Learning Visual Generative Priors without Text

Supplementary Material

F. Supplementary Experiments

F.1. Motivation experiments FID results

As shown in Figure S1, we provide the FID metric results of the motivation experiment (Figure 1b). Although the quality measurement standard of image-text pairs is customized based on clip score, our Lumos training framework also shows significant advantages in FID metric.

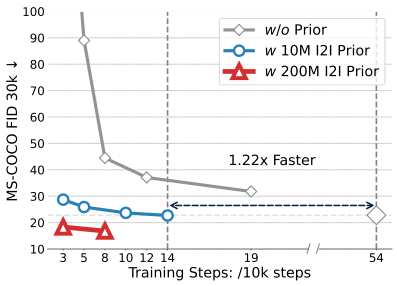


Figure S1. Supplementary FID results for motivation experiments.

F.2. Ablate the number of condition tokens for I2I framework

We explore this issue by setting the same number of condition tokens during the I2I framework as in the downstream text-to-image tasks. For the text-to-image task, we used T5 [40]. For the I2I training, we set the number of condition tokens to be the same as in the downstream task, which is 120. We conduct experiments using DINO-B, leveraging the excellent ability of the DINO [CLS] token to focus on foreground tokens. We select the top 119 tokens with a high correlation with the class token. Additionally, we design a set of comparative experiments, randomly selecting 119 tokens from patch tokens to introduce noise perturbation. The experimental results in Figure S2 (a) and (b) indicate that more fine-grained local information accelerates the training of I2I generation. However, the standalone global class token still exhibits superior transfer performance for downstream tasks.

F.3. Ablating the Scale of Vision Encoder

The comparison results in Figure S3 (a) and (b) demonstrate that larger and better vision encoders can provide a higher performance ceiling for downstream tasks. Therefore, for a better vision encoder, Lumos-I2I framework has greater potential.

G. Dataset

This section supplements a detailed introduction of training data and implementation details for the models.

G.1. Image-to-Image Generation

The training data of Lumos-I2I can be expanded indefinitely. In this paper, we curate and construct a pure image dataset totaling 190 million from existing open-source data. This includes 120 million images filtered from LAION-5B [46] with a resolution greater than 512 and an aesthetic score greater than 5.0, as well as 55 million images selected from COYO-700M [5] with the same resolution and aesthetic score criteria. Additionally, we include 10 million high-quality segmented scene data from the SAM [27] dataset, 4 million high aesthetic score images from JourneyDB [35], and 1 million classic natural scene images from the ImageNet-1K [14] dataset.

G.2. Text-to-Image Generation

We construct a dataset of 30 million text-image pairs, all of which are sourced from easily accessible open datasets. This includes 10 million and 5 million images selected from LAION-5B [46] and COYO-700M [5] respectively, based on a standard resolution greater than 512 and an aesthetic score higher than 5.5. Additionally, the dataset contains 10 million from SAM [27], 4 million from JourneyDB [35], and 1 million from Imagenet-1K [14]. For the text caption, we use the state-of-the-art multi-modal large language model (*i.e.*, InternVL [11]) to generate detailed long captions. Following DALLE-3 [2], we incorporate raw captions into the training process.

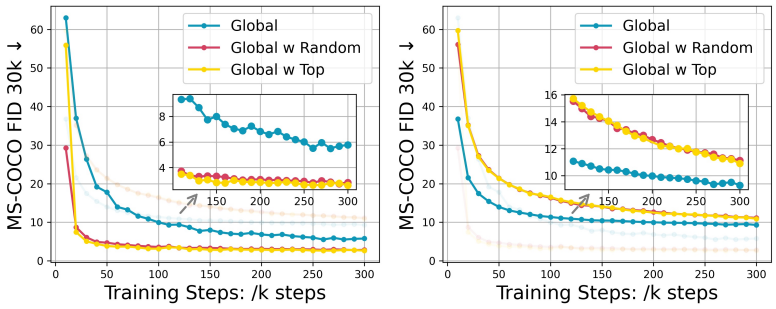


Figure S2. (a) Image-to-Image Generation, (b) Text-to-Image Generation.

