Supplementary material: OSEG: Improving Diffusion sampling through Orthogonal Smoothed Energy Guidance

Masud An Nur Islam Fahim
University of Vaasa,
Finland.
masud.fahim@uwasa.fi

Nazmus Saqib
Jeju National University,
Korea.
nsaqib1995@gmail.com

Joon-Min Gil
Jeju National University,
Korea.
jmgil@jejunu.ac.kr

1. Proof

Orthogonal Query Leads to Uniform Attention

Here, we present complete proof for the uniform attention when the query appears to be orthogonal to the key vectors. If the query vector $\mathbf{Q} \in \mathbb{R}^d$ is orthogonal to all key vectors $\mathbf{K}_i \in \mathbb{R}^d$ (i.e., $\mathbf{Q}\mathbf{K}_i^\top = 0; \forall i$), then the attention weights a_i are uniform:

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

Proof. Given $\mathbf{A}$ is the attention weights, $\mathbf{Q}$ is the query vector, $\mathbf{K}$ is the key vector, and d is the dimensionality of the vectors. For a single instance A_i in the attention weights given by:

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

Given the orthogonality condition $\mathbf{Q}\mathbf{K}_i^\top = 0$ for all i, substitute into the exponent:

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

Thus, the upper part of the 1 simplifies to:

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

Substitute this into 1 and we get:

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

For the denominator, the summation becomes:

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

Therefore, the attention weight for each i is:

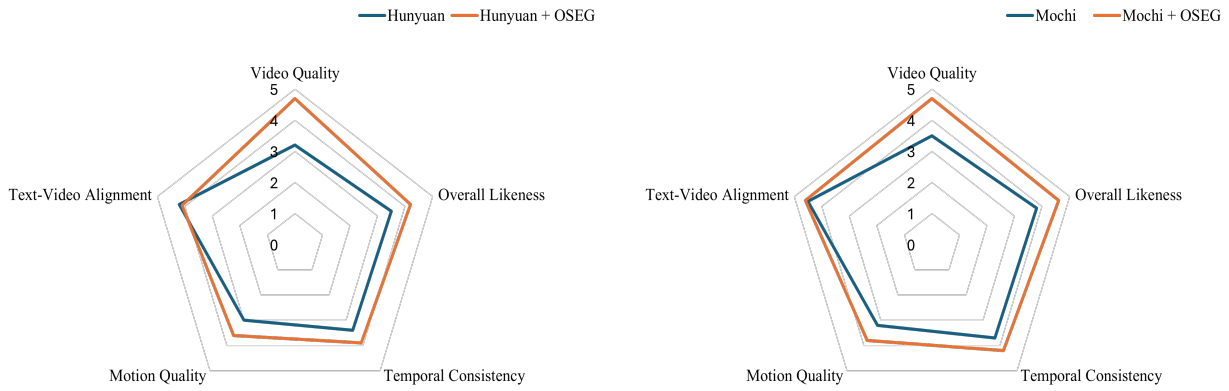
$$A_i = \frac{1}{n}.$$

Thus, we conclude the proof.

2. Human Evaluation

We follow the same procedure of [6] to evaluate subjective opinions from human interpretation which are designed across five key aspects: a) Video quality, where the higher scores indicates that the generated video consumes less blurriness, noise or unwanted artifacts; b) Text-video alignment, which highlights the alignment between the input text prompt and the generated video; c) Motion quality, which evaluates the correctness and realism of the motions depicted in the video; d) Temporal Consistency, measuring coherence between the frame-to-frame of the generated video to assess the smoothness of the movement, and e) Subjective Likeness which is an aesthetic score for human preferences.

For each metric of this experiment, we have collected feedback from five users who are privileged to rate the generalized video between scale 1 to 5, where higher scores represent better alignment. We use 100 prompts from Evalcrafter for T2V generation with Mochi [7] and Hunyuan [5].



