QWEN2VL-FLUX: UNIFYING IMAGE AND TEXT GUIDANCE FOR CONTROLLABLE IMAGE GENERATION

TECHNICAL REPORT

Pengqi Lu
erwold901@gmail.com

ABSTRACT

Recent advances in text-to-image diffusion models have demonstrated remarkable capabilities in generating high-quality images. However, these models often struggle with maintaining semantic consistency and style fidelity when incorporating reference images, primarily due to the limited multi-modal understanding of their text encoders. In this work, a novel approach is proposed that replaces the conventional T5-XXL text encoder in the Flux architecture with a vision-language model, enabling sophisticated multi-modal understanding and generation capabilities. This architectural innovation yields three significant advantages: (1) zero-shot style transfer ability, where the model can generate diverse images that preserve key features and styles from a single reference image without text prompts; (2) semantic-aware generation through the proposed GridDot Panel mechanism, which allows users to dynamically control attention weights on specific regions of the reference image, enabling fine-grained style and semantic transfer; and (3) cross-modal style-content fusion, where the model can seamlessly blend semantic elements from text prompts with visual styles from reference images. Specifically, the Qwen2VL-7B model is utilized as the vision-language model to replace the T5-XXL text encoder in FLUX.1-dev, successfully achieving enhanced multi-modal understanding and generation capabilities. The finetuned model is available at `https://huggingface.co/Djrango/Qwen2vl-Flux`, and the inference code can be accessed at `https://github.com/erwold/qwen2vl-flux`.

1 Introduction

Recent years have witnessed remarkable progress in text-to-image generation, primarily driven by the advancement of diffusion models Ramesh et al. [2021, 2022], Saharia et al. [2022], Esser et al. [2021], Rombach et al. [2022] and the emergence of diffusion transformers Peebles and Xie [2022]. These models have demonstrated unprecedented capabilities in generating high-quality images from textual descriptions. However, despite their success, they face two significant challenges: (1) maintaining semantic consistency and style fidelity when incorporating reference images, and (2) providing fine-grained control over the generation process. These limitations stem primarily from the restricted multi-modal understanding capabilities of their text encoders, which struggle to effectively bridge the semantic gap between textual descriptions and visual features.

Several approaches have been proposed to address these challenges. One line of research focuses on improving the text encoder's capabilities, with models like Stable Diffusion 3 Esser et al. [2024] and FLUX.1 Labs [2024] incorporating T5-XXL alongside CLIP for better text understanding. Another direction explores reference-based generation through various control mechanisms Ye et al. [2023], Zhang et al. [2023], Huang et al. [2023], such as spatial conditions or attention manipulation. However, these solutions either require extensive training of additional control modules or fail to achieve precise semantic transfer while maintaining style consistency.

In this work, a fundamentally different approach is proposed by replacing the conventional text encoder with a vision-language model (VLM) in the FLUX architecture. This architectural change is motivated by two key insights: (1) VLMs possess inherent capabilities for understanding both textual and visual information in a unified semantic space, and (2) the attention mechanism in VLMs can be leveraged for fine-grained control over the generation process. This approach not only enhances the model's ability to understand and generate images but also enables precise control over both style and semantic elements.

A key innovation in this work is the GridDot Panel mechanism, which allows users to dynamically control attention weights on specific regions of the reference image. This mechanism operates by creating an interactive interface where users can manipulate attention maps through a grid of control points, effectively guiding the model's focus during generation. Unlike previous approaches that require complex spatial conditions or extensive training, this method provides intuitive, real-time control over the generation process while maintaining the coherence of the generated images.

Through extensive experiments, three significant capabilities of the model are demonstrated: (1) zero-shot style transfer, where the model can generate diverse images that preserve key features and styles from a reference image without any text prompts; (2) semantic-aware generation with precise control through the GridDot Panel; and (3) cross-modal style-content fusion, enabling seamless blending of semantic elements from text prompts with visual styles from reference images. These capabilities are achieved without any additional training or fine-tuning, making this approach both efficient and practical.

The main contributions can be summarized as follows:

- A novel architecture is proposed that replaces the traditional text encoder with a vision-language model in diffusion transformers, enabling unified multi-modal understanding and generation.
- The GridDot Panel mechanism is introduced for intuitive and precise control over attention weights, allowing fine-grained manipulation of style and semantic transfer.
- State-of-the-art performance is demonstrated in maintaining both semantic consistency and style fidelity across various generation tasks, without requiring additional training or control modules.

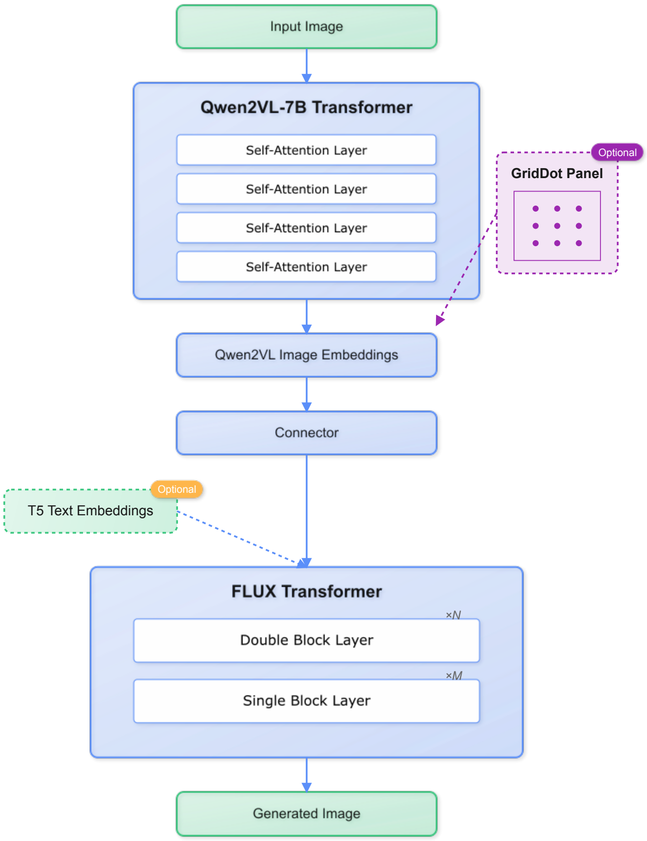


Figure 1: **Architecture Overview of Qwen2VL-Flux.** The model replaces the conventional T5-XXL text encoder in FLUX with Qwen2VL-7B, enabling sophisticated multi-modal understanding and generation capabilities through unified semantic space and attention mechanisms.

