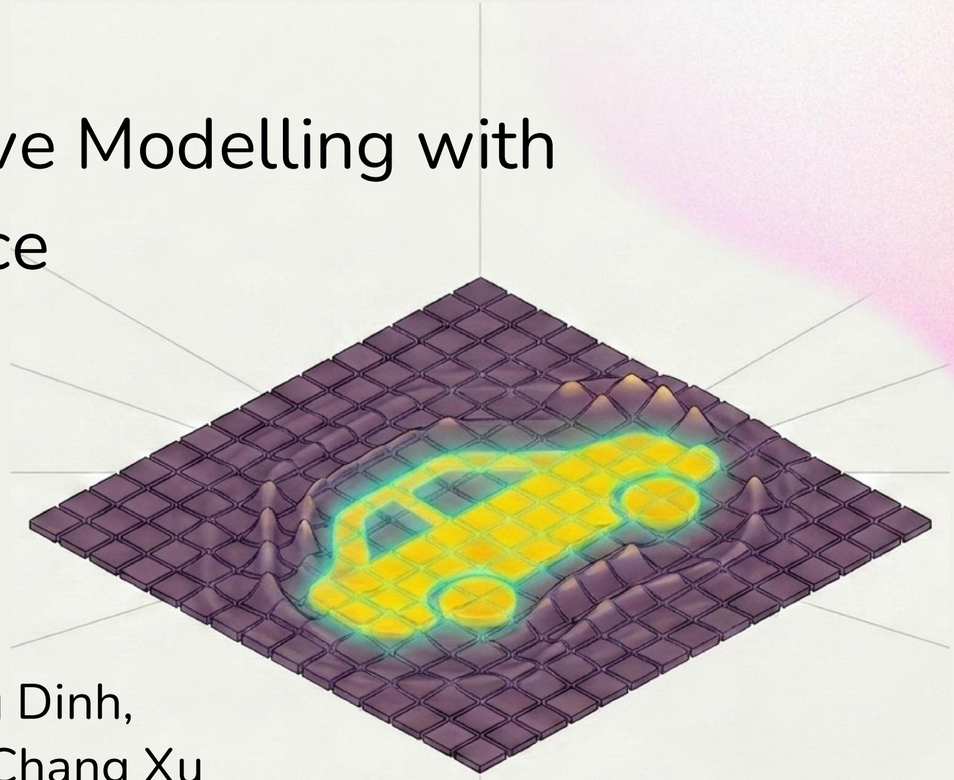
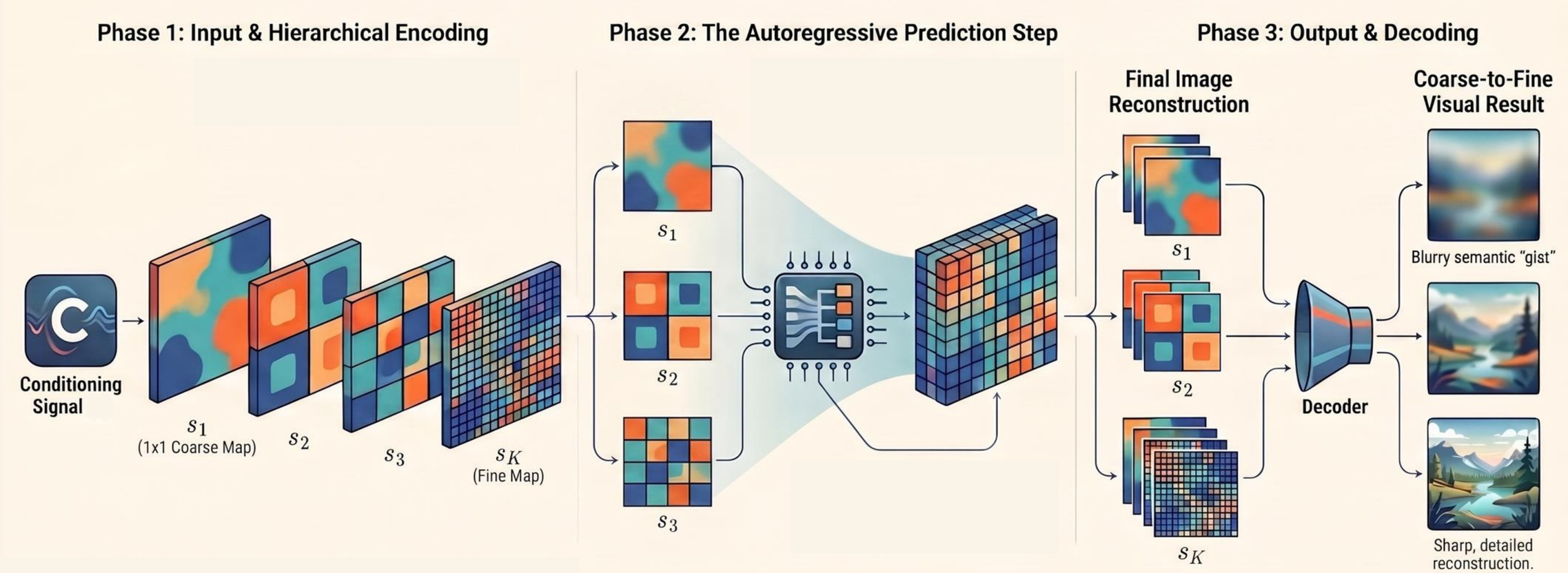
Not All Tokens are Guided Equal

