A. PanoLlama Generation Process

We provide detailed pseudocode in Alg. 1 to facilitate a better understanding of our framework. Aligned with LlamaGen [31], it incorporates FLAN-T5 XL [9] as the text encoder $f_{\mathcal{E}}$, VQVAE [35] as the image tokenizer $f_{\mathcal{T}}$, and Llama XL [34] as the token generator $f_{\mathcal{G}}$ to achieve text-guided panorama generation.

Algorithm 1 PanoLlama Generation Process

Input: $f_{\mathcal{E}}, f_{\mathcal{G}}, f_{\mathcal{T}d}$ ▷ pre-trained models: text encoder, token generator, decoder of image tokenizer
p ▷ max token limit of image tokenizer
y ▷ textual prompt
$mode$ ▷ direction of expansion
n ▷ expansion iterations
r, c ▷ rows and columns expanded per iteration
Output: x' ▷ panorama

function GENERATETOKENS-VERTICAL(v_{i-1}, r)
  $v_i \leftarrow f_{\mathcal{G}}(v_{i-1,r\sqrt{p}}, \ldots, v_{i-1,p})$
  return v_i
end function

function GENERATETOKENS-HORIZONTAL(v_{i-1}, c)
  for $j = 1, 2, \ldots, \sqrt{p}$ **do**
    $v_i^j \leftarrow f_{\mathcal{G}}(v_{i-1,\epsilon(v_{i-1}^j)-\sqrt{p}+c}, \ldots, v_{i-1,\epsilon(v_{i-1}^j)})$
  end for
  return v_i
end function

$s \leftarrow f_{\mathcal{E}}(y)$ // (i) Textual Conditioning

$v_1 \leftarrow f_{\mathcal{G}}(s)$ // (ii) Next-Crop Prediction
if $mode ==$ 'Vertical' **then**
  for $i = 2, 3, \ldots, n$ **do**
    $v_i \leftarrow$ GENERATETOKENS-VERTICAL(v_{i-1}, r)
  end for
  $V \leftarrow v_1 \oplus_{i=2}^{n} v_i$
else if $mode ==$ 'Horizontal' **then**
  for $i = 2, 3, \ldots, n$ **do**
    $v_i \leftarrow$ GENERATETOKENS-HORIZONTAL(v_{i-1}, c)
  end for
  $V \leftarrow v_1 \cup_{i=2}^{n} v_i$
end if

$x' \leftarrow f_{\mathcal{T}d}(V)$ // (iii) Decoding Tokens into Panorama
Return x'

B. More Qualitative Comparison Results

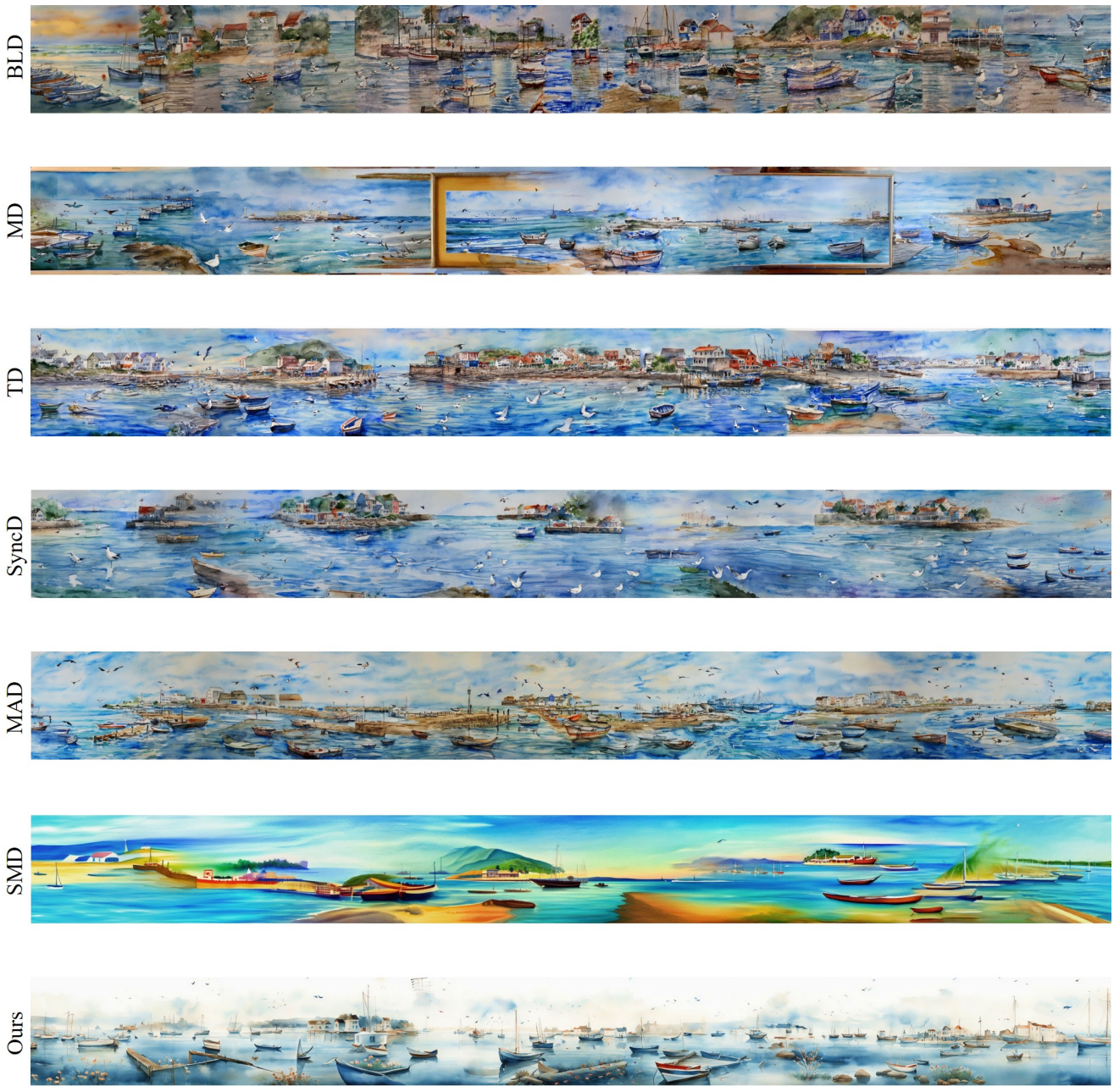
Figs. A1 to A13 gives more qualitative comparisons on panoramic image generation of 512×5120, highlighting the areas where improvements have been made.

