8. More Quantitive Experiments

8.1. Comparison with Training-Based Methods

We further compare our method with more training-based long audio generation models, including both diffusion models and language models. Although strictly speaking, the absolute performance between models of varying sizes and trained on different datasets seems incomparable, the relative performance degradation of each model with increasing audio generation length can highlight the strengths and weaknesses of these methods for long-generation tasks.

Baselines The training-based baselines include: (1) *AudioGen* [25] : An autoregressive model based on learned discrete audio representations, inherently supporting ultra-long audio generation. (2) *Stable Diffusion Audio* (SD-audio) [14]: A diffusion model trained on a fixed 96-second window size with long audio, generating variable-length outputs through end-cutting. Although the open-source version is trained on a 47-second window, longer audio can still be generated by customizing the initial noise size. (3) *Make-An-Audio2* (Make2) [13]: A diffusion model trained on variable-length window sizes, with audio lengths ranging from 0 to 20 seconds. It supports a maximum length of 27 seconds, constrained by the learnable positional encoding limit. For SaFa, we implement it on Make-An-Audio2 for a clearer comparison, following the settings described in Section 6.1.

Evaluation Settings We evaluate these four methods using a large-scale benchmark, AudioCaps [54], whose test set includes 880 ground-truth samples collected from YouTube videos. The target generation lengths are set to 32, 64, and 96 seconds. For SaFa, these outputs are formed by concatenating 4, 8, and 12 audio clips of 10 seconds, respectively. As in Section 6.1, we use FD, FAD, KL, and mCLAP to assess the generation quality and semantic alignment of the generated audio. Following previous work [14]], we apply a 10-second sliding window operation with an 8-second step on long audio samples and further evaluate them with AudioCaps test set.

Results As shown in Table 5, Make2, the SOTA diffusion-based audio generation model, demonstrates excellent performance in 10-second audio generation. However, it shows significant performance degradation when generating its maximum-length output of 27 seconds, as most training audio clips are under 20 seconds, and it lacks adaptation to longer unseen lengths. In contrast, our SaFa (32s) method maintains high performance in terms of KL and mCLAP, with only minor degradation observed in FD and FAD compared with the reference model Make2 (10s). Moreover, SaFa consistently delivers strong performance for 32-, 64-, and 96-second generation tasks with minimal degradation. As for AudioGen, the SOTA LM audio generation model, its architecture is inherently suited for generating longer audio compared to diffusion models, its performance degrades significantly as the generation length increases from 10 seconds to 96 seconds, accompanied by substantial increases in memory and time costs. For SD-audio, improved FD performance is observed when increasing the generation length from 10 to 32 seconds, likely due to the majority of its training data being focused on longer durations. However, other metrics consistently decline from 10 to 96 seconds, although the degradation is less pronounced compared to AudioGen. This highlights the robustness of diffusion models in generating longer outputs within their maximum training window.

Method	FD↓	FAD↓	KL↓	mCLAP↑
SD-audio (10s)	38.23	6.20	2.19	0.40
SD-audio (32s)	25.52	6.43	2.24	0.37
SD-audio (64s)	25.82	6.12	2.25	0.35
SD-audio (96s)	30.11	6.54	2.38	0.33
AudioGen (10s)	16.88	4.36	1.52	0.55
AudioGen (32s)	18.54	4.81	1.71	0.50
AudioGen (64s)	19.53	5.02	1.76	0.50
AudioGen (96s)	18.88	5.44	1.78	0.49
Make2 (10s)	14.37	1.12	1.28	0.57
Make2 (27s)	18.49	2.26	1.55	0.49
SaFa (32s)	15.21	1.45	1.25	0.57
SaFa (64s)	15.14	1.25	1.24	0.57
SaFa (96s)	15.36	1.33	1.25	0.57

Table 5. Quantitative Comparison with Training-Based Variable-Length Audio Generation Models.

8.2. Joint Diffusion on Open-Source Checkpoint

In this subsection, we discuss several design flaws in existing open-source audio generation models that limit the application of training-free methods, such as the joint diffusion method. In this way, we show our audio generation model as a potential contribution to advancing training-free approaches in audio generation.

Adaptation on Existing T2A Models Specifically, AudioLDM [32] and Tango [33] are trained with a fixed 10.24-second window, padding shorter clips with zeros or truncating longer clips. This flexible training pipeline causes unexpected end silence in generated audios. Consequently, implementing joint diffusion methods with these models often results in sudden silence in the overlap regions. Stable Diffusion Audio [14] is also trained with a fixed 96-second window and generates variable-length outputs by truncation, making it similarly challenging to adapt for joint diffusion methods. In comparison, Make-An-Audio2 follows a training pipeline similar to ours, using variable-length audio without excessive padding. It organizes samples into different buckets based on the length during training, randomly selecting samples from the same bucket within each batch. However, we observe some anomalous phenomena when

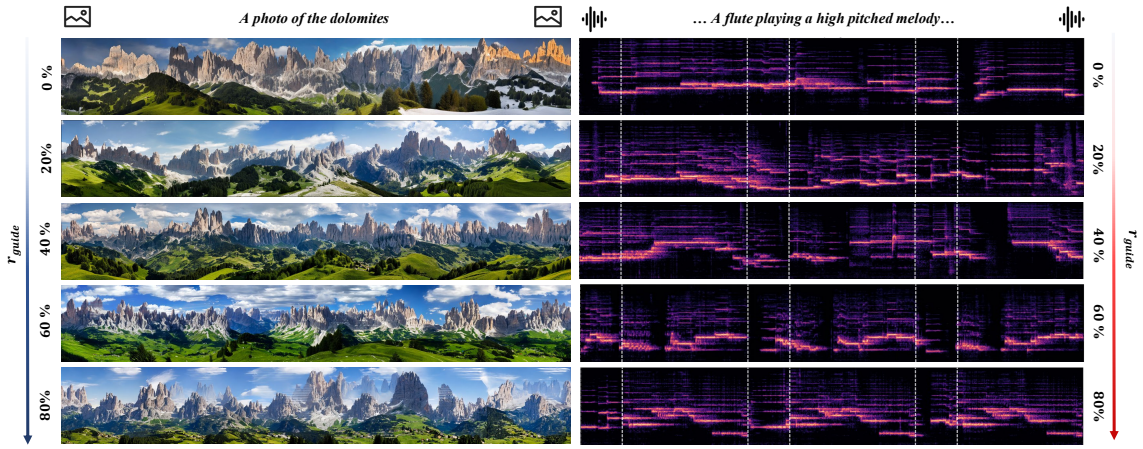


Figure 6. The effect of the trajectory guidance rate r_{guide} in Reference-Guided Swap on long spectrum and panorama generation.

Method	FD↓	FAD↓	KL↓	mCLAP↑
Make2	18.01	2.01	1.49	0.50
MD	65.28	17.70	3.22	0.24
MAD	62.53	16.88	3.05	0.26
SaFa	**15.36**	**1.32**	**1.27**	**0.57**

Table 6. Quantitative comparisons of joint diffusion on 24-second audio generation on Make-An-Audio2 [13].

Method	FAD↓	FD↓	KL↓	mCLAP↑
SD-audio	6.44	25.26	2.18	0.37
MD	7.06	38.78	2.24	0.35
MAD	7.56	38.9	2.21	0.35
SaFa	**4.19**	**24.40**	**1.96**	**0.41**

Table 8. Quantitative comparisons of join diffusion on 24-second audio generation on Stable Diffusion Audio [14].

applying Make2 with joint diffusion methods.

Comparison on 1D Convolution VAE Latent As shown in Figure 7, when applying the joint diffusion method to Make2, short abrupt transitions appear at the end of each overlap region. Although SaFa significantly improves blending and generation quality compared to MD and MAD, these abrupt transitions still persist. Through experiments, we identify two main causes of this issue: (1) The VAE latent map of Make-An-Audio2 is sensitive to the last token from an adjacent subview. To mitigate this, we apply Self-loop Swap with a five-token forward shift on the overlap regions. (2) Its VAE model is less robust to linear operations on the latent map compared to AudioLDM [32]. By performing concatenation at $t = 0$ on the mel-spectrogram rather than on the latent map, we effectively resolve this issue. As a result, the improved method, SaFa+, performs well in Figure 7.

For quantitative comparison in Table 6, our method, SaFa, significantly outperforms Make2 and other joint diffusion methods across all metrics for 24-second generation tasks.

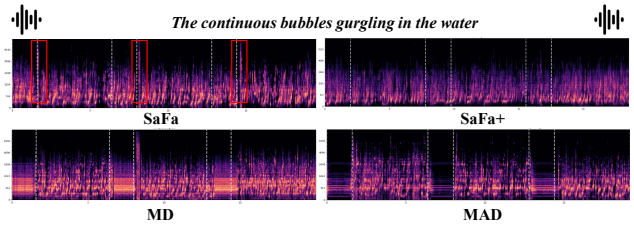


Figure 7. The long-form spectrum generated by various joint diffusion methods based on Make-An-Audio2.