2 Method

In this section, Qwen2VL-Flux is presented, a novel approach that enhances the generation capabilities of the FLUX architecture by integrating a vision-language model (VLM). This method consists of three key components: (1) a unified multi-modal encoder that replaces the traditional T5-XXL text encoder, (2) a GridDot Panel mechanism for fine-grained attention control, and (3) a cross-modal style-content fusion approach. Each component is detailed below.

2.1 Unified Multi-modal Architecture

2.1.1 Vision-Language Model Integration

Unlike conventional text-to-image models that rely on separate text encoders (e.g., T5-XXL) and vision encoders (e.g., CLIP), the conventional text encoder is replaced with a single vision-language model, specifically Qwen2VL-7B. This architectural change is motivated by two key insights:

- VLMs possess inherent capabilities for understanding both textual and visual information in a unified semantic space, enabling better alignment between image features and textual descriptions.
- The attention mechanisms in VLMs can be leveraged to provide fine-grained control over the generation process, allowing for more precise style and semantic transfer.

The model architecture consists of three main components (Figure 1):

- A Qwen2VL encoder that processes both images and text prompts
- A connector network that bridges the semantic gap between VLM and image generation

- A FLUX transformer that generates images based on the unified representations

The Qwen2VL encoder processes input images and optional text prompts through a series of transformer layers, producing rich multi-modal representations. These representations are then processed by a lightweight connector network that projects them into the FLUX transformer's input space. This design allows seamless integration of visual and textual features while maintaining the powerful generation capabilities of the FLUX architecture.

2.1.2 Connector Network

To bridge the semantic gap between the VLM's output space and the FLUX transformer's input space, a connector network is introduced. This network consists of a linear projection layer followed by normalization:

$$h_{flux} = \text{Norm}(W \cdot h_{vlm}) \tag{1}$$

where $h_{vlm} \in \mathbb{R}^{b \times n \times d_{vlm}}$ is the VLM's output, $W \in \mathbb{R}^{d_{vlm} \times d_{flux}}$ is a learnable projection matrix, and $h_{flux} \in \mathbb{R}^{b \times n \times d_{flux}}$ is the transformed representation compatible with the FLUX transformer.

2.2 Zero-shot Style Variation

A key advantage of the unified architecture is its ability to perform zero-shot style transfer without explicit text prompts. When presented with a reference image, the model:

1. Processes the image through the Qwen2VL encoder to extract style-relevant features
2. Uses the connector network to project these features into the FLUX transformer's space
3. Guides the generation process using the extracted style information while varying content through different noise seeds

This allows the model to generate diverse variations that preserve key stylistic elements of the reference image, such as color palette, lighting, and artistic technique, while exploring different compositional possibilities.

2.3 GridDot Panel for Semantic-Aware Generation

The GridDot Panel mechanism is introduced, which provides intuitive control over the attention weights applied to different regions of the reference image. This mechanism consists of:

2.3.1 Spatial Hidden State Recovery

Given the image hidden states $H \in \mathbb{R}^{B \times N \times D}$ from the VLM, where B is the batch size, N is the number of tokens, and D is the hidden dimension, the spatial hidden states are first recovered. For an input image with grid dimensions $h \times w$, the hidden states are reshaped to:

$$H_{spatial} \in \mathbb{R}^{B \times (h/2) \times (w/2) \times D} \tag{2}$$

where the spatial dimensions are downsampled by a factor of 2 to match the VLM's internal representation.

2.3.2 Radial Attention Generation

For each control point (x_c, y_c) specified through the GridDot Panel interface, a radial attention matrix $A \in \mathbb{R}^{h/2 \times w/2}$ is generated where each element a_{ij} is computed as:

$$a_{ij} = \max(0, 1 - \sqrt{(i - y_c)^2 + (j - x_c)^2}/(r \cdot \min(h, w))) \tag{3}$$

where:

- (i, j) are spatial coordinates in the attention grid
- r is a radius parameter controlling the extent of attention falloff
- The coordinates are normalized by the minimum dimension $\min(h, w)$

2.3.3 Attention Application

The attention weights are applied to the spatial hidden states through element-wise multiplication:

$$H_{attended} = H_{spatial} \odot A_{expanded} \tag{4}$$

where $A_{expanded} \in \mathbb{R}^{1\times(h/2)\times(w/2)\times 1}$ is the attention matrix expanded to match the hidden state dimensions. The attended hidden states are then reshaped back to the sequence format:

$$H_{out} \in \mathbb{R}^{B\times N\times D} \tag{5}$$

This mechanism enables precise spatial control over which regions of the reference image influence the generation process.

2.4 Cross-modal Feature Integration

The architecture achieves cross-modal generation through concatenative feature fusion of visual and textual embeddings in a shared latent space.

2.4.1 Feature Concatenation

Given image embeddings $E_{img} \in \mathbb{R}^{B\times N_i\times D}$ from Qwen2VL and text embeddings $E_{txt} \in \mathbb{R}^{B\times N_t\times D}$ from T5-XXL, where N_i and N_t are the respective sequence lengths, feature concatenation is performed along the sequence dimension:

$$E_{combined} = [E_{img}; E_{txt}] \in \mathbb{R}^{B\times(N_i+N_t)\times D} \tag{6}$$

2.4.2 Unified Processing

The FLUX transformer processes $E_{combined}$ through its self-attention layers:

$$h_l = \text{TransformerLayer}_l(h_{l-1}), h_0 = E_{combined} \tag{7}$$

where h_l represents the hidden states at layer l. This allows the model to learn cross-modal interactions through its attention mechanisms, naturally balancing visual style from the reference image with semantic content from the text prompt.

