Demystifying Classifier-Free Guidance for Auto-Regressive Image Generation

Zhiling Zhou[1] Jiachun Pan[2] Fengzhuo Zhang[1] Dirk Bergemann[1] Zhuoran Yang[1]

[1]Yale University [2]National University of Singapore

Summary of Results

- Classifier-Free Guidance (CFG) is effective because it subtracts a shared repetition bias. The guidance direction $\text{logit}^{\text{cond}} - \text{logit}^{\text{uncond}}$ carries texture and prompt-specific details.
- The bias is not a pretraining failure. A shallow-transformer analysis shows that sparse texture regions naturally make repeated-token dynamics dominate top-ranked conditional logits.
- The unconditional pass is partly redundant. AttnReuse masks text-token attention and reuses conditional attention weights, reducing attention FLOPs from 100% to about 75% for AR models.

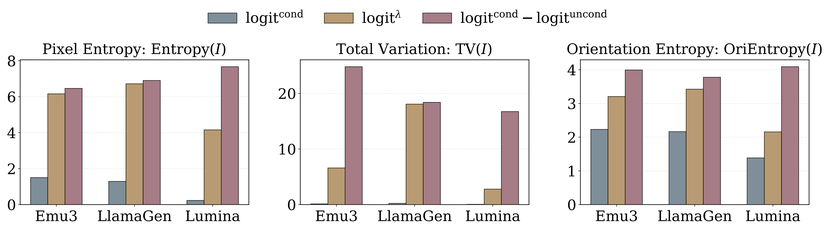
Main Question: What Does CFG Change?

CFG: $\text{logit}^{\lambda} = \text{logit}^{\text{cond}} + \lambda\left(\text{logit}^{\text{cond}} - \text{logit}^{\text{uncond}}\right)$

- AR image generators usually decode with CFG even though text generation often works without it.
- We inspect the top-ranked semantics of $\text{logit}^{\text{cond}}$, CFG logits, and $\text{logit}^{\text{diff}} = \text{logit}^{\text{cond}} - \text{logit}^{\text{uncond}}$ in the generated images below.
- The conditional logits alone often decode to nearly pure-color or low-texture images; the difference logits recover localized texture.

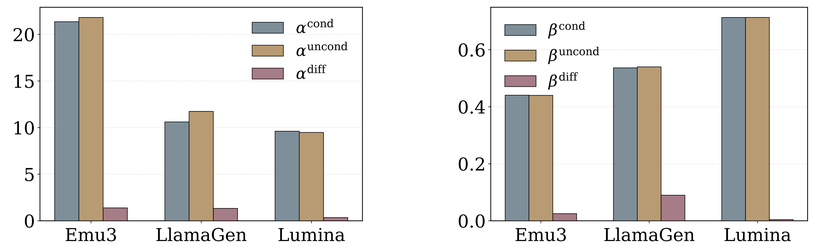
Takeaway: CFG is not just sharpening. It exposes a texture-rich direction hidden beneath a dominant repeated-token component.

Texture Lives in the Guidance Direction



- Texture-complexity metrics compare images decoded from $\text{logit}^{\text{cond}}$, CFG logits, and $\text{logit}^{\text{diff}}$; $\text{logit}^{\text{diff}}$ produces the most texture-rich decoded images.
- $\text{logit}^{\text{cond}}$ has the lowest texture complexity, explaining why greedy decoding from it alone can collapse to simple color fields.

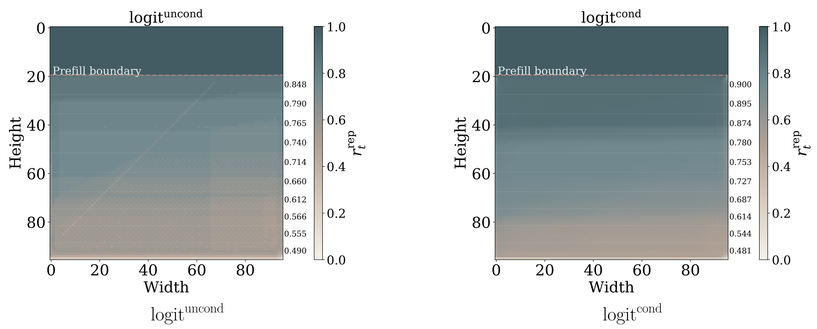
Conditional and Unconditional Logits Share Semantics



- Similarity under image prefill: α^{diff} is the normalized top-k logit ℓ_1 gap, and β^{diff} is the normalized pixel ℓ_2 gap after greedy top-token decoding; $\alpha^{\text{cond}}, \alpha^{\text{uncond}}, \beta^{\text{cond}}, \beta^{\text{uncond}}$ are scale baselines.
- In both logit space and pixel space, $\text{logit}^{\text{cond}}$ and $\text{logit}^{\text{uncond}}$ are close.
- The highest-ranked coordinates of $\text{logit}^{\text{cond}}$ are largely dominated by the same semantics as $\text{logit}^{\text{uncond}}$.