"A nature scenery featuring herds of wild animals grazing and wandering across a vast green grassland, with a backdrop of distant mountains under an endless, open sky, capturing the essence of untouched natural beauty."

Figure A1. Additional qualitative comparison results on 512×5120 panorama generation.

"A magical winter landscape featuring a cozy cabin surrounded by snow-covered trees, with warm light glowing from the windows, as snowflakes gently fall and a serene atmosphere prevails, evoking a sense of warmth and comfort."

Figure A2. Additional qualitative comparison results on 512×5120 panorama generation.

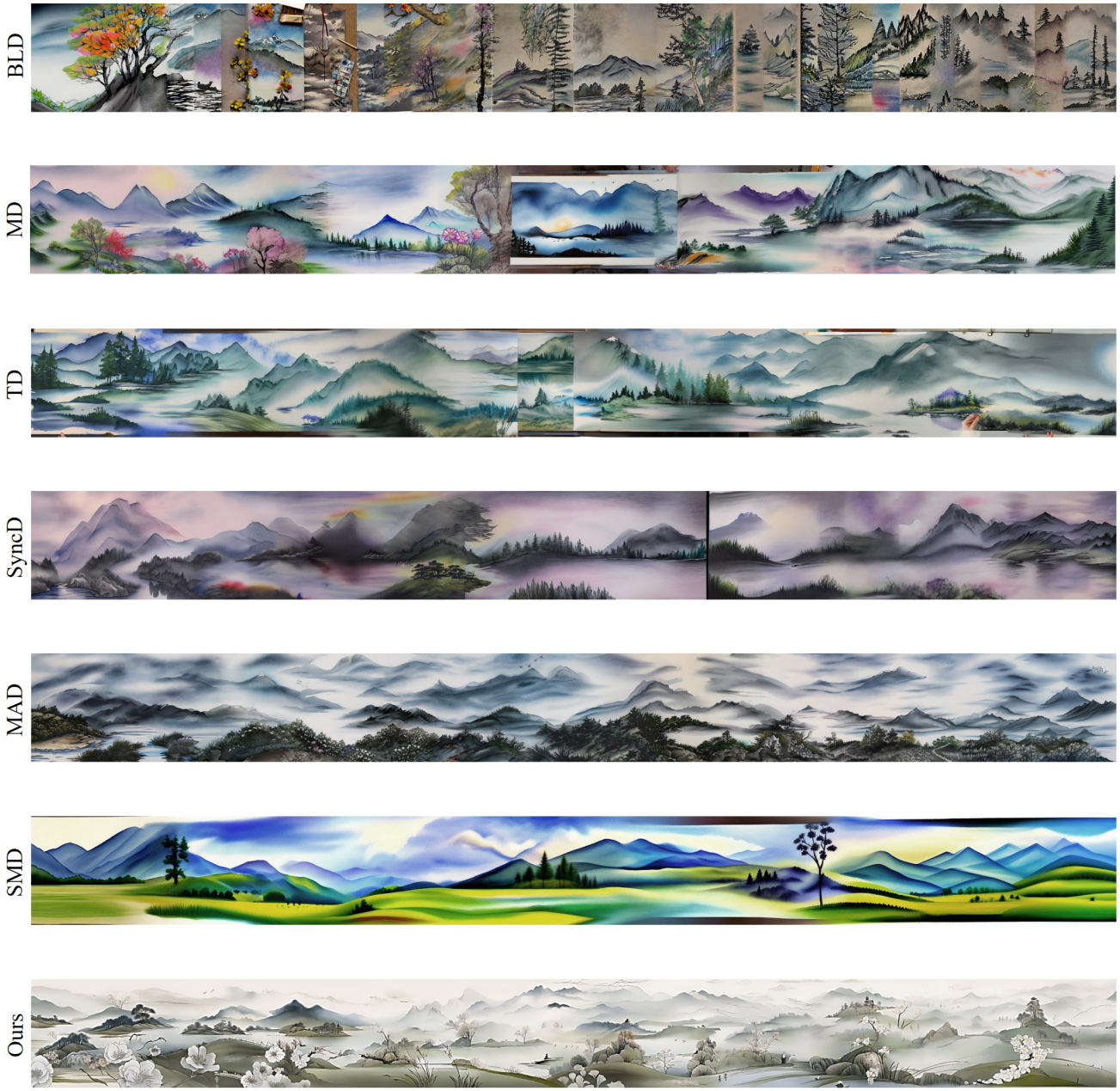


"A serene watercolor painting of a seaside harbor, with fishing boats gently bobbing in the water, gentle waves crashing on the shore, seagulls flying overhead, creating a sense of peace and nostalgia."

Figure A3. Additional qualitative comparison results on 512×5120 panorama generation.