3 Experiments

3.1 Implementation Details

3.1.1 Model Configuration

The model is built upon FLUX.1-dev as the base architecture, replacing its T5-XXL text encoder with Qwen2VL-7B. The connector network consists of a linear projection layer that maps from Qwen2VL's 3584-dimensional hidden space to FLUX's 4096-dimensional space. A staged training strategy is employed for different components:

1. **Qwen2VL**: The last layer of Qwen2VL (k=1) is made trainable to fine-tune high-level visual understanding while keeping most of the pre-trained weights frozen.
2. **FLUX Transformer**: A three-stage progressive training approach is adopted for the FLUX transformer blocks:
   Stage 1 - Flux Transformer: Trainable layers: [1, 4, 7, 12, 16, 19, 24, 28, 32, 37, 40, 44, 48, 52, 56]
   Stage 2 - Flux Transformer: Trainable layers: [0, 3, 6, 9, 14, 16, 21, 26, 30, 34, 39, 42, 46, 50, 54, 56]
   Stage 3 - Flux Transformer: Trainable layers: [2, 5, 8, 11, 15, 18, 23, 25, 29, 33, 36, 41, 45, 49, 53, 55]
3. **Connector Network**: The entire connector network remains trainable throughout all stages to maintain effective feature space bridging.

3.1.2 Training Procedure

Data Processing: Following FLUX, multiple aspect ratios are supported during training, including 16:9, 2:1, 2.35:1, 1:1, and 1:2. Images are automatically resized while maintaining their aspect ratios, with a maximum dimension of 1024 pixels. For each training step, a batch of reference images is processed through both Qwen2VL and the FLUX pipeline.

Optimization: The model is trained using AdamW optimizer with the following hyperparameters:

- Learning rates: 1e-5 for FLUX layers, connector network, and Qwen2VL layers
- Batch size: 128 (distributed across 8 GPUs)
- Gradient accumulation steps: 2
- Weight decay: 0.1
- Adam β_1, β_2 = (0.9, 0.95)

The learning rate follows a trapezoidal schedule with a brief warmup period and linear decay.

Training Objective: The original FLUX training objective based on rectified flow matching is maintained. For a noisy input x_t and target x_0, the loss is computed as:

L = $\|f_\theta(x_t, t) - (x_0 - x_t)\|^2$

where f_θ is the model's denoising prediction and t is sampled from the noise schedule. Importantly, no additional loss terms are introduced, as the unified architecture naturally learns to balance style and semantic information through the self-attention mechanisms.

Multi-GPU Training: Distributed training is implemented using DistributedDataParallel (DDP) across 8 GPUs. The effective batch size is achieved through a combination of per-GPU batches and gradient accumulation:

Effective batch size = num_gpus * per_gpu_batch_size * grad_accum_steps = 8 * 8 * 2 = 128

3.1.3 Training Infrastructure

The model was trained on 8 NVIDIA A100 GPUs with 80GB memory each. To optimize memory usage, the following techniques are employed:

- Mixed precision training with bfloat16
- Gradient checkpointing in both Qwen2VL and FLUX

Training runs for 100K updates on a dataset of 3 million high-quality images curated by the author, taking approximately 1 month on the hardware configuration.

4 Experimental Results

Acknowledgments

I would like to express my sincere gratitude to the black-forest-lab team for open-sourcing their high-quality base model, which has been instrumental in this research. I also extend my appreciation to the Hugging Face community for their valuable contributions in adapting inference code and exploring training methodologies. Special thanks to Claude-3.5-Sonnet-New for the assistance in completing this technical report.

The author is currently seeking employment opportunities. Interested parties can reach out via email at erwold901@gmail.com.