(a) Human evaluation using Hunyuan [5]

(b) Human evaluation using Mochi [7]

Figure 1. User study results for OSEG on Mochi Hunyuan [5], and [7] using 100 prompts from EvalCrafter [6]. The results demonstrate that incorporating OSEG leads to improved video quality across all evaluated aspects.

3. Additional qualitative results

Figure 2 provides more visual comparison results of T2I generation.

3.1. Conditional Image generation

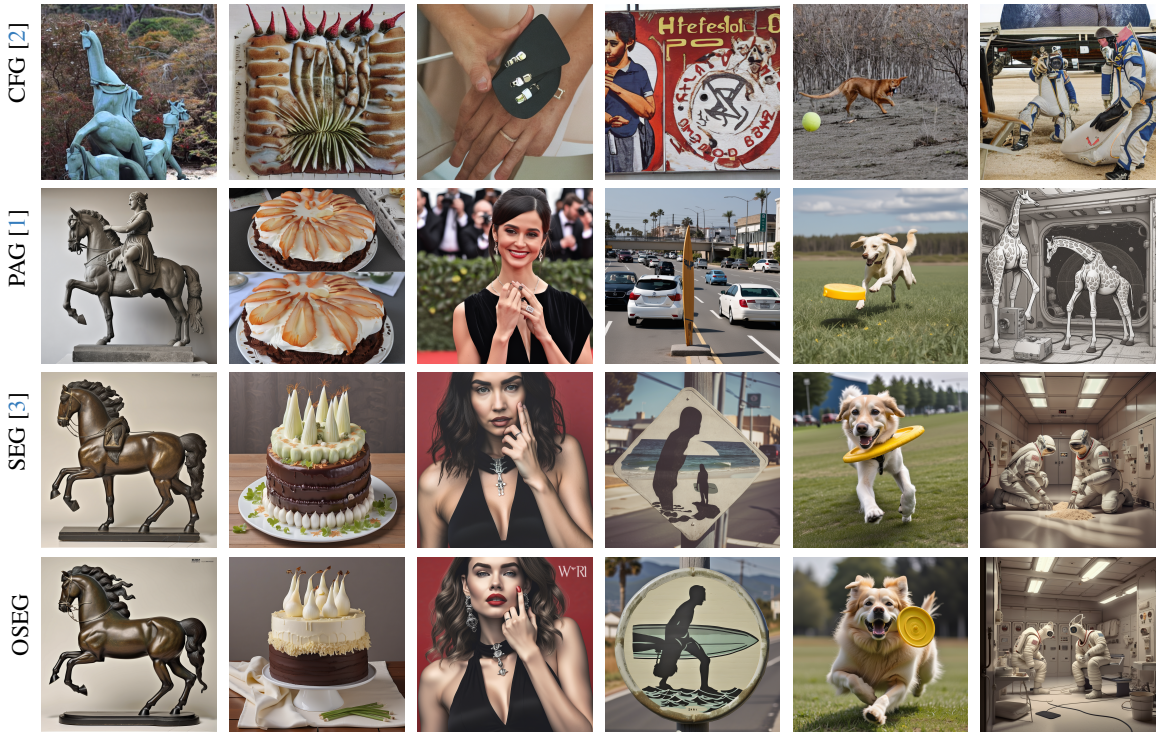


Figure 2. **Qualitative comparison under conditional generation between CFG, PAG, SEG and OSEG.**

4. Conditional Video generation

Figure 3 and Figure 4 provide additive T2V results generated by Mochi and Hunyuan, respectively, with the guidance of OSEG. Each figure provides a single frame of the videos, clicking on which will direct you to the original video link.

"Close-up of the front, the man's head is surrounded by flames, his terrified face under the fire, his face is covered with dust and stains, his deep blue eyes are tearful, his eyes are wide open with fear. Water flows down from the top of his head, the flames turn into white smoke, he slowly turns his head back, the dark background creates a tense atmosphere, movie-level image quality, super detail performance, perfect light and shadow effects."

