OSEG

Nazmus Saqib
Jeju National University

Simple Recap!!!!!

This was my main paper….

SEG has defined the Self-attention map from energy landscape perspective.

Smoothed Energy Guidance: Guiding Diffusion Models with Reduced Energy Curvature of Attention

Susung Hong
Korea University

Abstract

Conditional diffusion models have shown remarkable success in visual content generation, producing high-quality samples across various domains, largely due to classifier-free guidance (CFG). Recent attempts to extend guidance to unconditional models have relied on heuristic techniques, resulting in suboptimal generation quality and unintended effects. In this work, we propose Smoothed Energy Guidance (SEG), a novel training- and condition-free approach that leverages the energy-based perspective of the self-attention mechanism to enhance image generation. By defining the energy of self-attention, we introduce a method to reduce the curvature of the energy landscape of attention and use the output as the unconditional prediction. Practically, we control the curvature of the energy landscape by adjusting the Gaussian kernel parameter while keeping the guidance scale parameter fixed. Additionally, we present a query blurring method that is equivalent to blurring the entire attention weights without incurring quadratic complexity in the number of tokens. In our experiments, SEG achieves a Pareto improvement in both quality and the reduction of side effects. The code is available at https://github.com/SusungHong/SEG-SDXL.

1 Introduction

Diffusion models [12, 45, 46] have emerged as a promising tool for visual content generation, producing high-quality and diverse samples across various domains, including image [38, 40, 42, 8, 13, 30, 2, 24, 9, 29, 34, 33, 4, 41, 5, 20, 22], video [11, 50, 23, 18, 15, 3, 19, 44], and 3D generation [36, 27, 6, 26, 49, 43, 48, 16]. The success of these models can be largely attributed to the use of classifier-free guidance (CFG) [14], which enables sampling from a sharper distribution, resulting in improved sample quality. However, CFG is not applicable to unconditional image generation, where no specific conditions are provided, creating a disparity between the capabilities of text-conditioned sampling and sampling without text. This disparity results in a restriction in application, *e.g.*, synthesizing images with ControlNet[51] without a text prompt (see the last two columns of Fig. 1).

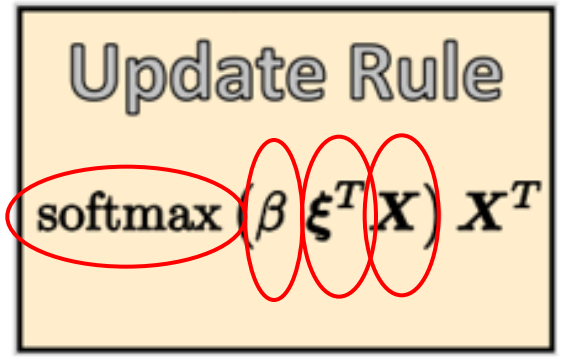
Recent literature [17, 1] has attempted to decouple CFG and image quality by extending guidance to general diffusion models, leveraging their inherent representations [25, 32, 17]. Self-attention guidance (SAG) [17] proposes leveraging the intermediate self-attention map of diffusion models to blur the input pixels and provide guidance, while perturbed attention guidance (PAG) [1] perturbs the attention map itself by replacing it with an identity attention map. Despite these efforts, these methods rely on heuristics to make perturbed predictions, resulting in unintended effects such as smoothed-out details, saturation, color shifts, and significant changes in the image structure when given a large guidance scale. Notably, the mathematical underpinnings of these unconditional guidance approaches are not well elucidated.

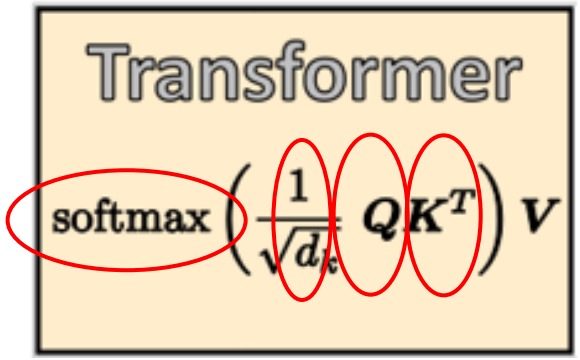
In this work, we approach the objective from an energy-based perspective of the self-attention mechanism, which has been previously explored based on its close connection to the Hopfield energy [39, 31, 7]. Specifically, we start from the definition of the energy of self-attention, where

Simple Recap!!!!!

Energy as a window offers an interesting view of the self-attention mechanism, which is closely associated with Hopfield energy.

Self-attention mechanism is equivalent to the updating rule of Hopfield network.





HOPFIELD NETWORKS IS ALL YOU NEED

Hubert Ramsauer* **Bernhard Schäfl*** **Johannes Lehner*** **Philipp Seidl***
Michael Widrich* **Thomas Adler*** **Lukas Gruber*** **Markus Holzleitner***
Milena Pavlović‡,§ **Geir Kjetil Sandve§** **Victor Greiff‡** **David Kreil†**
Michael Kopp† **Günter Klambauer*** **Johannes Brandstetter*** **Sepp Hochreiter*,†**
*ELLIS Unit Linz, LIT AI Lab, Institute for Machine Learning,
Johannes Kepler University Linz, Austria
†Institute of Advanced Research in Artificial Intelligence (IARAI)
‡Department of Immunology, University of Oslo, Norway
§Department of Informatics, University of Oslo, Norway

ABSTRACT

We introduce a modern Hopfield network with continuous states and a corresponding update rule. The new Hopfield network can store exponentially (with the dimension of the associative space) many patterns, retrieves the pattern with one update, and has exponentially small retrieval errors. It has three types of energy minima (fixed points of the update): (1) global fixed point averaging over all patterns, (2) metastable states averaging over a subset of patterns, and (3) fixed points which store a single pattern. The new update rule is equivalent to the attention mechanism used in transformers. This equivalence enables a characterization of the heads of transformer models. These heads perform in the first layers preferably global averaging and in higher layers partial averaging via metastable states. The new modern Hopfield network can be integrated into deep learning architectures as layers to allow the storage of and access to raw input data, intermediate results, or learned prototypes. These Hopfield layers enable new ways of deep learning, beyond fully-connected, convolutional, or recurrent networks, and provide pooling, memory, association, and attention mechanisms. We demonstrate the broad applicability of the Hopfield layers across various domains. Hopfield layers improved state-of-the-art on three out of four considered multiple instance learning problems as well as on immune repertoire classification with several hundreds of thousands of instances. On the UCI benchmark collections of small classification tasks, where deep learning methods typically struggle, Hopfield layers yielded a new state-of-the-art when compared to different machine learning methods. Finally, Hopfield layers achieved state-of-the-art on two drug design datasets. The implementation is available at: https://github.com/ml-jku/hopfield-layers