Comparison on Waveform VAE Latent As shown in Table 8, we also compare SaFa with other joint diffusion methods on waveform VAE latents using the open-source checkpoint from SD-Audio. We note that although we restrict the initial latent maps on 10 seconds to adapt joint diffusion method (while the model was trained on fixed 96-second latent maps), this unoptimized setting does not compromise fairness. As a result, SaFa also significantly outperforms existing methods when evaluated on AudioCaps with SD-audio.

8.3. Effect of Guidance Rate and Swap Interval

In Figure6, we further demonstrate the progressive transition from cross-view diversity to similarity by varying r_{guide} in both mel-spectrum and panorama generation using Reference-Guided Latent Swap. All other settings for SaFa remain consistent with Section 6. As shown in Figure 6, using an appropriate trajectory guidance rate r_{guide}, 20% to 40%, results in unified cross-view coherence while preserving the diversity of local subviews. However, as the guidance rate r_{guide} increases beyond 60%, excessive repetition and artifacts begin to appear. This occurs because Reference-Guided Swap is a unidirectional operation, where the denoising process of the reference view is independent and unaffected by each subview. Consequently, it does not adapt as seamlessly to subviews in the later stages as the bidirectional Self-Loop Swap operation does. This is also one of the reasons why we restrict Reference-Guided Swap

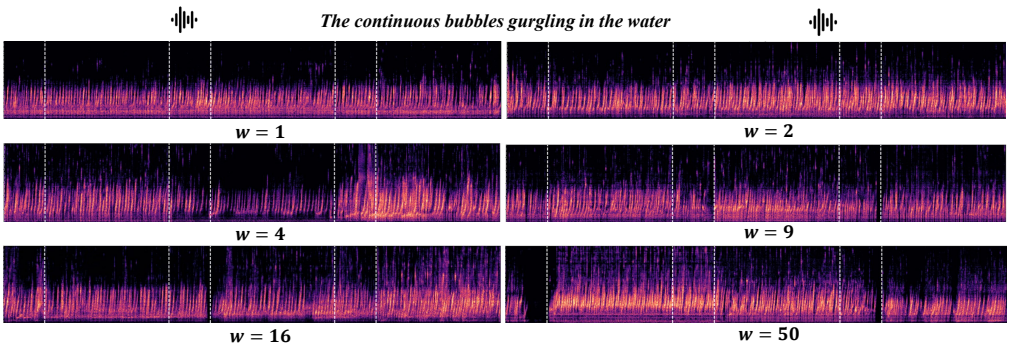


Figure 8. The effect of the swap interval w (Eq. 10) of Self-Loop Latent Swap on spectrum generation. Better transition is achieved with lower values of w, 1 or 2, which indicate a high swap frequency between two step-wise differential trajectories to enhance the high-frequency component in the denoised mel-spectrum with better-blender transitions.

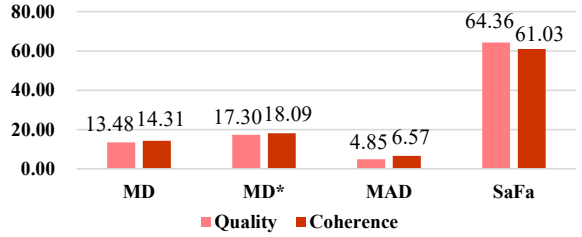


Figure 9. User study results on audio generation.

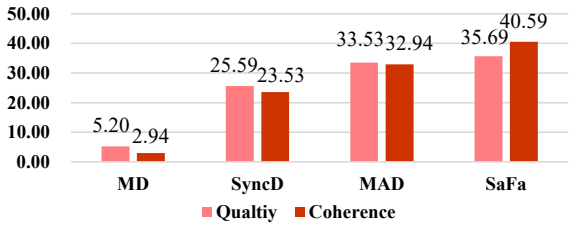


Figure 10. User study results on panorama generation.

to the early denoising stages.

To further explore the effects of the swap interval w (in Eq. 10), we apply the Self-Loop Latent Swap with various w values in spectrum generation, as shown in Figure 8. We observe that using a small swap interval (1 or 2), corresponding to higher swap frequencies, produces smoother transitions. Conversely, larger w values indicate larger swap units, resulting in less seamless transitions between subviews. This outcome aligns with the high-frequency variability of mel tokens, leading us to default the Self-Loop Latent Swap to frame-level operations with $w = 1$ for optimal performance.

8.4. Length Adaptation on Panorama Generation

In Table 9, We utilize SD 2.0 model to estimate performance of SaFa on panorama images with resolutions of 512×1600, 512×3200, and 512×4800. As a result, SaFa maintains stable and great performance across all evaluated metrics in different length output.

9. User Study

For subjective evaluation, we randomly select samples from the qualitative results of the top four methods in audio and panorama generation for user studies. We use the same notation as in Section 6.1. Specifically, SaFa is compared with MD, MD*, and MAD for audio generation, while for panorama generation, SaFa is compared with MD, MAD, and SyncD. For each task, we randomly select 30 parallel comparison groups (plus 2 additional pairs as a vigilance group) from the four compared methods, evenly distributed

Method	CLIP ↑	FID ↓	KID ↓	I-LPIPS ↓	I-StyleL ↓
SaFa (1600)	31.88	34.47	9.71	0.59	1.67
SaFa (3200)	31.84	34.71	9.91	0.61	1.74
SaFa (4800)	31.88	34.97	10.68	0.62	1.78

Table 9. Length adaptation of SaFa on panorama generation.

across six prompts. We recruit 39 participants with basic machine learning knowledge but no prior familiarity with the research presented in this paper. Each participant is required to select the best sample from each of the 32 groups based on two evaluation dimensions: generation quality and global coherence. Semantic alignment is not considered, as most samples align well with the prompts semantically and cannot be easily distinguished in this regard. We ultimately collect 34 valid responses out of 39 participants. The results indicate that SaFa consistently outperforms the baseline methods, achieving superior human preference scores across both evaluation dimensions. Figure 9 highlight the significant preference of human evaluators for SaFa in both quality and coherence assessments of audio generation. This preference stems from the swap operator's enhanced adaptability to the inherent characteristics of spectral data, which lacks the typical global structural features or contours present in images. Meanwhile in Figure 10, with significantly faster inference speeds and relying solely on fixed self-attention windows, SaFa achieves comparable performance to SyncD and MAD in the subjective evaluation of panorama generation.

10. Theoretical Analysis of Refer-Guided Swap for Cross-View Similarity-Diversity Balance

Reference-Guided Latent Swap improves cross-view consistency comparing with independent denoising process with reference model directly. When SD-2.0 [41] is employed as the reference model Φ, we have the following proposition, which describes the difference between two updated samples from arbitrary starting points $\mathbf{x}_{t_2}^{(1)}, \mathbf{x}_{t_2}^{(2)} \in \mathcal{X}$:

Proposition Recall that the approximated reversed VP-SDE [5] used for conditionally generation in SD-2.0 is:

$$\mathrm{d}\mathbf{x} = \left[-\frac{1}{2}\beta(t)\mathbf{x} - \beta(t)s_\theta(\mathbf{x}, t, y)\right]\mathrm{d}t + \sqrt{\beta(t)}\mathrm{d}\tilde{\mathbf{w}}, \tag{12}$$

where $s_\theta(\mathbf{x}, t, y)$ is a estimation for $\nabla_{\mathbf{x}} \log p_t(\mathbf{x}|y)$, and $\tilde{\mathbf{w}}$ is a Wiener process when time flows backwards from $t = 1$ to $t = 0$. Denote that $\Phi_{t_2 \to t_1}(\cdot|y)$ is the the sampling procedure from t_2 to t_1 condition on y in SD-2.0, and $\sigma_{t_2 \to t_1}^2 = -\int_{t_2}^{t_1} \beta(u)\mathrm{d}u$. Assume that $\forall \mathbf{x} \in \mathcal{X}, \forall t \in [0, 1], \forall y \in \mathcal{Y}, \|s_\theta(\mathbf{x}, t, y)\|_2 \leq C$, then $\forall 0 \leq t_1 < t_2 \leq 1$, $\forall \mathbf{x}_{t_2}^{(1)}, \mathbf{x}_{t_2}^{(2)} \in \mathcal{X}$, $\forall y \in \mathcal{Y}$, $\forall \delta \in (0, 1)$, with probability at least $(1 - \delta)$,

$$\begin{aligned}\left\|\Phi_{t_2 \to t_1}\left(\mathbf{x}_{t_2}^{(1)}|y\right) - \Phi_{t_2 \to t_1}\left(\mathbf{x}_{t_2}^{(2)}|y\right)\right\|_2^2 \leq \exp(\sigma_{t_2 \to t_1}^2)\left[\left\|\mathbf{x}_{t_2}^{(1)} - \mathbf{x}_{t_2}^{(2)}\right\|_2 + 2C\left\|\int_{t_2}^{t_1} \exp\left(-\frac{1}{2}\sigma_{t_2 \to s}^2\right)\beta(s)\,\mathrm{d}s\right\|_2\right]^2 \\ + 2\sigma_{t_2 \to t_1}^2\left(d + 2\sqrt{d \cdot (-\log\delta)} + 2 \cdot (-\log\delta)\right).\end{aligned} \tag{13}$$

where d is the number of dimensions of $\mathbf{x}_{t_2}^{(1)}, \mathbf{x}_{t_2}^{(2)}, \mathbf{x}_{t_2}^{(ref)}$.

