GASS: Geometry-Aware Spherical Sampling for Disentangled Diversity Enhancement in Text-to-Image Generation

Ye Zhu [1 2], Kaleb S. Newman [2], Johannes F. Lutzeyer [1],
Adriana Romero-Soriano [3 4 5 6], Michal Drozdzal [3], Olga Russakovsky [2]

[1] LIX, CNRS, Ecole Polytechnique, IP Paris, Fance
[2] Department of Computer Science, Princton University, USA
[3] FAIR at Meta – Montreal, Canada [4] McGill University, Canada
[5] Mila, Quebec AI Institute, Canada [6] CIFAR AI Chair, Canada

ICML International Conference On Machine Learning

Challenge: Limited Sample Diversity in Text-to-Image Generation

The lack of diversity not only restricts user choice, but also risks amplifying societal biases.

"A photo of papillon."

"A red colored car."

Images generated using SD3-M with classifier-free guidance (CFG).

Motivation: Identify Different Sources of Image Variations

Existing major paradigm:

Increase the entropy defined by the batch of generated samples.

Motivation:

The “diversity” can be further attributed to more fine-grained components, sometimes unspecified by the given text prompt.

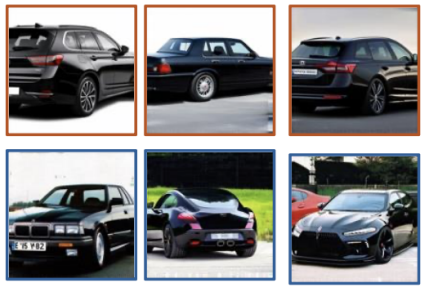
Our approach:

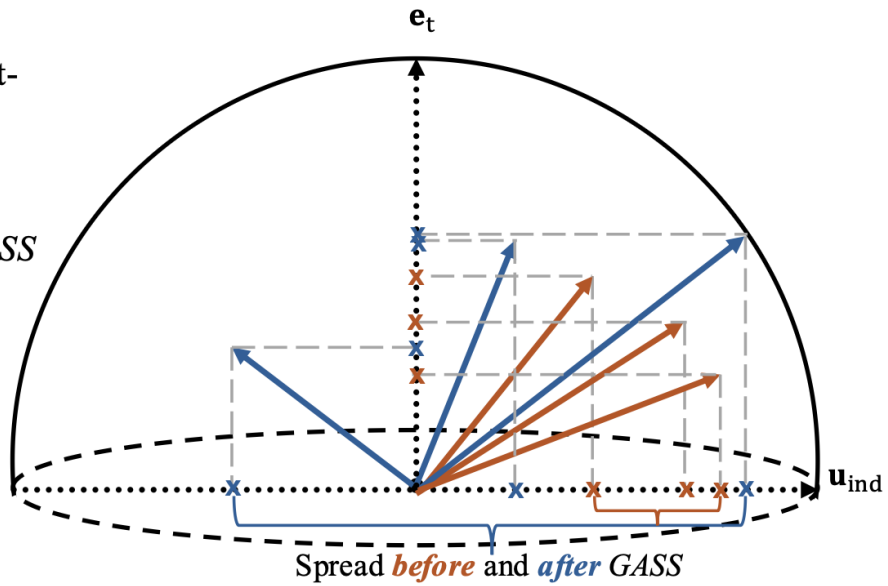
Define and measure the image diversity from different sources of variations, specifically through “prompt-dependent” and “prompt-independent” angles.

Geometric Analysis: Decomposition and Measure in CLIP Spherical Basis

Disentangle the diversity in two orthogonal directions: the text embedding (prompt-dependent) and an identified orthogonal dominant residual direction (prompt-independent).