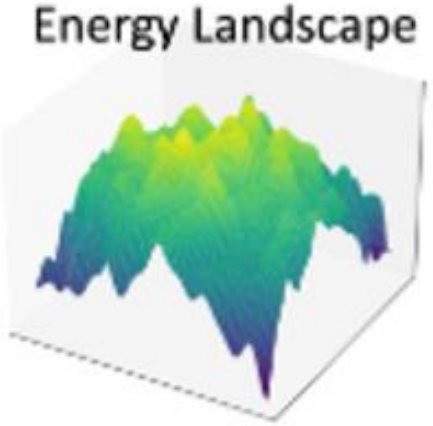
1 INTRODUCTION

The deep learning community has been looking for alternatives to recurrent neural networks (RNNs) for storing information. For example, linear memory networks use a linear autoencoder for sequences as a memory (Carta et al., 2020). Additional memories for RNNs like holographic reduced representations (Danihelka et al., 2016), tensor product representations (Schlag & Schmidhuber, 2018; Schlag et al., 2019) and classical associative memories (extended to fast weight approaches) (Schmidhuber, 1992; Ba et al., 2016a;b; Zhang & Zhou, 2017; Schlag et al., 2021) have been suggested. Most approaches to new memories are based on attention. The neural Turing machine (NTM) is equipped with an external memory and an attention process (Graves et al., 2014). Memory networks (Weston et al., 2014) use an arg max attention by first mapping a query and patterns into a space and then retrieving the pattern with the largest dot product. End to end memory networks (EMN) make this attention scheme differentiable by replacing arg max through a softmax (Sukhbaatar et al., 2015a;b). EMN with dot products became very popular and implement a key-value attention (Daniluk et al., 2017) for self-attention. An enhancement of EMN is the transformer (Vaswani et al., 2017a;b) and its

Simple Recap!!!!!

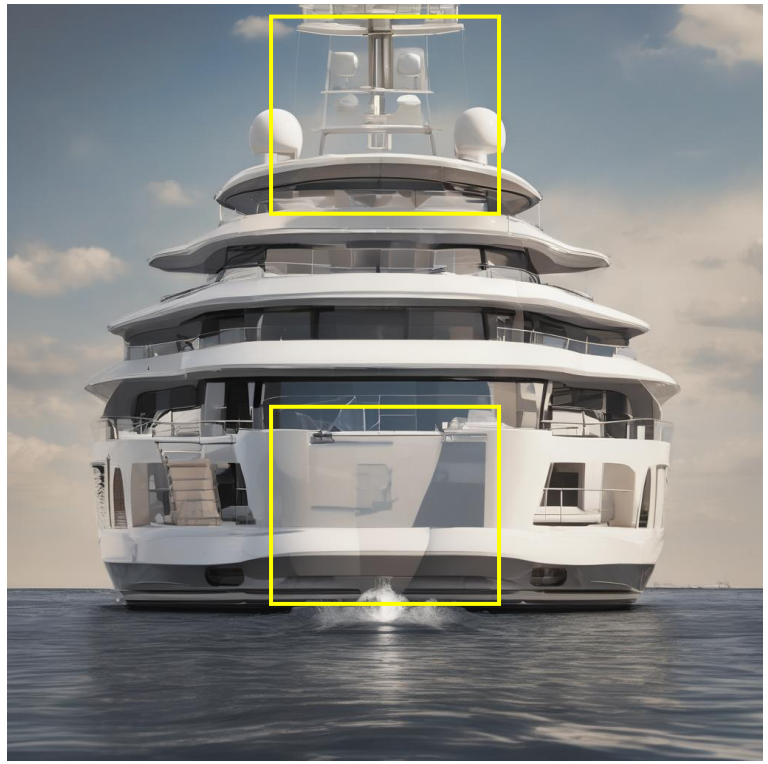
What about the energy function for the self-attention from the Hopfield energy:

$$E(\mathbf{A}) = -\log\left(\sum_{i=1}^{n}\exp\left(\frac{\mathbf{Q}^\top\mathbf{K}}{\sqrt{d}}\right)\right)$$



→ CFG →

"A cruise ship in the sea, front view, high-quality"

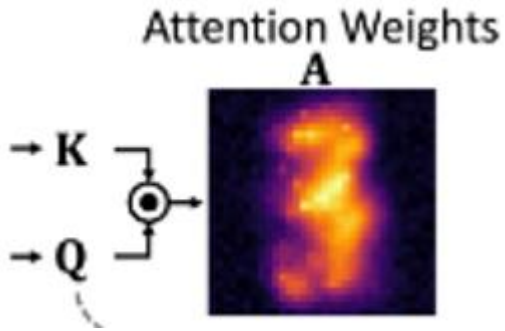


Simple Recap!!!!!

What does SEG interpret the energy landscape to be? SEG focuses on the attention weights……

$$E(\mathbf{A}) = -\log\left(\sum_{i=1}^{n}\exp\left(\frac{\mathbf{Q}^\top\mathbf{K}}{\sqrt{d}}\right)\right)$$

$$A_i = \frac{\exp\left(\frac{\mathbf{Q}^\top\mathbf{K}_i}{\sqrt{d}}\right)}{\sum_{j=1}^{n}\exp\left(\frac{\mathbf{Q}^\top\mathbf{K}_j}{\sqrt{d}}\right)}$$

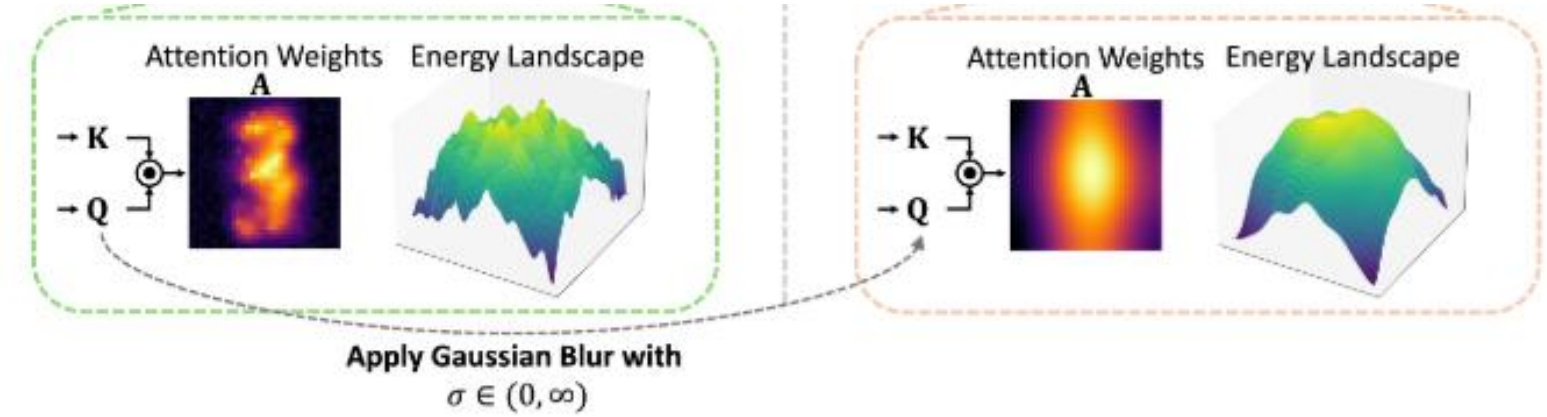


Simple Recap!!!