Proof Using method of variation of parameters, solution for pre-mentioned SDE (1), $\Phi_{t_2 \to t_1}(\mathbf{x}_{t_2}|y)$, can be written as

$$\Phi_{t_2 \to t_1}(\mathbf{x}_{t_2}|y) = \exp(\frac{1}{2}\sigma_{t_2 \to t_1}^2)\left[\mathbf{x}_{t_2} - \int_{t_2}^{t_1} \exp(-\frac{1}{2}\sigma_{t_2 \to s}^2)\beta(s)s_\theta(\mathbf{x}_s, s, y)\mathrm{d}s\right] + \int_{t_2}^{t_1} \sqrt{\beta(t)}\mathrm{d}\tilde{\mathbf{w}}, \tag{14}$$

so for $\forall \mathbf{x}_{t_2}^{(1)}, \mathbf{x}_{t_2}^{(2)} \in \mathcal{X}$, we have:

$$\begin{aligned}&\left\|\Phi_{t_2 \to t_1}\left(\mathbf{x}_{t_2}^{(1)}|y\right) - \Phi_{t_2 \to t_1}\left(\mathbf{x}_{t_2}^{(2)}|y\right)\right\|_2^2 \\ =&\|\exp(\frac{1}{2}\sigma_{t_2 \to t_1}^2)\Big[(\mathbf{x}_{t_2}^{(1)} - \mathbf{x}_{t_2}^{(2)}) - \int_{t_2}^{t_1} \exp(-\frac{1}{2}\sigma_{t_2 \to s}^2)\beta(s)(s_\theta(\mathbf{x}_s^{(1)}, s, y) - s_\theta(\mathbf{x}_s^{(2)}, s, y))\mathrm{d}s\|_2^2\Big] \\ &+ \int_{t_2}^{t_1} \sqrt{\beta(t)}\mathrm{d}\tilde{\mathbf{w}}_1 - \int_{t_2}^{t_1} \sqrt{\beta(t)}\mathrm{d}\tilde{\mathbf{w}}_2\|_2^2 \\ =&\exp(\sigma_{t_2 \to t_1}^2)\left\|\mathbf{x}_{t_2}^{(1)} - \mathbf{x}_{t_2}^{(2)} + \int_{t_2}^{t_1} \exp(-\frac{1}{2}\sigma_{t_2 \to s}^2)\beta(s)(s_\theta(\mathbf{x}_s^{(1)}, s, y) - s_\theta(\mathbf{x}_s^{(2)}, s, y))\mathrm{d}s\right\|_2^2 \\ &+ \left\|\int_{t_2}^{t_1} \sqrt{\beta(t)}\mathrm{d}\tilde{\mathbf{w}}_1 - \int_{t_2}^{t_1} \sqrt{\beta(t)}\mathrm{d}\tilde{\mathbf{w}}_2\right\|_2^2 \\ \leq&\exp(\sigma_{t_2 \to t_1}^2)\left[\left\|\mathbf{x}_{t_2}^{(1)} - \mathbf{x}_{t_2}^{(2)}\right\|_2 + \left\|\int_{t_2}^{t_1} \exp(-\frac{1}{2}\sigma_{t_2 \to s}^2)\beta(s)(s_\theta(\mathbf{x}_s^{(1)}, s, y) - s_\theta(\mathbf{x}_s^{(2)}, s, y))\mathrm{d}s\right\|_2\right]^2 \\ &+ 2\left\|\int_{t_2}^{t_1} \sqrt{\beta(t)}\mathrm{d}\tilde{\mathbf{w}}\right\|_2^2.\end{aligned} \tag{15}$$

From the assumption over $s_\theta(\mathbf{x}, t, y)$, we have:

$$
\begin{aligned}
&\left\|\Phi_{t_2\to t_1}\left(\mathbf{x}_{t_2}^{(1)}|y\right)-\Phi_{t_2\to t_1}\left(\mathbf{x}_{t_2}^{(2)}|y\right)\right\|_2^2 \\
=&\left\|\exp\left(\frac{1}{2}\sigma_{t_2\to t_1}^2\right)\left[(\mathbf{x}_{t_2}^{(1)}-\mathbf{x}_{t_2}^{(2)})-\int_{t_2}^{t_1}\exp\left(-\frac{1}{2}\sigma_{t_2\to s}^2\right)\beta(s)\big(s_\theta(\mathbf{x}_s^{(1)},s,y)-s_\theta(\mathbf{x}_s^{(2)},s,y)\big)\mathrm{d}s\right]\right. \\
&\left.+\int_{t_2}^{t_1}\sqrt{\beta(t)}\,\mathrm{d}\tilde{\mathbf{w}}_1-\int_{t_2}^{t_1}\sqrt{\beta(t)}\,\mathrm{d}\tilde{\mathbf{w}}_2\right\|_2^2 \\
=&\exp\left(\sigma_{t_2\to t_1}^2\right)\left\|\mathbf{x}_{t_2}^{(1)}-\mathbf{x}_{t_2}^{(2)}+\int_{t_2}^{t_1}\exp\left(-\frac{1}{2}\sigma_{t_2\to s}^2\right)\beta(s)\big(s_\theta(\mathbf{x}_s^{(1)},s,y)-s_\theta(\mathbf{x}_s^{(2)},s,y)\big)\mathrm{d}s\right\|_2^2 \\
&+\left\|\int_{t_2}^{t_1}\sqrt{\beta(t)}\,\mathrm{d}\tilde{\mathbf{w}}_1-\int_{t_2}^{t_1}\sqrt{\beta(t)}\,\mathrm{d}\tilde{\mathbf{w}}_2\right\|_2^2 \\
\le&\exp\left(\sigma_{t_2\to t_1}^2\right)\left[\left\|\mathbf{x}_{t_2}^{(1)}-\mathbf{x}_{t_2}^{(2)}\right\|_2+\left\|\int_{t_2}^{t_1}\exp\left(-\frac{1}{2}\sigma_{t_2\to s}^2\right)\beta(s)\big(s_\theta(\mathbf{x}_s^{(1)},s,y)-s_\theta(\mathbf{x}_s^{(2)},s,y)\big)\mathrm{d}s\right\|_2\right]^2 \\
&+2\left\|\int_{t_2}^{t_1}\sqrt{\beta(t)}\,\mathrm{d}\tilde{\mathbf{w}}\right\|_2^2.
\end{aligned}
\tag{16}
$$

Then we complete the proof for Proposition 10.

Corollary If we use introduce reference-guided latent swap operation before updating, by the definition of Swap$(\cdot)$, we have $\forall 0 \le t_1 < t_2 \le 1, \forall \mathbf{x}_{t_2}^{(1)}, \mathbf{x}_{t_2}^{(2)}, \mathbf{x}_{t_2}^{(ref)} \in \mathcal{X}, \forall y \in \mathcal{Y}, \forall \delta \in (0,1)$, with probability at least $(1-\delta)$,

$$
\begin{aligned}
&\left\|\Phi_{t_2\to t_1}\left(\mathrm{Swap}(\mathbf{x}_{t_2}^{(ref)},\mathbf{x}_{t_2}^{(1)})|y\right)-\Phi_{t_2\to t_1}\left(\mathrm{Swap}(\mathbf{x}_{t_2}^{(ref)},\mathbf{x}_{t_2}^{(2)})|y\right)\right\|_2^2 \\
&\le\exp\left(\sigma_{t_2\to t_1}^2\right)\left[\left\|(1-W_{\text{swap}})\odot(\mathbf{x}_{t_2}^{(1)}-\mathbf{x}_{t_2}^{(2)})\right\|_2+2C\left\|\int_{t_2}^{t_1}\exp\left(-\frac{1}{2}\sigma_{t_2\to s}^2\right)\beta(s)\,\mathrm{d}s\right\|_2\right]^2 \\
&+2\sigma_{t_2\to t_1}^2\left(d+2\sqrt{d\cdot(-\log\delta)}+2\cdot(-\log\delta)\right).
\end{aligned}
\tag{17}
$$

Comparing with Eq.13, Eq.17 have a tighter upper bound, since $\mathrm{Swap}(\mathbf{x}_{t_2}^{(ref)},\mathbf{x}_{t_2}^{(1)})$, $\mathrm{Swap}(\mathbf{x}_{t_2}^{(ref)},\mathbf{x}_{t_2}^{(2)})$ shares the same part $W_{\text{swap}}\odot\mathbf{x}_{t_2}^{(ref)}$. This indicates that within a fixed time interval $[t_1,t_2]$, performing a reference-guided swap operation on the initial points before updating the sample points helps improve the similarity of the results.