Rethinking Visual Autoregressive Modelling with Information-Grounding Guidance



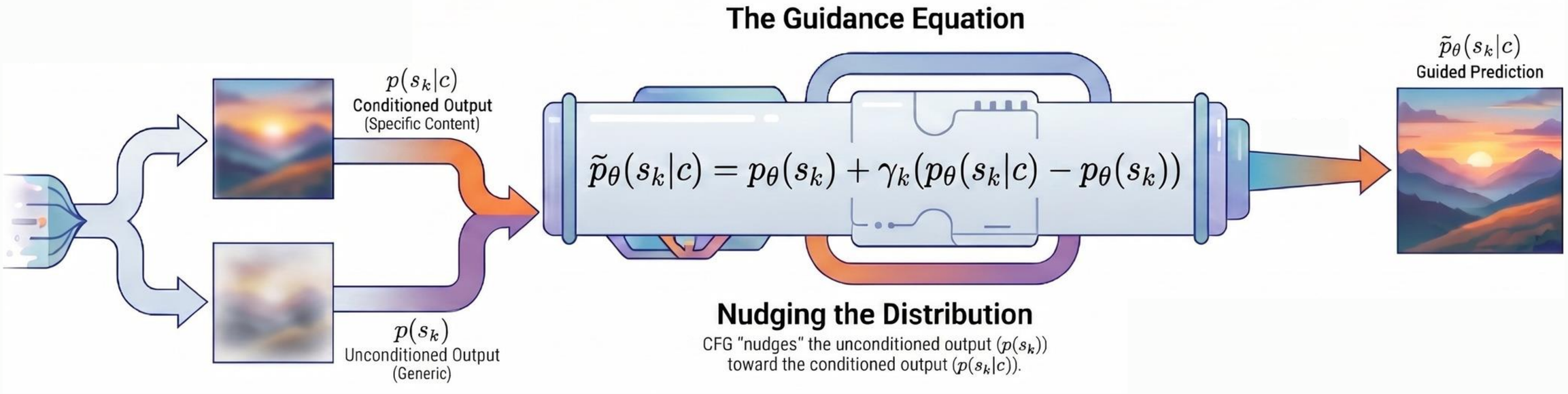
Ky Dan Nguyen, Hoang Lam Tran, Anh-Dung Dinh,
Daochang Liu, Weidong Cai, Xiuying Wang, Chang Xu

Scale-wise Autoregressive (SwAR) modelling



$$p(s|c) = \prod p(s_k|s_{<k}, c)$$

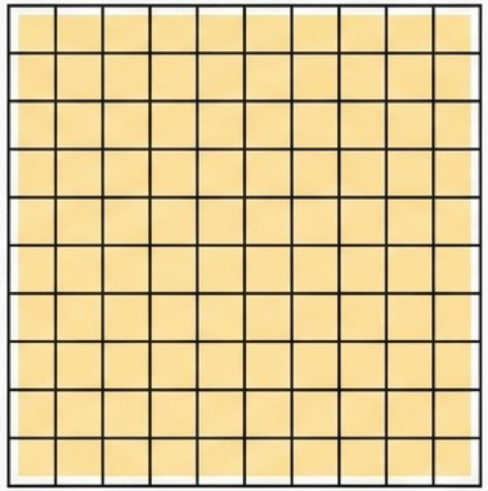
Classifier-free Guidance (CFG) applied to SwAR



What is wrong with CFG in SwAR modelling?

