A. Training Details

Multi-stage Training. Directly optimizing joint image-and-video training poses significant challenges, as the network must simultaneously learn spatial semantics critical for images and temporal motion dynamics essential for videos. To tackle this complexity, we introduce a decomposed, multi-stage training strategy that progressively enhances the model's capabilities, ensuring effective and robust learning across both modalities.

- **Stage-1: Text-Semantic Pairing.** In the initial stage, we focus on establishing a solid understanding of text-to-image relationships by pretraining Goku on text-to-image tasks. This step is critical for grounding the model in basic semantic comprehension, enabling it to learn to associate textual prompts with high-level visual semantics. Through this process, the model develops a reliable capacity for representing visual concepts essential for both image and video generation, such as object attributes, spatial configurations, and contextual coherence.
- **Stage-2: Image-and-video joint learning.** Building on the foundational capabilities of text-to-semantic pairing, we extend Goku to joint learning across both image and video data. This stage leverages the unified framework of Goku, which employs a global attention mechanism adaptable to both images and videos. Besides, acquiring a substantial volume of high-quality video data is generally more resource-intensive compared to obtaining a similar amount of high-quality image data. To address this disparity, our framework integrates images and videos into unified token sequences during training, enabling the rich information inherent in high-quality images to enhance the generation of video frames [12]. By curating a carefully balanced dataset of images and videos, Goku not only gains the capability to generate both high-quality images and videos but also enhances the visual quality of videos by leveraging the rich information from high-quality image data.
- **Stage-3: Modality-specific finetuning.** In the final stage, we fine-tune Goku for each specific modality to further enhance its output quality. For text-to-image generation, we implement image-centric adjustments aimed at producing more visually compelling images. For text-to-video generation, we focus on adjustments that improve temporal smoothness, motion continuity, and stability across frames, resulting in realistic and fluid video outputs.

Cascaded Resolution Training. In the second stage of joint training, we adopt a cascade resolution strategy to optimize the learning process. Initially, training is conducted on low-resolution image and video data (288×512), enabling the model to efficiently focus on fundamental text-semantic-motion relationships at reduced computational costs. Once these core interactions are well-established, the resolution of the training data is progressively increased, transitioning from 480×864 to 720×1280. This stepwise resolution enhancement allows Goku to refine its understanding of intricate details and improve overall image fidelity, ultimately leading to superior generation quality for both images and videos.

B. Benchmark Configurations

T2I-Compbench [39] We evaluate the alignment between the generated images and text conditions using T2I-Compbench, a comprehensive benchmark for assessing compositional text-to-image generation capabilities. Specifically, we report scores for color binding, shape binding, and texture binding. To evaluate these results, we employ the Disentangled BLIP-VQA model. For each attribute, we generate 10 images per prompt, with a total of 300 prompts in each category.

GenEval [28] GenEval is an object-focused framework designed to evaluate compositional image properties, such as object co-occurrence, position, count, and color. For evaluation, we generate a total of 2,212 images across 553 prompts. The final score is reported as the average across tasks.

DPG-Bench [38] Compared to the aforementioned benchmarks, DPGBench offers longer prompts with more detailed information, making it effective for evaluating compositional generation in text-to-image models. For this evaluation, we generate a total of 4,260 images across 1,065 prompts, with the final score reported as the average across tasks.

VBench [40] VBench is a benchmark suite for evaluating video generative models. It provides a structured Evaluation Dimension Suite that breaks down "video generation quality" into precise dimensions for detailed assessment. Each dimension and content category includes a carefully crafted Prompt Suite and samples Generated Videos from various models.

C. Data

C.1. Data Processing and Filtering

To construct a high-quality video dataset, we implement a comprehensive processing pipeline comprising several key stages. Raw videos are first preprocessed and standardized to address inconsistencies in encoding formats, durations, and frame rates. Next, a two-stage video clipping method segments videos into meaningful and diverse clips of consistent length. Additional filtering processes are applied, including visual aesthetic filtering to retain photorealistic and visually rich clips, OCR filtering to exclude videos with excessive text, and motion filtering to ensure balanced motion dynamics. In addition, the multi-level training data is segmented based on resolution and corresponding filtering thresholds for DINO similarity, aesthetic score, OCR text coverage, and motion score, as summarized in Table 3. We provide the details of each processing step as follows.