According to Eq.13 and Eq.17, we can trade-off between similarity and diversity by tuning r_{guide}. As r_{guide} increases, the swap operation is employed more frequently applied during the sampling process, leading to higher similarity across subviews. Conversely, an increase in the L_2 distance between the final subview images signifies enhanced sample diversity.

11. Further Qualitative Comparison

More qualitative results on the audio generation are in Fig. 11 to 19 and panorama generation are in Fig. 20 to 30.

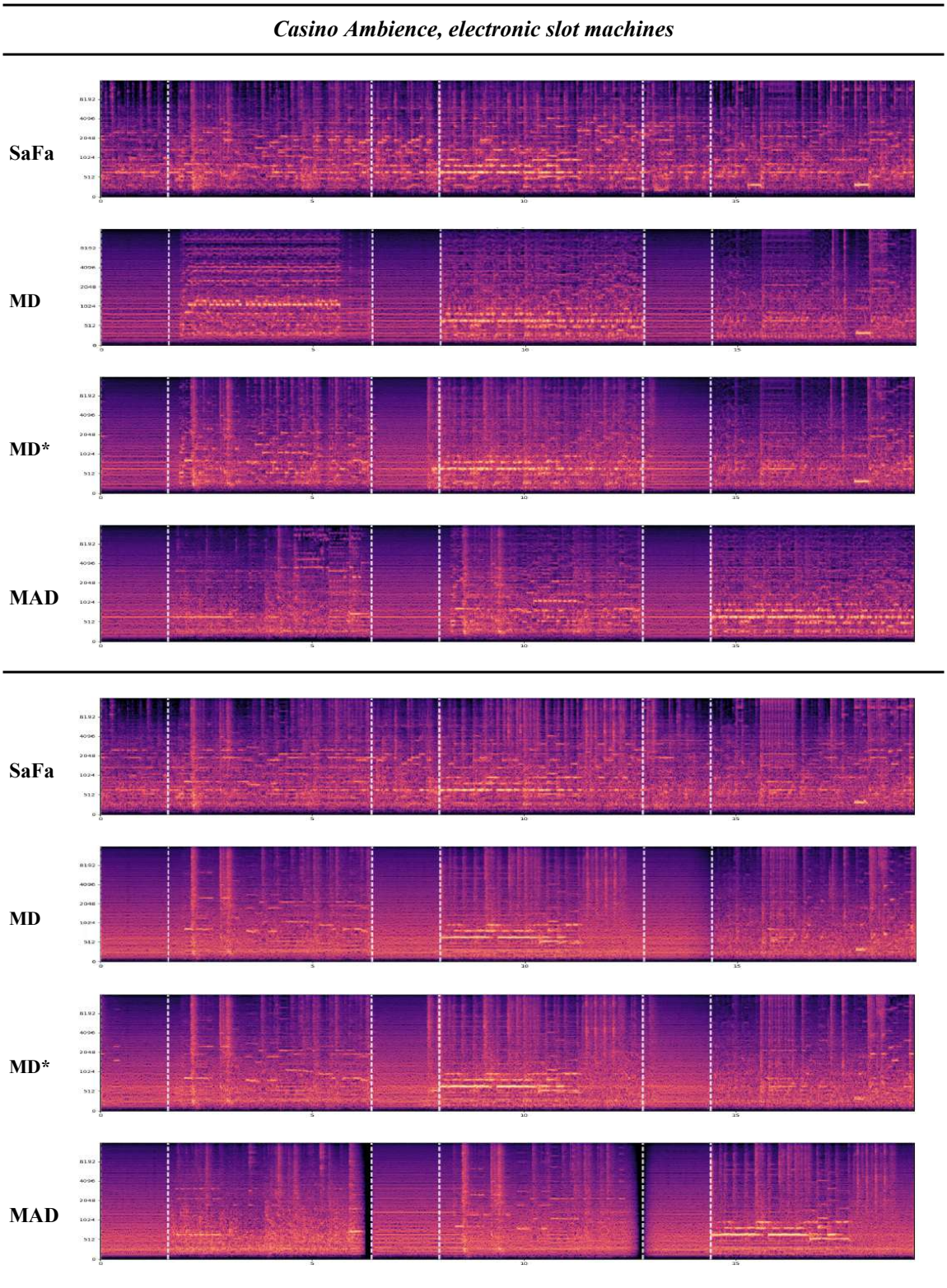


Figure 11. Qualitative comparison on soundscape generation. MD* represent an enhanced MD method with triangular windows.

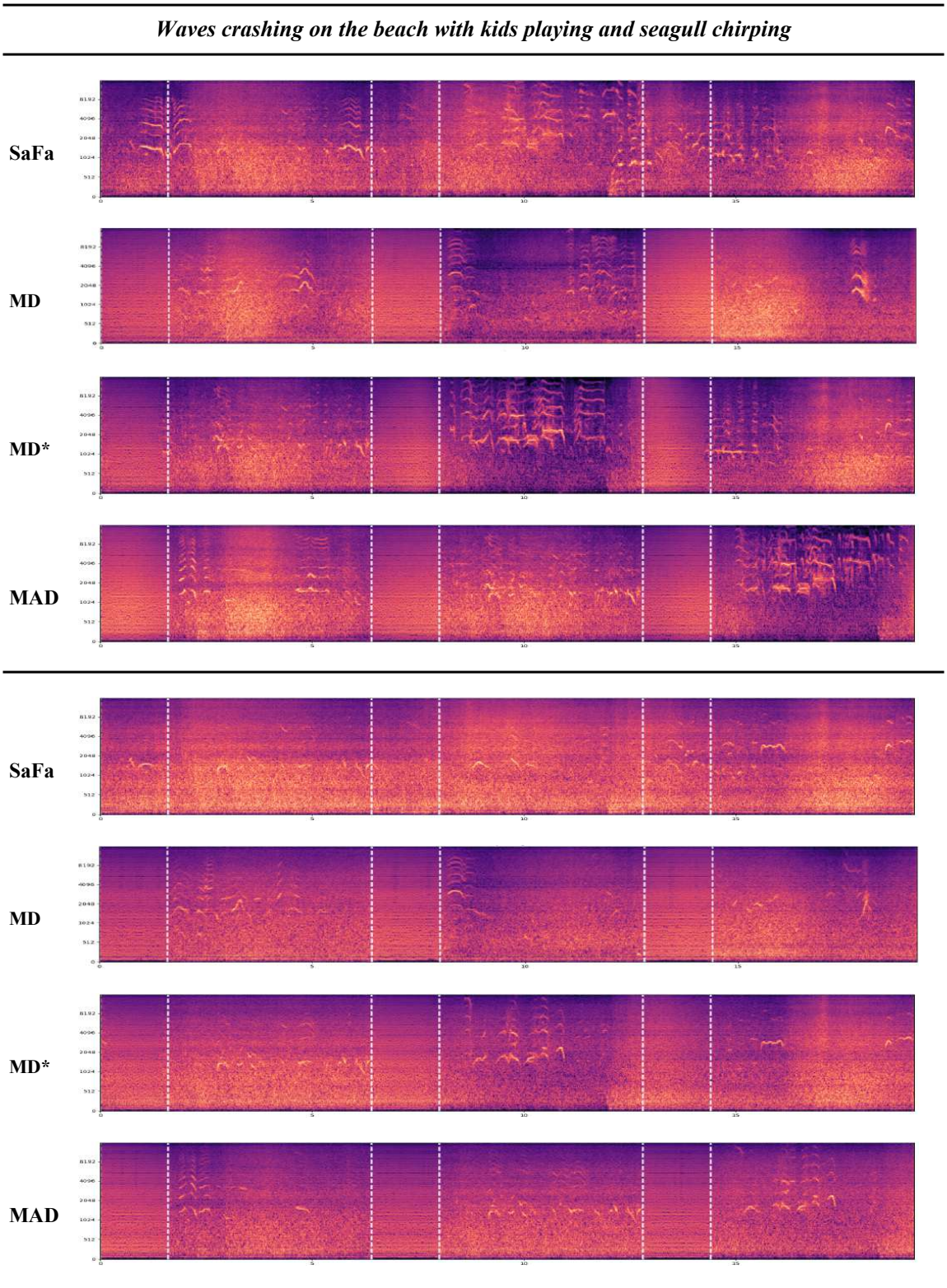


Figure 12. Qualitative comparison on soundscape generation. MD* represent an enhanced MD method with triangular windows.