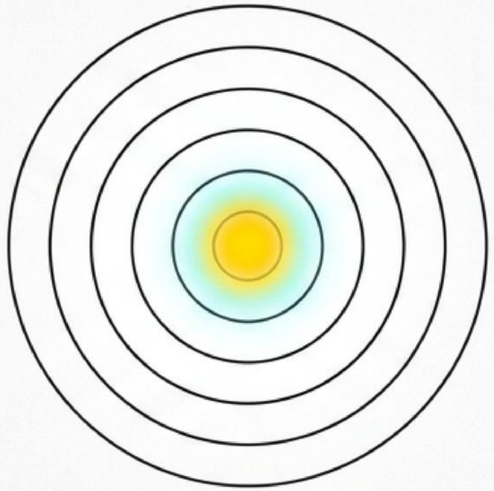
1. Guidance tends to **treat all tokens with equal importance**.
2. Guidance tends to be **misaligned w.r.t. foreground tokens**.

1. Evenness (PEI)



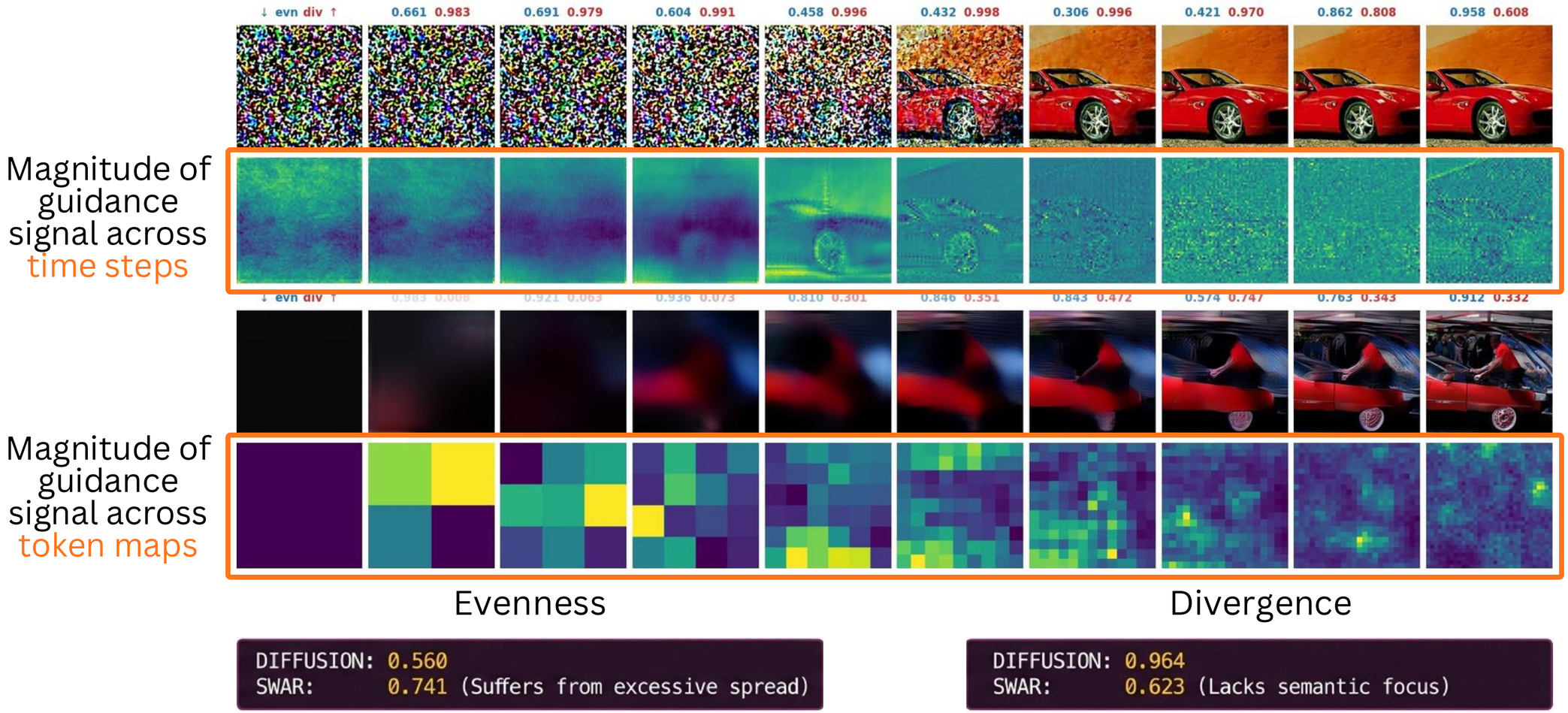
Normalised Shannon entropy. Measures how evenly guidance is distributed across all tokens. Lower is better (indicates concentration).