Table 7 presents the key parameters and their corresponding thresholds used for video quality assessment. Each parameter is essential in ensuring the generation and evaluation of high-quality videos. The Duration parameter specifies that raw video lengths should be at least 4 seconds to capture meaningful temporal dynamics. The Resolution criterion ensures that the minimum dimension (either height or width) of the video is no less than 480 pixels, maintaining adequate visual clarity. The Bitrate, which determines the amount of data processed per second during playback, requires a minimum of 500 kbps to ensure sufficient quality, clarity, and manageable file size. Videos with low bitrate typically correspond to content with low complexity, such as static videos or those featuring pure color backgrounds. Finally, the Frame Rate enforces a standard of at least 24 frames per second (film standard) or 23.976 frames per second (NTSC standard) to guarantee smooth motion and prevent visual artifacts. These thresholds collectively establish a baseline for evaluating and generating high-quality video content.

- **Preprocessing and Standardization of Raw Videos.** Videos collected from the internet often require extensive preprocessing to address variations in encoding formats, durations, and frame rates. Initially, we perform a primary filtering step based on fundamental video attributes such as duration, resolution, bitrate. The specific filtering criteria and corresponding thresholds are detailed in Table 7. This initial filtering step is computationally efficient compared to more advanced, model-based filtering approaches, such as aesthetic [67] evaluation models. Following this stage, the raw videos are standardized to a consistent coding format, H.264 [84], ensuring uniformity across the dataset and facilitating subsequent processing stages.
- **Video Clips Extraction.** We employ a two-stage video clipping method for this stage. First, we use PySceneDetect [10] for shot boundary detection, resulting coarse-grained video clips from raw videos. Next, we further refine the video clips by sampling one frame per second, generating DINOv2 [58] features and calculating cosine similarity between adjacent frames. When similarity falls below a set threshold, we mark a shot change and further divide the clip. Specifically, as shown in Table 3, for video resolutions around 480×864, we segmented the video clips where the DINO similarity between adjacent frames exceeds 0.85. For resolutions greater than 720×1280, the threshold is set at 0.9. Besides, to standardize length, we limit clips to a maximum of 10 seconds. Furthermore, we consider the similarity between different clips derived from the same source video to ensure diversity and maintain quality. Specifically, we compute the perceptual hashing [17] values of keyframes from each clip and compare them. If two clips have similar hash values, indicating significant overlap, we retain the clip with a higher aesthetic score. This ensures that the final dataset includes diverse and high-quality video clips.
- **Visual Aesthetic Filtering.** To assess the visual quality of the videos, we utilize aesthetic models [67] to evaluate the keyframes. The aesthetic scores of the keyframes are averaged to obtain an overall aesthetic score for each video. For videos with resolutions around 480×864, those with an aesthetic score below 4.3 are discarded, while for resolutions exceeding 720×1280, the threshold is raised to 4.5. This filtering process ensures that the selected clips are photorealistic, visually rich, and of high aesthetic quality.
- **OCR Filtering.** To exclude videos with excessive text, we employ an internal OCR model to detect text within the keyframes. The OCR model identifies text regions, and we calculate the text coverage ratio by dividing the area of the largest bounding box detected by the total area of the keyframe. Videos with a text coverage ratio exceeding predefined thresholds are discarded. Specifically, for videos with resolutions around 480×864, the threshold is set at 0.02, while for resolutions exceeding 720×1280, the threshold is reduced to 0.01. This process effectively filters out videos with excessive text content.
- **Motion Filtering.** Unlike images, videos require additional filtering based on motion characteristics. To achieve this, we utilize RAFT [75] to compute the mean optical flow of video clips, which is then used to derive a motion score. For videos with resolutions around 480×864, clips with motion scores below 0.3 (indicating low motion) or above 20.0 (indicating excessive motion) are excluded. For resolutions exceeding 720×1280, the thresholds are adjusted to 0.5 and 15.0, respectively. Furthermore, to enhance motion control, the motion score is appended to each caption.

Parameter	Description	Threshold
Duration	Raw video length	$\geq$ 4 seconds
Resolution	Width and height of the video	$min\{$ height, width$\} \geq 480$
Bitrate	Amount of data processed per second during playback, which impacts the video's quality, clarity, and file size	$\geq$ 500 kbps
Frame Rate	Frames displayed per second	$\geq$ 24 FPS (Film Standard) / 23.976 FPS (NTSC Standard)

Table 7. **Summary of video quality parameters and their thresholds for preprocessing.** The table outlines the criteria used to filter and standardize raw videos based on essential attributes, ensuring uniformity and compatibility in the dataset.

C.2. Training Data Balancing

The model's performance are significantly influenced by the data distribution, especially for video data. To balance the video training data, we first use an internal video classification model to generate semantic tags for the videos. We then adjust the data distribution based on these semantic tags to ensure a balanced representation across categories.

