9.69)	261.78 (+233.97)
(j)	full, no timestep-layer modulation	28.38 (–0.81)	33.90 (+6.09)

20k steps on a 55-video pool, the same 82 held-out frames. Deltas are relative to row (g). Rows (a), (b), (e) are in the full table (backup).

- **No RQS**, row (i): PSNR falls 9.69 dB.
- A linear spline, row (h), costs 0.45 dB.
- The spectral gate, (c) to (d), adds 0.98 dB.
- Rows (f) and (g) are one configuration, 0.52 dB apart, so smaller gaps are noise.

Video without temporal training



Per-frame mean PU21 luma, first 120 frames of four clips, relative to each reference mean. Stabilized LumaFlux only.

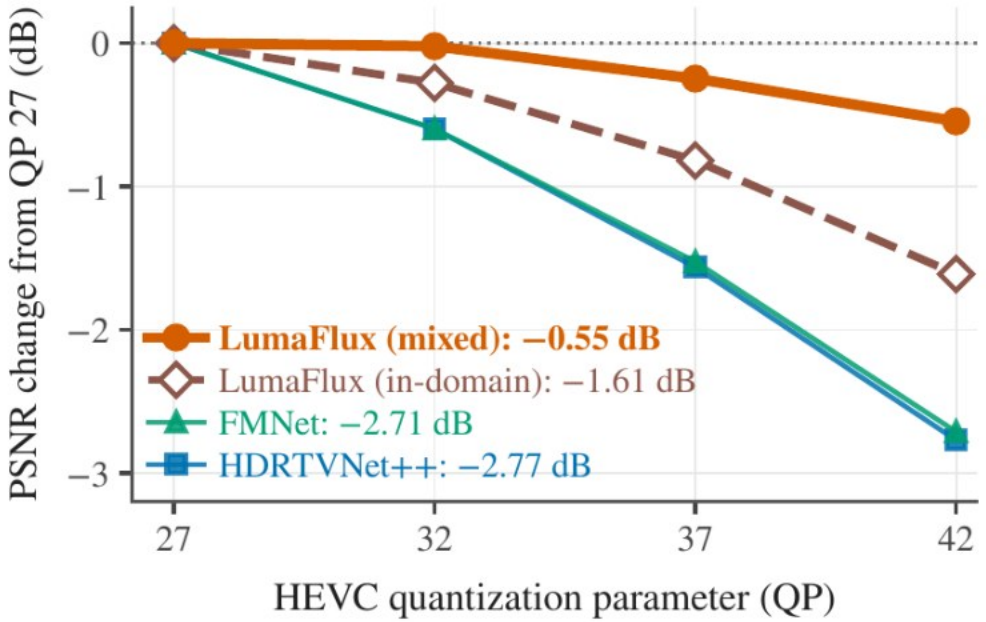
Method	Bonfire	02c3	055e	0392	0237	Mean
Reference	0.000	0.000	0.000	0.000	0.000	0.000
BT.2446c	7.813	1.411	0.450	**0.063**	**0.124**	1.972
HDRTVNet++	3.005	0.207	0.147	0.442	1.257	**1.012**
FMNet	3.853	**0.202**	**0.074**	0.579	1.124	1.166
LumaFlux	**1.992**	0.691	2.500	0.968	0.569	1.344
LumaFlux + stabilization	2.109	0.469	1.676	0.879	0.362	1.099

$|\text{flicker}_{\text{pred}} - \text{flicker}_{\text{ref}}|$, 10^6 PU21-energy units, lower is closer. 02c3 to 0237 are CHUG clips (Saini et al., ICIP 2025).

- Draw the bridge noise once and reuse it for every frame.
- Smooth the spline with an EMA, $\bar{\theta}_f = 0.8\ \bar{\theta}_{f-1} + 0.2\ \theta_f$.
- Signed excess flicker falls 53.3%, from 547.4 to 255.5 ($\times 10^3$).
- Raw flicker energy stays within 3.27% of the reference.

What holds up, and what falls short

Holds up under compression



PSNR change from QP 27 on the 117 HDRTV1K pairs, HEVC intra coding of the SDR input. Flatter is better.

From QP 27 to 42 the main model loses 0.55 dB; HDRTVNet++ loses 2.77 and FMNet 2.71.

Falls short

	LumaFlux	Reference point
HDRTV1K, in-domain checkpoint	33.34 dB	38.60 dB, best published
HDRTV1K, main model zero-shot	27.52 dB	same protocol
Mean ITP chroma, five scenes	60.65	49.80, the references
Seconds per 1080p frame	5.282	0.367 to 1.708, single pass

- The VAE is an SDR codec, so heavily clipped detail stays lost.
- Colors run oversaturated, and in-domain CNNs stay ahead on HDRTV1K.