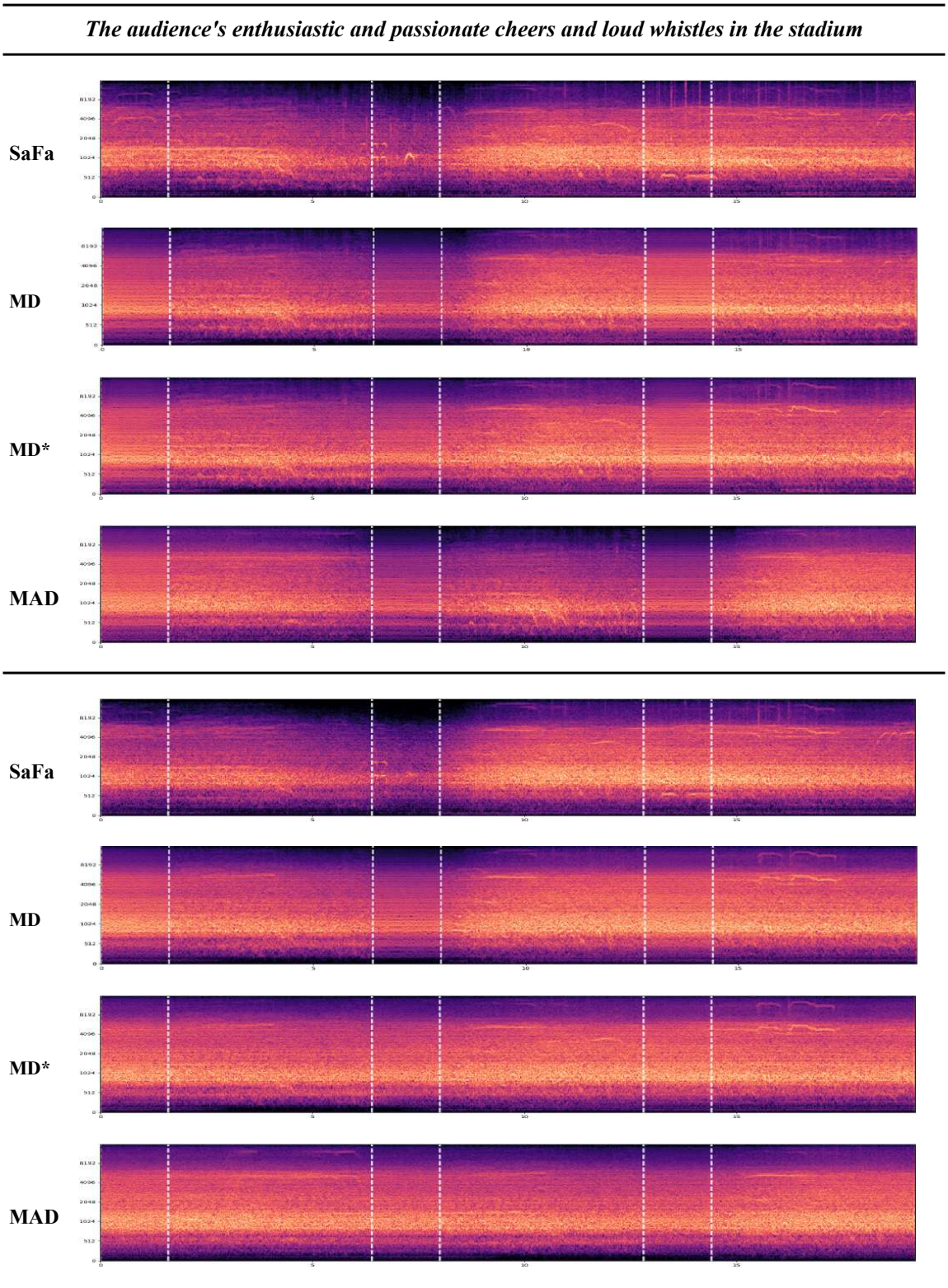


Figure 13. Qualitative comparison on soundscape generation. MD* represent an enhanced MD method with triangular windows.

Low fidelity audio from a live performance featuring a solo direct input acoustic guitar strumming airy, suspended open chords. Also present are occasional ambient sounds, perhaps papers being shuffled.

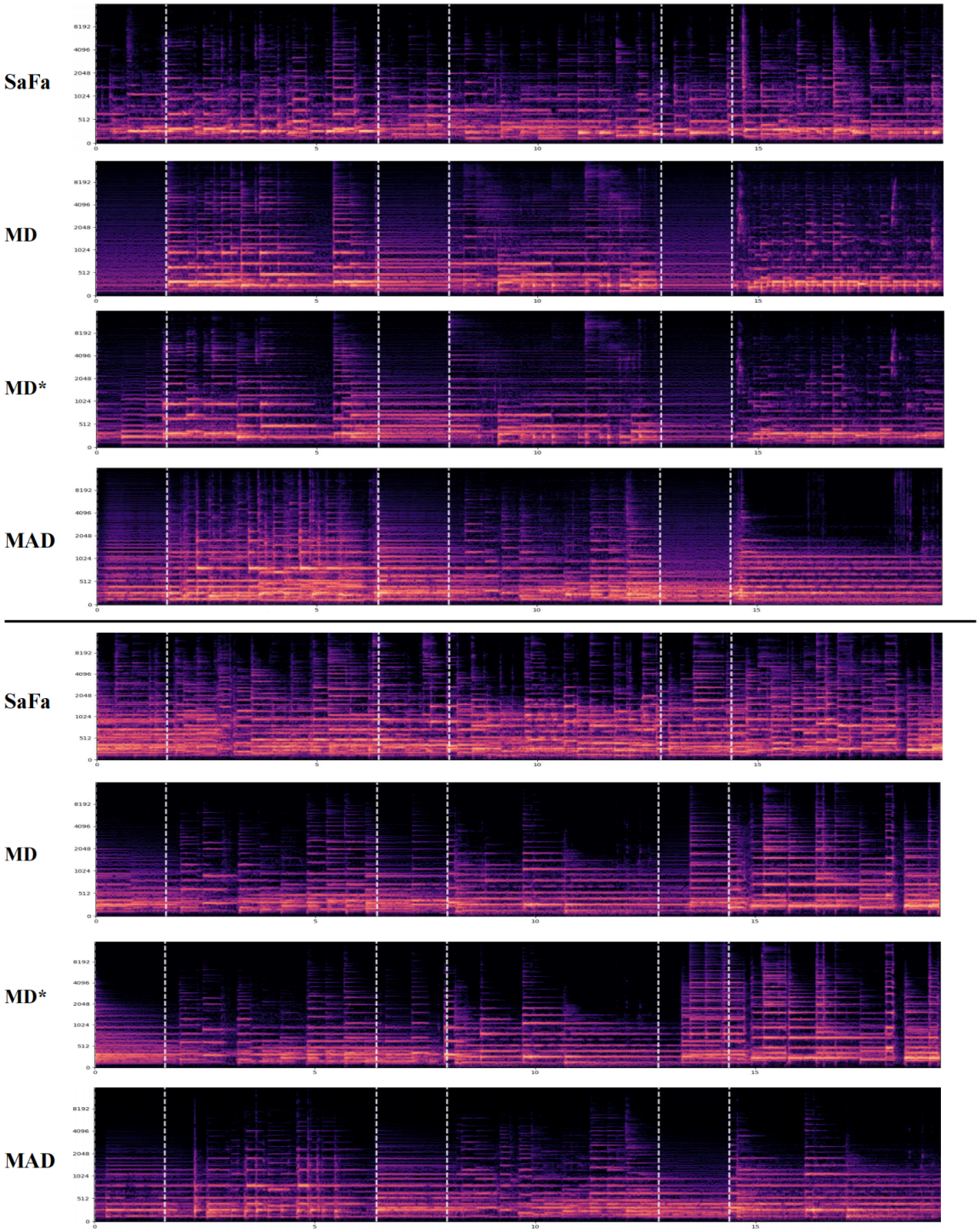


Figure 14. Qualitative comparison on music generation. MD* represent an enhanced MD method with triangular windows.

The instrumental music features an ensemble that resembles the orchestra.The melody is being played by a brass section while strings provide harmonic accompaniment.At the endof the music excerpt one can hear a double bass playing a long note and then a percussive noise.