2. Divergence (JSD)



Jensen-Shannon distance. Evaluates if guidance preferentially targets semantic foreground over the background. Higher is better.

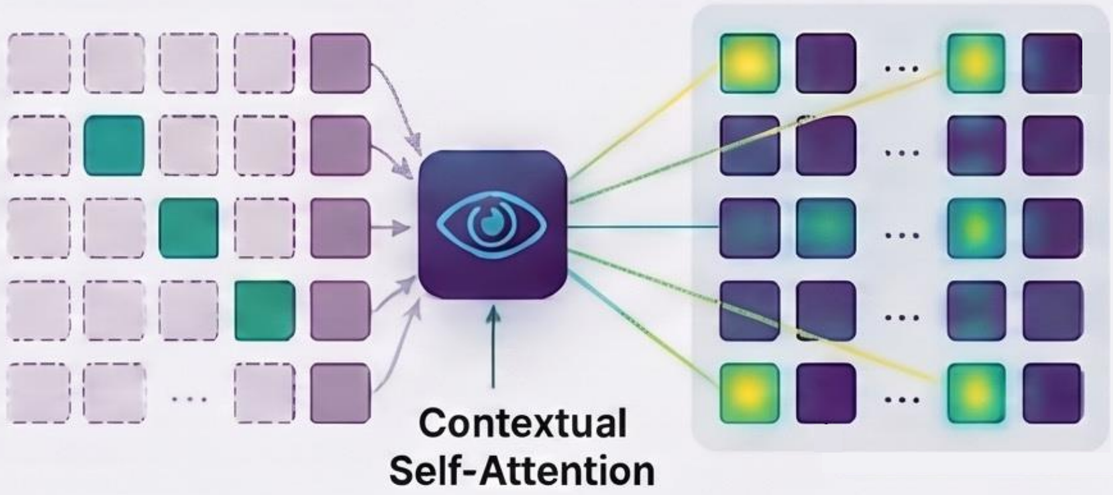
What is wrong with CFG in SwAR modelling?



Information-Grounding Guidance (IGG)

$$\tilde{p}_\theta(s_k|c) = p_\theta(s_k) + f_k(s_k|c) \cdot \big(p_\theta(s_k|c) - p_\theta(s_k)\big)$$