Theorem 3.1

H Hessian of the Attention weights

$\breve{H}$ Hessian of the blurred Attention weights



Lemma 3.1 & 3.2

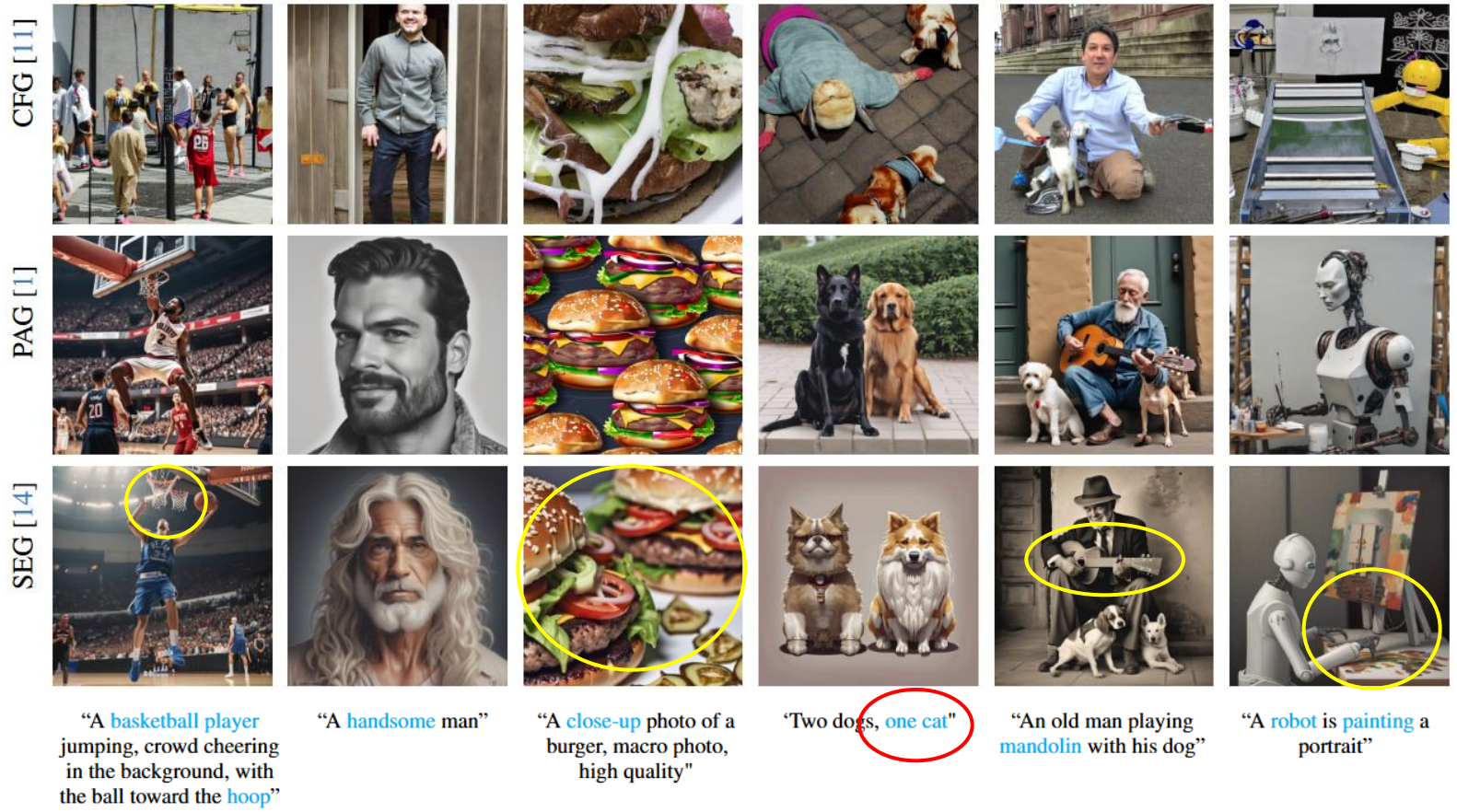
$$|\det(\mathbf{H})| > |\det(\tilde{\mathbf{H}})|$$

Lower variance

the minimization step is performed on a smoother energy landscape with attenuated Gaussian curvature

Simple Recap!!!

What that occurs????



So, what are the challenges still?

- Limited generation quality.
- Prompt alignment.
- Detail suppression.

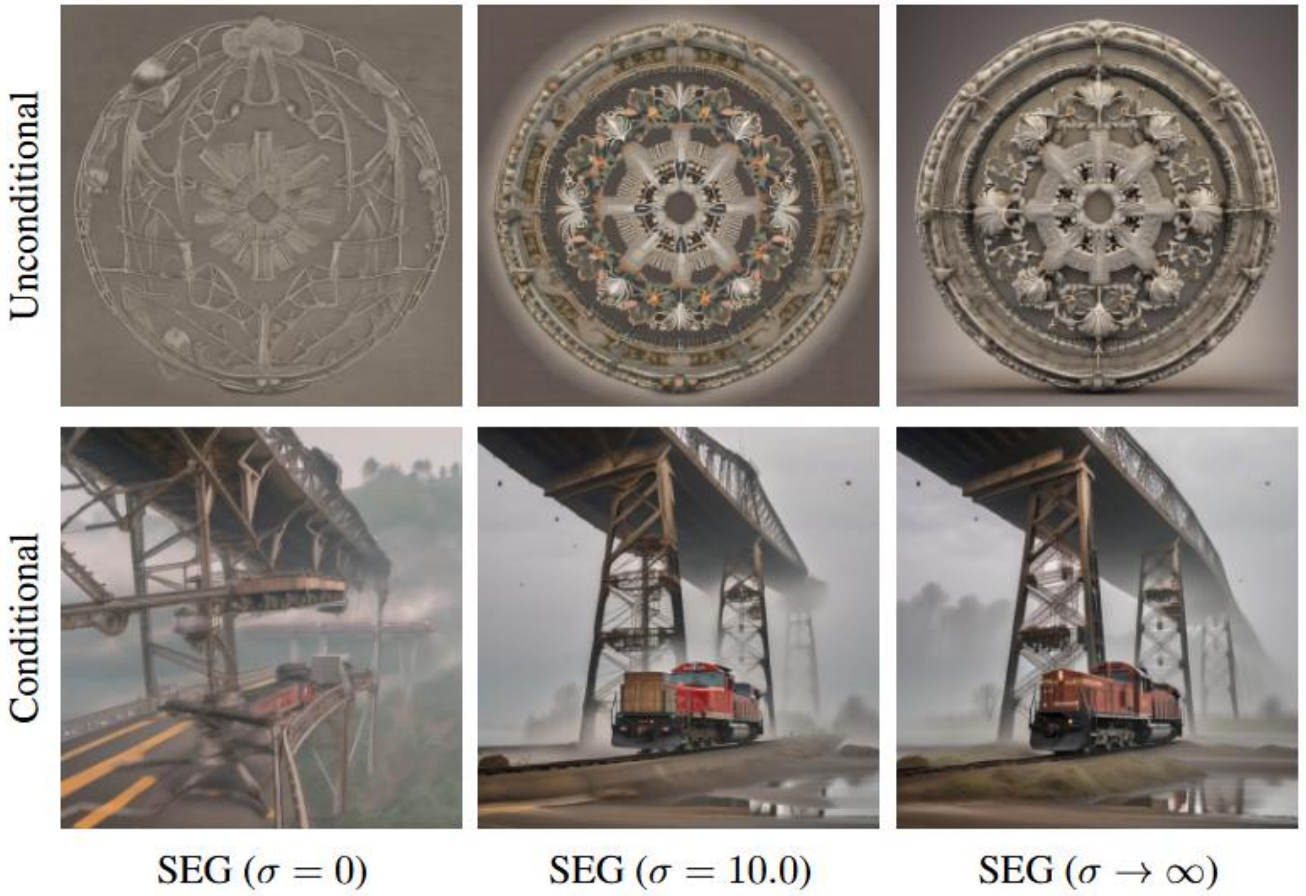
Moreover…..

```
#Smoother Energy Guidance (SEG) Pipeline
from pipeline_seg import StableDiffusionXLSEGPipeline
pipe = StableDiffusionXLSEGPipeline.from_pretrained("stabilityai/stable-diffusion-xl-base-1.0",torch_dtype=torch.float16)
pipe = pipe.to(device)
generator = torch.Generator(device="cuda").manual_seed(seed)
output_seg = pipe(prompts,guidance_scale=3.0,seg_scale=3.0,generator=generator,num_inference_steps=num_inference_steps).images[0]
```