G.3. Novel View Synthesis

We finetune Lumos-I2I for novel view synthesis task using a subset (750k) of the released Objaverse [13] dataset, a large-scale open-source collection comprising over 800K 3D models created by more than 100K artists. We randomly sample 32 camera extrinsic matrices $\mathcal{M}_1$ which are oriented towards the center of the object, followed by rendering 32 views using a ray tracing engine.

G.4. Image-to-Video Generation

For the image-to-video generation task, Lumos-I2V is initialized from Lumos-I2I and trained on WebVid10M [1] dataset by sampling 16 frames with 3 frames interval.

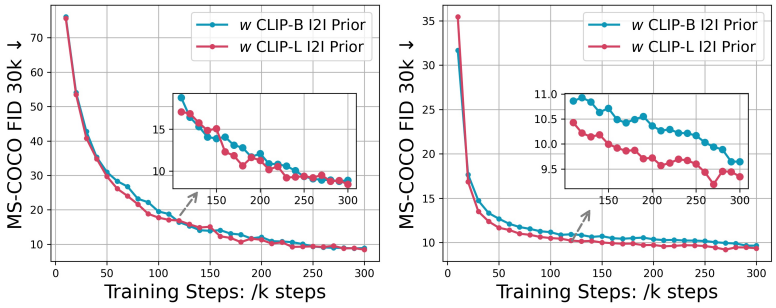


Figure S3. (a) Image-to-Image Generation, (b) Text-to-Image Generation.

H. Model and Implementation Details

H.1. Lumos-I2I Model

Implementation and Training Details. We train the Lumos-I2I (as shown in Figure 3 (a)) on 64 A100 GPUs with the total batch size of 16384. For saving memory, we use the mixed $fp16$ format with gradient checkpointing. The AdamW optimizer is utilized with a weight decay of 0.03 and a constant 1.6×10^{-4} learning rate. In the initial phase of training, we set a warm-up of 1000 steps for stable training.

H.2. Lumos-T2I Model

Implementation and Training Details. Lumos-T2I (as shown in Figure 3 (b)) uses the T5 [40] large language model (specifically 4.3B Flan-T5-XXL) as the text encoder for conditional feature extraction, with the text condition length set to 120. Inspired by SDXL [37] and Pixart-α [7], Lumos-T2I adopts the progressive resolution training strategy and provide four resolution versions of the text-to-image model, which are 256×256, 512×512, 1024×1024, and an arbitrary aspect ratio version at 1024 resolution. Lumos-T2I is trained on 64 A100 GPUs. Following the training setting of Lumos-I2I, we use the mixed $fp16$ format with gradient checkpointing with the AdamW optimizer and the warm-up setting for stable training. Details of the training information are shown in Table S1.

Table S1. **Detailed training information about every Lumos-T2I training stage.**

Stage	Image Resolution	Training Steps(K)	Batch Size	Learning Rate	weight decay	Warm Up Steps(K)
1	256×256	65	256×64	1.6×10^{-4}	0.03	1
2	512×512	60	64×64	8×10^{-5}	0.03	1
3	1024×1024	20	16×64	4×10^{-5}	0.03	1
4	Multi-scale 1024	20	16×64	4×10^{-5}	0.03	1

H.3. Lumos-NVS Model

Training Objective. Inspired by the definition of the task in Zero-1-to-3 [31], we tune Lumos-I2I to the novel view synthesis task. As shown in Figure 3 (c), the task is to synthesize an image of an object from a new camera viewpoint. The training data consists of image-viewpoint pairs, where each pair includes a single image $x \in \mathbb{R}^{h\times w\times 3}$ and its corresponding condition $c = (R, T)$. In detail, the relative camera rotation $R \in \mathbb{R}^{3\times3}$ and translation $T \in \mathbb{R}^3$ determine the desired viewpoint. Lumos-NVS is trained via