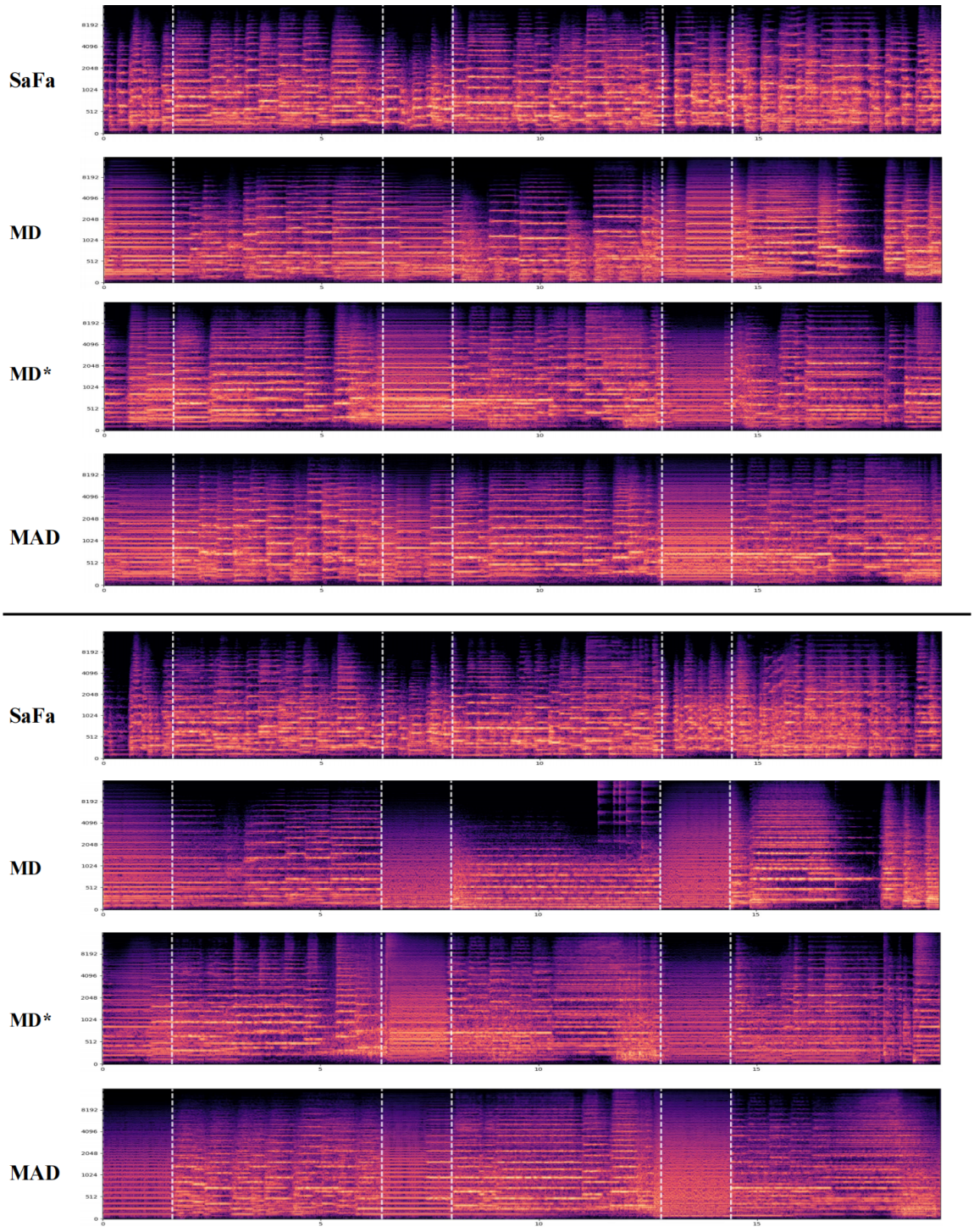
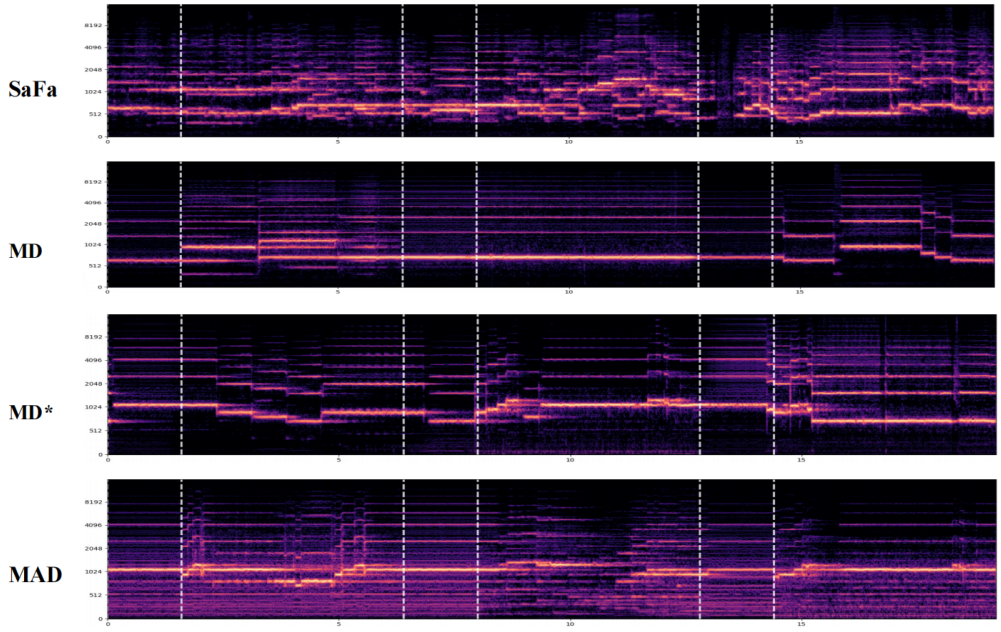


Figure 15. Qualitative comparison on music generation. MD* represent an enhanced MD method with triangular windows.

This instrumental song features a flute playing a high pitched melody. The melody starts off with one high pitched staccato note. After a brief pause, the flute plays two ascending notes in which the second note is sustained. Then a third higher note is played followed by a descending run of four more notes and ending on one higher note. There are no other instruments in this song. There is no percussion in this song. This song has a relaxing mood. This song can be played at a meditation center.



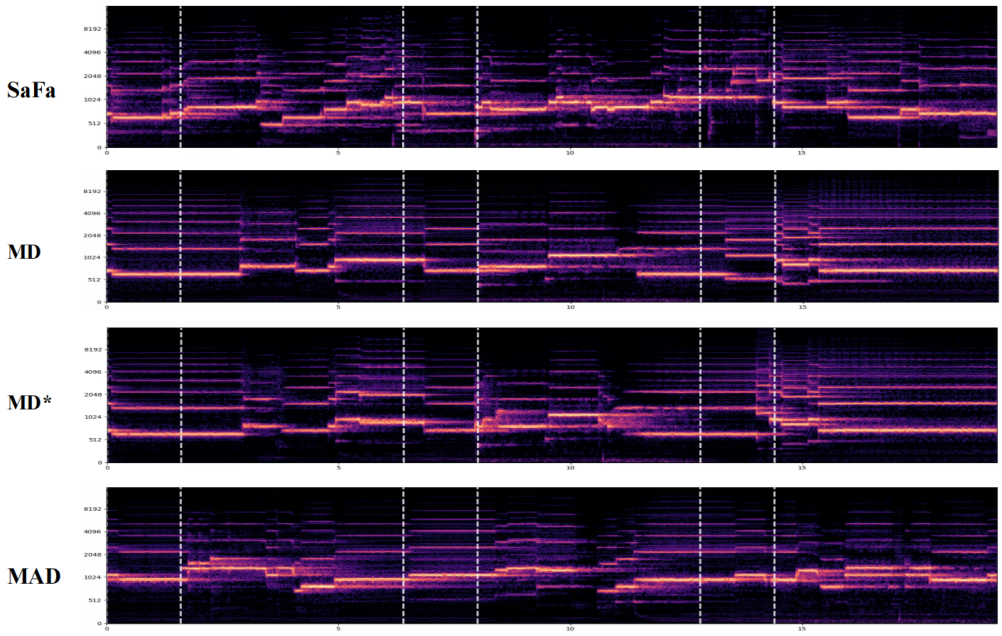


Figure 16. Qualitative comparison on music generation. MD* represent an enhanced MD method with triangular windows.