✓ 17.1s

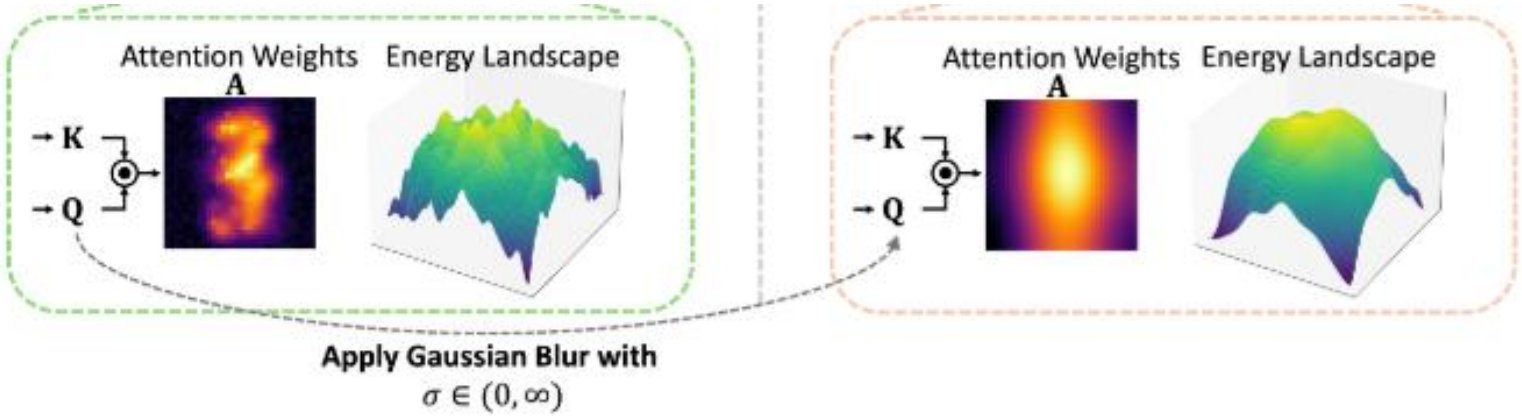
```
Loading pipeline components...: 100%|██████████| 7/7 [00:03<00:00,  2.2
100%|██████████| 50/50 [00:09<00:00,  5.09it/s]
```

1. Self-attention is equivalent to a gradient step on the energy function, where uniform attention yields an optimal gradient direction, reducing curvature.
2. This occurs blur kernel variance $\sigma \to \infty$, enhancing sample quality, but introducing drawbacks.



"A cargo train on a steel bridge in the fog"

Why are these happening?

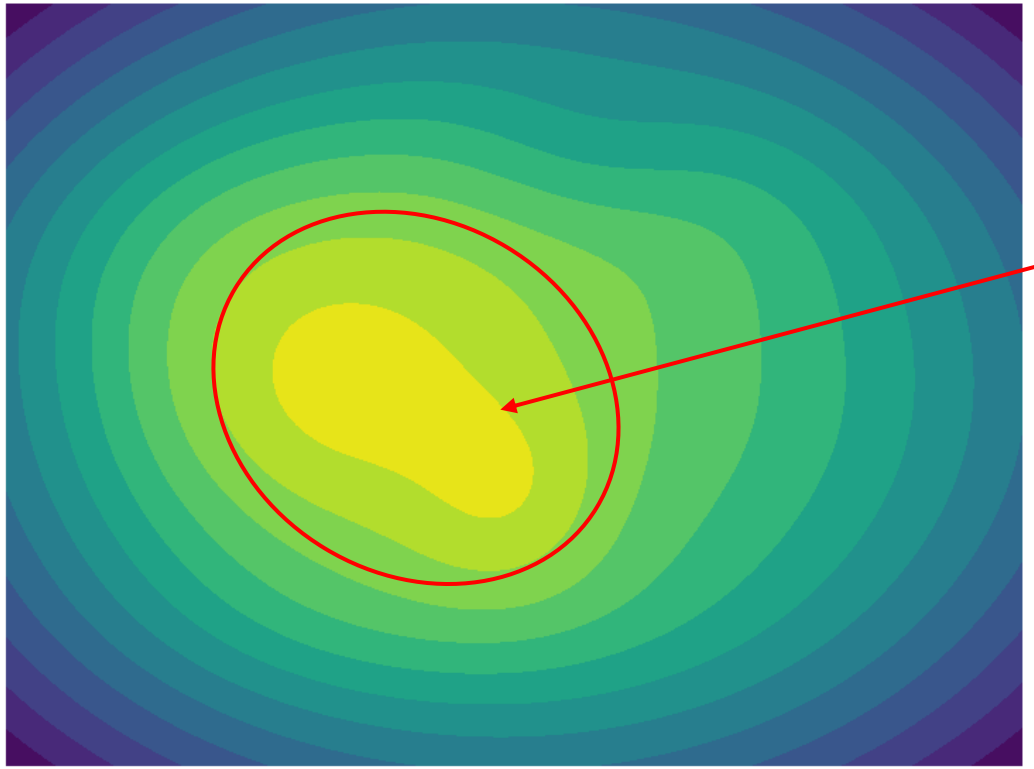


Indiscriminate blurring of query tokens….

So, our concern is are all components of the query tokens are responsible for better generation?

No large variance……

- So, we will not rely on large variance σ.
- We will perform the minimization step by precisely targeting the uniform attention.
- For that purpose, we first know about the curvature bound of the energy landscape.