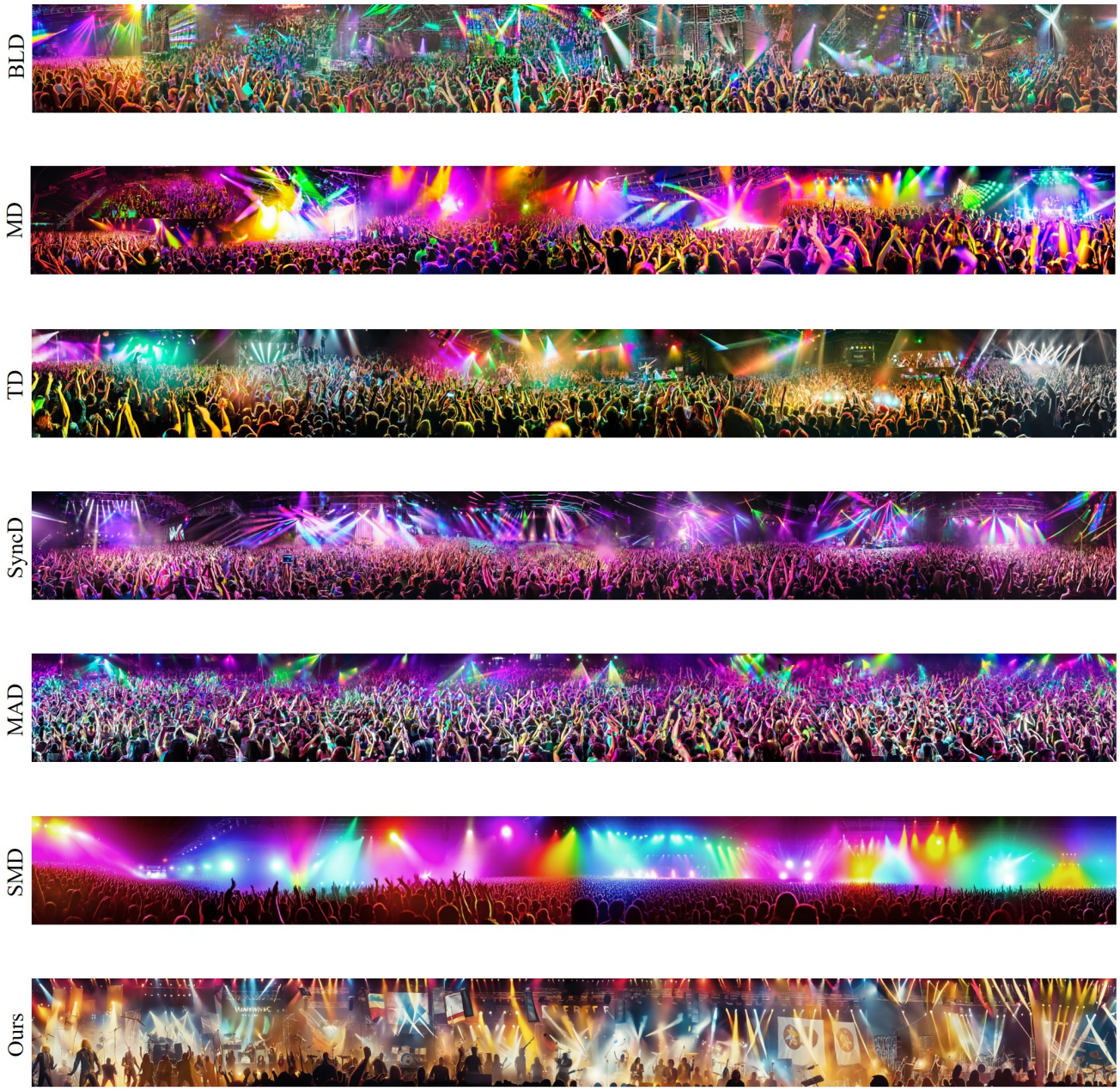
"A classic oil painting of blooming flowers, featuring clusters of deep red roses and cheerful yellow sunflowers. Delicate white daisies peek through lush green leaves, creating a vibrant scene. The artist's textured brushstrokes add depth, with soft light casting gentle shadows that enhance the beauty of each petal, celebrating the romance of nature in full bloom."

Figure A4. Additional qualitative comparison results on 512×5120 panorama generation.