Takeaways

1. **Measurements can steer a frozen flow transformer.**
   0.57% of weights trained, no prompt.

2. **A monotone tone field keeps luminance ordering.**
   Worth 9.69 dB.

3. **One checkpoint leads every primary metric and does video.**
   24.23 dB on LumaEval, flicker down 53.3%.

Saini et al., arXiv:2604.02787 · github.com/shreshthsaini/LumaFlux · shreshthsaini.github.io/LumaFlux

Next, Rectified-CFG++ (Saini et al., NeurIPS 2025) **changes how guidance is applied.**

Part 3 of 3 · Generative guidance

Rectified-CFG++, geometry-aware flow guidance

How do we guide a rectified flow model without pushing it off its manifold?

Saini, Gupta and Bovik, NeurIPS 2025

Why does guidance break on the best image generators?

Flux, SD3 and SD3.5 sample a rectified flow with a deterministic ODE. Classifier-free guidance (CFG) pushes their trajectories off the manifold, and nothing brings them back.

Why we need guidance at all

Without it, a flow model gives generic images that only loosely match the prompt

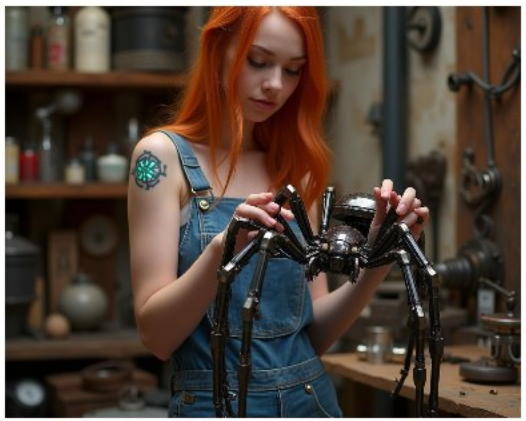
- **No guidance.** Dark, muddy, weak on the prompt.
- **CFG.** Follows the prompt, but oversaturates.
- **Ours.** Follows the prompt and stays natural.

> CFG was built for stochastic diffusion samplers (DDPM, SDXL). Flow models sample a deterministic ODE.