Curvature bound of the attention……

1. Hessian of the energy function w.r.t. Q: $\nabla^2 E(\mathbf{A}) = \frac{1}{d}\mathbf{K}\left(\operatorname{diag}(\mathbf{A}) - \mathbf{A}\mathbf{A}^\top\right)\mathbf{K}^\top$

Jacobian

Covariance Matrix

2. From Hessian, we can obtain curvature information:

$$\lambda_{\max}(\nabla^2 E(\mathbf{A})) = \frac{1}{d}\lambda_{\max}\left(\mathbf{K}\left(\operatorname{diag}(\mathbf{A}) - \mathbf{A}\mathbf{A}^\top\right)\mathbf{K}^\top\right)$$

3. From spectral norm property, we can put the following bound:

$$\begin{aligned}
&\lambda_{\max}(\nabla^2 E(\mathbf{A})) \leq \\
&\left\|\frac{1}{d}\mathbf{K}\left(\operatorname{diag}(\mathbf{A}) - \mathbf{A}\mathbf{A}^\top\right)\mathbf{K}^\top\right\|_2 \leq \\
&\frac{1}{d}\|\mathbf{K}\|_2\left\|\operatorname{diag}(\mathbf{A}) - \mathbf{A}\mathbf{A}^\top\right\|_2\|\mathbf{K}^\top\|_2 \leq \\
&\frac{1}{d}\|\mathbf{K}\|_2^2\left\|\operatorname{diag}(\mathbf{A}) - \mathbf{A}\mathbf{A}^\top\right\|_2 \leq \\
&C\left\|\operatorname{diag}(\mathbf{A}) - \mathbf{A}\mathbf{A}^\top\right\|_2
\end{aligned}$$

Curvature bound of the attention……

A $C \left\| \mathrm{diag}(\mathbf{A}) - \mathbf{A}\mathbf{A}^\top \right\|_2$

~~QK^T~~ ~~K^T~~ $\left\| \mathrm{diag}(\mathbf{A}) - \mathbf{A}\mathbf{A}^\top \right\|_2$

The spectral norm bound for $||diag(A) - AA^T||$ depends on the nature of incoming query Q.

So, we will focus on:

How Q aligns with K?	What is the effect of Attention weights then?	What will be the Hessian of the energy function then?
$Q \leftrightarrow K$	A_i	$\nabla^2 E(A)$

From SEG to OSEG

Q

- Orthogonal($Q_\perp$)
  - $Q_\perp K^T = 0$
  - $\mathrm{A_i} = \frac{1}{\mathrm{n}}$ for all i
  - $\nabla^2 E(A) = \frac{1}{\mathrm{d}} \mathrm{K}(\frac{1}{\mathrm{n}} \mathrm{I} - \frac{1}{\mathrm{n}^2} 11^{\mathrm{T}}) \mathrm{K}^{\mathrm{T}}$
  - $\lambda_{\max}(\nabla^2 E(\mathrm{A}_\perp)) = \frac{1}{\mathrm{n}} \frac{||\mathrm{K}||_2^2}{\mathrm{d}}$
- Parallel($\mathrm{Q}_\parallel$)

From SEG to OSEG

Q

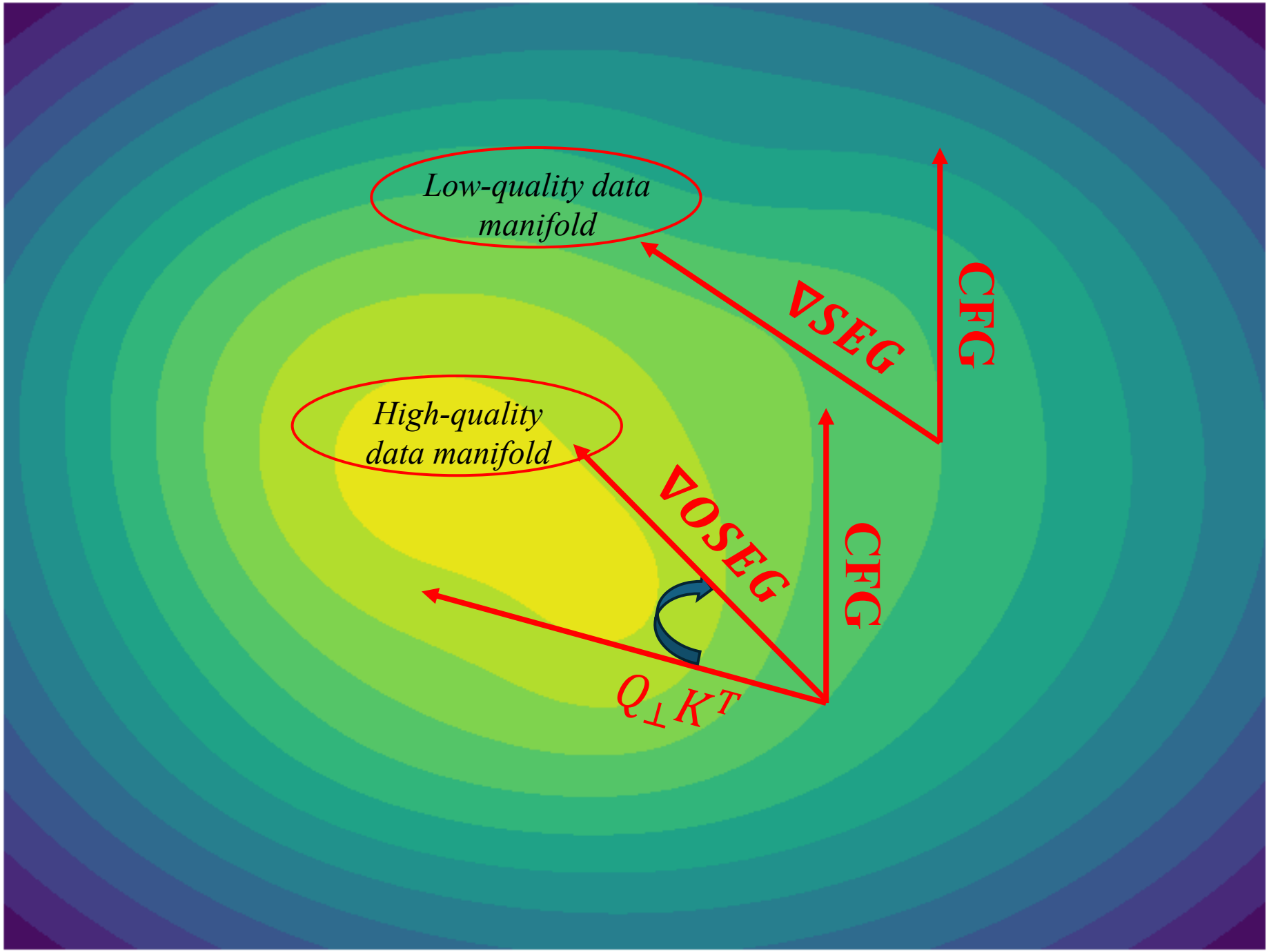
- Orthogonal($Q_\perp$)
- Parallel($Q_\parallel$)