The IGG Mechanism: Selective Weighting

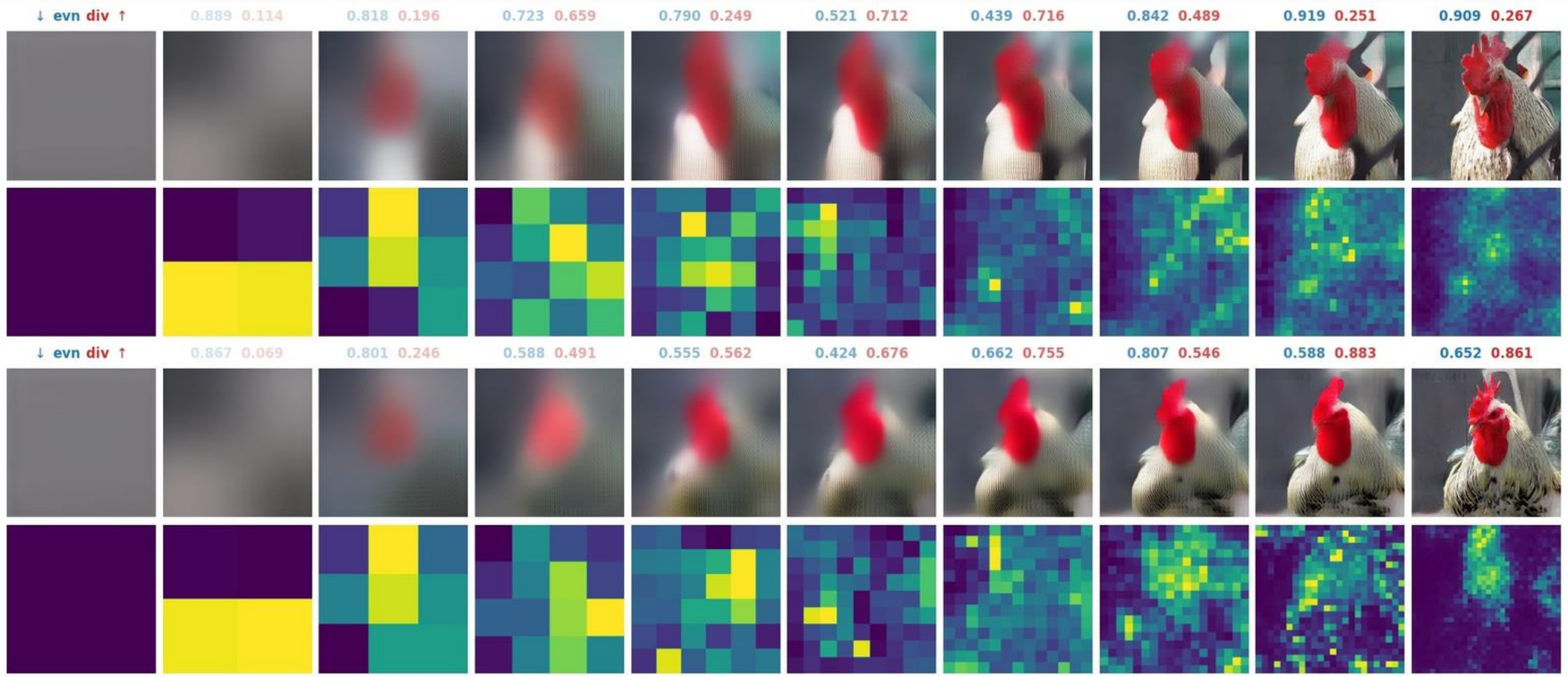


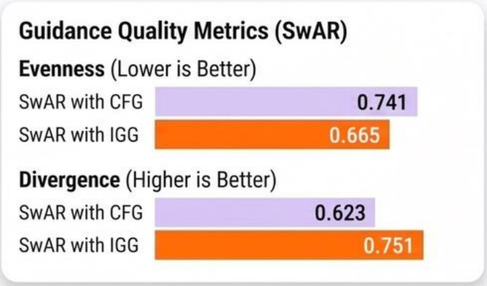
Replaces uniform guidance with a weighted approach. The weight function f_k uses a softmax-based self-attention operation over guidance signals.

CFG becomes a special case of IGG where

$$f_k(s_k|c) = \gamma_k$$

What are the results?





- IGG scales up the foreground signals, while tuning down the background ones.

What are the results?

Class-conditioned (ImageNet 512 w/ VAR-d36)

Guidance Method	FID (Lower Is Better)	IS (Higher is Better)	Rec (Higher is Better)	Pre (Higher is Better)
CFG (Standard)	2.61	293.7	0.56	0.82
IGG (Ours)	**2.56**	**314.3**	**0.57**	**0.82**

Text-to-image (GenEval w/ Switti)

Exaluation Category	Switti + CFG	Switti + IGG
Single-object	0.957	1.000
Two-object	0.753	0.808
Counting	0.556	0.609
Colour attribution	0.398	0.430

Computational cost

Minimal Impact on Inference Time

<0.01 seconds

increase per image for a VAR-d36 model compared to standard CFG

Zero Training Cost Increase

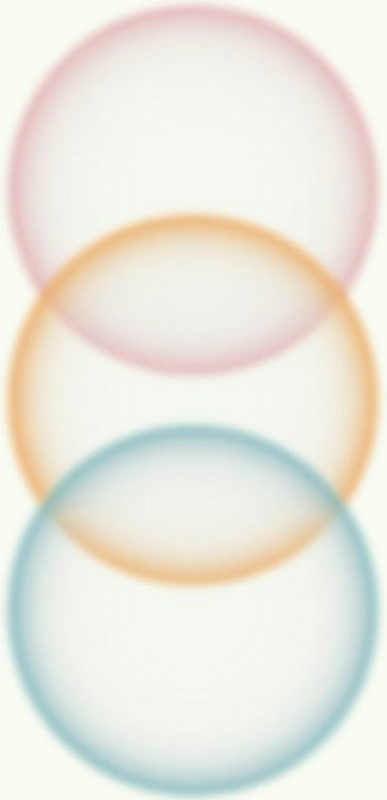
0 GFLOP Increase

CFG

IGG

"An astronaut riding a horse on the moon, oil painting by Van Gogh"

"A small cactus with a happy face in the Sahara desert"



Thank you!

Summary of contributions:

- Analyse the behaviour of CFG in SwAR modelling, revealing that guidance tends to be uniform and misaligned;
- Propose IGG, a novel technique that adaptively concentrates guidance towards semantically important tokens; and
- Validate IGG through extensive experiments on class-conditioned and text-to-image generation.

Graphics (NOT images) were generated using NotebookLM.