Flux-dev, same prompt and seed



No guidance



CFG, ω = 3.5



Ours, λ = 3.5

SD3.5, same prompt and seed



No guidance



CFG, ω = 3.5

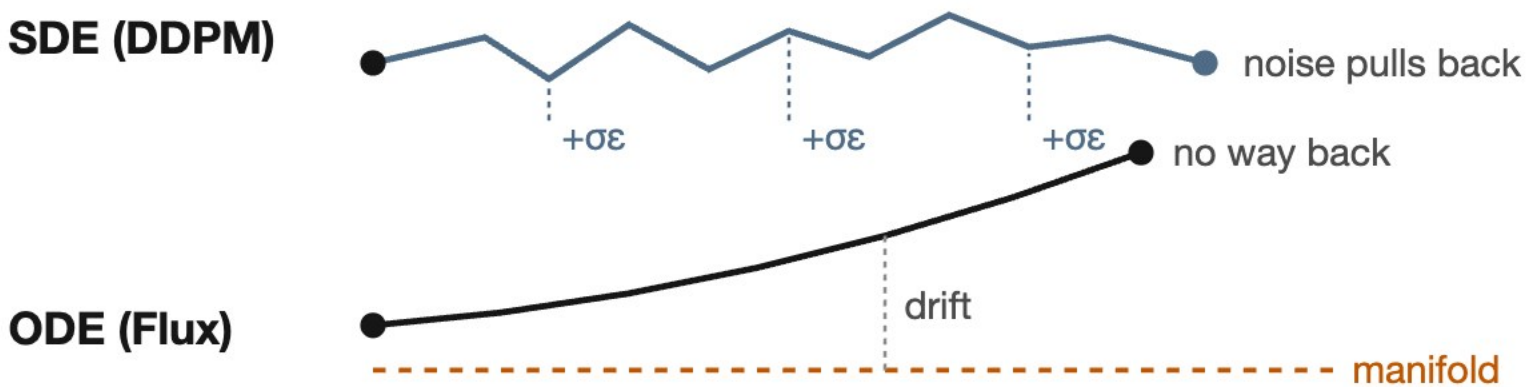


Ours, λ = 3

Why CFG works on diffusion but breaks on flows

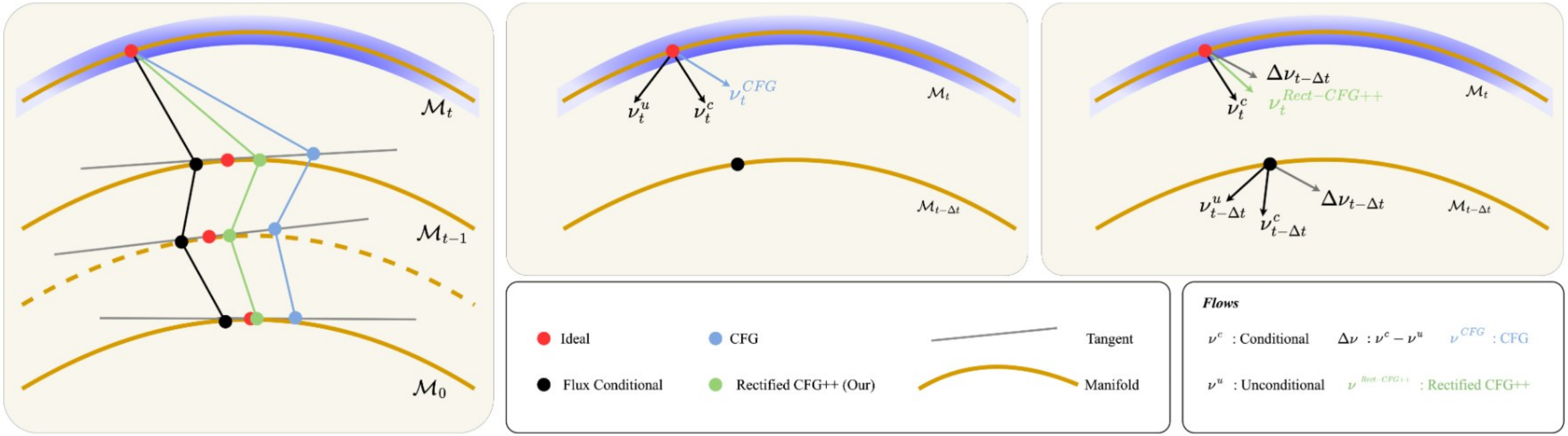
Same extrapolation trick, different samplers

Sampler	One step	After CFG overshoots
Diffusion SDE DDPM, SDXL	$\mathrm{d}x = [f(x,t) - g(t)^2\,\nabla \log p_t(x)]\,\mathrm{d}t + g(t)\,\mathrm{d}\bar{w}$ DDPM: $x_{t-1} = \mu_\theta(x_t, t) + \sigma_t\, z,\ z \sim \mathrm{N}(0, \mathrm{I})$	Fresh noise at every step pulls the sample back toward the data
Flow ODE Flux, SD3, SD3.5	$\mathrm{d}x = v_\theta(x, t)\,\mathrm{d}t$	Nothing pulls it back. Step t+1 starts off the manifold and the error compounds



So we need guidance designed for deterministic ODE samplers.

Why CFG fails and how we fix it, geometrically



Ideal flow. The velocity stays tangent to the manifold.

CFG. With ω > 1 the step lands off the manifold, for good.

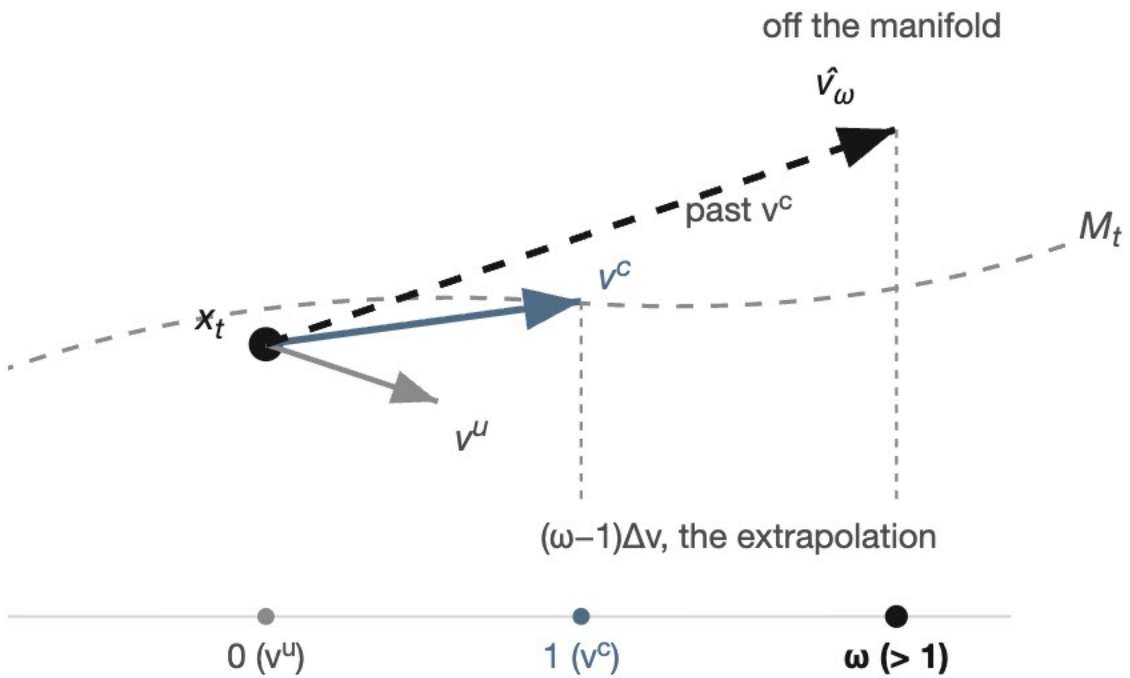
Ours. Predict a midpoint with v^c, evaluate guidance there, add a bounded correction.