$Q_\parallel K^T \neq 0$

$$\lambda_{max}(\nabla^2 E(A_\parallel)) < \frac{1}{n}\frac{||K||_2^2}{d}$$

$$A_i < \frac{1}{n} \text{ for all } i$$

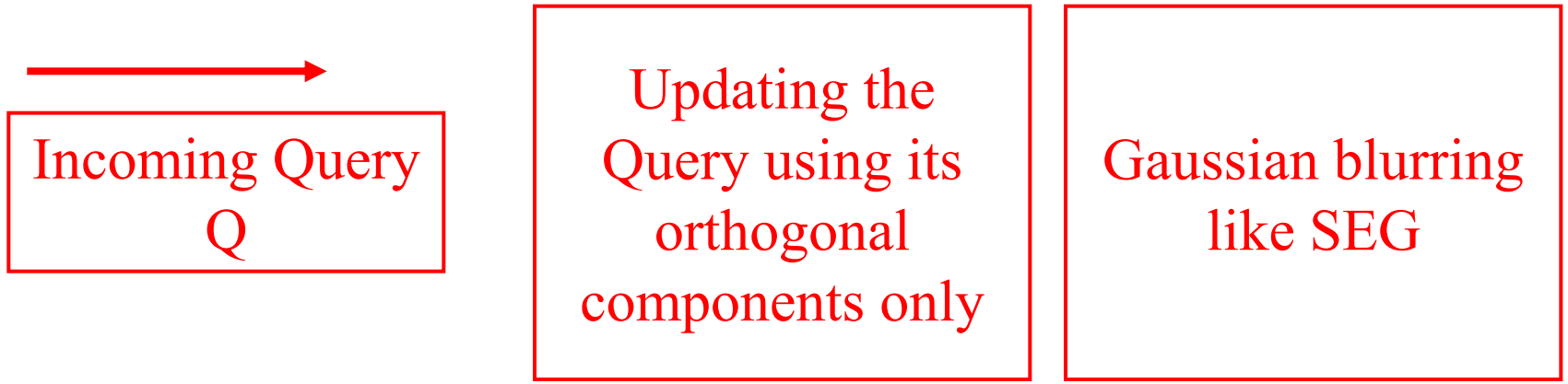
From SEG to OSEG



From SEG to OSEG

From SEG to OSEG

So, our study reveals that only the orthogonal component of a query predominantly drives the curvature.



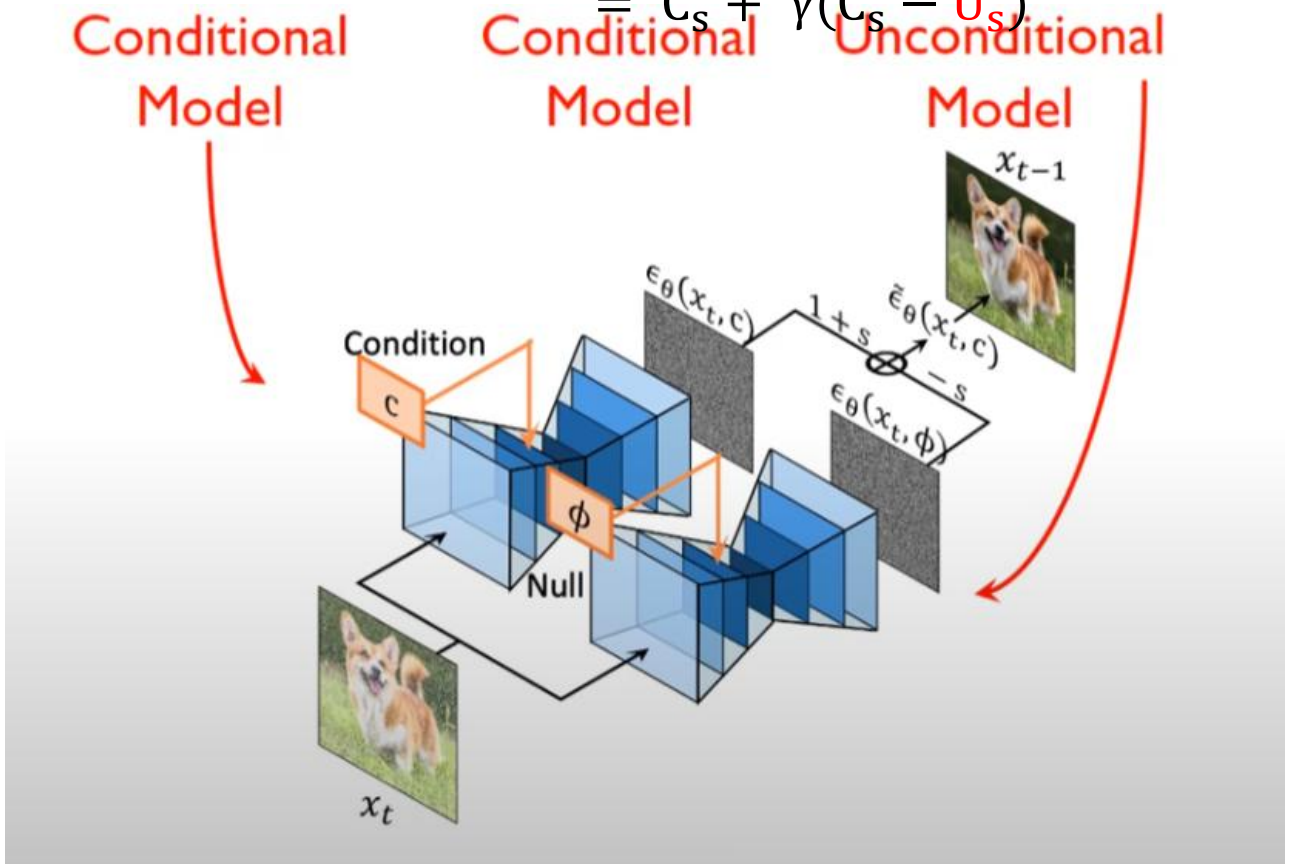
This enables our self-attention mechanism to perform targeted blurring, not random blurring like SEG.