$$L_{\theta_{\text{NVS}}} := \mathbb{E}_{\mathcal{E}(\hat{x}_{R,T}),\mathcal{E}(x),x,\epsilon\sim\mathcal{N}(0,1),t}\left[\left\|\epsilon - \epsilon_\theta\left(\langle z_t, \mathcal{E}(x)\rangle, t, \tau^{img}(x), R, T\right)\right\|_2^2\right], \tag{S1}$$

where $\langle\cdot,\cdot\rangle$ represents the concatenate operation and $\hat{x}_{R,T}$ denotes the synthesized image. Meanwhile, based on the generated novel view image list, we can directly use the off-the-shelf sparse views 3D Reconstructor (*e.g.*, LGM [48] and GRM [56]) to handle the Single View 3D Reconstruction task. In this paper, we mainly use the open-source LGM to reconstruct the new perspective of the sparse view generated by our model into a 3D Gaussian [25].

Implementation and Training Details. We train Lumos-NVS on 64 A100 GPUs with a total batch size of 16384. The mixed $fp16$ format with gradient checkpointing is utilized for saving memory. The AdamW optimizer is utilized with a weight decay of 0.03 and a constant 1.6×10^{-4} learning rate. We set a warm-up of 1000 steps for stable training in the initial phase of training.

H.4. Lumos-I2V Model

Training Objective. We finetune our Lumos-I2I for the image-to-video generation task, where the video model receives a still input image as the condition. Following stable video diffusion [4], we use a 2D VAE encoder $\mathcal{E}$ to compress each frame of an n-frame video $v = [f^0, \ldots, f^n]$ into latent representation $z^{[1\ldots n]} = [z^0, \ldots, z^n]$, where $z^i = \mathcal{E}(f^i)$. Besides the diffusion transformer initialized from Lumos-I2I, we attach the temporal module to the Lumos-I2V as shown in Fig. 3 (d). Transformers can be easily extended to support image-to-image and video-to-video tasks due to the macro modeling ability. Different from SVD, which concatenates the condition frame to the latent noise of all generation video frames, we leverage mask strategy [58]) to support image conditioning. Meanwhile, we maintain the original image condition control method of Lumos-I2I for the image-to-video task. The training objective is as follows:

$$L_{\theta_{\text{I2V}}} := \mathbb{E}_{[\mathcal{E}(f^0),\mathcal{E}(f^1),\ldots,\mathcal{E}(f^n)],f_0,\epsilon\sim\mathcal{N}(0,1),t}\left[\left\|\epsilon - \epsilon_\theta\left([z_0^0, z_t^1, \ldots, z_t^n], t, \tau^{img}(f_0)\right)\right\|_2^2\right]. \tag{S2}$$

where c_{in} includes the latent representation of the first frame z^0 and the extracted semantic information of the first frame $\tau^{img}(f^0)$. During the inference phase, we unmask the conditional frame. Moreover, the unmasked frame is assigned timestep 0, while others remain the same t.

Implementation and Training Details. We train Lumos-I2V on 64 A100 GPUs with a total batch size of 4096. We use $bf16$ format with gradient checkpointing and ZeRO stage-2 optimizer. The HybridAdam optimizer is set with 2×10^{-5} learning rate.

I. Qualitative Results

This section provides more qualitative results of text-to-image generation, novel view synthesis, and image-to-video generation tasks. Moreover, we exhibit the generative capabilities of our Lumos-I2I for image interpolation.

I.1. Text-to-Image Generation

As shown in Figure S4 and Figure S5, we provide more generated images and their corresponding text prompts. Lumos-T2I can generate high-quality aesthetic photos while maintaining image and text alignment.

I.2. Novel View Synthesis

We provide more examples and novel views generated by Lumos-NVS in Figure S6.

I.3. Image-to-Video Generation

More videos generated by Lumos-I2V from the input frame are provided in Figure S8.

A peaceful mountain lake reflecting the surrounding pine trees and snowy peaks, photorealistic, tranquil

Two female rabbit adventurers dressed in a fancy velvet coats next to a Christmas tree, Christmas theme, on an antique opulent background , jean - baptiste monge , smooth, anthropomorphic photorealistic, photography, lifelike, high resolution, smooth

A peaceful forest in autumn, with golden leaves falling and a stream running through it, illuminated by soft sunlight.