- **Data Semantic Distribution.** The video classification model assigns a semantic tag to each video based on four evenly sampled keyframes. The model categorizes videos into 9 primary classes (*e.g.*, human, scenery, animals, food) and 86 subcategories (*e.g.*, half-selfie, kid, dinner, wedding). Figure 5a presents the semantic distribution across our filtered training clips, with humans, scenery, food, urban life, and animals as the predominant categories.
- **Data Balancing.** The quality of the generated videos is closely tied to the semantic distribution of the training data. Videos involving humans pose greater modeling challenges due to the extensive diversity in appearances, whereas animals and landscapes exhibit more visual consistency and are relatively easier to model. To address this disparity, we implement a data-balancing strategy that emphasizes human-related content while ensuring equitable representation across subcategories within each primary category. Overrepresented subcategories are selectively down-sampled, whereas underrepresented ones are augmented through artificial data generation and oversampling techniques. Balanced data distribution is shown in Figure 5b.

D. More Visualization Examples

D.1. Goku-T2I Samples Visualization

We present more generated image samples with their text prompts in Figure 6. The prompts are randomly selected from the Internet [1]. Goku-T2I achieves strong performance in both visual quality and text-image alignment. It can interpret visual elements and their interactions from complex natural language descriptions. Notably, in Figure 7, Goku-T2I exhibits impressive abilities on generating images with rich details, for example, the clear textures of leaves and berries.

D.2. Goku-T2V Samples Visualization

In Figure 8 we show more examples generated by Goku-T2V, in both *landscape* (*e.g.*, rows one through five) and *portrait* mode (*e.g.*, the last row). Goku-T2V is capable of generating high-motion videos (*e.g.*, skiing) and realistic scenes (*e.g.*, forests). All videos are configured with a duration of 4 seconds, a frame rate of 24 FPS, and a resolution of 720p. For visualization, we uniformly sample five frames in temporal sequence.

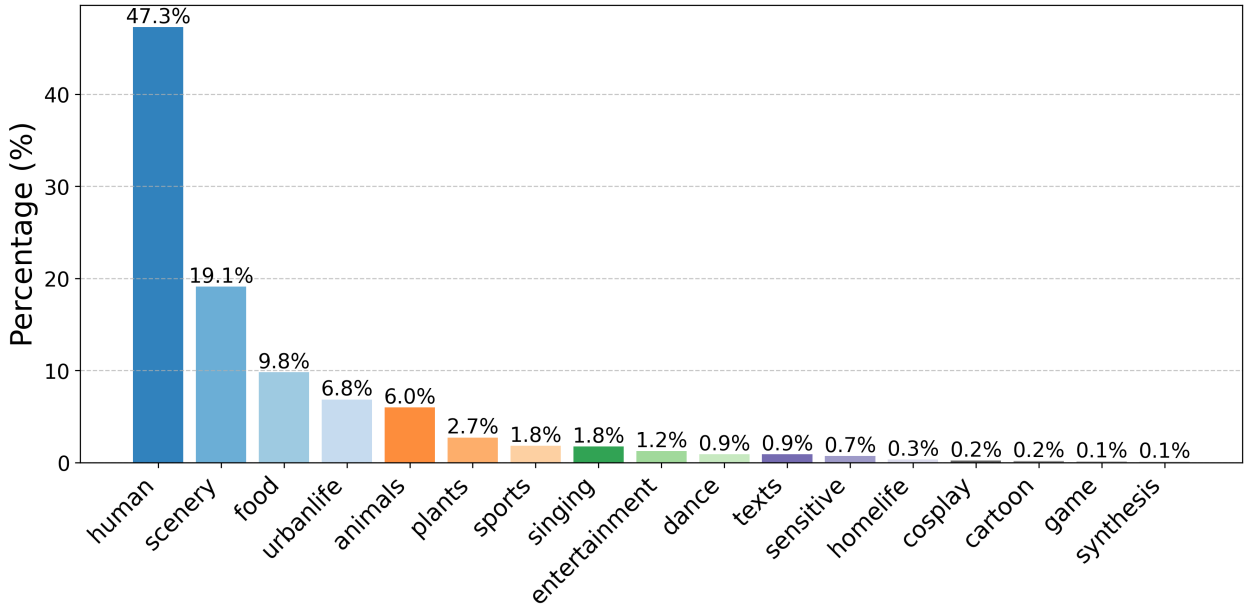
D.3. Goku-T2V Comparisons with Prior Arts

Additional comparisons with state-of-the-art text-to-video generation models are presented in Figure 9 and Figure 10. These results demonstrate the strong performance of Goku when evaluated against both open-source models [89, 95] and commercial products [4, 49, 55, 60]. For instance, in Figure 10, Goku successfully generates smooth motion and accurately incorporates the specified low-angle shot. In contrast, other models, such as CogVideoX [89], Vidu [4], and Kling [49], often produce incorrect objects or improper camera views.