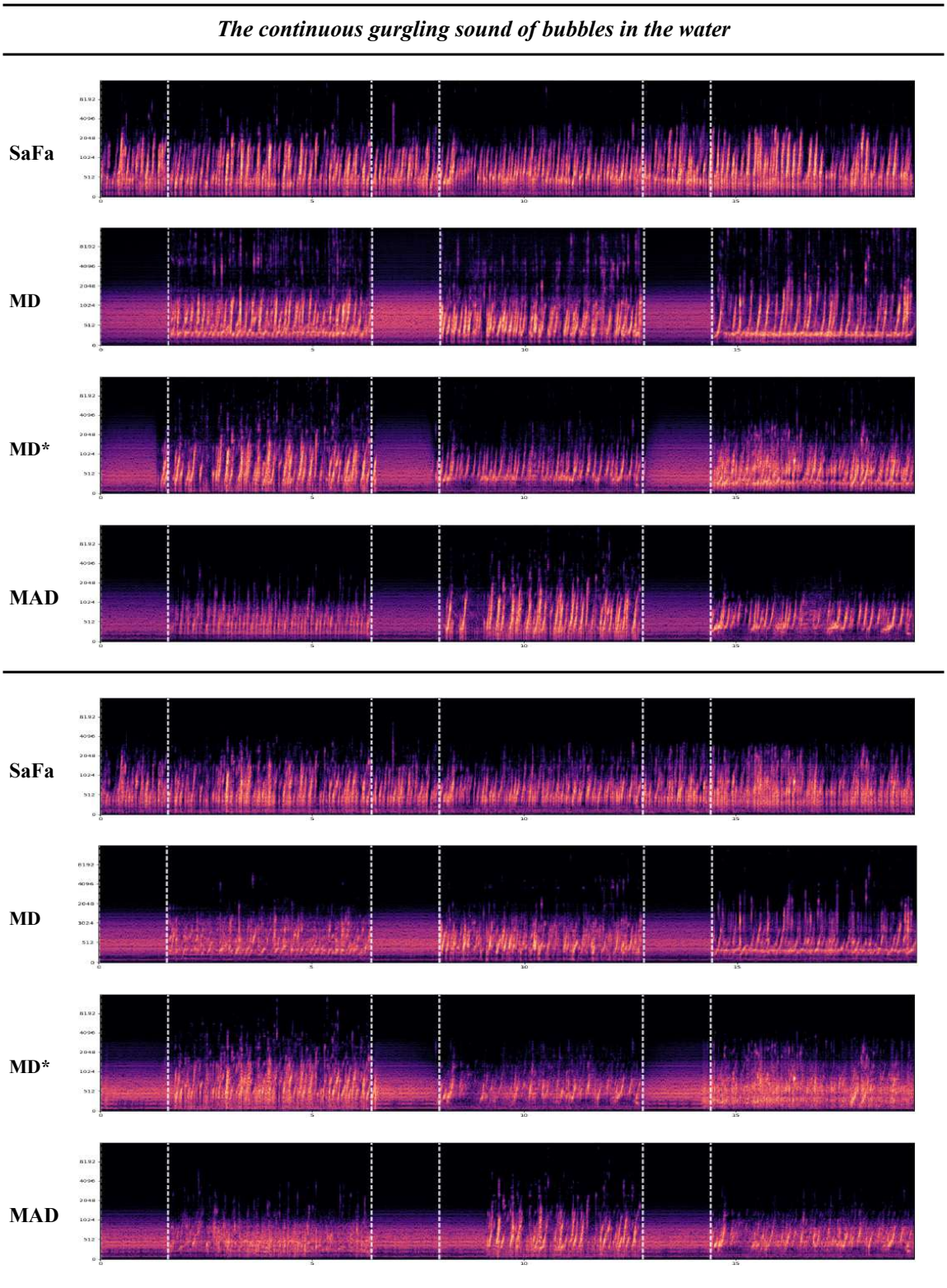


Figure 17. Qualitative comparison on audio effect generation. MD* represent an enhanced MD method with triangular windows.

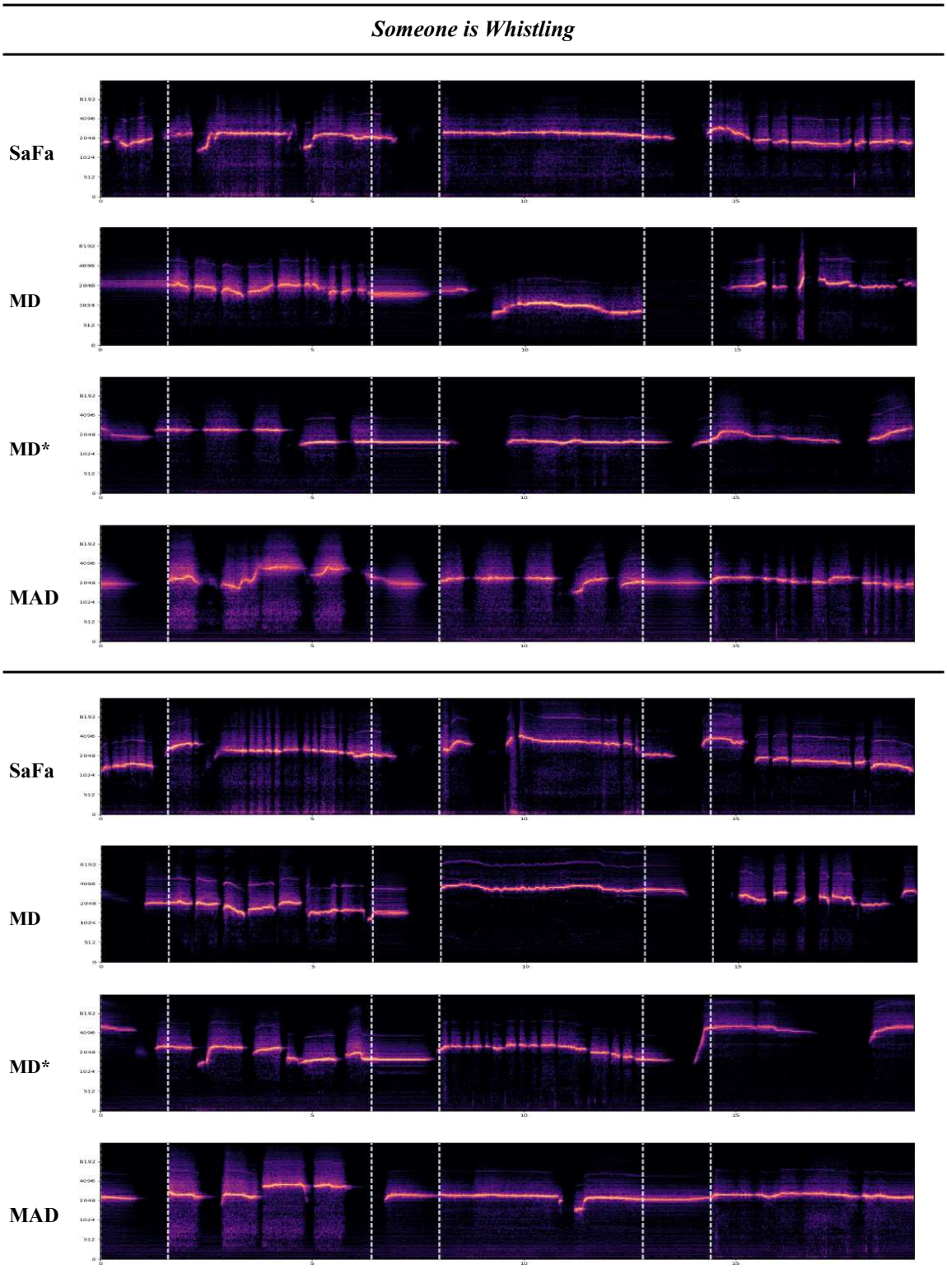


Figure 18. Qualitative comparison on audio effect generation. MD* represent an enhanced MD method with triangular windows.

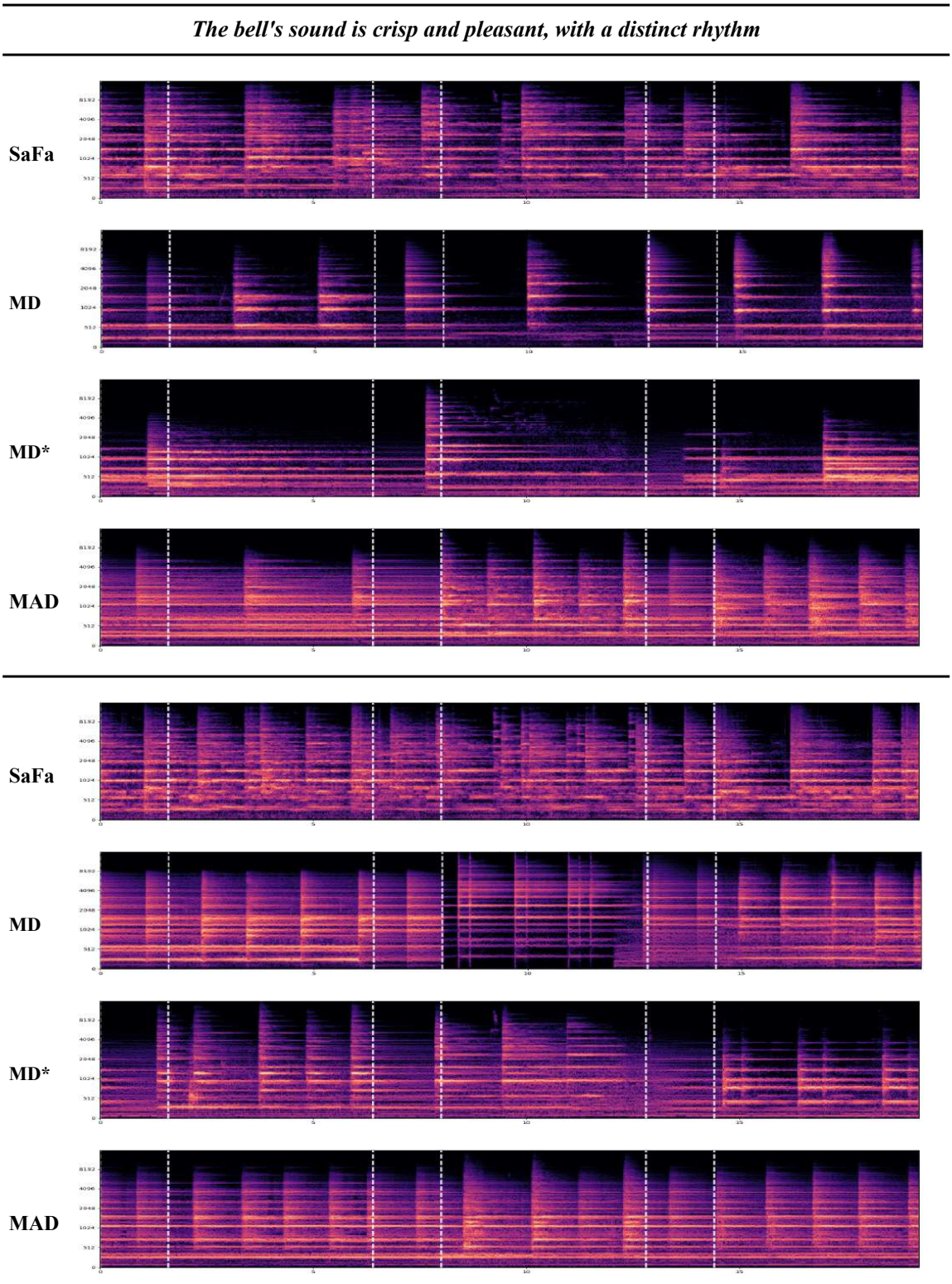


Figure 19. Qualitative comparison on audio effect generation. MD* represent an enhanced MD method with triangular windows.