References

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

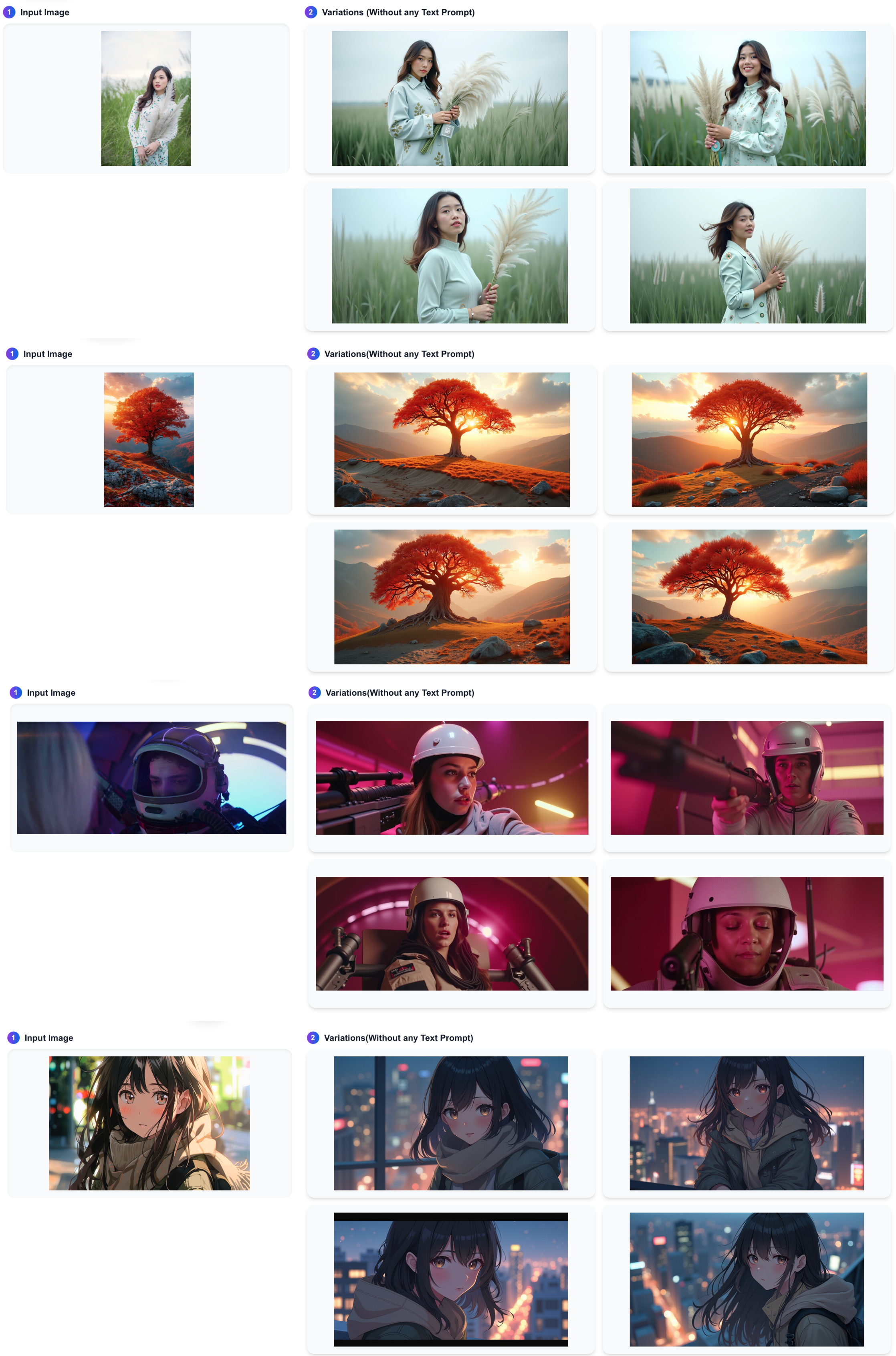
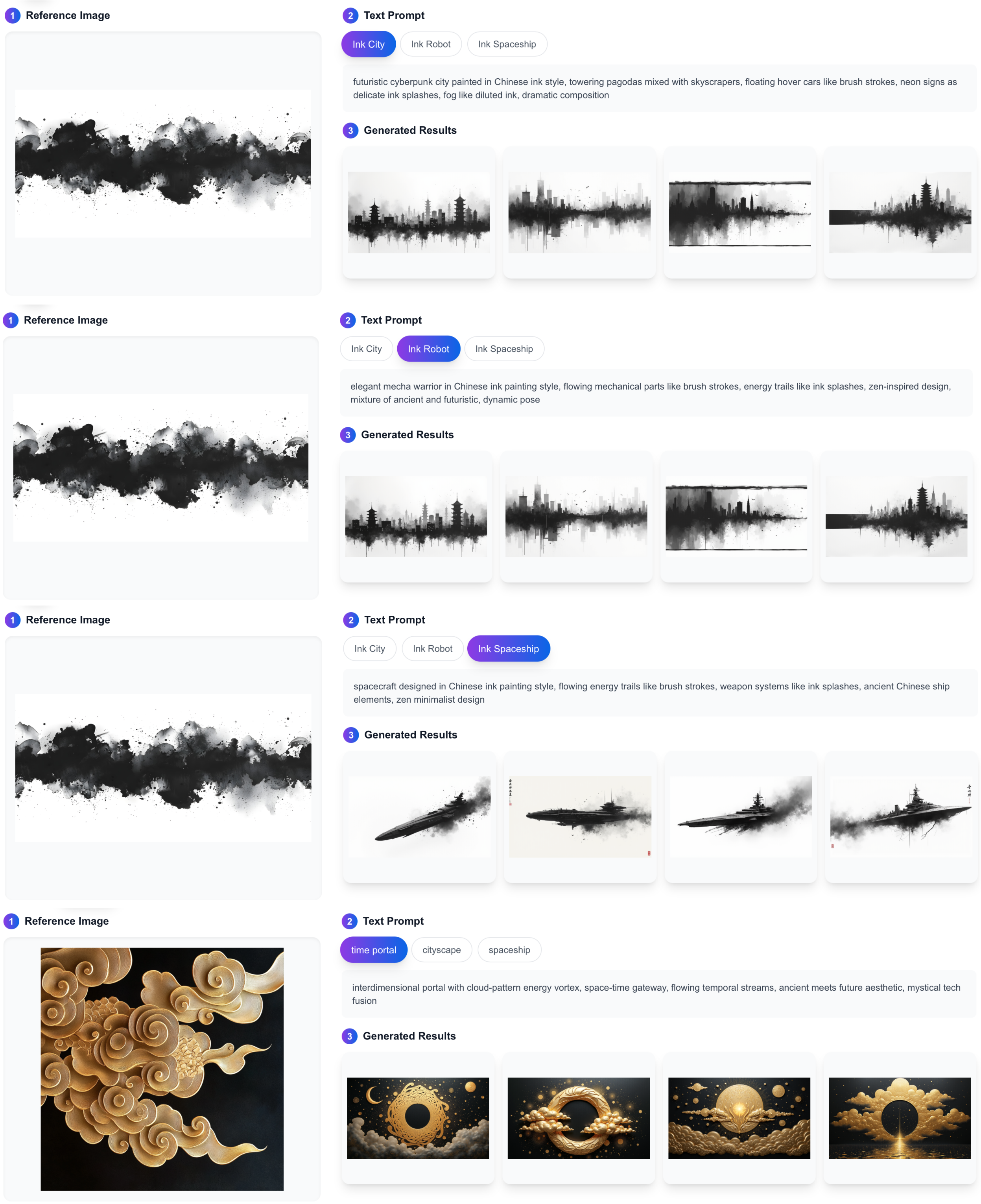

1 Input Image
2 Variations (Without any Text Prompt)
1 Input Image
2 Variations(Without any Text Prompt)
1 Input Image
2 Variations(Without any Text Prompt)
1 Input Image
2 Variations(Without any Text Prompt)

Figure 2: Image variation results using only Qwen2VL-7B image embeddings without text prompts. The model demonstrates strong capability in preserving key features and style of the input images while generating diverse variations.