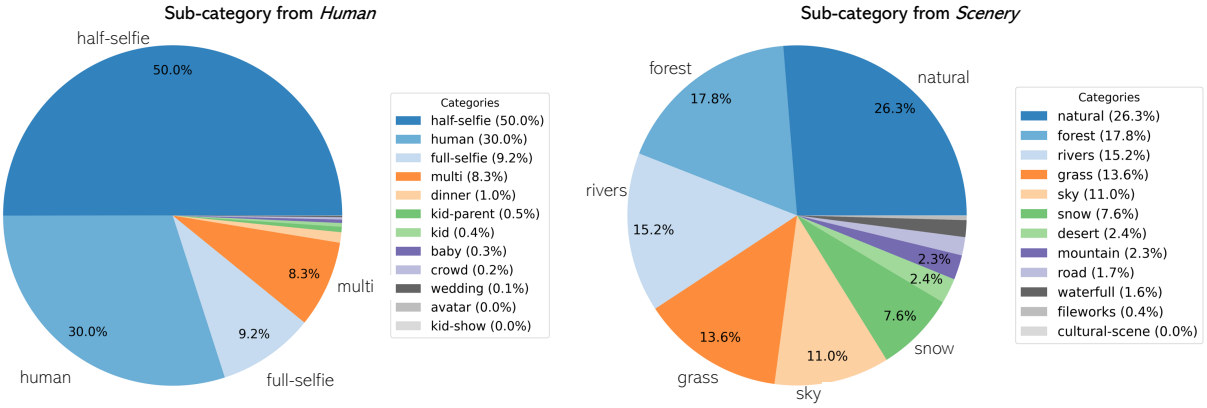
[1] https://promptlibrary.org/

Method	*Total Score*	*Quality Score*	*Semantic Score*	*subject consistency*	*background consistency*	*temporal flickering*	*motion smoothness*	*dynamic degree*	*aesthetic quality*	*imaging quality*	*object class*	*multiple objects*	*human action*	*color*	*spatial relationship*	*scene*	*appearance style*	*temporal style*	*overall consistency*
AnimateDiff-V2	80.27	82.90	69.75	95.30	97.68	98.75	97.76	40.83	67.16	70.10	90.90	36.88	92.60	87.47	34.60	50.19	22.42	26.03	27.04
VideoCrafter-2.0	80.44	82.20	73.42	96.85	**98.22**	98.41	97.73	42.50	63.13	67.22	92.55	40.66	95.00	**92.92**	35.86	55.29	**25.13**	25.84	**28.23**
OpenSora V1.2	79.23	80.71	73.30	94.45	97.90	99.47	98.20	47.22	56.18	60.94	83.37	58.41	85.80	87.49	67.51	42.47	23.89	24.55	27.07
Show-1	78.93	80.42	72.98	95.53	98.02	99.12	98.24	44.44	57.35	58.66	93.07	45.47	95.60	86.35	53.50	47.03	23.06	25.28	27.46
Gen-3	82.32	84.11	75.17	97.10	96.62	98.61	99.23	60.14	63.34	66.82	87.81	53.64	96.40	80.90	65.09	54.57	24.31	24.71	26.69
Pika-1.0	80.69	82.92	71.77	96.94	97.36	**99.74**	**99.50**	47.50	62.04	61.87	88.72	43.08	86.20	90.57	61.03	49.83	22.26	24.22	25.94
CogVideoX-5B	81.61	82.75	77.04	96.23	96.52	98.66	96.92	70.97	61.98	62.90	85.23	62.11	99.40	82.81	66.35	53.20	24.91	25.38	27.59
Kling	81.85	83.39	75.68	**98.33**	97.60	99.30	99.40	46.94	61.21	65.62	87.24	68.05	93.40	89.90	73.03	50.86	19.62	24.17	26.42
Mira	71.87	78.78	44.21	96.23	96.92	98.29	97.54	60.33	42.51	60.16	52.06	12.52	63.80	42.24	27.83	16.34	21.89	18.77	18.72
CausVid	84.27	**85.65**	78.75	97.53	97.19	96.24	98.05	**92.69**	64.15	68.88	92.99	72.15	**99.80**	80.17	64.65	56.58	24.27	25.33	27.51
Luma	83.61	83.47	**84.17**	97.33	97.43	98.64	99.35	44.26	65.51	66.55	**94.95**	**82.63**	96.40	92.33	83.67	**58.98**	24.66	**26.29**	28.13
HunyuanVideo	83.24	85.09	75.82	97.37	97.76	99.44	98.99	70.83	60.36	67.56	86.10	68.55	94.40	91.60	68.68	53.88	19.80	23.89	26.44
Goku	**84.85**	85.60	81.87	95.55	96.67	97.71	98.50	76.11	**67.22**	**71.29**	94.40	79.48	97.60	83.81	**85.72**	57.08	23.08	25.64	27.35

