Qualcomm

ICCV OCT 19-23, 2025 HONOLULU HAWAII

# Supercharged One-Step Text-to-Image Diffusion Models with Negative Prompts


Viet Nguyen*

Anh Nguyen*

Trung Dao

Khoi Nguyen

Cuong Pham

Toan Tran

Anh Tran

* Equal Contribution

# Community Need: Negative Prompts for Controllability and Quality

Vintage photo of a faceless entity with a blue orb head, wearing a Hylian robe and holding absinthe in a smoky setting.


A deep-fried, crispy image of a kitten holding a sign that says "i want buzz".

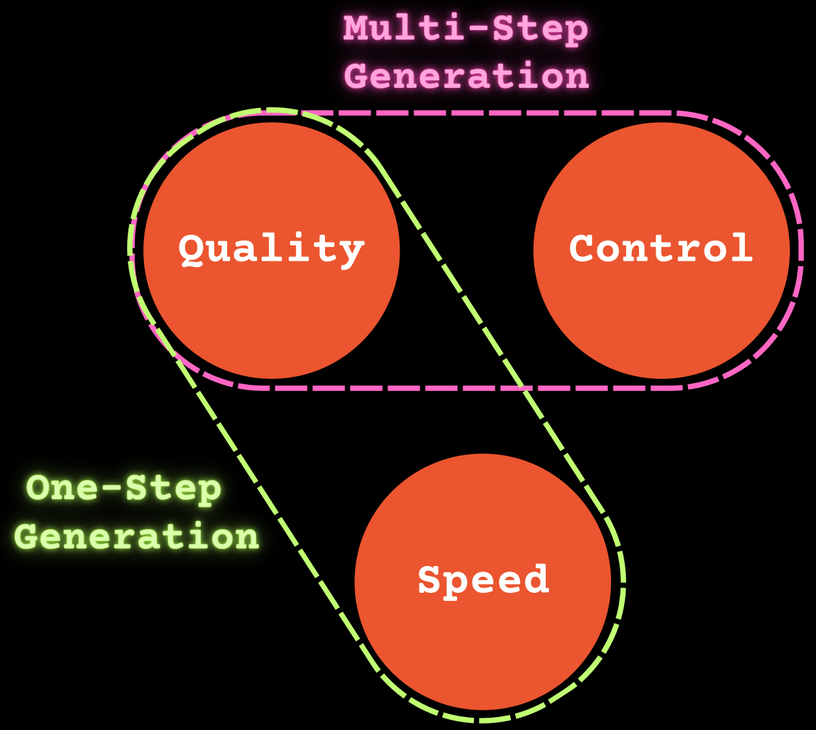
8k macro illustration of a magical forest creature in a wicked hat with a magic wand, in a mossy landscape at night.

Minimalistic 8k illustration of a majestic rabbit-hare with dragon wings and a scorpion tail.

```
bad quality, worst quality, low quality, jpeg artifacts, grainy, blurry, ugly, tiling, draft, signature, watermark, text, logo, author name, patreon, bad anatomy, poorly drawn, distorted, mutated, twisted, extra limbs, body out of frame, out of frame, cut off, poorly drawn face, closed eyes, poorly drawn hands, bad hands, extra digits, missing fingers, poorly drawn feet, censored, child, loli, underage, nsfw, nudity, cleavage
```

On platforms like CivitAI, the highest-quality images are almost universally created using extensive negative prompts to refine **controllability** and **quality**.