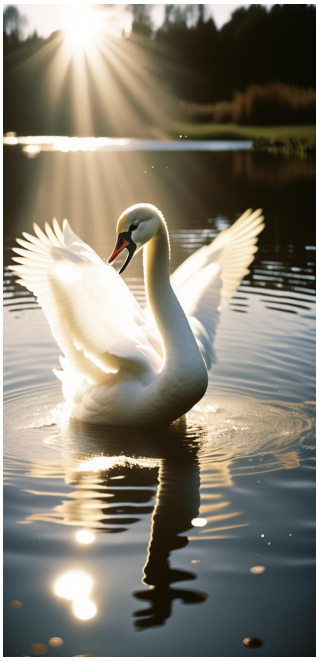
A bear with fur made of chocolate shavings, standing in a clearing filled with marshmallow mushrooms

A close-up of a sunlit butterfly resting on a flower in a garden

An owl constructed from layers of caramel popcorn and hazelnut chocolate, perched on a pretzel branch

Figure S4. The samples generated by Lumos-T2I exhibit remarkable quality, characterized by exceptional fidelity and precise alignment with the provided textual descriptions.

A dramatic mountain range during a thunderstorm, with dark clouds, lightning strikes, and rugged terrain



a close up portrait of a swan flapping its wings in a lake, sunlight pouring in, strong sunlight, strong shadow, kodak portra 800 film, cinematic, film grain

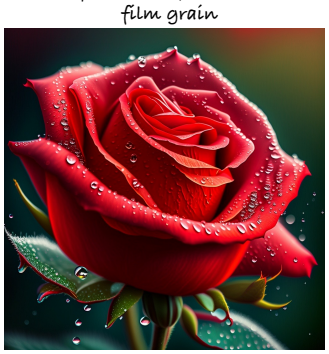
A majestic bald eagle soaring over a snowy mountain range

A cyborg superhero with a robotic arm and high-tech gadgets, standing atop a skyscraper

A hippopotamus with a body of jelly-like translucent gelatin, lounging in a pool of liquid sherbet

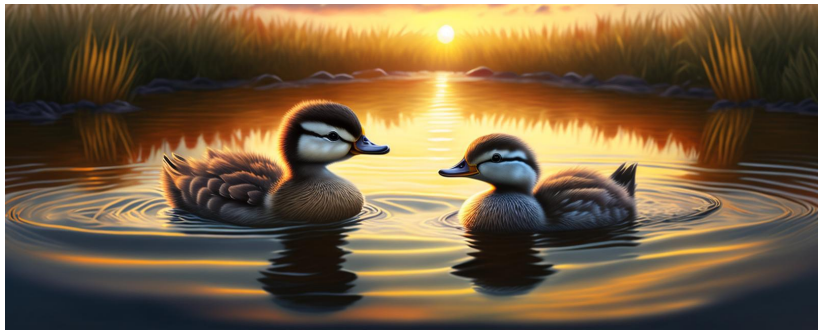
A baby rabbit wearing a tiny knitted hat, ultra-detailed, photorealistic

A close-up photograph of a lion with its mane blowing in the wind against the savanna backdrop



A close-up of a vibrant, fully bloomed red rose with dew drops on its petals

A group of astronauts standing on the surface of Mars, with Earth visible in the distant sky.



Two baby ducks swimming in a pond at sunset, highly detailed, hyper-realistic

Figure S5. The samples generated by Lumos-T2I exhibit remarkable quality, characterized by exceptional fidelity and precise alignment with the provided textual descriptions.

Figure S6. **Qualitative results of Lumos-NVS**, where the leftmost one is the input view, the middle ones are the randomly sampled generated views, and the rightmost one is the reconstructed Gaussian.