"Front view of a giant angel flapping its wings and slowly walking forward. The angel in white is wearing a pure white robe. The ground is covered with flying gravel and the fog is thick. The angel's feathers flutter gently in the wind, and every feather on the wings is clearly visible. The background is a rocky wasteland, with clouds floating in the gray-blue sky and light shining through the clouds. Realistic style, super clear details."

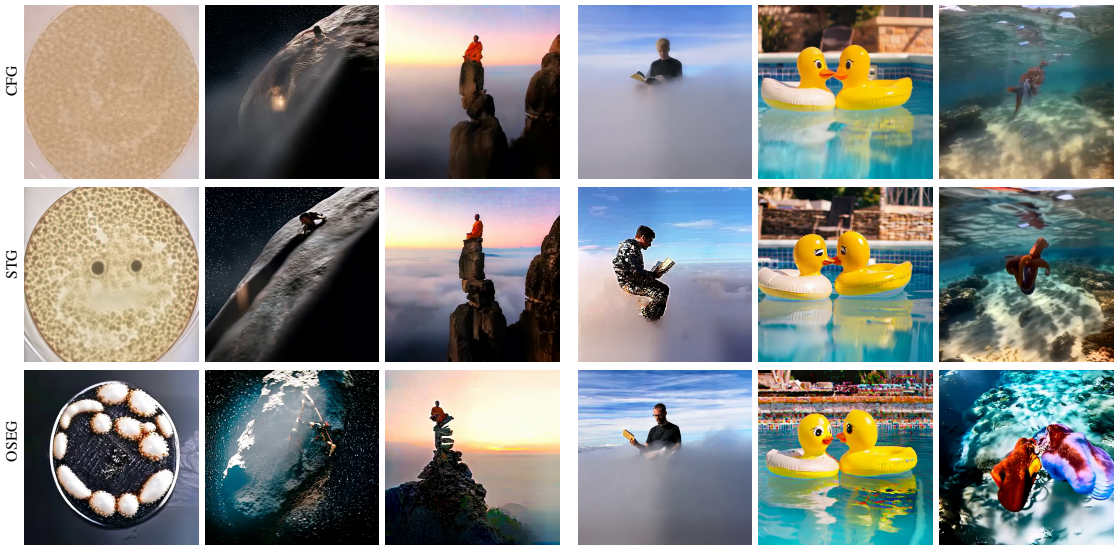
"A food kingdom, where houses are made of bread and pastries, and roofs are made of golden biscuits."

"A fierce dragon roars slowly, spew fire on the men ,steam rising from nostrils , cinematic camera shake, wings spread wide,glowing eyes fixed on warrior with intense rage, A man witnesses a dragon spewing flames, instantly casting a protective magical barrier,many fire sparks and debris floating in the foregroundDark mood"

"An anthropomorphic kitten is lying on a ball of kneaded dough, with its two front paws resting on the edge of the dough. Some kitchen facilities can be seen in the background."

"A car changes from golden to white."

Figure 3. Conditional video generation comparison between CFG[2], STG[4], and OSEG by Hunyuan[5]. The first video produces a better scale of the human face, while the second video provides a clearer representation of the giant angel in thick fog. The third video demonstrates improved content preservation of golden biscuits. Similarly, OSEG generates the cat on the dough with better prompt alignment. Finally, both OSEG and CFG show similar appearance in the car video.



"A top-down microscopic view reveals a petri dish teeming with large cells undergoing mitosis, forming the shape of a smiley face"

"A low-angle, long exposure shot of a lone female climber, wearing shorts and tank top rock climbing on a massive asteroid in deep space. The climber is suspended against a star-filled void. dramatic shadows across the asteroid's rugged surface, emphasizing the climber's isolation and the scale of the space rock. Dust particles float in the light beams, catching the light. The climber moves methodically, with focused determination."