From SEG to OSEG

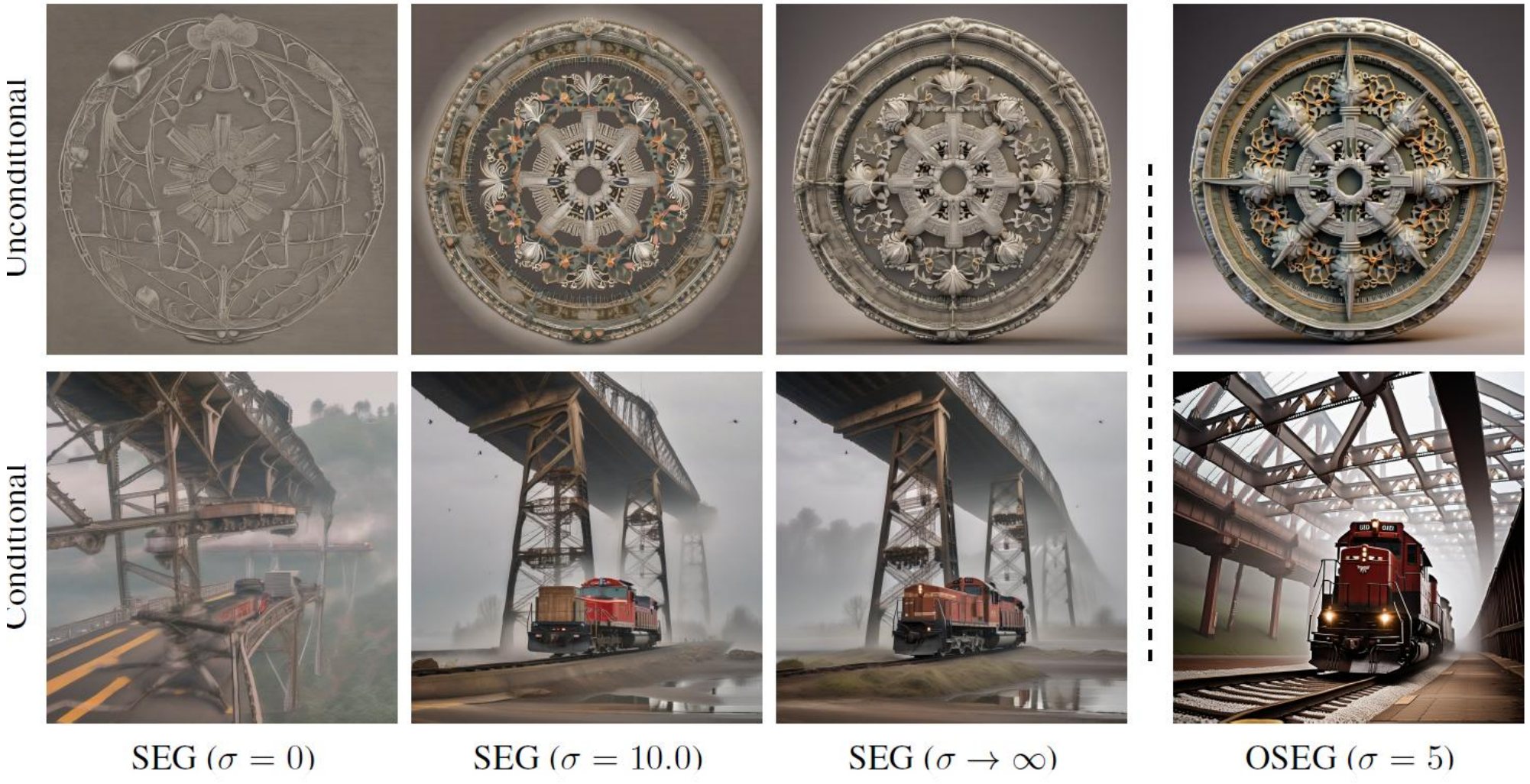
Conditional Conditional-Unconditional

$$\tilde{\epsilon}_\theta(x_t, c, t) = C_s + \gamma(C_s - U_s)$$

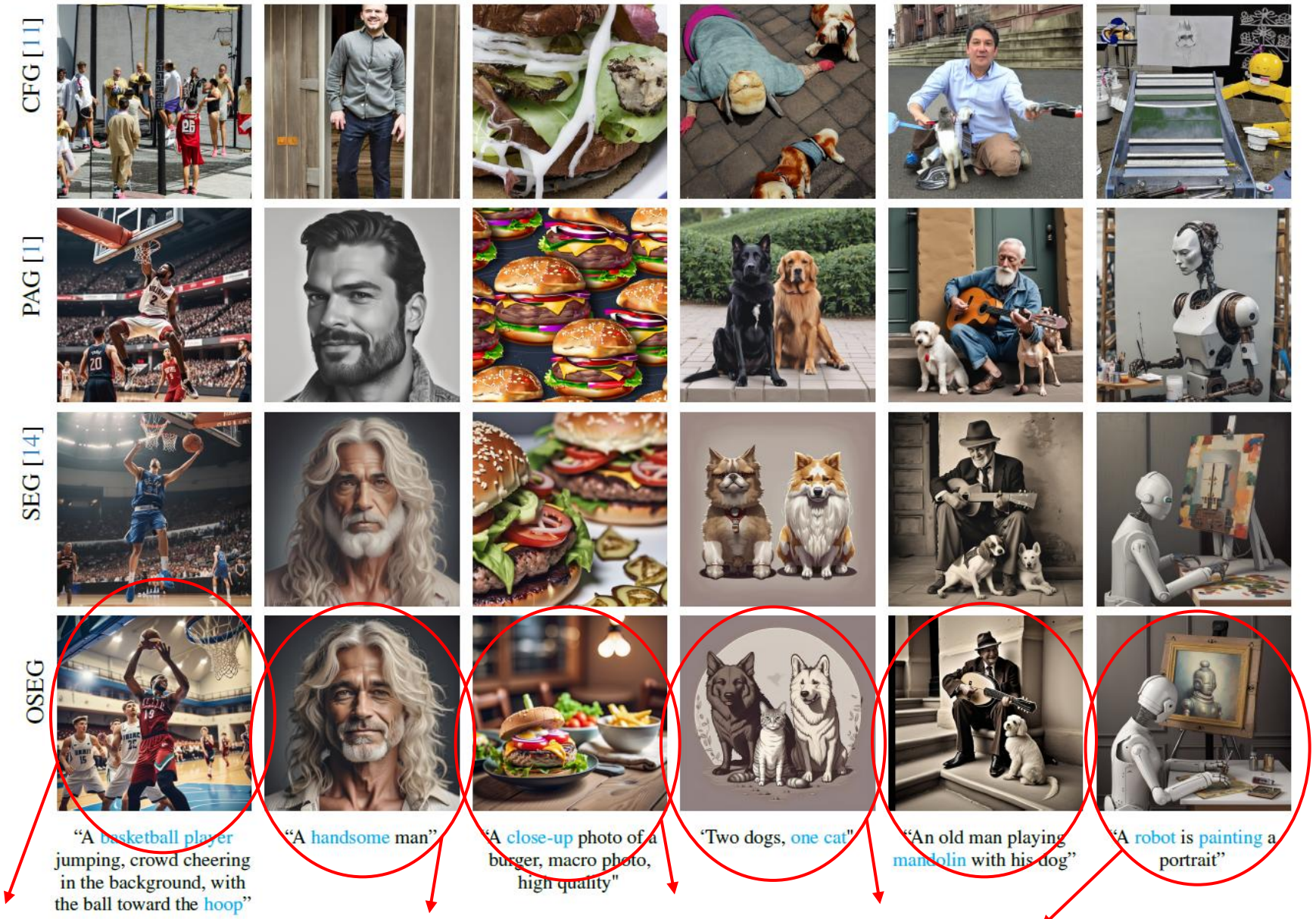
$$\tilde{\epsilon}_\theta(x_t, c, t) = C_s + \gamma(C_s - \check{U}_s)$$



$$\tilde{\epsilon}_\theta(x_t, c, t) = C_s + \gamma(C_s - \check{U}_s) + \lambda(C_s - \check{C}_s)$$



No Large variance….