Left to right: ideal flow, CFG, Rectified-CFG++. The step-by-step version is in the backup slides.

Extrapolation versus interpolation

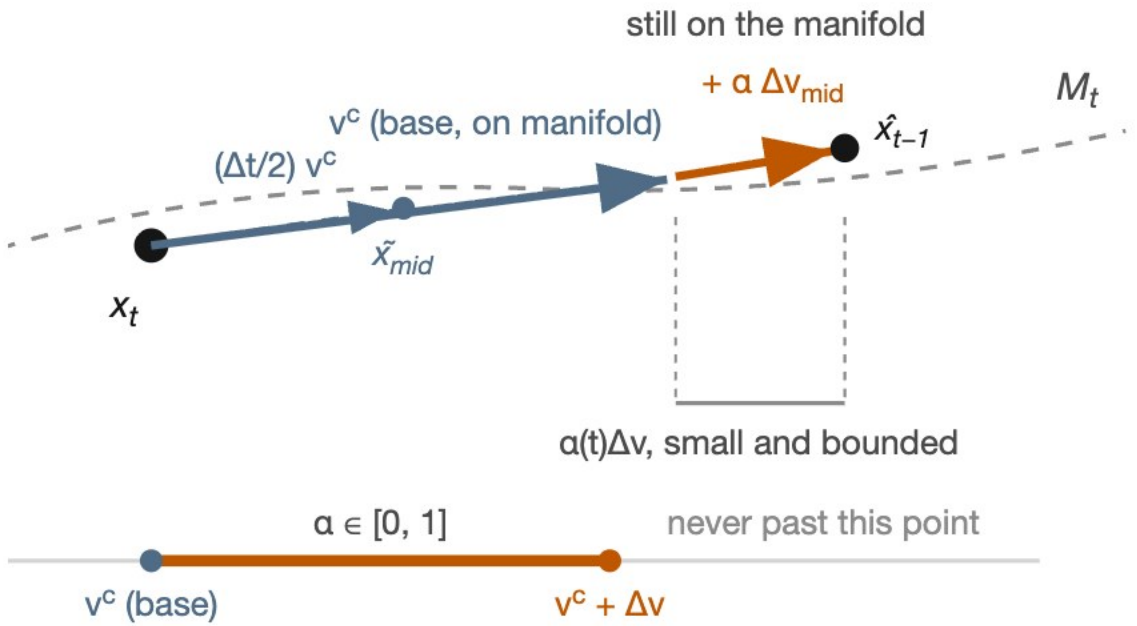
This one difference explains why CFG fails and Rectified-CFG++ works

CFG extrapolates (ω > 1)



ω = 3 means three times the distance from v^u to v^c, into territory the model never learned.

Ours interpolates (0 ≤ α ≤ 1)



The base is always v^c, on the manifold. The correction α(t) Δv_{mid} is bounded and decays to 0 near the data.

CFG runs ω from 3 to 9. Ours keeps α between 0 and 1.

The two algorithms side by side

Highlighted lines are what changes

Standard CFG

```
for t = T down to 1:
  v^c = v_θ(x_t, t, y)
  v^u = v_θ(x_t, t, ∅)
  v̂ = v^u + ω(v^c − v^u)          extrapolation
  x_{t−1} = x_t + Δt v̂
```

2 network evaluations (NFE) per step. Nothing corrects the overshoot.