Observed decomposition: $\text{logit}^{\text{cond}} \approx \text{logit}^{\text{uncond}} + \text{texture/prompt residual}$

Both Branches Tend to Repeat Previous Tokens



- Repeated-prefix test: set $\hat{x}_{1:L} = (s, \ldots, s)$, continue generation, and measure position-wise repetition $r_t^{\text{rep}} = \mathbf{1}\{x_t = s\}$.

Model	$\text{logit}^{\text{uncond}}$	$\text{logit}^{\text{cond}}$
Emu3	0.5168	0.5479
LlamaGen	0.7457	0.7517
Lumina-mGPT	0.6804	0.7333

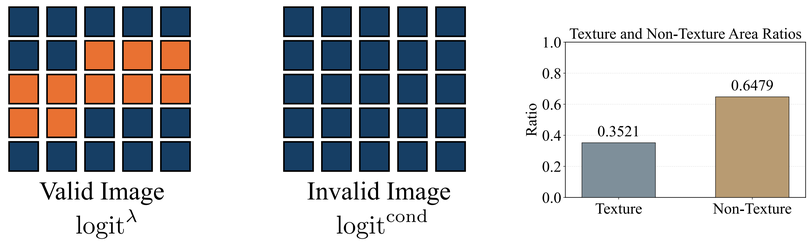
- The table reports the average repetition ratio $r^{\text{rep}} = (T_{\text{vis}} - L)^{-1} \sum_{t=L+1}^{T_{\text{vis}}} r_t^{\text{rep}}$ across model backbones.
- After injecting a repeated prefix token, both branches continue repeating that token with high probability.
- CFG improves generation by subtracting this shared repetition tendency and amplifying the residual texture signal.

Mechanistic Picture

unconditional logits $\approx$ repeated visual-token dynamics

guidance direction $\approx$ sparse texture and prompt detail

Why Does Pretraining Learn This Bias?



- The toy image model and ImageNet statistics show that texture regions are sparse relative to repeated non-texture/background tokens.
- We model images as mostly non-texture tokens plus sparse feature/texture blocks.
- Texture sparsity means the cross-entropy objective sees many more non-texture events than texture events.
- Training dynamics of shallow transformers predict that conditional logits first learn the dominant repeated-token behavior.

Informal theorem. Under texture sparsity and orthonormal embeddings, the pretrained shallow transformer satisfies. Here $l_{k,j}$ is the target token image with feature class k and background class j, and $\hat{l}^{\lambda}_{k,j}$ is the image greedily generated with guidance scale λ:

1. **(Minimal Training loss.)** Small population loss: $\mathcal{L}_{\text{vis}}(\hat{\theta}) = O(T_{\text{vis}}^{-1}(\log T_{\text{vis}})^2)$.
2. **(Alignment of $\text{logit}^{\text{cond}}$ and $\text{logit}^{\text{uncond}}$.)** On target-image prefixes, $\text{logit}^{\text{cond}}$ and $\text{logit}^{\text{uncond}}$ have the same top prediction at every visual position.
3. **(Conditional generation collapses to repeated background tokens.)** Pure conditional greedy generation collapses to background repetition: $\hat{l}^{0}_{k,j} = (K+j, \ldots, K+j) \neq l_{k,j}$.
4. **(CFG recovers valid images.)** For some constant $\lambda^* = \Theta(1)$, CFG exactly recovers the target: $\hat{l}^{\lambda^*}_{k,j} = l_{k,j}$.

AttnReuse: Reusing the Conditional Forward Pass

$$\text{Attn}(q_t, K, V) = \gamma_t \frac{\sum_{l>T_{\text{text}}} \exp(q_t^\top k_l) v_l}{\sum_{l>T_{\text{text}}} \exp(q_t^\top k_l)} + (1-\gamma_t) \frac{\sum_{l\le T_{\text{text}}} \exp(q_t^\top k_l) v_l}{\sum_{l\le T_{\text{text}}} \exp(q_t^\top k_l)}.$$

- Visual-token attention captures the shared conditional/unconditional trajectory.
- Text-token attention contributes the prompt-specific deviation.
- AttnReuse approximates the unconditional attention by masking text-token entries in conditional attention weights and renormalizing over visual keys.

Efficiency With Little Degradation

Model	Setting	Overall ↑	Attn. FLOPs ↓
Emu3	CFG	76.7	100%
Emu3 + AttnReuse	$\tau = 0$	74.7	**75%**
Janus-Pro 7B	CFG	83.3	100%
Janus-Pro + AttnReuse	$\tau = 0$	82.3	**75%**
Lumina-GPT 2.0	CFG	84.3	100%
Lumina + AttnReuse	$\tau = 1$	83.7	**75%**
SD3-Medium	CFG	84.1	100%
SD3 + AttnReuse	diffusion	82.5	**87%**

- DPG-Bench comparison of standard CFG and AttnReuse; FLOPs are normalized to CFG.
- Reduces attention-related computation by about 25% for AR models.
- Preserves fine-grained semantic alignment on DPG-Bench across multiple backbones.

One-Sentence Takeaway

CFG cancels shared repeated-token semantics and amplifies sparse texture; the same shared structure enables efficient attention reuse.