$\mathbf{e}_t$: text embedding
$\mathbf{u}_{ind}$: our identified unit vector orthogonal to $\mathbf{e}_t$ captures prompt-independent image diversity
x: projections on unit vectors
→: original image embeddings
→: image embeddings after *GASS*





Algorithm 1 Dominant Residual Basis Identification

Input: Text embedding $\mathbf{e}_t$, image batch embeddings $\mathcal{P} = \{\mathbf{e}_i\}_{i=1}^{B}$, number of candidate directions N
Generate N direction vectors $\{\mathbf{r}_k\}_{k=1}^{N}$ orthogonal to $\mathbf{e}_t$ via Gram-Schmidt (Leon et al., 2013).
for $k = 1$ to N **do**
  $E_k \leftarrow \frac{1}{B}\sum_{i=1}^{B} |\mathbf{e}_i^\top \mathbf{r}_k|$
end for
$k^* \leftarrow \arg\max_k E_k$
return $\mathbf{u}_{ind} \leftarrow \mathbf{r}_{k^*}$

Img Source	ClipScore	$\mathcal{D}_{dep}$	$\mathcal{D}_{ind}$	SPP
SD2.1	**0.303**±0.03	0.071±0.02	0.075±0.02	0.146±0.04
SD3-M	0.302±0.02	0.060±0.03	0.065±0.03	0.126±0.05
Real	0.293±0.02	**0.110**±0.03	**0.110**±0.02	**0.220**±0.05

GASS: Geometry-Aware Spherical Sampling for Increased Spread

Step 1: Latent dynamic spherical guidance

- Expansion shit δ_i^k to increase the spread

$$\delta_i^k \sim \mathcal{U}[-r_k, r_k]$$

- Renormalization after GASS perturbation

$$\tilde{e}_i \leftarrow \frac{\tilde{e}_i}{||\tilde{e}_i||_2}$$

Step 2: Optimization based on CLIP gradient

Algorithm 2 Optimization based on CLIP Gradient

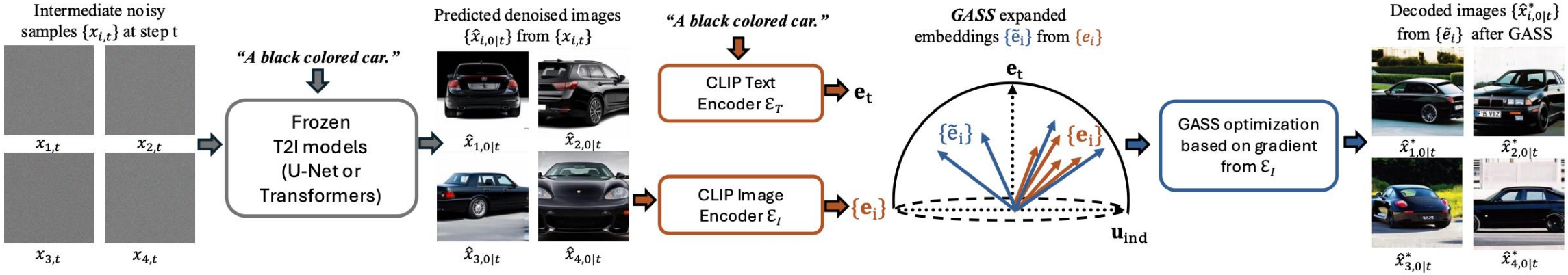
Input: Current batch estimates $\{\hat{x}_{i,0|t}\}_{i=1}^B$, target embeddings $\tilde{\mathcal{P}} = \{\tilde{\mathbf{e}}_i\}_{i=1}^B$, CLIP encoder $\mathcal{E}_I$, step size η

Encode batch estimates: $\{\mathbf{e}_i\}_{i=1}^B = \mathcal{E}_I(\{\hat{x}_{i,0|t}\}_{i=1}^B)$

$\mathcal{L}_{\text{spp}} = \sum_{i=1}^B (1 - \mathbf{e}_i^\top \tilde{\mathbf{e}}_i)$ {spherical spread loss}