Table 8. **Comparison with state-of-the-art models on video generation benchmarks.** We evaluate on VBench [40] and compare with Gen-3 [66], Vchitect-2.0 [74], VEnhancer [32], Kling [49], LaVie-2 [82], CogVideoX [89], Emu3 [81].



(a) Semantic distribution of video clips.



(b) The balanced semantic distribution of subcategories.

Figure 5. **Training data distributions.** The balanced semantic distribution of primary categories and subcategories are shown in (a) and (b), respectively.

D.4. Goku-I2V Samples Visualization

We present additional visualization of generated samples from Goku-I2V in Figure 11, which further validate the effectiveness and versatility of our approach. As shown in the figure, Goku-I2V demonstrates an impressive ability to synthesize coherent and visually compelling videos from diverse reference images, maintaining consistency in motion and scene semantics.

For instance, in the first row, the model successfully captures the dynamic and high-energy nature of water boxing, generating fluid and natural movements of splashes synchronized with the subject's motions. In the second row, the sequence of a child riding a bike through a park illustrates the model's proficiency in creating smooth and realistic forward motion while preserving environmental consistency. Finally, the third row showcases the model's ability to handle creative and imaginative scenarios, as seen in the detailed depiction of pirate ships battling atop a swirling coffee cup. The photorealistic rendering and accurate motion trajectories underscore the model's robustness in both realism and creativity.

These examples highlight Goku-I2V's capacity to generalize across a wide range of inputs, reinforcing its potential for applications in video generation tasks requiring high fidelity and adaptability.

E. Acknowledgements

We sincerely appreciate the support of our collaborators at ByteDance who contributed to this work. Xibin Wu, Chongxi Wang, Yina Tang, Fangzhou Ai, Yi Ren, Wei Wang, Chen Chen, Colin Young, Bobo Zeng, Ge Bai, Yi Fu, Ruoyu Guo, Prasanna Raghav, Weiguo Feng, Xugang Ye, Adithya Sampath, Aaron Shen, Da Tang, Yuan Fang, Qijun Gan, Chen Zhang, Zhenhui Ye, Pan Xie, Houmin Wei, Gaohong Liu, Zherui Liu, Chenyuan Wang, Yun Zhang, Kaihua Jiang, Zhuo Jiang, Yang Bai, Weiqiang Lou, Hongkai Li, Xi Yang, Shuguang Wang, Junru Zheng, Zuquan Song, Zixian Du, Jingzhe Tang, Yongqiang Zhang, Mingji Han, Heng Zhang, Li Han, Sophie Xie, Shuo Li, Xinzhi Yao, Peng Li, Lianke Qin, Dongyang Wang, Yang Cheng, Chundian Liu, Wenhao Hao, Haibin Lin, Xin Liu

A portrait featuring a 26-year-old Chinese male model in a six-grid layout. He has a sleek, naturally layered Korean hairstyle with subtly drooping bangs. Each panel shows him wearing modern

An embroidered sweater with an anatomical illustration of the human torso and chest, the skin is open to reveal the internal anatomy.

Prototype flying fox made from blown glass, Lino Tagliapietra style Muranese glassmaking, intricate details.

3d cube woman underwater, iridescent water, dreamlike

Close up shot of hand of a woman touching oats in oat farm. Shot from behind.

3D illustration of the chip with text "AI" floating above it, with a blue color scheme.

A simple design in black on a white background. The word "VINTAGE" is at the bottom.