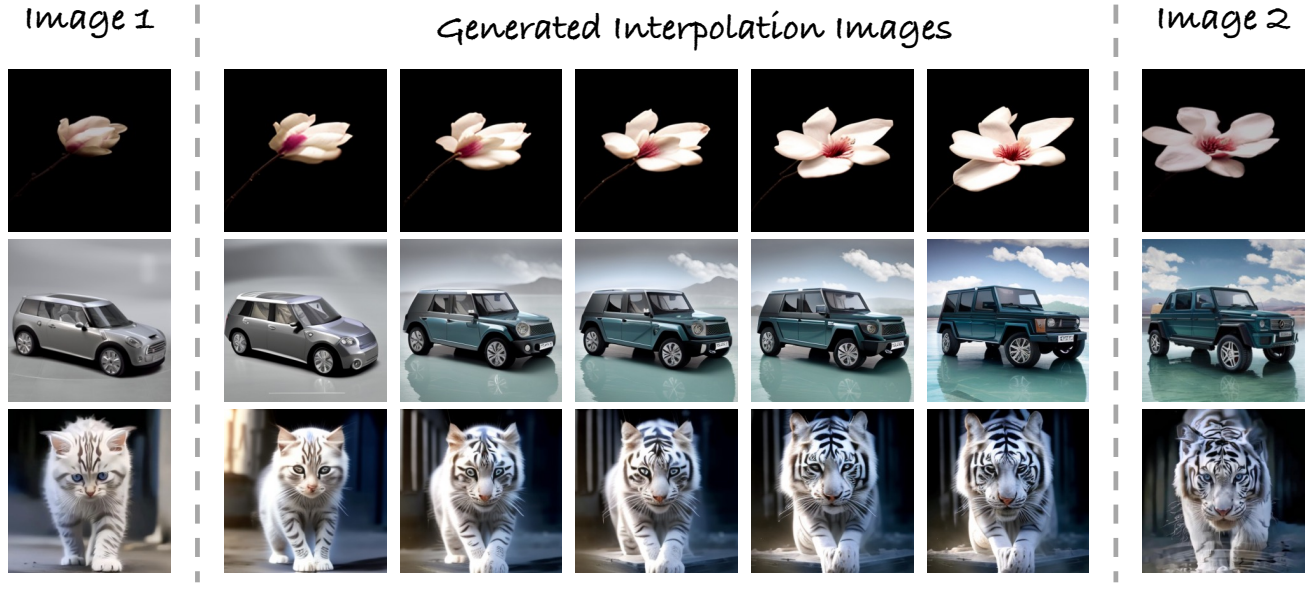


Figure S7. **Qualitative results of Lumos-I2I Interpolation**, where the leftmost and the rightmost ones are the input images and the middle ones are generated interpolation images.