# Speed–Quality–Control in One-Step: Why CFG Does Not Transfer




**Visual Generative Trilemma:**
Speed, Quality, Control. Pick Two.

A photo of livestock in a farm

w/o negative guidance

w/ negative guidance

cow

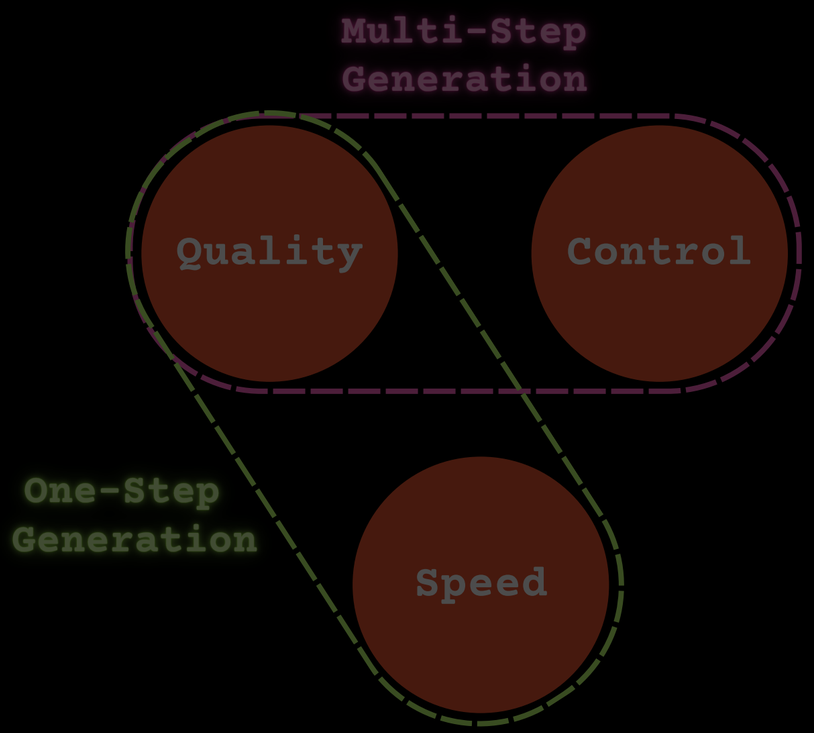
Opps! 🤧
CFG fails 🤦‍♂️

Blending artifacts ! 🤦‍♂️

Applying multi-step **CFG** strategy to a single-step process induces **blending artifacts**.

This flaw shows that a **fundamentally different** approach is needed.