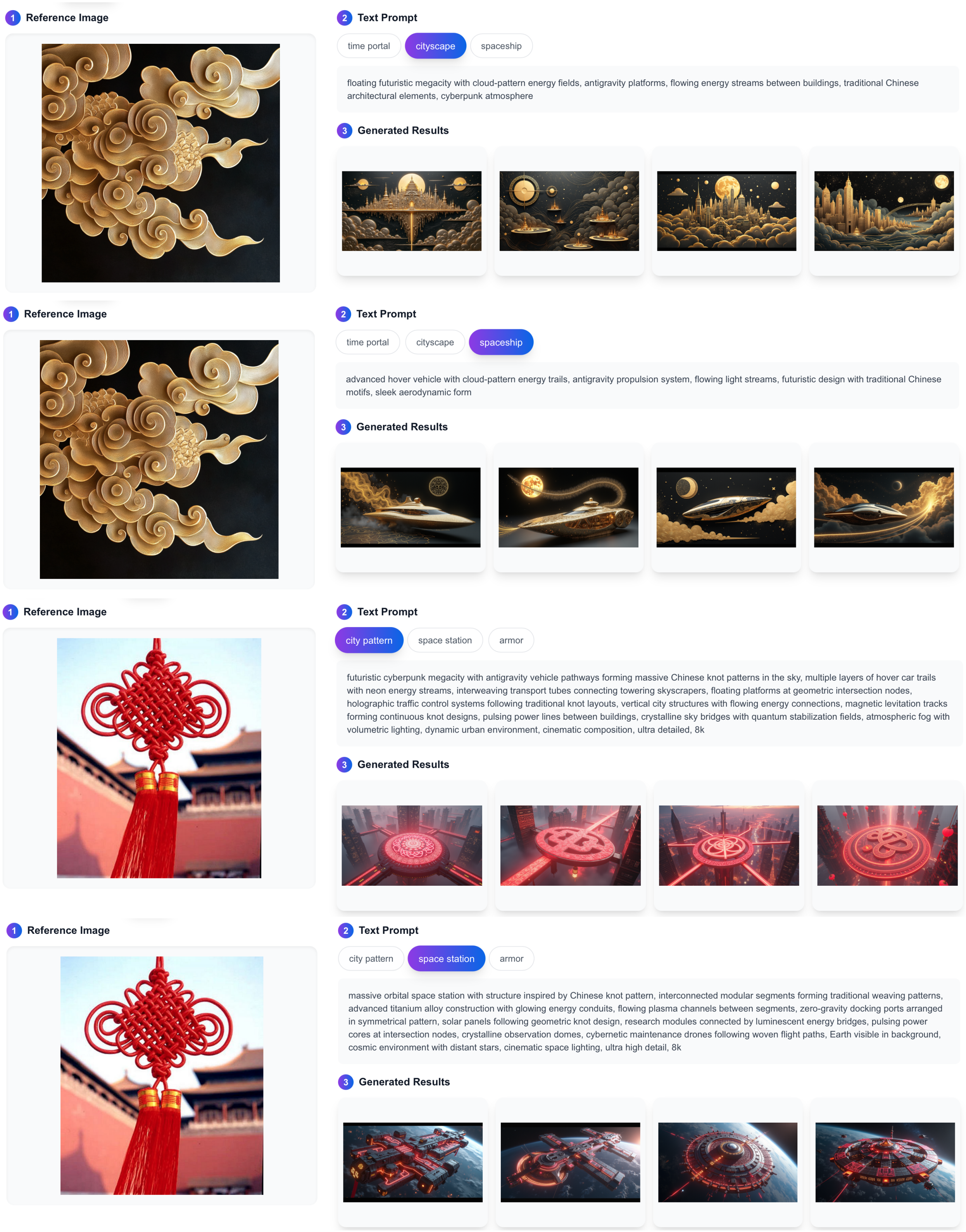
futuristic cyberpunk city painted in Chinese ink style, towering pagodas mixed with skyscrapers, floating hover cars like brush strokes, neon signs as delicate ink splashes, fog like diluted ink, dramatic composition

elegant mecha warrior in Chinese ink painting style, flowing mechanical parts like brush strokes, energy trails like ink splashes, zen-inspired design, mixture of ancient and futuristic, dynamic pose

spacecraft designed in Chinese ink painting style, flowing energy trails like brush strokes, weapon systems like ink splashes, ancient Chinese ship elements, zen minimalist design

interdimensional portal with cloud-pattern energy vortex, space-time gateway, flowing temporal streams, ancient meets future aesthetic, mystical tech fusion

Figure 3: Text-image blend results (Part 1). Generated images demonstrate the model's ability to combine visual features from reference images with semantic guidance from text prompts.



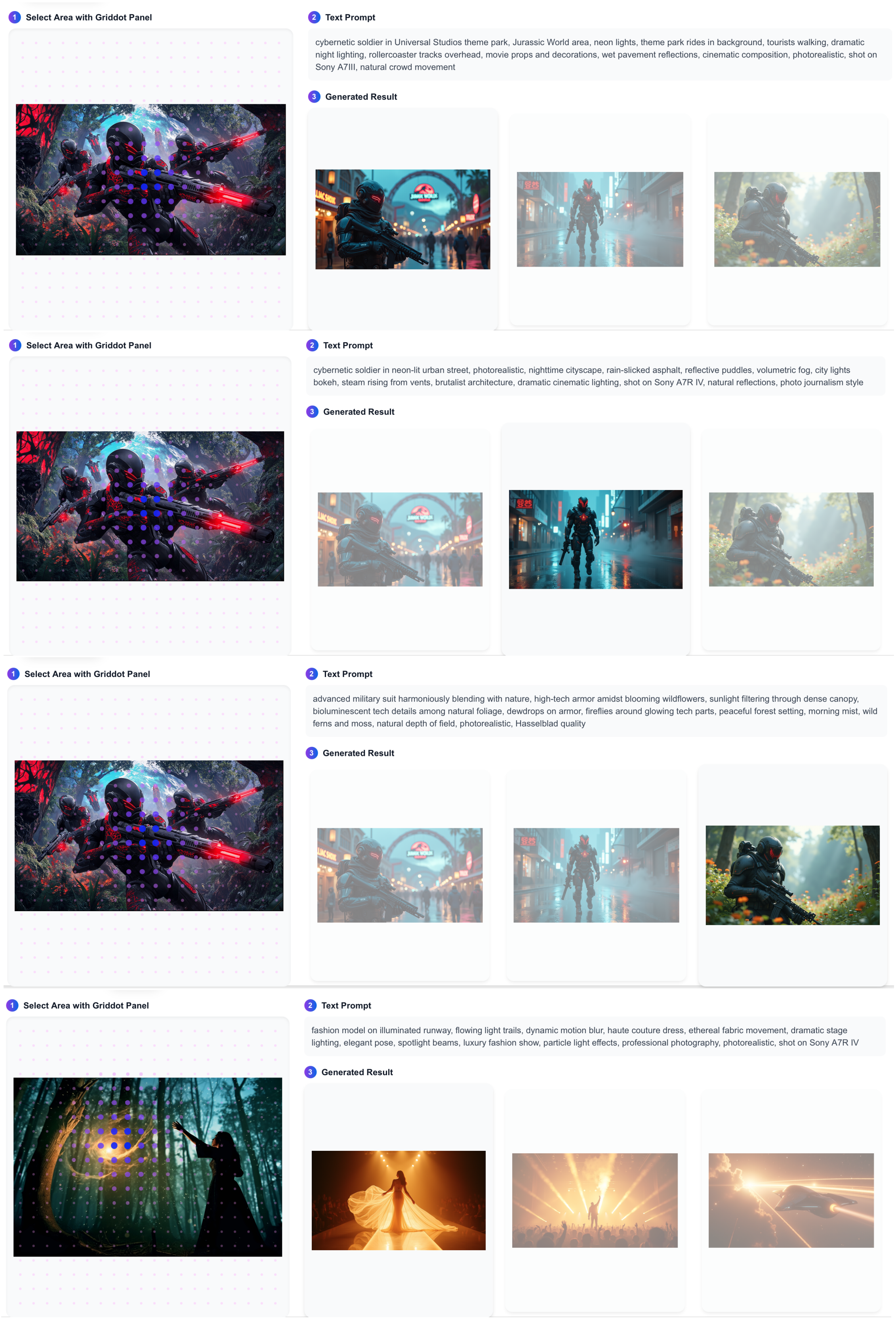
floating futuristic megacity with cloud-pattern energy fields, antigravity platforms, flowing energy streams between buildings, traditional Chinese architectural elements, cyberpunk atmosphere