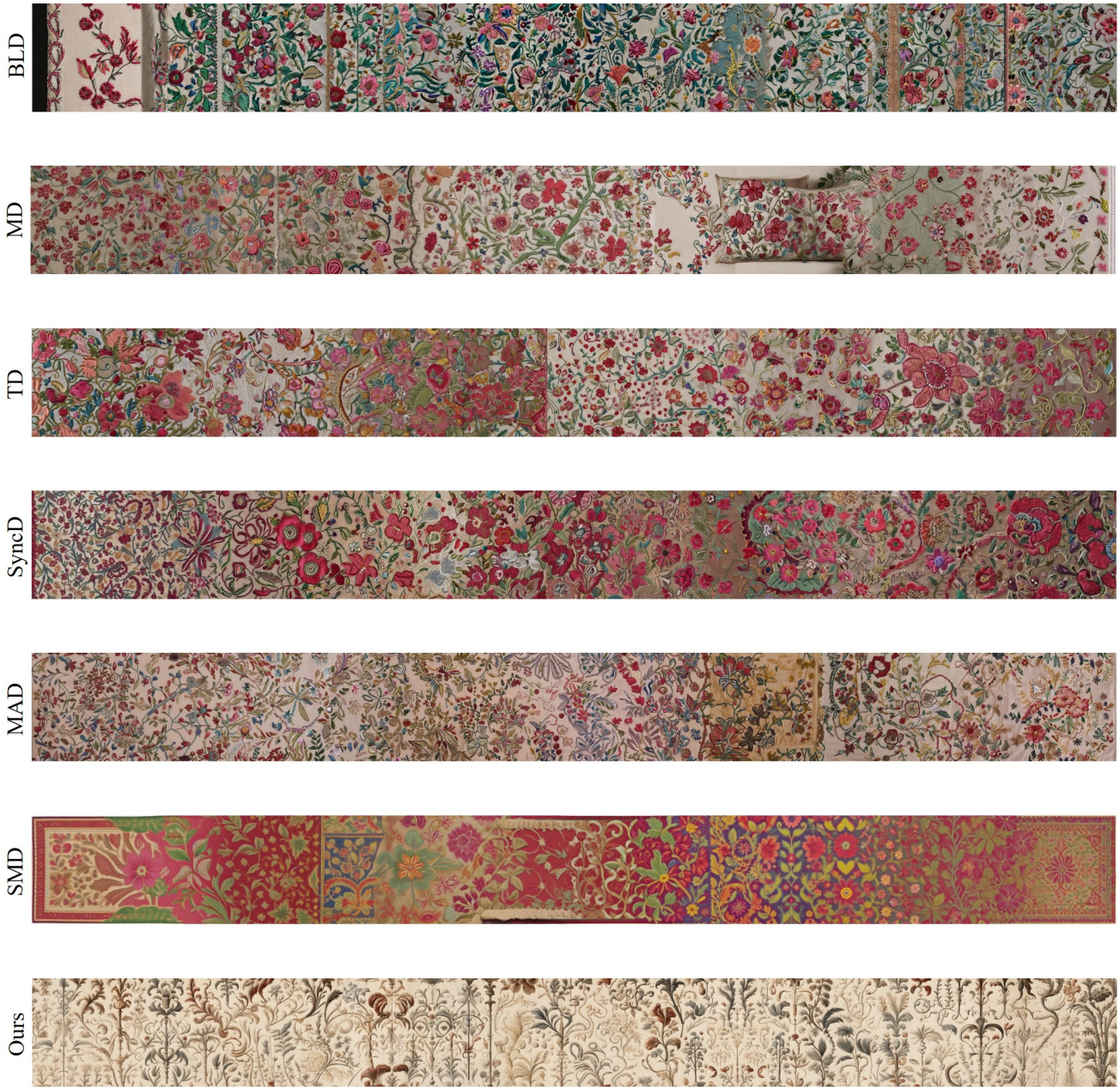
"Generate a serene landscape ink painting featuring rolling hills, a tranquil lake, and majestic mountains. Include blooming flowers and a misty sky to evoke tranquility and harmony. Emphasize traditional ink painting techniques with delicate brush strokes."

Figure A5. Additional qualitative comparison results on 512×5120 panorama generation.



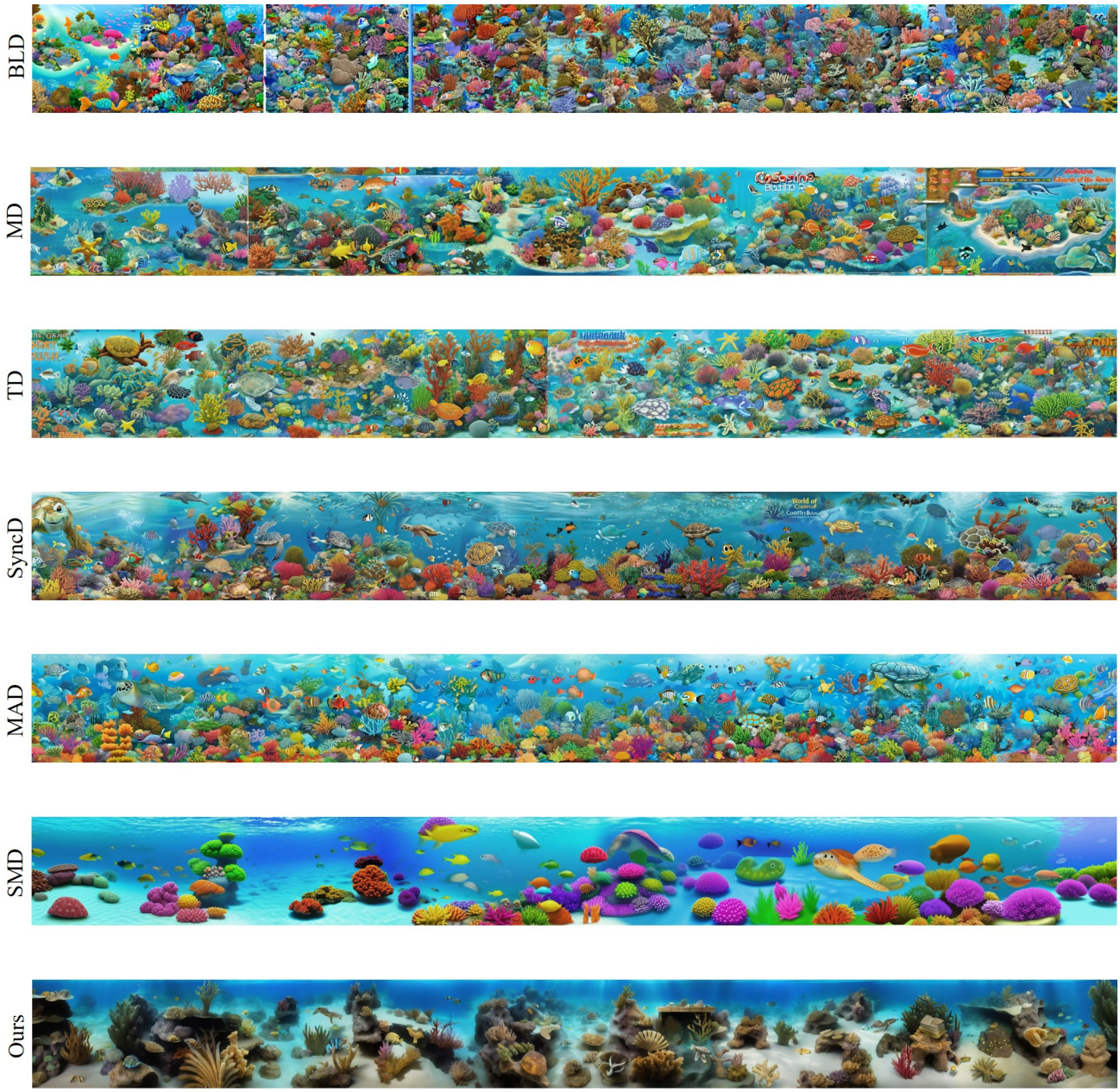
"A photo of a rock concert featuring an excited crowd of fans singing along, with colorful stage lights illuminating the scene. The lead singer engages the audience, while band members play energetically. Banners wave and speakers pulse, capturing the vibrant atmosphere of this real-life event."