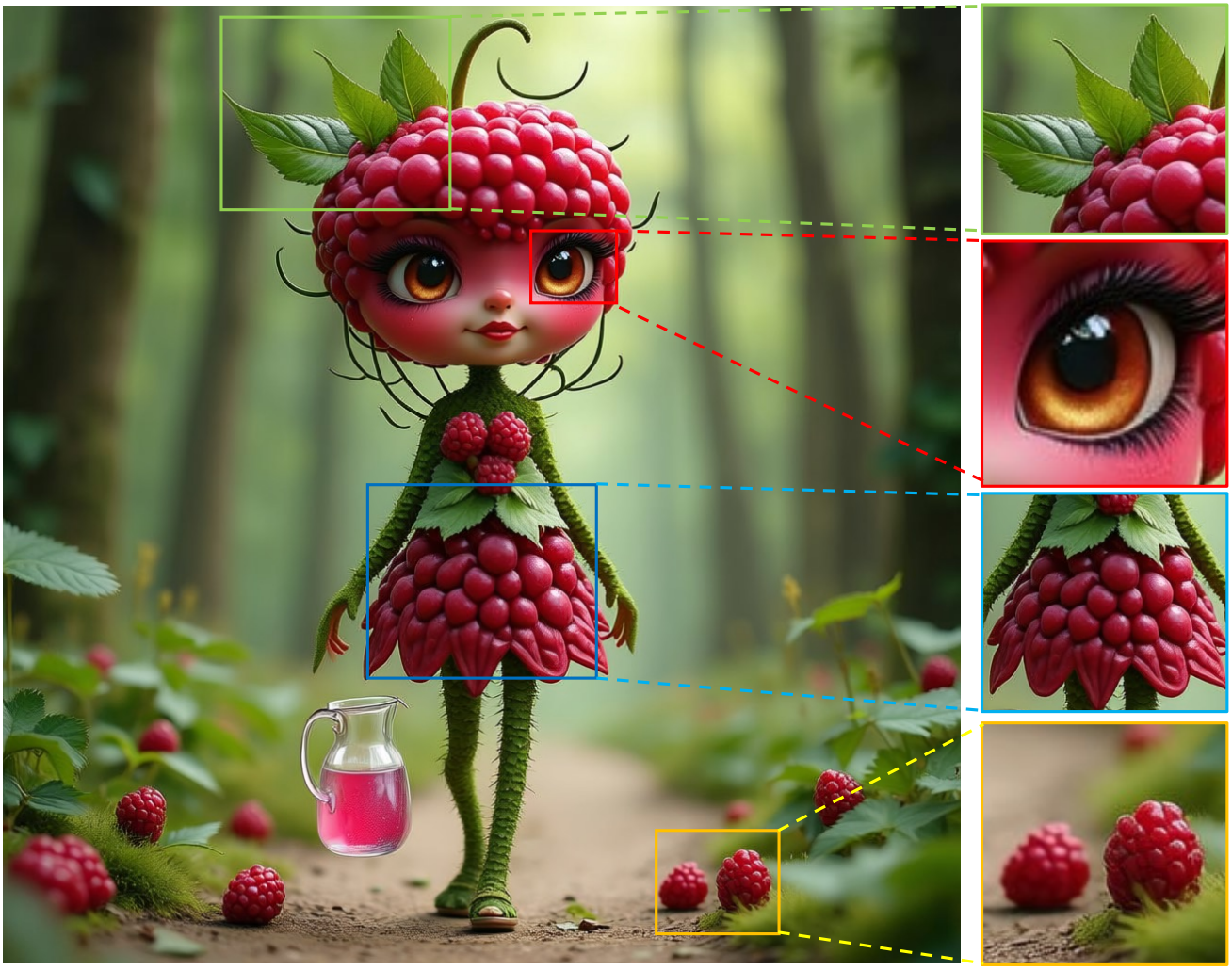
Great Dane Dog sitting on a toilet bowl in wide bathroom, reading a large double page spread newspaper, sit like human. The background is in a white room.

Full body shot of balenciaga fashion model and parrot hybrid with a human body and the head of the parrot. He is walking through a podium like a model.

Full body photo of a screaming cauliflower monster roaring towards the viewer, very detailed textures. The background is clean and blue.

Create realistic playing cards on fire. The playing cards are presented with 4A. The fire is red and intense. The background is black.

Figure 6. **Qualitative samples of Goku-T2I.** Key words are highlighted in RED.



Prompt : Raspberry in the form of women walk along the path of a fairy tale forest. She carries a jug of water with her. Her head is made of one big raspberry on which she has big and beautiful eyes, as well as nose and mouth. The skin of the face has a raspberry color. She has very beautiful hair which consists of raspberry, leaves and thin stems. Her arms and legs are made entirely of intertwined stems. She also wears a skirt with raspberry leaves and small raspberries and she looks very delicate and feminine.

Figure 7. **Qualitative samples of Goku-T2I.** Key words are highlighted in RED. For clarity, we zoom in on specific regions to enhance visualization.

At an aquarium, a diver in a yellow wetsuit is feeding tropical fish in a large tank.