$\hat{x}^*_{i,0|t} \leftarrow \hat{x}_{i,0|t} - \eta \cdot \nabla_{\hat{x}_{i,0|t}} \mathcal{L}_{\text{spp}}$ for $i = 1 \dots B$

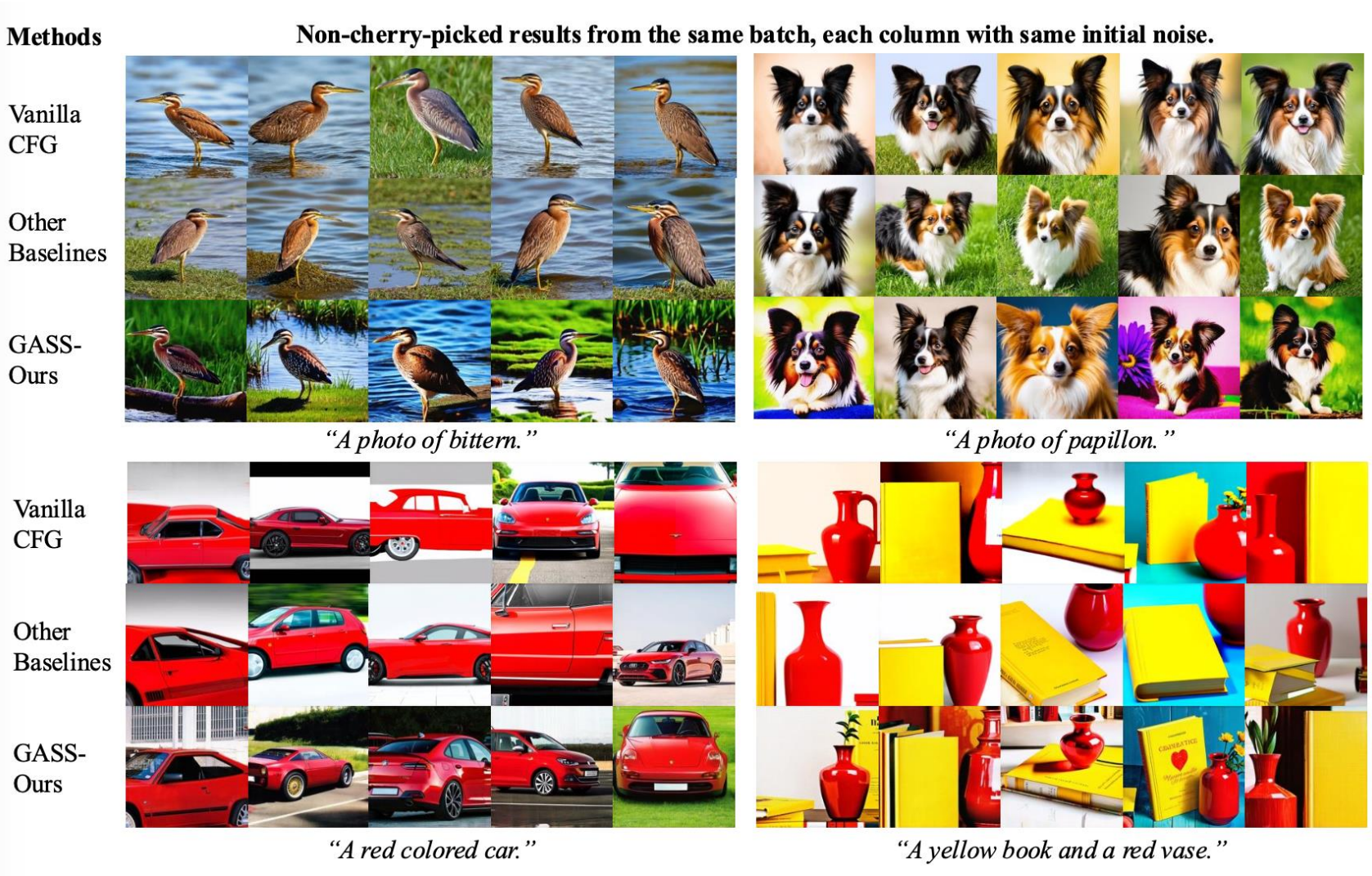
return $\{\hat{x}^*_{i,0|t}\}_{i=1}^B$



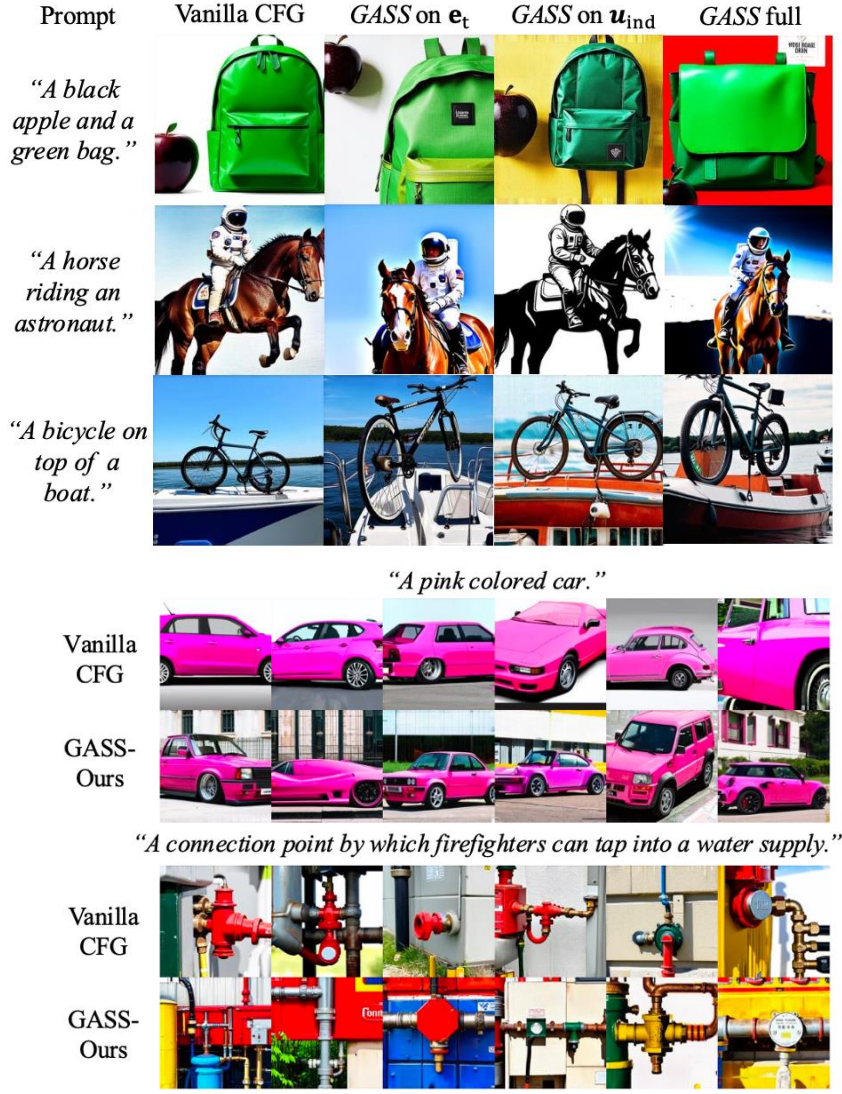
GASS alters the predicted clean image through the geometric expansion, and thus guide the iterative sampling process with frozen generative backbones.

Enhanced Diversity with Disentangled Sources

Good diversity enhancement performance, with both richer semantic variation and more detailed and diverse backgrounds.



Disentangled sources and diversity under complex prompts.



A training-free, plug-and-play method that enhances diversity in T2I through geometry-aware sampling.

Paper

Code