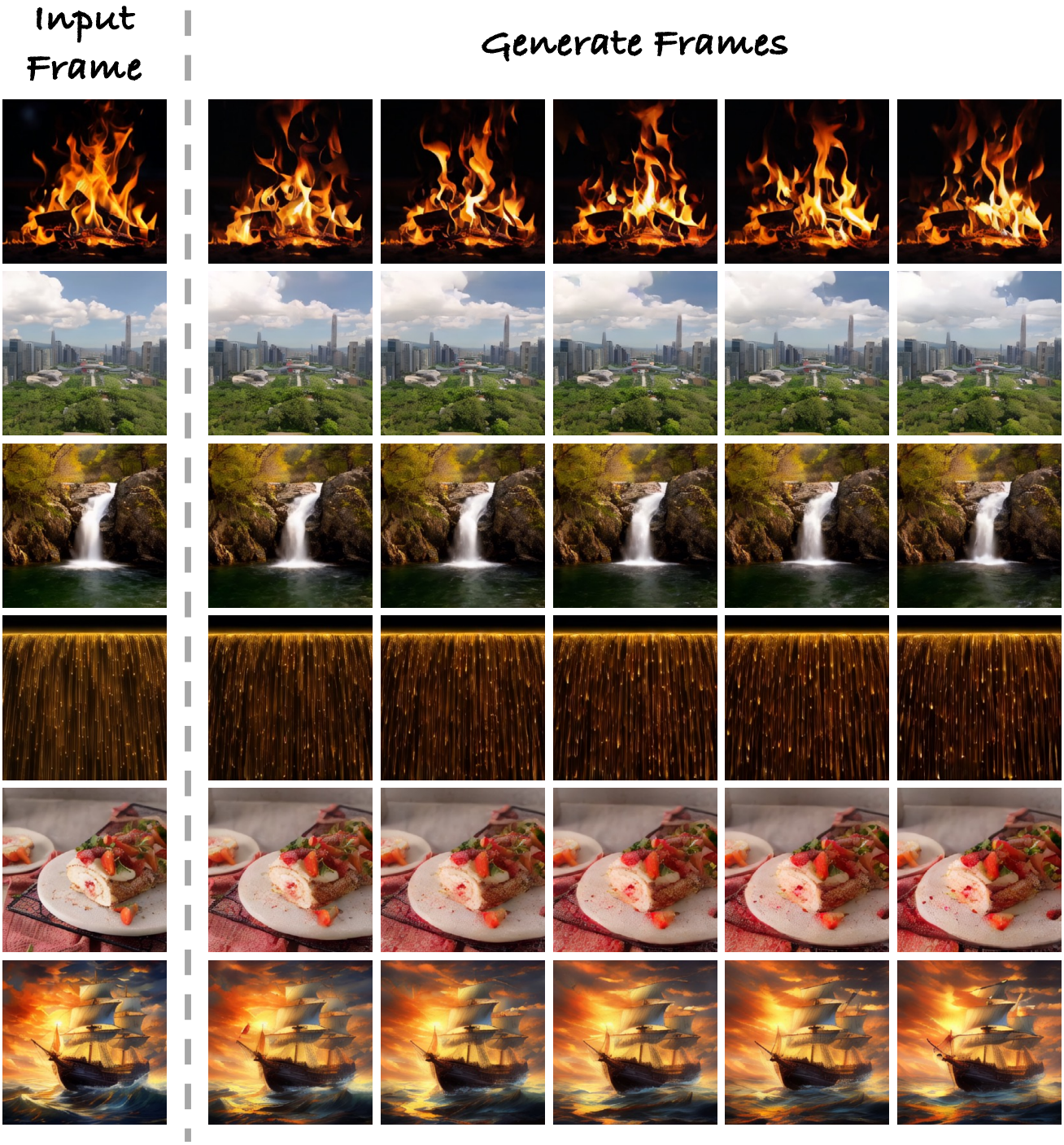


Figure S8. **Qualitative results of Lumos-I2V**, where the leftmost one is the input frame, right ones are generated frames.

I.4. Image Interpolation

We provide an application example of Lumos-I2I in image interpolation, as shown in Figure S7. Since Lumos-I2I can generate images that highly retain the original image information, it has a good effect on the image interpolation task.

J. Prompts in Figure 1a

We provide the text prompts adopted to generate images in Figure 1a. The prompts are arranged from top to bottom, left to right.

- *"golden sunset shines on the top of snow-capped mountains, with small villages at its foot and surrounding buildings."*
- *"A rustic bedroom showcasing a round bed, earth-toned decor, and a cluttered, yet charming ambiance."*
- *"Documentary-style photography of a bustling marketplace in Marrakech, with spices and textiles."*
- *"group characters from fantasy myth in the style of ori and the blind forest, riot games, ghibli, ori environment."*
- *"Post-Apocalyptic Wanderer, character design, style by kim jung gi, zabrocki, karlkka, jayison devadas, 8k."*
- *"The picture shows a cute little tiger, wearing a blue hoodie and hat, sitting on a small cardboard boat on calm water."*
- *"A dragon made of molten chocolate, with scales that glisten like gold leaf and eyes of crystalline sugar."*
- *"This professional photo from National Geography shows the subtleties in a erased face of god in the shape of the subtle cloud but we can clearly see the face of almighty god with this stormy atmosphere that is brewing in this Nevada desert, volumetric lighting, high contrast."*