Rectified-CFG++

```
for t = T down to 1:
  v^c = v_θ(x_t, t, y)
  x̃_mid = x_t + (Δt/2) v^c                predictor
  v^c_mid = v_θ(x̃_mid, t − Δt/2, y)
  v^u_mid = v_θ(x̃_mid, t − Δt/2, ∅)
  v̂ = v^c + α(t)(v^c_mid − v^u_mid)       interpolation
  x_{t−1} = x_t + Δt v̂
```

3 NFE per step. $\alpha(t) = \lambda(1-t)^{\gamma}$ decays to 0 near the data.

The base is always v^c, guidance comes from the midpoint, and $\alpha(t)$ fades out.

Theoretical guarantees

Four mild assumptions give three bounds. No earlier guidance method has any.

Result	Bound	In words
Lemma 1 guidance stability	$\|\Delta v^{\theta}_{t-\Delta t/2} - \Delta v^{\theta}_{t}(x_t)\| \le L\, V_{max}\, \Delta t$	Guidance at the midpoint differs from guidance at the current point by only O(Δt)
Proposition 1 bounded single-step perturbation	$\|\hat{x}_{t-1} - \tilde{x}_{t-1}\| \le \alpha(t)\, B\, \Delta t$	Deviation from the conditional path is set by α(t), so it vanishes near the data
Proposition 2 distributional deviation	$D_{\mathrm{KL}}(\hat{p}_0 \,\|\, p_0) \le C_s\, (B + L\, V_{max}\, \Delta t) \int \alpha(\tau)\, d\tau$	Total KL divergence is set by the integrated guidance ∫α(τ)dτ

Assumptions. (A1) $v_\theta(x,t,y)$ and $v_\theta(x,t,\varnothing)$ are L-Lipschitz in x. (A2) Bounded guidance direction, $\|\Delta v_t^{\theta}(x)\| \le B$. (A3) The schedule α($t$) is bounded and integrable. (A4) Bounded conditional velocity, $\|v^c_t\| \le V_{max}$.

One knob, ∫α(τ)dτ, controls how far the output can drift.

CFG versus ours, property by property

Lemma 2, manifold-faithful corrector

$$\operatorname{dist}(\hat{x}, M_{t-\Delta t}) \le C\,\varepsilon\,\Delta t$$

Distance to the manifold is bounded by the training error ε alone.

Integrated guidance

CFG: ∫(ω−1)dt = ω−1 ≈ 2 to 8



Ours: ∫α(τ)dτ = λ/(γ+1) ≈ 0.5



0.5

4 to 16× lower KL divergence

Property	CFG	Ours
Per-step perturbation	(ω−1) $B\,\Delta t$ constant, never vanishes	**α(t) $B\,\Delta t$** goes to 0 near the data
Manifold distance	Unbounded errors accumulate	**O(ε Δt)** training error only
Integrated guidance	ω−1 ≈ 2 to 8	**λ/(γ+1) ≈ 0.5**
KL($\hat{p}_0 \,\|\, p_0$)	O(ω−1)	**O(0.5)**
Late-step behavior	Same ω everywhere, destroys fine detail	α(t) goes to 0, preserves fine detail
Guidance evaluation	At the current point (possibly off the manifold)	At the predicted midpoint (on the manifold)

Watching the artifacts grow, or not

CFG

ω = 2.5



Ours

λ = 0.5





CFG, ω = 2.5



Ours, λ = 0.5

Each strip shows 7 denoising steps, from noise on the left to the final image on the right.

CFG drifts by step 2 or 3 and every later step amplifies it. Ours looks natural throughout.

Baseline, CFG and Rectified-CFG++ on three Flux prompts



Rows, top to bottom: Flux baseline with no guidance, CFG, Rectified-CFG++.

- **Cactus.** CFG draws a cartoon with an emoji face.
- **Dog and moon gate.** CFG loses the stone arch and paints mountains.
- **Parrot.** CFG turns the bird into a psychedelic abstraction.

Ours keeps the realism and the structure in all three.

CFG versus ours on Flux-dev, eight pairs



CFG



Ours



CFG



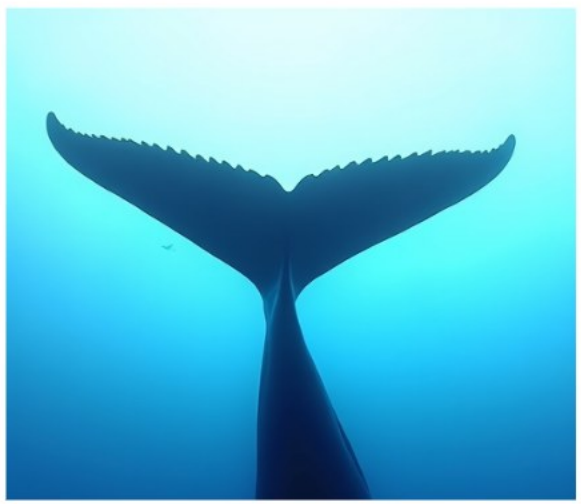
Ours



CFG



Ours



CFG



Ours

Text rendering, the stop sign test

Text needs pixel-precise strokes in the last denoising steps



CFG



Rectified-CFG++

Prompt: "A stop sign with 'ALL WAY' written below it."