Figure A6. Additional qualitative comparison results on 512×5120 panorama generation.



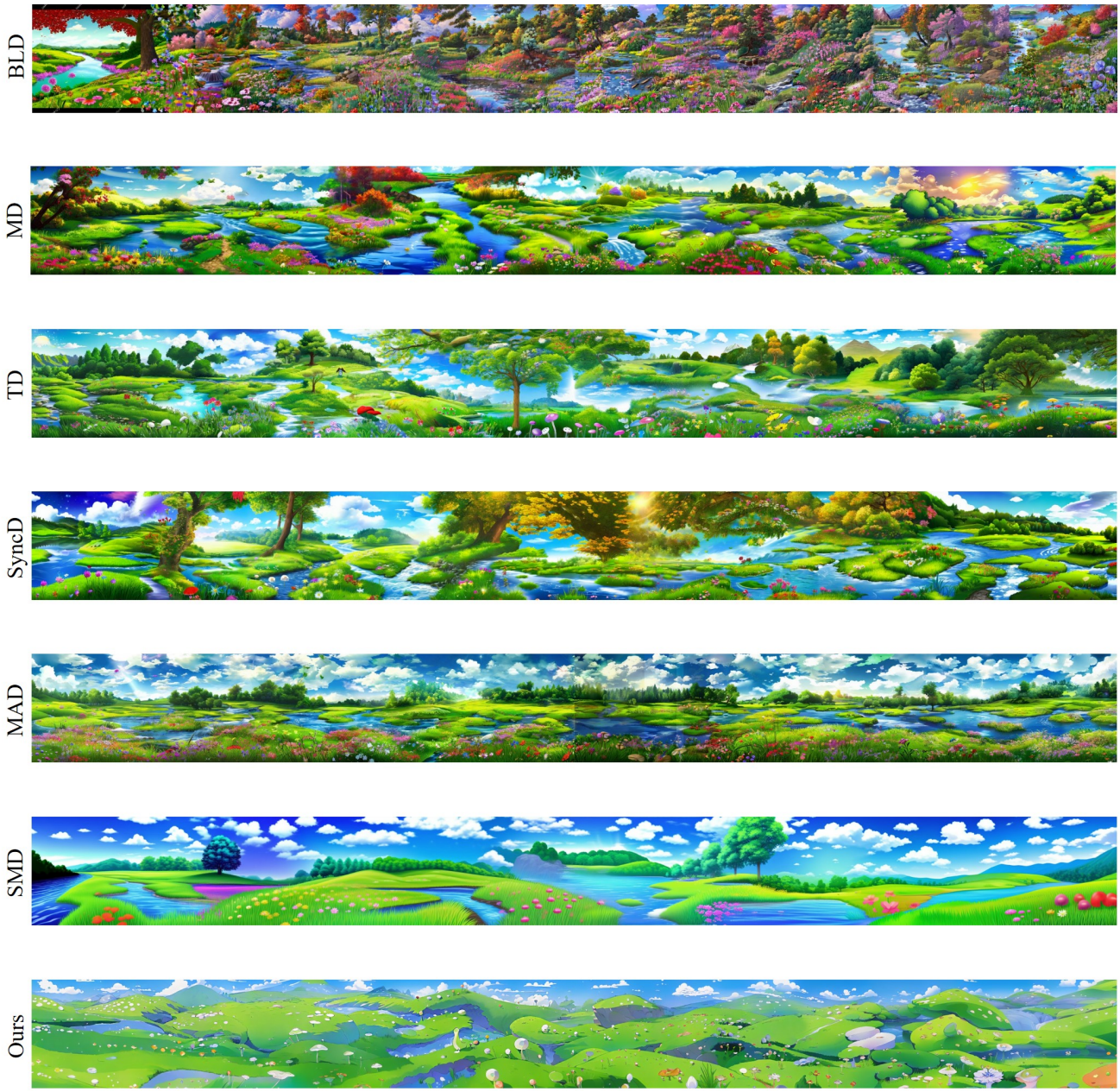
"An intricate hand-embroidered piece showcasing antique floral patterns, featuring delicately stitched petals, leaves, and vines, with rich, vibrant colors and textures that add a sense of elegance and artistry to the design."

Figure A7. Additional qualitative comparison results on 512×5120 panorama generation.