# Our Method: Recovering Control in One-Step Without Sacrificing Speed

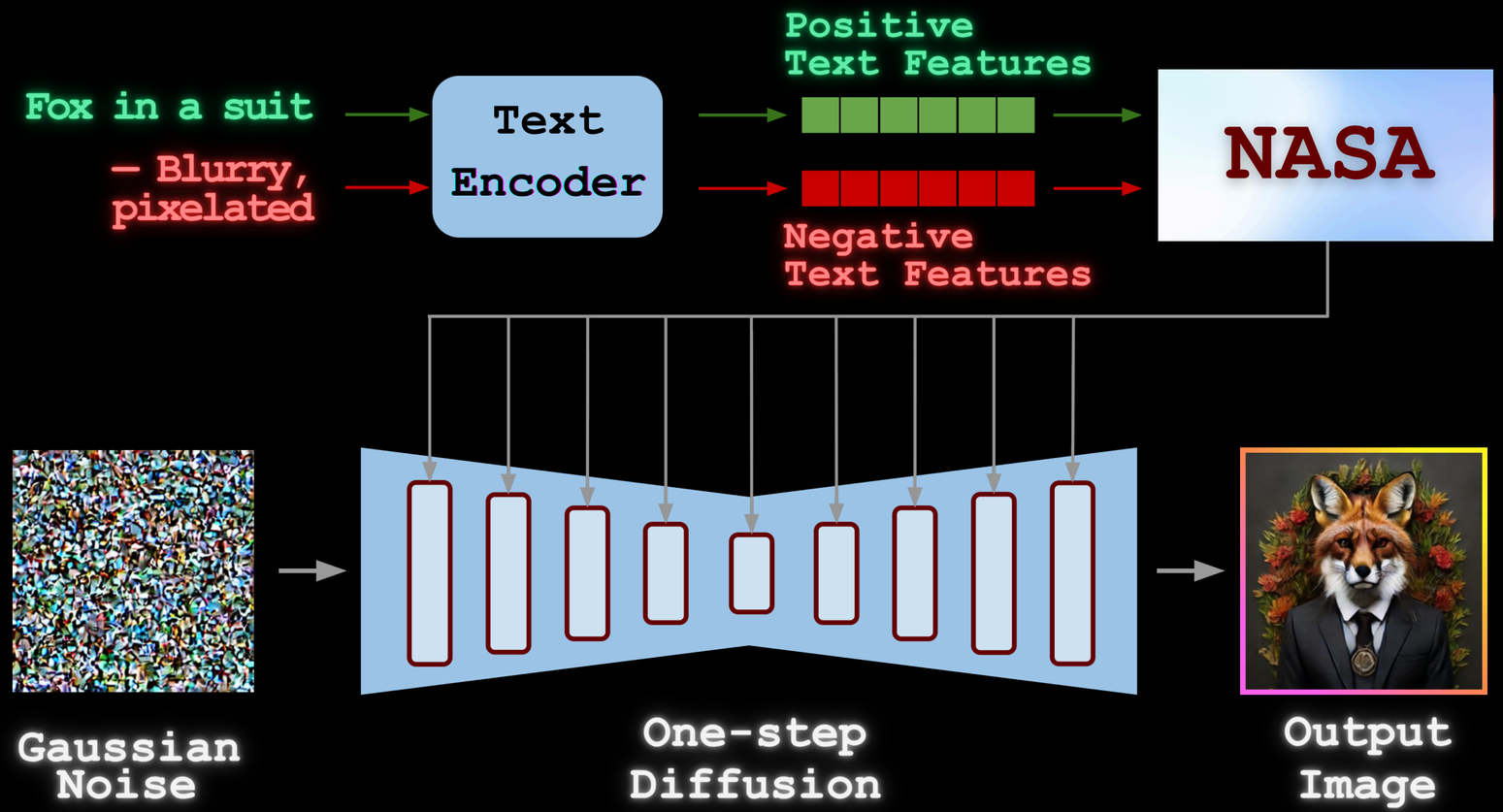


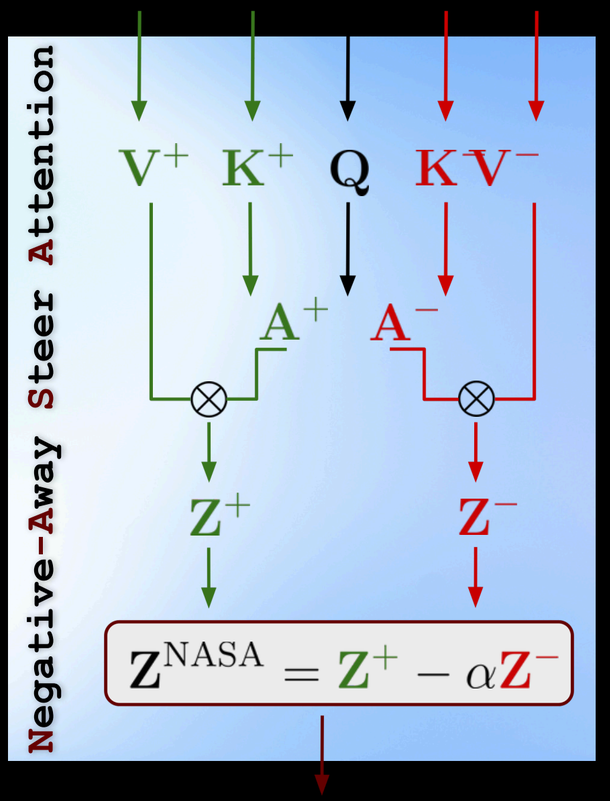

**Visual Generative Trilemma:**
Speed, Quality, Control. Pick Two.


With **NASA**, **speed**, **quality**, and **control** are achievable together —**without compromise**.

One-step models are **fast**, but **lack control**. Multi-step models offer **control** but are **slow**. We want the **best of both worlds**.

# NASA: Guidance in the Representation Space




Larger α → Stronger Suppression

A **Simple** yet **Powerful Subtraction** in Attention Outputs

We intervene at the semantic level, BEFORE pixels are ever generated.