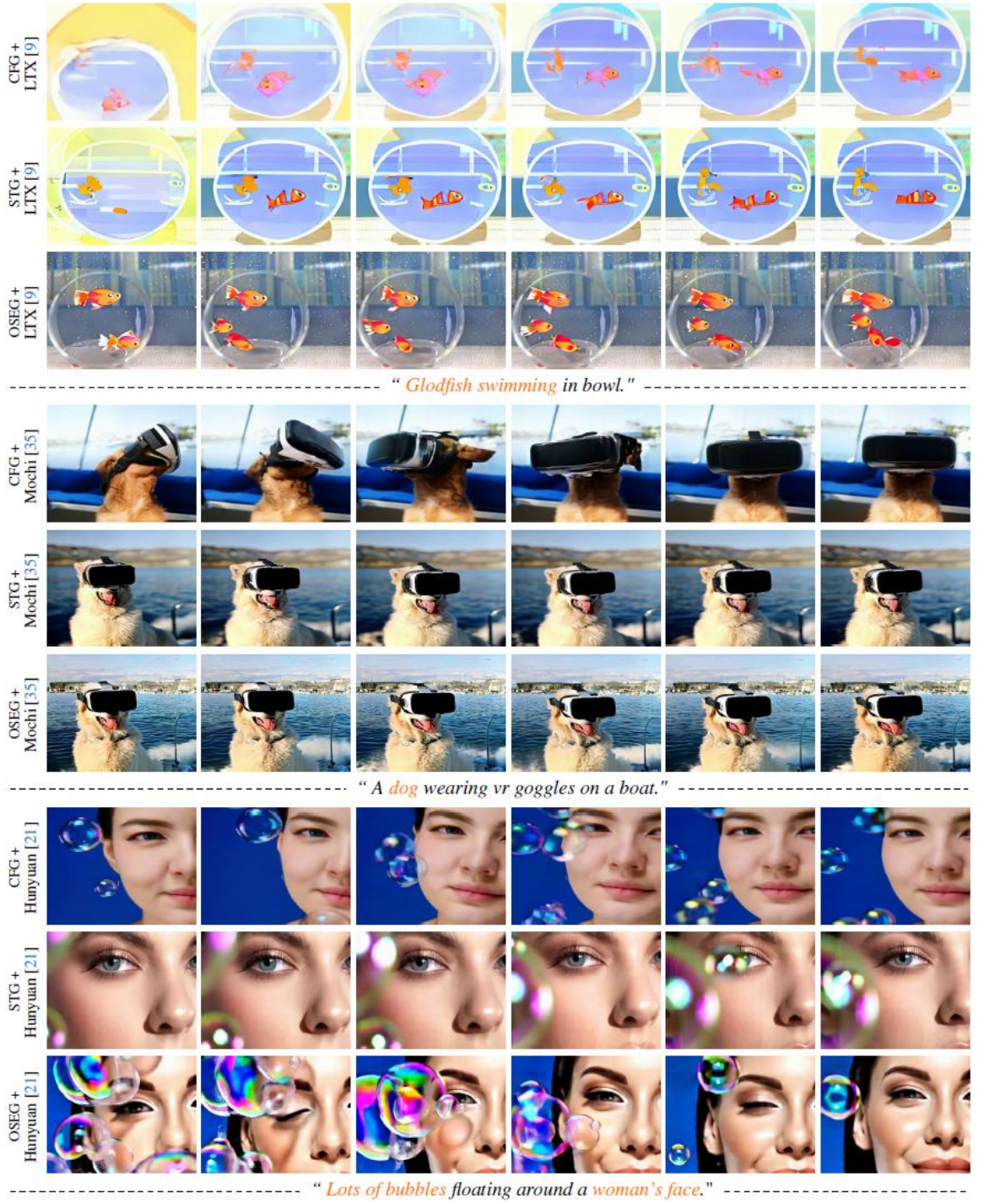

CFG [11]
PAG [1]
SEG [14]
OSEG
"A basketball player jumping, crowd cheering in the background, with the ball toward the hoop"
"A handsome man"
"A close-up photo of a burger, macro photo, high quality"
"Two dogs, one cat"
"An old man playing mandolin with his dog"
"A robot is painting a portrait"
Error correction
Better Quality
Better Scale
Better prompt alignment

OSEG for T2V

LTX

Mochi

Hunyuan



What about Quantitative analysis??(T2I)

Metrics	SAG [16]	PAG [1]	SEG [14]	OSEG
FID ↓	106.683	105.271	95.316	**88.014**
$LPIPS_{vgg}$ ↓	0.706	0.542	0.522	**0.506**
$LPIPS_{alex}$ ↓	0.644	0.472	0.454	**0.422**

Table 1. Quantitative performance comparison with different guidances under unconditional generation of Vanilla SDXL.

Metrics	SAG [16]	PAG [1]	SEG [14]	OSEG
FID ↓	72.08	65.45	<u>59.49</u>	**56.94**
CLIPS ↑	0.364	0.301	<u>0.312</u>	**0.363**
$LPIPS_{vgg}$ ↓	0.912	0.743	<u>0.729</u>	**0.713**
$LPIPS_{alex}$ ↓	0.905	0.715	<u>0.704</u>	**0.686**

Table 2. Quantitative performance comparison with different guidances under text-conditional generation of Vanilla SDXL.

Unconditional generation(for null prompt)

Conditional generation

What about Quantitative analysis??(T2V)

Models	Imaging Quality	Aesthetic Quality	Motion Smoothness	Dynamic Degree	Temporal Flickering
Mochi (CFG)	0.524	0.507	0.985	**0.87**	0.976
Mochi (STG)	0.628	0.554	0.988	0.86	0.978
Mochi (OSEG)	**0.644**	**0.567**	**0.992**	**0.86**	**0.981**
Hunyuan (CFG)	0.533	0.482	0.977	0.913	0.965
Hunyuan (STG)	0.598	0.495	0.963	0.872	0.961
Hunyuan (OSEG)	**0.613**	**0.502**	**0.975**	**0.896**	**0.983**

Table 3. Quantitative results for Mochi [35] and Hunyuan [21] on VBench [17] T2V benchmarks.

Models	FVD (↓)	IS	Imaging Quality	Aesthetic Quality	Motion Smoothness	Dynamic Degree
SVD (CFG)	151.3	38.0	0.687	0.637	0.966	0.562
SVD (STG)	128.7	38.5	0.694	0.639	0.968	0.694
SVD (OSEG)	**122.3**	**39.2**	**0.698**	**0.646**	**0.972**	**0.703**

Table 4. Quantitative results for SVD [5] on FVD, IS, and VBench [17] I2V benchmarks.

Let's have some mathematical proof!!!!

Why only orthogonal components lead to a Uniform Attention?

$$A_i = \frac{1}{n}; \forall i$$

$$A_i = \frac{\exp\left(\frac{\mathbf{Q}^\top \mathbf{K}_i}{\sqrt{d}}\right)}{\sum_{j=1}^{n} \exp\left(\frac{\mathbf{Q}^\top \mathbf{K}_j}{\sqrt{d}}\right)}$$

$$A_i = \frac{1}{\sum_{j=1}^{n} \exp\left(\frac{\mathbf{Q}^\top \mathbf{K}_j}{\sqrt{d}}\right)}.$$

$$A_i = \frac{1}{n}.$$

$$\mathbf{Q}^\top \mathbf{K}_i = 0$$

$$\frac{\mathbf{Q}^\top \mathbf{K}_i}{\sqrt{d}} = \frac{0}{\sqrt{d}} = 0.$$

$$\exp\left(\frac{\mathbf{Q}^\top \mathbf{k}_i}{\sqrt{d}}\right) = \exp(0) = 1.$$

$$\sum_{j=1}^{n} \exp\left(\frac{\mathbf{Q}^\top \mathbf{K}_j}{\sqrt{d}}\right) = \sum_{j=1}^{n} 1 = n.$$

Some Ablations…….

Impact of the signal scale….

"A beautiful motorbike, front view, high quality"

$\sigma = 3$ $\sigma = 15$ $\sigma = \infty$

Some Ablations…

….

Some Unconditional generations….

For better experience of videos, you can visit the project website.

https://oseg-guidance.github.io/

Thank You....