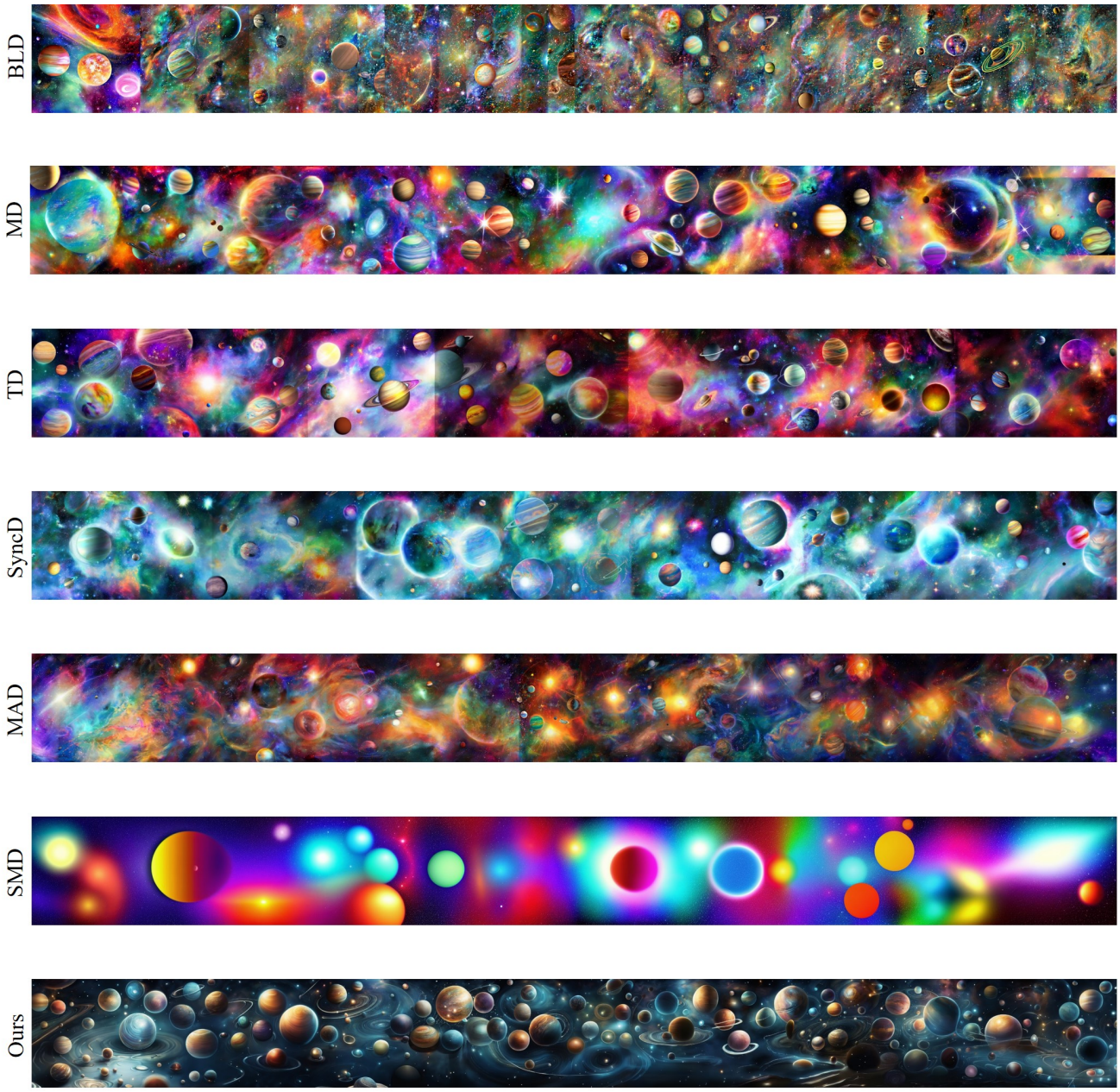
"Create a world of a coastline with short coral bushes in various types and hues. Include schools of tropical fish swimming among the corals, a graceful sea turtle gliding by, and perhaps hidden treasure chests among the seaweeds, illuminating the lively ocean life."

Figure A8. Additional qualitative comparison results on 512×5120 panorama generation.



"An anime-style illustration of a beautiful natural landscape, with rolling green meadows, vivid wildflowers and small mushrooms, clear sky, few fluffy clouds float overhead, while a sparkling river winds through the scene."

Figure A9. Additional qualitative comparison results on 512×5120 panorama generation.