Zooming through a dense, lush rainforest at incredible speed, weaving between colossal trees, with rays of sunlight breaking through the canopy and exotic birds scattering in the distance.

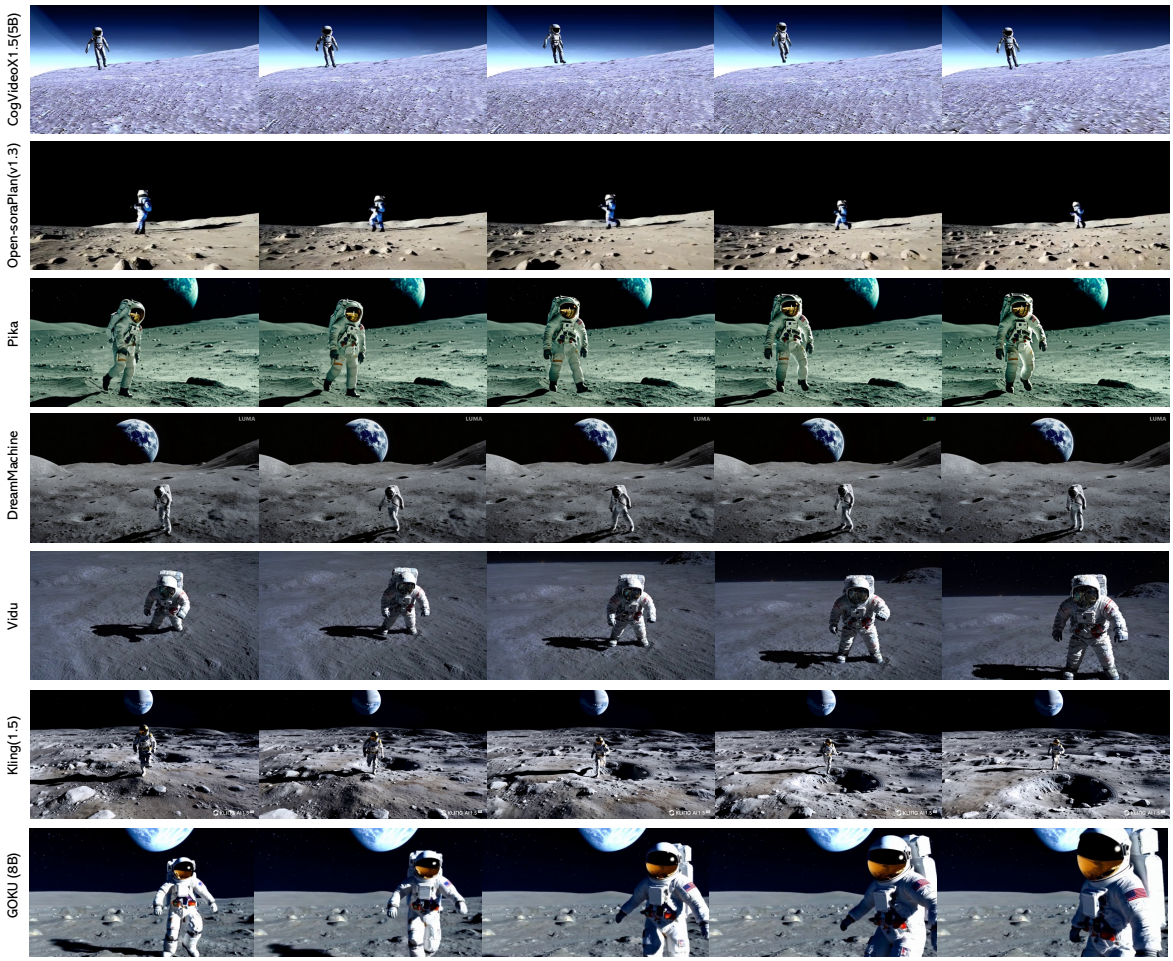
A snowboarder carves down a steep slope, their board cutting swiftly through the snow.

A boxer dances around the ring, fists raised and jabbing rapidly at their opponent.

A kung fu master swiftly maneuvers through a series of rapid punches and palm strikes, their arms blurring with speed.

In a cozy living room with a roaring fireplace and plush furniture, a dog with a shiny coat sits contentedly on a soft rug.

Figure 8. **Qualitative samples of Goku-T2V.** Key words are highlighted in **RED**.