CFG. STOP warps and ALL WAY is unreadable.

Ours. $\alpha(t)$ is 0 in the last steps, so the strokes stay crisp.

Results on MS-COCO 10K

Model	Method	FID ↓	CLIP ↑	ImgRwd ↑	Pick ↑	HPS ↑
Lumina	CFG	26.93	0.261	0.547	21.03	0.253
Lumina	Ours	**22.49**	**0.268**	**0.621**	**21.19**	**0.259**
SD3	CFG	26.33	0.273	0.684	21.48	0.264
SD3	Ours	**24.68**	**0.278**	**0.753**	**21.62**	**0.269**
Flux	CFG	37.86	0.285	0.892	22.04	0.279
Flux	Ours	**32.23**	**0.289**	**0.961**	**22.18**	**0.283**

CFG versus Rectified-CFG++ on three flow models. Bold is best. ImgRwd is ImageReward, Pick is PickScore.

Guidance (SD3.5)	FID ↓	ImgRwd ↑	CLIP ↑	HPSv2 ↑
CFG	67.71	1.053	**0.352**	0.294
CFG-Zero*	68.39	0.995	0.346	0.288
APG	67.23	1.075	0.351	0.294
Rectified-CFG++	**67.15**	**1.085**	0.351	**0.296**

Compared with the other guidance methods on SD3.5.

Best or tied on every metric, for every model.

–5.6
FID on Flux

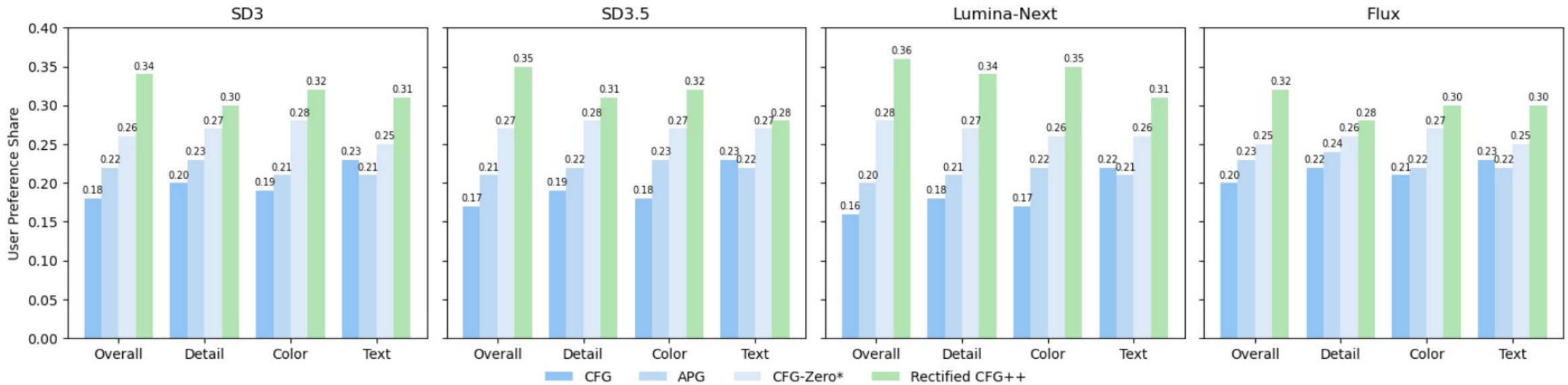
–4.4
FID on Lumina

4
models tested

0
extra training

User study

30 experts × 32 prompts × 4 models × 4 criteria = 15,360 responses



Four-way forced choice: CFG (dark blue), APG (medium blue), CFG-Zero* (light blue), Rectified-CFG++ (green). Green is tallest in every comparison.