# Setting 1 (NASA-I): Instant Control, Zero Training



- ✓ **Training-Free:** Works with any pre-trained one-step model.
- ✓ **Near-Perfect Control:** 97-100% success rate in removing unwanted attributes.
- ✓ **Ultra-Efficient:** 1.89% FLOPs overhead vs. 2x for classic CFG (dual pass).
- ✓ **Generalizable:** Extends to modern backbone like FLUX and text-to-video.

| Method | VBench-Long benchmark | |
|---|---|---|
| | Aesthetic Quality ↑ | Imaging Quality ↑ |
| None | 61.98 | 67.12 |
| **NASA–I** | 63.33 | 67.36 |

(A) Applied to the **CausVid** video model, NASA-I **improves both** aesthetic and imaging quality.

| Method | FLUX.1-schnell | | SDXL-LCM | | SDXL-DMD2 | |
|---|---|---|---|---|---|---|
| | 4 steps | 1 step | 4 steps | 1 step | 4 steps | 1 step |
| None | 23% | 44% | 43% | - | 27% | 25% |
| CFG | 30% | 0% | 14% | - | 25% | 0% |
| **NASA–I** | 100% | 99% | 97% | - | 100% | 100% |

(B) NASA-I achieves **97-100% success** in removing unwanted features where **CFG fails** in few- & one-step models.

# Setting 2 (NASA-T): Integrating Guidance into Distillation

| Dataset | NegOpt | | HPSv2 | | | | |
|---|---|---|---|---|---|---|---|
| **Method** | **CLIP$^+$ ↑** | **CLIP$^-$ ↓** | **Anime ↑** | **Photo ↑** | **CA ↑** | **Paintings ↑** | **Average ↑** |
| PixArt-$\alpha$-based backbone | | | | | | | |
| PixArt-$\alpha$ [Teacher] | 0.35 | 0.05 | 29.62 | 29.17 | 28.79 | 28.69 | 29.07 |
| YOSO | 0.36 | 0.08 | 28.75 | 28.06 | 28.52 | 28.57 | 28.48 |
| + **NASA–I** | 0.36 | 0.06 | 28.74 | 28.05 | 28.56 | 28.60 | 28.49 (+0.01) |
| DMD | 0.35 | 0.08 | 29.31 | 28.67 | 28.46 | 28.41 | 28.71 |
| + CFG = 1.5 | 0.34 | 0.09 | 30.02 | 27.07 | 28.36 | 28.07 | 28.38 (-0.33) |
| + CFG = 2.5 | 0.31 | 0.13 | 26.74 | 23.86 | 25.13 | 24.66 | 25.10 (-3.61) |
| + **NASA–I** | 0.35 | 0.05 | 29.33 | 28.71 | 28.49 | 28.53 | 28.77 (+0.06) |
| SBv2* | 0.36 | 0.09 | 32.19 | 29.09 | 30.39 | 29.69 | 30.34 |
| + **NASA–I** | 0.36 | 0.06 | 32.60 | 29.58 | 31.09 | 30.65 | 30.98 (+0.64) |
| + **NASA–T** | 0.35 | 0.08 | 32.33 | 29.26 | 30.75 | 30.10 | 30.61 (+0.27) |
| + **NASA–T** + CFG = 1.5 | 0.34 | 0.10 | 29.47 | 26.50 | 28.22 | 27.68 | 27.97 (-2.37) |
| + **NASA–T** + **NASA–I** | 0.35 | 0.05 | <u>32.65</u> | <u>29.65</u> | **31.45** | **31.06** | **31.21** (+0.87) |

(C) On the **NegOpt** benchmark (30k prompt pairs).

**CLIP+/-** measures **positive**/**negative** prompt alignment.

✓ **Deep Integration:** The student model is explicitly trained to handle negative prompts.

✓ **Improved Representation:** The student learns to disentangle positive and negative concepts.

✓ **SOTA Results:** This approach yields the highest performance on human preference benchmarks.

# Conclusion: Toward Speed, Quality, and Control—Together

1. **First** 🚀: Introduced NASA - the first method support negative guidance  for one-step diffusion models.
    - In inference, enhance image **quality** and **controllability** .
    - In distillation, improve model performance → **Sets a new SOTA** 🚀

2. **Highly Efficient** 🚀: The training-free variant adds only ~1.9% FLOPs overhead.

3. **Generalizable** 🚀: Works across **modern backbones** (e.g., FLUX) and also extends to **video generation**.