advanced hover vehicle with cloud-pattern energy trails, antigravity propulsion system, flowing light streams, futuristic design with traditional Chinese motifs, sleek aerodynamic form

futuristic cyberpunk megacity with antigravity vehicle pathways forming massive Chinese knot patterns in the sky, multiple layers of hover car trails with neon energy streams, interweaving transport tubes connecting towering skyscrapers, floating platforms at geometric intersection nodes, holographic traffic control systems following traditional knot layouts, vertical city structures with flowing energy connections, magnetic levitation tracks forming continuous knot designs, pulsing power lines between buildings, crystalline sky bridges with quantum stabilization fields, atmospheric fog with volumetric lighting, dynamic urban environment, cinematic composition, ultra detailed, 8k

massive orbital space station with structure inspired by Chinese knot pattern, interconnected modular segments forming traditional weaving patterns, advanced titanium alloy construction with glowing energy conduits, flowing plasma channels between segments, zero-gravity docking ports arranged in symmetrical pattern, solar panels following geometric knot design, research modules connected by luminescent energy bridges, pulsing power cores at intersection nodes, crystalline observation domes, cybernetic maintenance drones following woven flight paths, Earth visible in background, cosmic environment with distant stars, cinematic space lighting, ultra high detail, 8k

Figure 4: Text-image blend results (Part 2). Additional examples showing the model's multimodal generation capabilities.


1 Select Area with Griddot Panel
2 Text Prompt
3 Generated Result

cybernetic soldier in Universal Studios theme park, Jurassic World area, neon lights, theme park rides in background, tourists walking, dramatic night lighting, rollercoaster tracks overhead, movie props and decorations, wet pavement reflections, cinematic composition, photorealistic, shot on Sony A7III, natural crowd movement

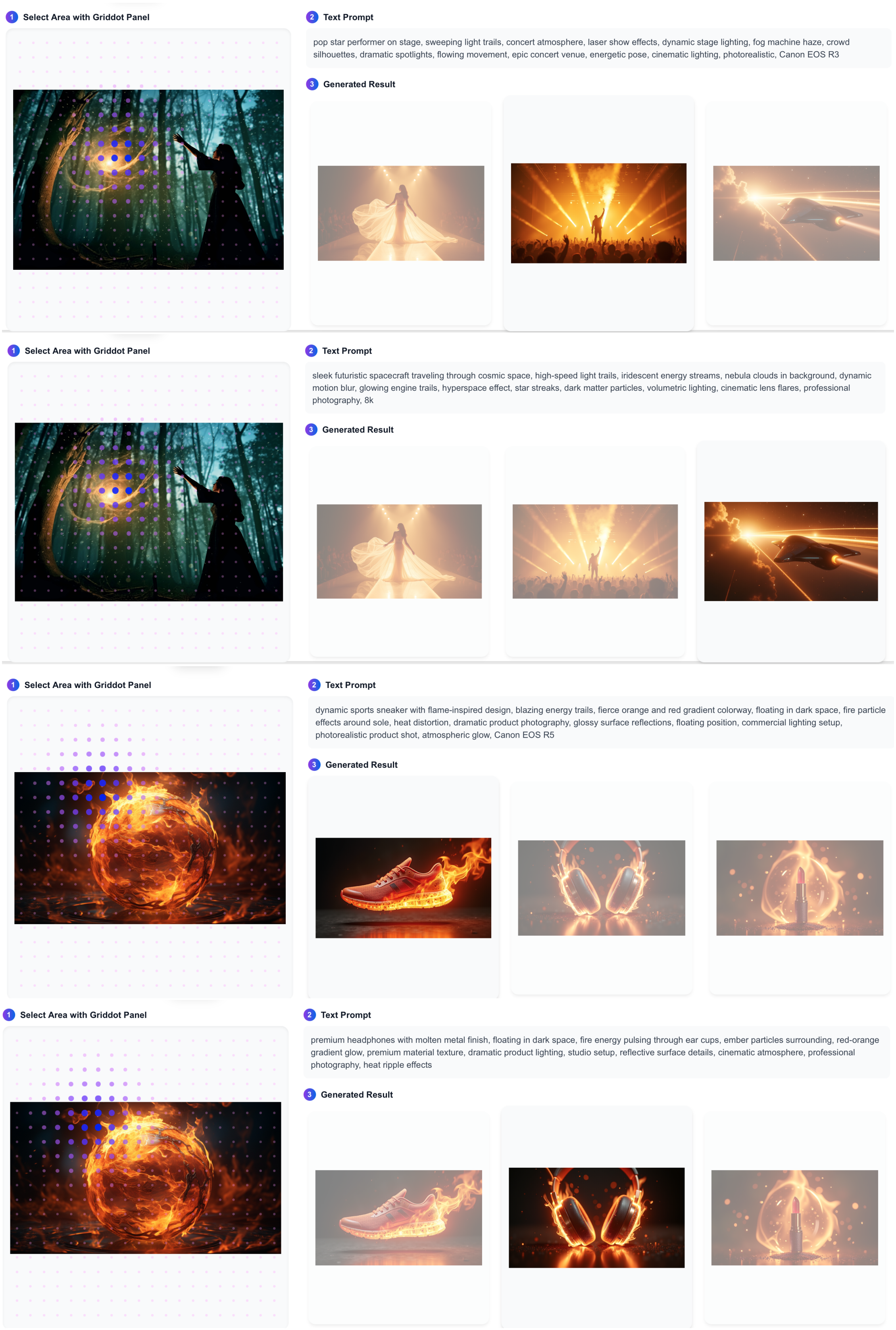
1 Select Area with Griddot Panel
2 Text Prompt
3 Generated Result