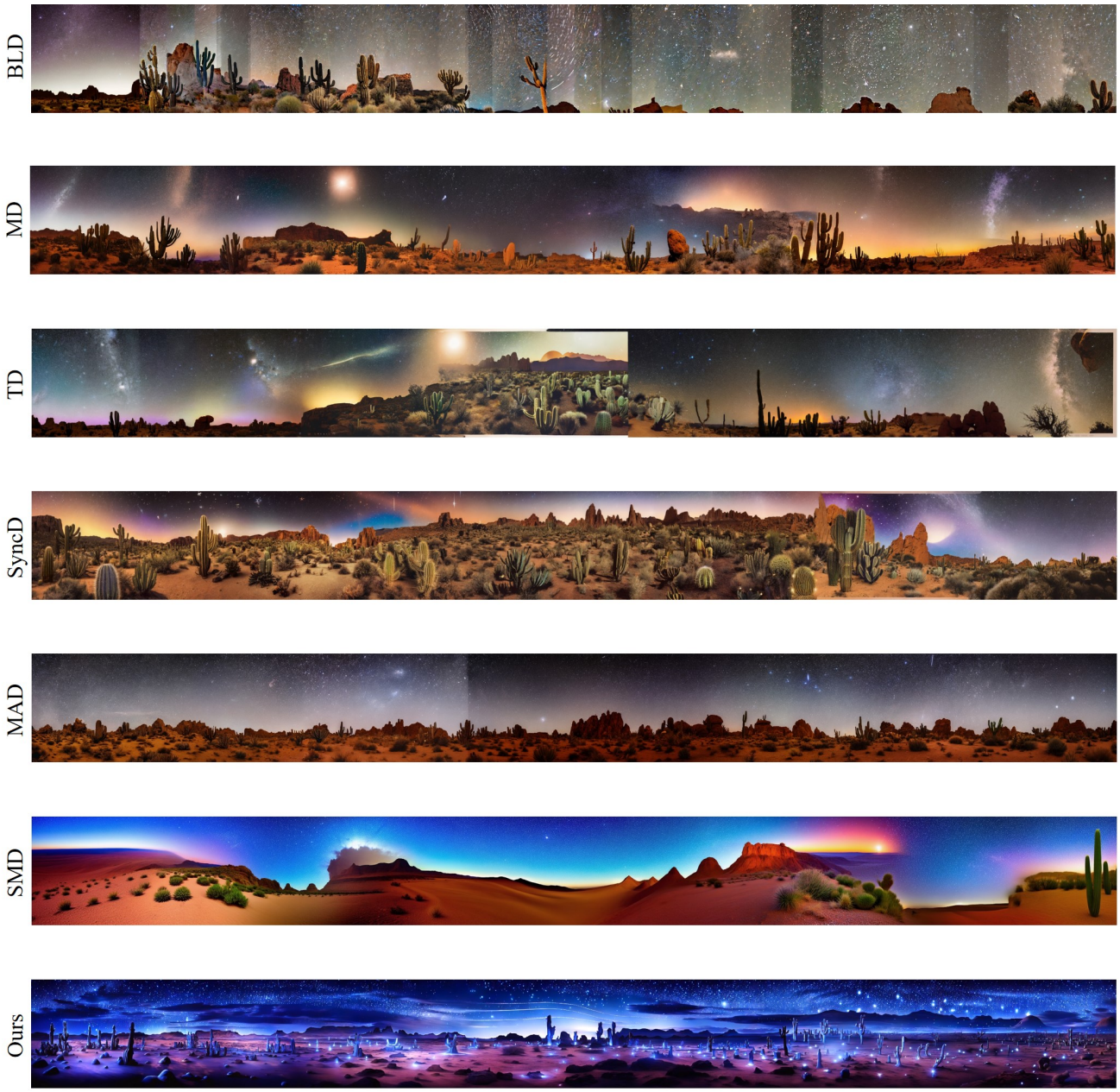


"A digital painting of planets floating in a dreamy universe, surrounded by colorful nebulae and sparkling stars. The planets should vary in size and color, with intricate details on their surfaces, reflecting the ethereal glow of distant cosmic light."

Figure A10. Additional qualitative comparison results on 512×5120 panorama generation.

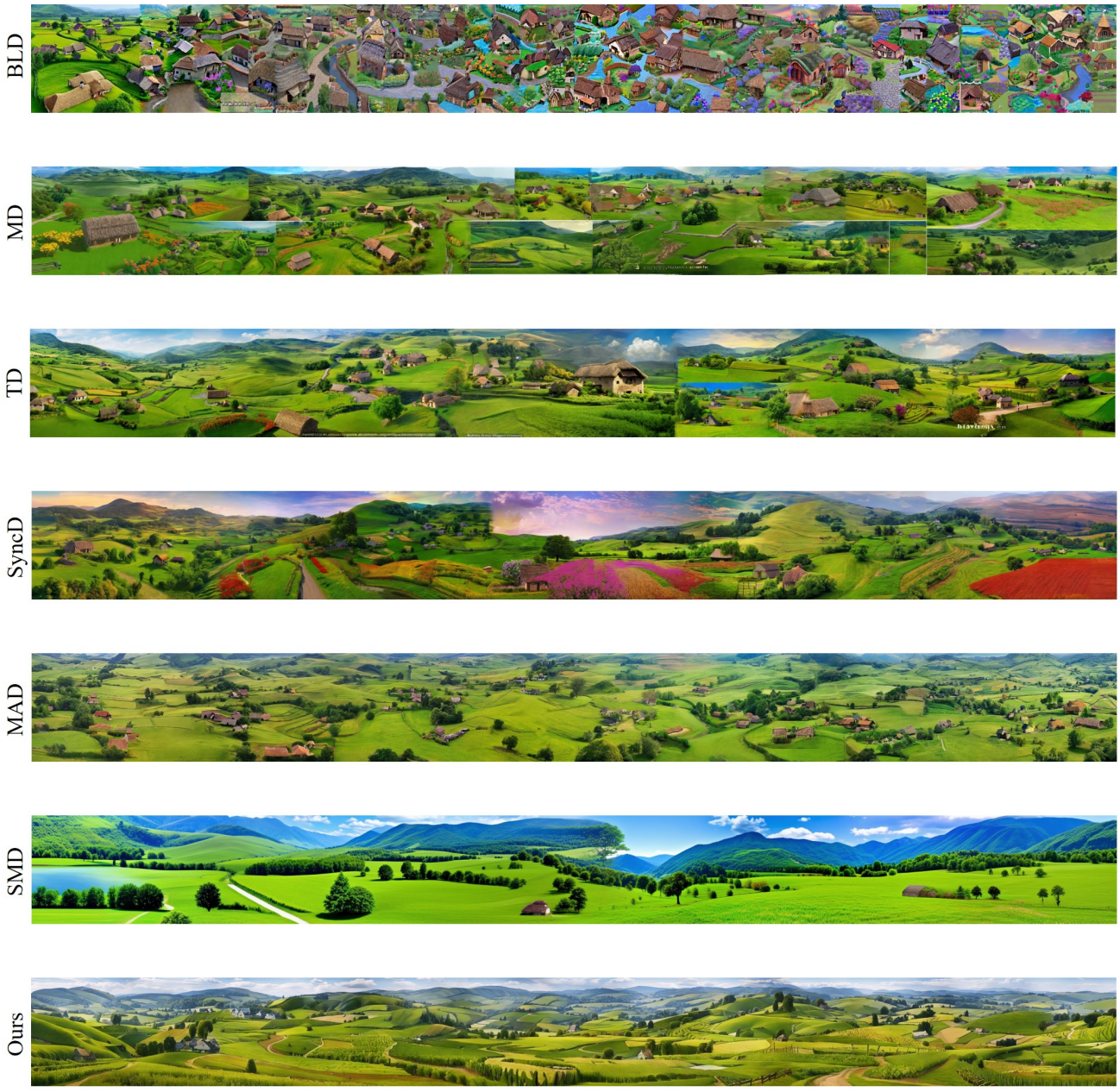
"A paper-cut style portrayal of a Chinese festival celebration, crowd of people wearing traditional clothing, lanterns hanging and glowing in the background, and dancers and musicians adding energy to the festive atmosphere."