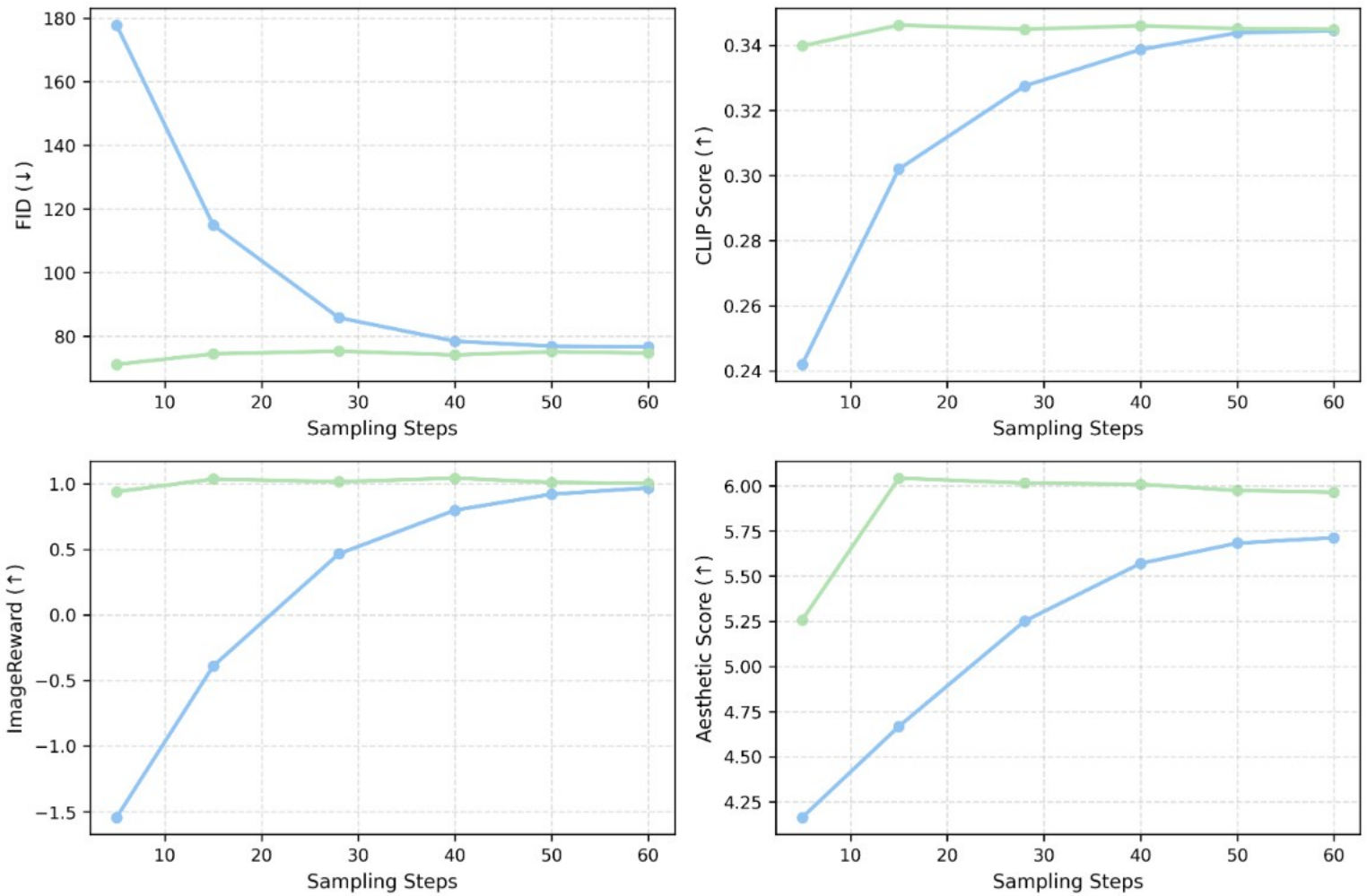
Overall. 34 to 36% of votes on every model. Chance is 25%. $p < 0.001$ against APG.

Text legibility. The largest gain, 28 to 32% against 22 to 25%.

Agreement. Fleiss' $\kappa = 0.61$, with image order randomized.

Sweeping the number of sampling steps

Better quality with fewer steps



- **5 steps.** FID 71 against 178 for CFG, 2.5× better. ImageReward 1.0 against −1.5.
- **15 of our steps match 28 of CFG's**, about 2× faster.

Component ablation on SD3.5

Configuration	FID ↓	CLIP ↑	HPSv2 ↑
Unconditional only	91.12	0.144	0.187
Without predictor	73.70	0.341	0.297
Without corrector	74.65	0.341	0.298
Full	**72.97**	**0.345**	**0.300**

Both the predictor and the corrector are needed.

x-axis: sampling steps. Blue is CFG, green is ours. Ours converges faster on every metric.

Takeaways

1. **On-manifold guidance for rectified flows.**
   Predictor-corrector, adaptive decay, bounded manifold distance.

2. **4 to 16× tighter distributional bound than CFG.**
   Integrated guidance $\lambda/(\gamma+1)$ against $\omega-1$.

3. **A drop-in for Flux, SD3, SD3.5 and Lumina.**
   No retraining; 15 steps match CFG's 28.