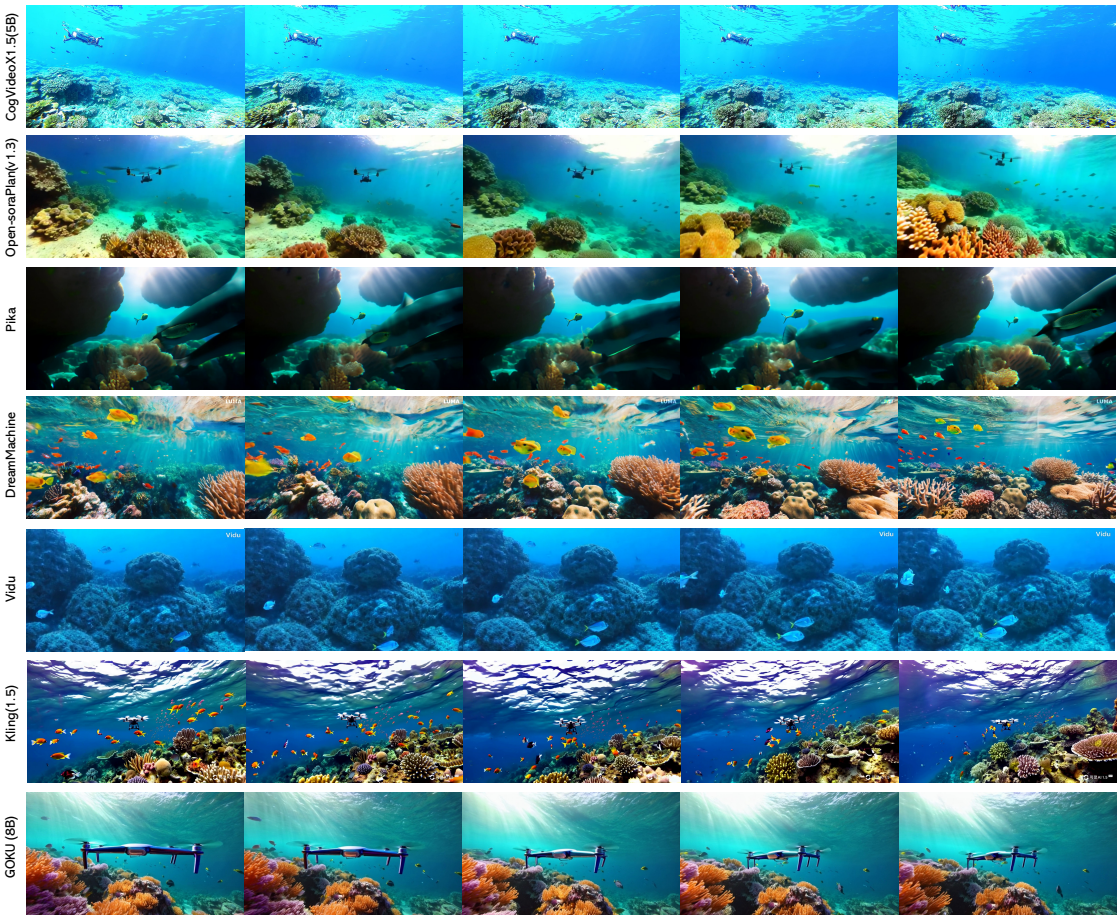
Prompt： An astronaut runs across the surface of the moon, with a low-angle shot showcasing the vast lunar landscape. The movements are smooth and light.

Figure 9. **Qualitative comparisons of Goku-T2V with SOTA video generation methods.** Key words are highlighted in **RED**.

Prompt： A man surfing on a wave, with the camera following his movement and focusing on his face. He is smiling and giving a thumbs-up to the camera, conveying a sense of enjoyment and excitement. The ocean waves are vibrant and dynamic around him, with sunlight glistening on the water. The background features a clear blue sky, enhancing the lively atmosphere of the scene as he rides the waves with confidence and enthusiasm.

Figure 10. **Qualitative comparisons of Goku-T2V with SOTA video generation methods.** Key words are highlighted in **RED**.

Figure 11. **Qualitative samples of Goku-I2V.** Key words are highlighted in **RED**.



Prompt : *Gliding through a crystal-clear coral reef, the drone skims just above the vibrant marine life below. Brightly colored corals, schools of fish, and rays of sunlight penetrating the water's surface all contribute to the serene yet fast-paced journey. The scene showcases the beauty of the underwater world, as the drone swiftly maneuvers through coral arches and narrow underwater channels.*

Figure 12. **Qualitative comparisons with state-of-the-art (SoTA) video generation models.** This figure showcases comparisons with leading models, including [89], Open-Sora Plan [50], Pika [60], DreamMachine [55], Vidu [4], and Kling v1.5 [49].