cybernetic soldier in neon-lit urban street, photorealistic, nighttime cityscape, rain-slicked asphalt, reflective puddles, volumetric fog, city lights bokeh, steam rising from vents, brutalist architecture, dramatic cinematic lighting, shot on Sony A7R IV, natural reflections, photo journalism style

advanced military suit harmoniously blending with nature, high-tech armor amidst blooming wildflowers, sunlight filtering through dense canopy, bioluminescent tech details among natural foliage, dewdrops on armor, fireflies around glowing tech parts, peaceful forest setting, morning mist, wild ferns and moss, natural depth of field, photorealistic, Hasselblad quality

fashion model on illuminated runway, flowing light trails, dynamic motion blur, haute couture dress, ethereal fabric movement, dramatic stage lighting, elegant pose, spotlight beams, luxury fashion show, particle light effects, professional photography, photorealistic, shot on Sony A7R IV

Figure 5: GridDot Panel results (Part 1). Demonstration of the novel GridDot Panel interface allowing users to control attention on specific regions of the reference image.


1 Select Area with Griddot Panel
2 Text Prompt
3 Generated Result

pop star performer on stage, sweeping light trails, concert atmosphere, laser show effects, dynamic stage lighting, fog machine haze, crowd silhouettes, dramatic spotlights, flowing movement, epic concert venue, energetic pose, cinematic lighting, photorealistic, Canon EOS R3

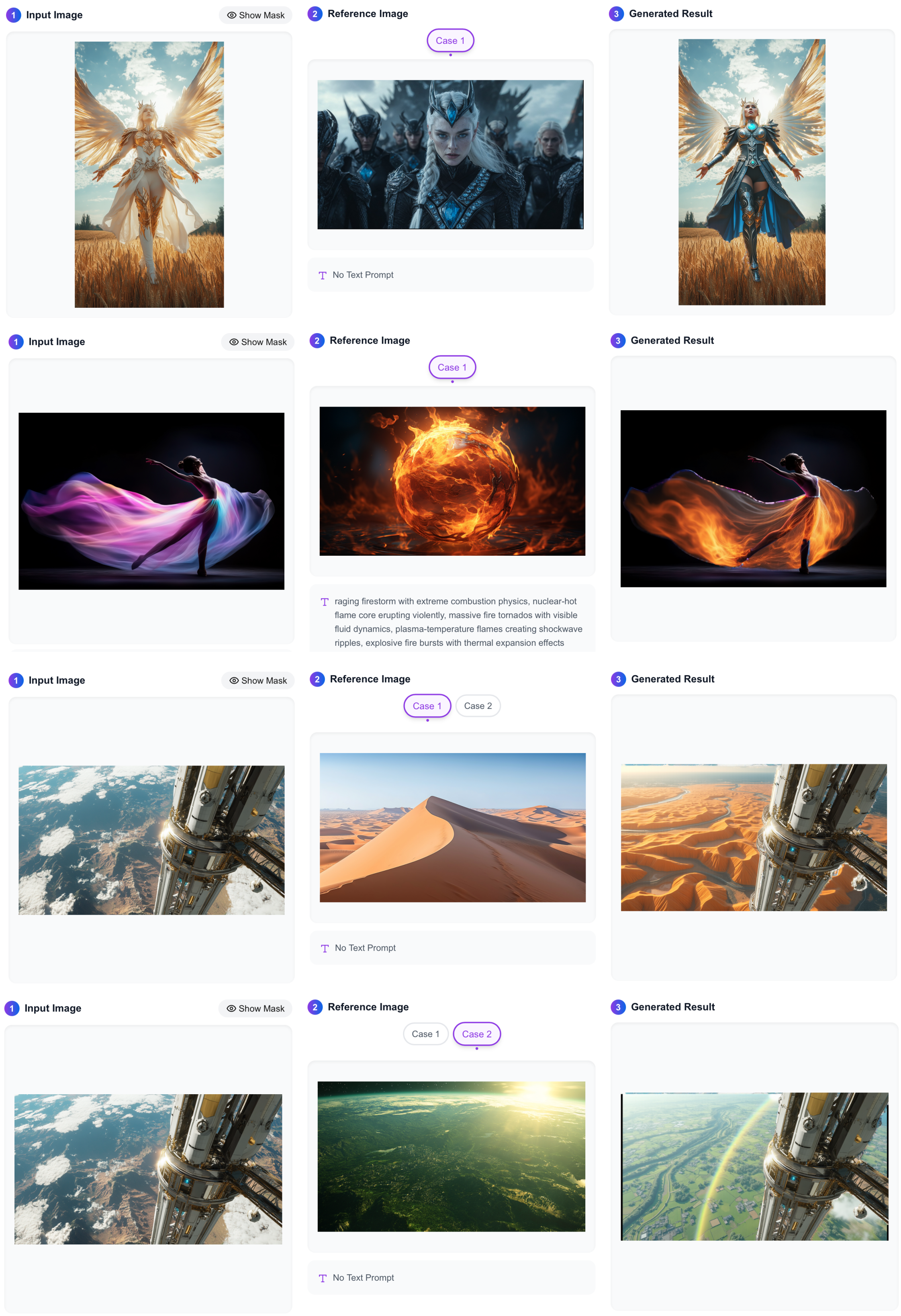
1 Select Area with Griddot Panel
2 Text Prompt
3 Generated Result