43.5%

expert preference, 15K responses
$p < 0.001$

Next, video models, distilled flow models and HDR generation inside LumaFlux.

Summary of contributions

1. HDR-Q

The first multimodal LLM for HDR video quality, trained with HAPO

Saini et al., CVPR 2026

0.921

SROCC, +8.2% over prior best

2. Rectified-CFG++

On-manifold guidance for rectified flows, provably bounded

Saini et al., NeurIPS 2025

43.5%

expert preference share

3. LumaFlux

Frozen FLUX.1-dev, monotone RQS tone field, 8 flow steps

Saini et al., arXiv:2604.02787

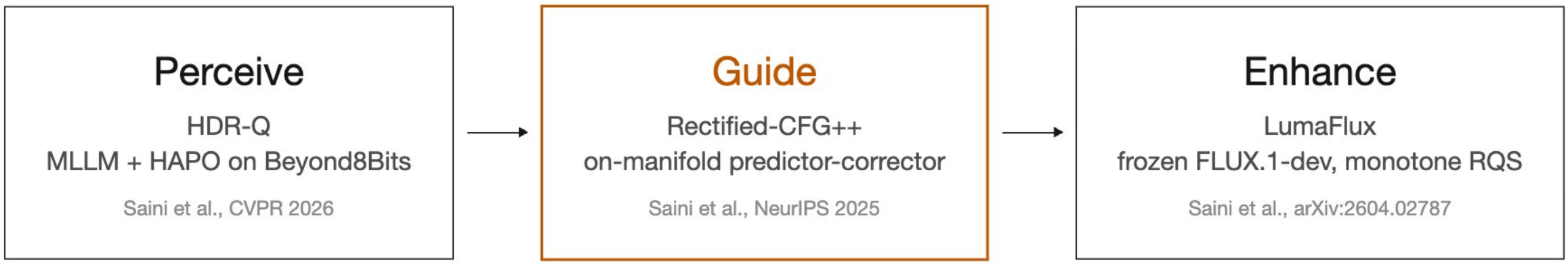
24.23 dB

PU21-PSNR, 56.54 ΔE_{ITP}, FR-HIDROVQA 0.631 (synthetic track)

Built on Beyond8Bits (44K HDR videos, 1.5M+ ratings; CVPR 2026), LGDM (zero-shot IQA, ICML 2025), HIDRO-VQA (WACV 2024) and BrightRate (WACV 2026).

Connecting the contributions

Understanding perception enables guiding generation, which enhances perception.



Generative priors are the common thread, in assessment and in enhancement.

Publications

Perceptual quality

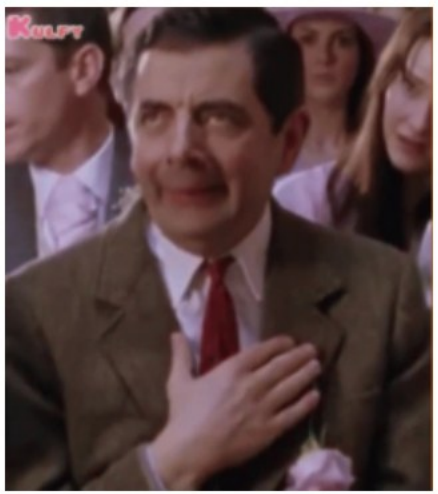
- BrightRate-LM. MVA, WACV 2026 Special Issue, in submission
- **Seeing Beyond 8bits** (HDR-Q). **CVPR 2026**
- BrightRate. WACV 2026, Oral, Best Student Paper Award
- GameScope. ICIP 2026
- LGDM. ICML 2025
- CHUG. ICIP 2025
- HIDRO-VQA. WACV 2024

Generation and enhancement

- CachedSearch. Preprint 2026
- LumaGuide. Preprint 2026
- **LumaFlux**: Structure-Aware SDR-to-HDR Conversion with Monotone Luminance Flows. **arXiv:2604.02787**

  Paper · Code · Project page · Weights on Hugging Face
- **Rectified CFG++** for Flow Based Models. **NeurIPS 2025**

Newest first. Bold marks the three dissertation chapters.



Questions?

Thank you

Perceptual quality assessment and enhancement of visual media using generative priors

Shreshth Saini, PhD candidate

shreshth@utexas.edu

shreshthsaini.github.io

Slides at shreshthsaini.github.io/slides/defense.html

Laboratory for Image and Video Engineering (LIVE), The University of Texas at Austin
Advised by Prof. Alan C. Bovik



TEXAS The University of Texas at Austin

The University of Texas at Austin LIVE Laboratory for Image & Video Engineering

YouTube