"A wide-angle shot shows a serene monk meditating perched a top of the letter E of a pile of weathered rocks that vertically spell out 'ZE'. The rock formation is perched atop a misty mountain peak at sunrise. The warm light bathes the monk in a gentle glow, highlighting the folds of his saffron robes. The sky behind him is a soft gradient of pink and orange, creating a tranquil backdrop. The camera slowly zooms in, capturing the monk's peaceful expression and the intricate details of the rocks. The scene is bathed in a soft, ethereal light, emphasizing the spiritual atmosphere."

"A young man at his 20s is sitting on a piece of cloud in the sky, reading a book."

"A high view wide angle shot of two octopuses gliding slowly through a sunlit patch of ocean. They float closer to one another in fluid, spiraling motion. The scene unfolds in a shallow, coral-framed cove where beams of light filter through the water and dance over soft sand. One octopus is a rich, reddish orange, the other a delicate blend of violet and pale blue. As they near, their arms begin to entwine in a gentle, synchronized movement—tactile, intimate, and unhurried. Around them, small tropical fish dart through the water, and as the octopuses embrace, the fish begin to shift, forming a loose, shimmering heart shape around the couple. The camera slowly pushes in as the pulsing bodies continue their dance, their tentacles crossing and curling. In the final moment, two arms intertwine at the center of the frame, forming a perfect heart shape in close-up, framed by the swirling school. The light catches every scale and ripple, holding the ocean in a breathless stillness. Filmed in 16mm film."

"A front-facing wide angle shot of two inflatable duck floaties. They drift toward each other until their beaks meet in a gentle kiss and then magically one of them closes his eyes. The moment takes place in a quiet backyard pool under warm afternoon sun, with calm water and tiled edges reflecting soft light. One floatie is slightly faded, with creases from long use; the other is bright yellow, glossy and new. As they float closer, their heads tilt subtly, and the space between their beaks, rounded heads, and arched necks forms an evident heart shape from this angle. Just as they touch, surprisingly each floatie magically winks—one eye narrowing in a playful, synchronized gesture. The water stays nearly still, disturbed only by their closeness. Soft highlights ripple across their surfaces, and the pool glows in a warm, summery tone. Filmed in 16mm film."

Figure 4. Conditional video generation comparison between CFG[2], STG[4], and OSEG by Mochi[7].Compared to the existing SOTA guidance methods, incorporating OSEG alongside CFG noticeably improves spatial resolution with natural semantic coherence of the structures within the samples. This combination effectively strengthens fine-grained details and overall harmony, leading to high-resolution generations such as the modern microscopic view (first video) and better exposure (second video). In the third video, OSEG creates the structure of **ZE**, which addresses better prompt alignment. The remaining videos also provide better background details with diversity. For better visualization, please visit the website: *https://oseg-guidance.github.io/*

References

[1] Donghoon Ahn, Hyoungwon Cho, Jaewon Min, Wooseok Jang, Jungwoo Kim, SeonHwa Kim, Hyun Hee Park, Kyong Hwan Jin, and Seungryong Kim. Self-rectifying diffusion sampling with perturbed-attention guidance. In *European Conference on Computer Vision*, pages 1–17. Springer, 2024. 3

[2] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 3, 4, 5

[3] Susung Hong. Smoothed energy guidance: Guiding diffusion models with reduced energy curvature of attention. *Advances in Neural Information Processing Systems*, 37:66743–66772, 2024. 3

[4] Junha Hyung, Kinam Kim, Susung Hong, Min-Jung Kim, and Jaegul Choo. Spatiotemporal skip guidance for enhanced video diffusion sampling. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 11006–11015, 2025. 4, 5

[5] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 2, 4

[6] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22139–22149, 2024. 2

[7] Genmo Team. Mochi 1. https://github.com/genmoai/models, 2024. 2, 5