Figure A11. Additional qualitative comparison results on 512×5120 panorama generation.



"A cyaneous night sky filled with stars shining and scattering around, illuminating a vast desert landscape, where silhouettes of cacti and rocky formations create a striking contrast against the celestial backdrop."

Figure A12. Additional qualitative comparison results on 512×5120 panorama generation.



"Generate a sprawling, verdant countryside landscape with rolling hills, winding streams, and distant mountains. The scene should have a classic, pastoral aesthetic, with lush green meadows dotted with vibrant wildflowers and clusters of deciduous trees. A small, rustic village with stone buildings and thatched roofs should sit nestled in a valley, surrounded by neatly tended farmland and orchards. A dirt road should wind its way through the landscape, flanked by picket fences and the occasional weathered barn or silo. The sky should be a warm, hazy blue, with fluffy white clouds drifting lazily overhead. The overall mood should be one of tranquility, timelessness, and the natural beauty of the countryside. The color palette should be muted and earthy, with plenty of greens, browns, grays, and ochres. The lighting should be soft and diffused, creating long shadows and highlights that give depth and dimensionality to the scene. Attention to detail is important, so include small touches like wildflowers, grazing livestock, and the smoke rising from chimneys in the village. This landscape should have a distinctly European, almost storybook-like quality, evoking the pastoral landscapes of Normandy, the English countryside, or the rolling hills of Tuscany. The entire scene should feel cohesive and harmonious, with a sense of balance and proportion that makes it feel like a complete, self-contained world."

Figure A13. Additional qualitative comparison results on 512×5120 panorama generation.

C. Comparison with SD Variants

We further conduct experiments on additional diffusion variants, evaluating baselines using Φ = SDXL. In line with Sec. 4.1, we continue to generate images at a resolution of 512×5120 and then crop them to create 512×512 image sets.

Tab. A1 presents the comparison results. Some baselines do not support SDXL and are therefore excluded from this table. Notably, even with SDXL, our approach excels at enhancing the multilevel coherence of generated panoramas, though to a lesser degree than the SD-based reference model. This is expected, given SDXL's advanced design as a diffusion model with doubled default resolution in both height and width.

Table A1. Comparisons between PanoLlama and baselines with another Φ variants. Our approach still stands out in improving the coherence of the generated panoramas while maintaining aesthetic quality and operational speed.

	Coherence				Fidelity & Diversity		Compatibility	
	LPIPS↓	DISTS↓	TV↓	SSIM↑	FID↓	IS↑	CLIP↑	CLIP-aesthetic↑
SD_{XL}	–	–	–	–	34.50	7.60	33.23	6.76
LlamaGen	–	–	–	–	37.82	6.43	31.62	6.74
BLD_{XL}	0.790	0.356	0.075	0.013	85.20 (+50.70)	5.97 (-1.63)	32.59 (-0.64)	5.80 (-0.96)
MD_{XL}	0.676	0.253	0.055	0.220	39.15 (+4.65)	6.42 (-1.18)	34.66 (+1.43)	6.88 (+0.12)
TD_{XL}	0.638	0.214	0.051	0.274	40.05 (+5.55)	6.48 (-1.12)	**34.79 (+1.56)**	6.89 (+0.13)
MAD_{XL}	0.517	0.208	0.032	0.296	56.55 (+22.05)	5.13 (-2.47)	32.07 (-1.16)	6.94 (+0.18)
Ours	**0.410**	**0.196**	**0.021**	**0.305**	**40.09 (+2.27)**	**5.97 (-0.46)**	31.56 (-0.06)	**6.97 (+0.23)**

D. Dataset Construction Details

To ensure a comprehensive and fair evaluation, our dataset construction focused on diversity and challenge. (i) Theme Selection: We select 25 common themes stratified by typical scene scale: from expansive scenes well-suited for panoramas (e.g., 'landscape') to medium-scale subjects (e.g., 'architecture') and challenging dense/small-object scenes (e.g., 'crowd'). This variety is designed to robustly test PIG methods across diverse content types, as explored in our prompt theme analysis in Sec. 4.2. (ii) Sub-Themes: Each theme is further diversified into 3-8 sub-themes (e.g., 'crowd' includes 'festival', 'market', 'concert', $\cdots$) to ensure broad conceptual coverage. (iii) Styles: We include a mix of styles (photorealistic, artistic, chosen randomly). (iv) Creation Process: The prompts are generated by an AI based on the aforementioned structured themes and styles, resulting in 400 prompts per theme. (v) Fairness: This random, broad design ensures no implicit bias toward our PanoLlama; all methods are evaluated fairly.

E. Implementation Details about Our Applications

Similar to joint diffusion methods, our approach can also introduce additional optimizations on top of next-token prediction to achieve smoother transitions. For instance, we can apply a basic blending function as follows:

$$\begin{aligned} \bar{e}_{i,1} &= \lambda e_{i,1} + (1-\lambda) e_{i-1,p} \\ v^*_{i,1} &= \arg\min_{v_{i,1}} \|\bar{e}_{i,1} - e\|^2 \end{aligned} \tag{13}$$

where $v_{i-1,p}$ and $v_{i,1}$ represent the boundary tokens between v_{i-1} and v_i, $e_{i-1,p}$ and $e_{i,1}$ refer to the embeddings indexed by these tokens, λ denotes the transition factor, e represents the trained embeddings in $f_{\mathcal{T}}$, $v^*_{i,1}$ is the resulting token after blending.

By employing Eq. (13) with $\lambda \in [0.5, 0.8]$, we can achieve smoother transitions under different textual conditions for multi-layout and multi-guidance applications. However, our experiments demonstrate that with a consistent prompt, the performance of PanoLlama peaks at $\lambda = 1.0$, declining at other values. This suggests that our method reaches its best in single-prompt scenarios without the need for blending operations.