sleek futuristic spacecraft traveling through cosmic space, high-speed light trails, iridescent energy streams, nebula clouds in background, dynamic motion blur, glowing engine trails, hyperspace effect, star streaks, dark matter particles, volumetric lighting, cinematic lens flares, professional photography, 8k

dynamic sports sneaker with flame-inspired design, blazing energy trails, fierce orange and red gradient colorway, floating in dark space, fire particle effects around sole, heat distortion, dramatic product photography, glossy surface reflections, floating position, commercial lighting setup, photorealistic product shot, atmospheric glow, Canon EOS R5

premium headphones with molten metal finish, floating in dark space, fire energy pulsing through ear cups, ember particles surrounding, red-orange gradient glow, premium material texture, dramatic product lighting, studio setup, reflective surface details, cinematic atmosphere, professional photography, heat ripple effects

Figure 6: GridDot Panel results (Part 2). Additional examples showing fine-grained control over image generation through attention manipulation.



raging firestorm with extreme combustion physics, nuclear-hot flame core erupting violently, massive fire tornados with visible fluid dynamics, plasma-temperature flames creating shockwave ripples, explosive fire bursts with thermal expansion effects

Figure 7: Scene blending results (Part 1). Examples of combining semantic elements and styles from two different reference images using Qwen2VL-7B embeddings and ControlNet.

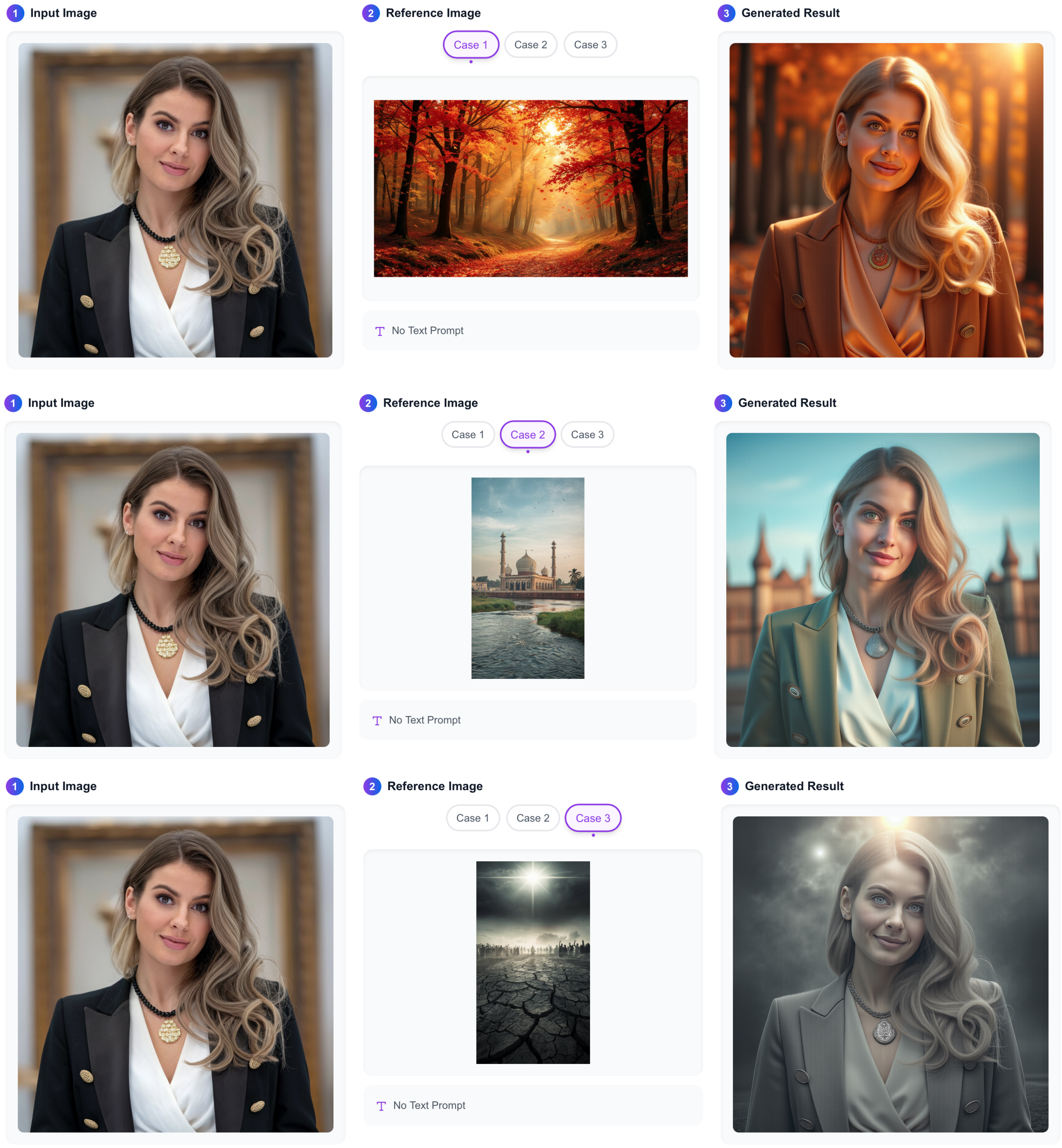


Figure 8: Scene blending results (Part 2). Additional examples demonstrating the model's ability to seamlessly merge different visual elements and styles.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

William S. Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4172–4182, 2022.

Patrick Esser, Sumith Kulal, A. Blattmann, Rahim Entezari, Jonas Muller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. *ArXiv*, abs/2403.03206, 2024.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

Black Forest Labs. Flux: Inference repository. `https://github.com/black-forest-labs/flux`, 2024. Accessed: 2024-10-25.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. *arXiv preprint arXiv:2302.09778*, 2023.