Figure 20. Qualitative comparison on panorama image generation. MD* represent an enhanced MD method with triangular windows.

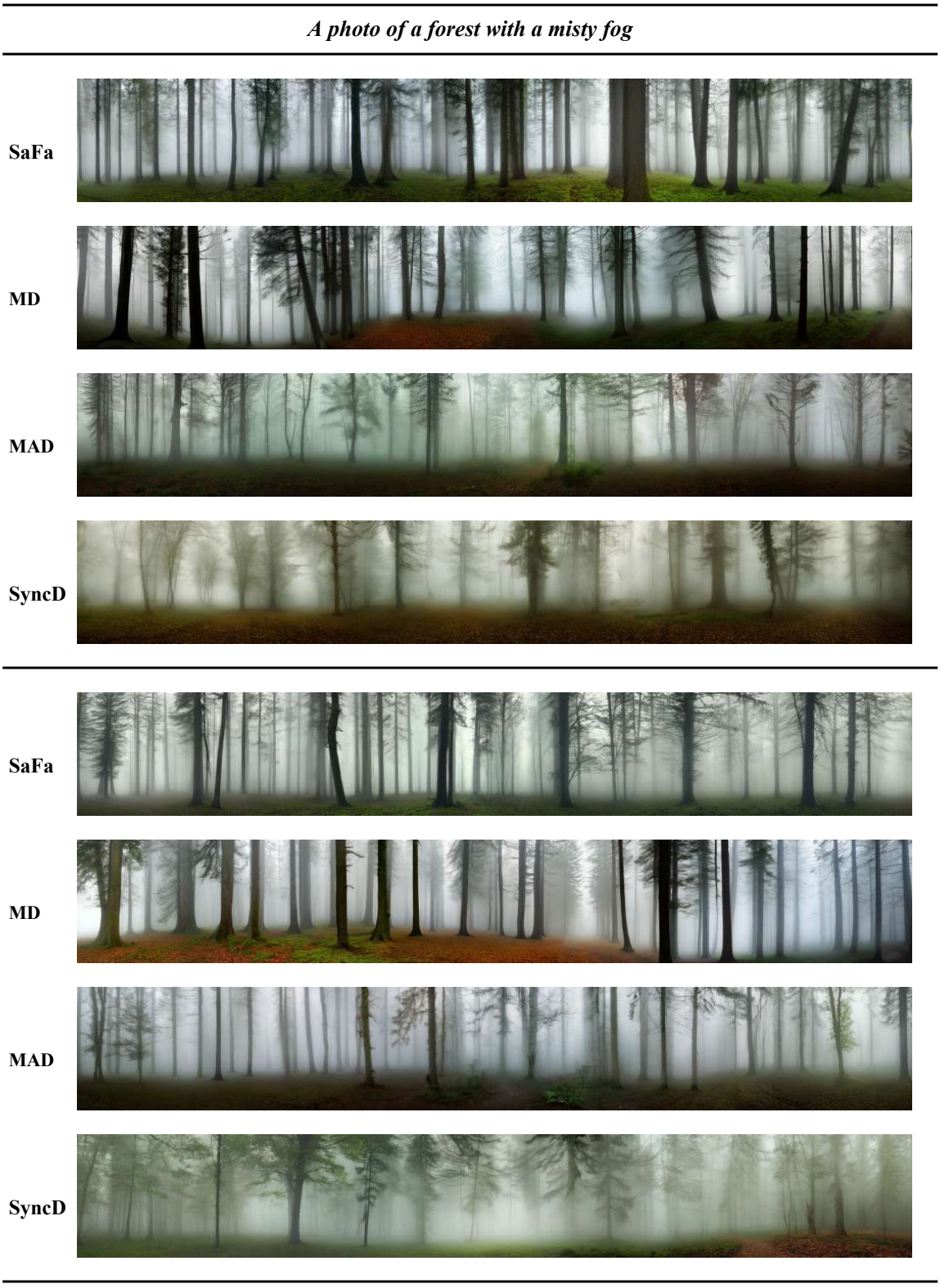


Figure 21. Qualitative comparison on panorama image generation. MD* represent an enhanced MD method with triangular windows.

Figure 22. Qualitative comparison on panorama image generation. MD* represent an enhanced MD method with triangular windows.

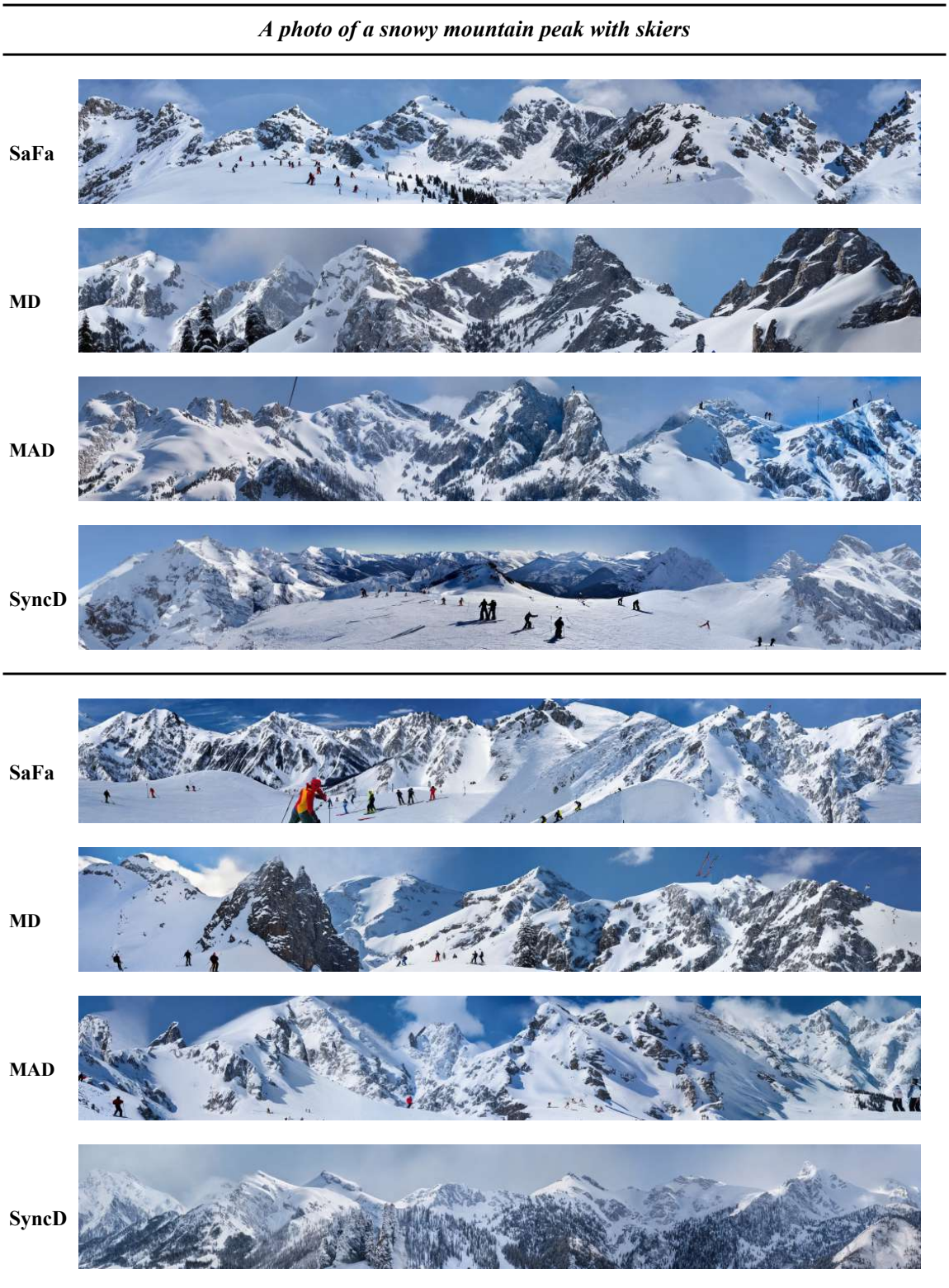


Figure 23. Qualitative comparison on panorama image generation. MD* represent an enhanced MD method with triangular windows.

Cartoon panorama of spring summer beautiful nature

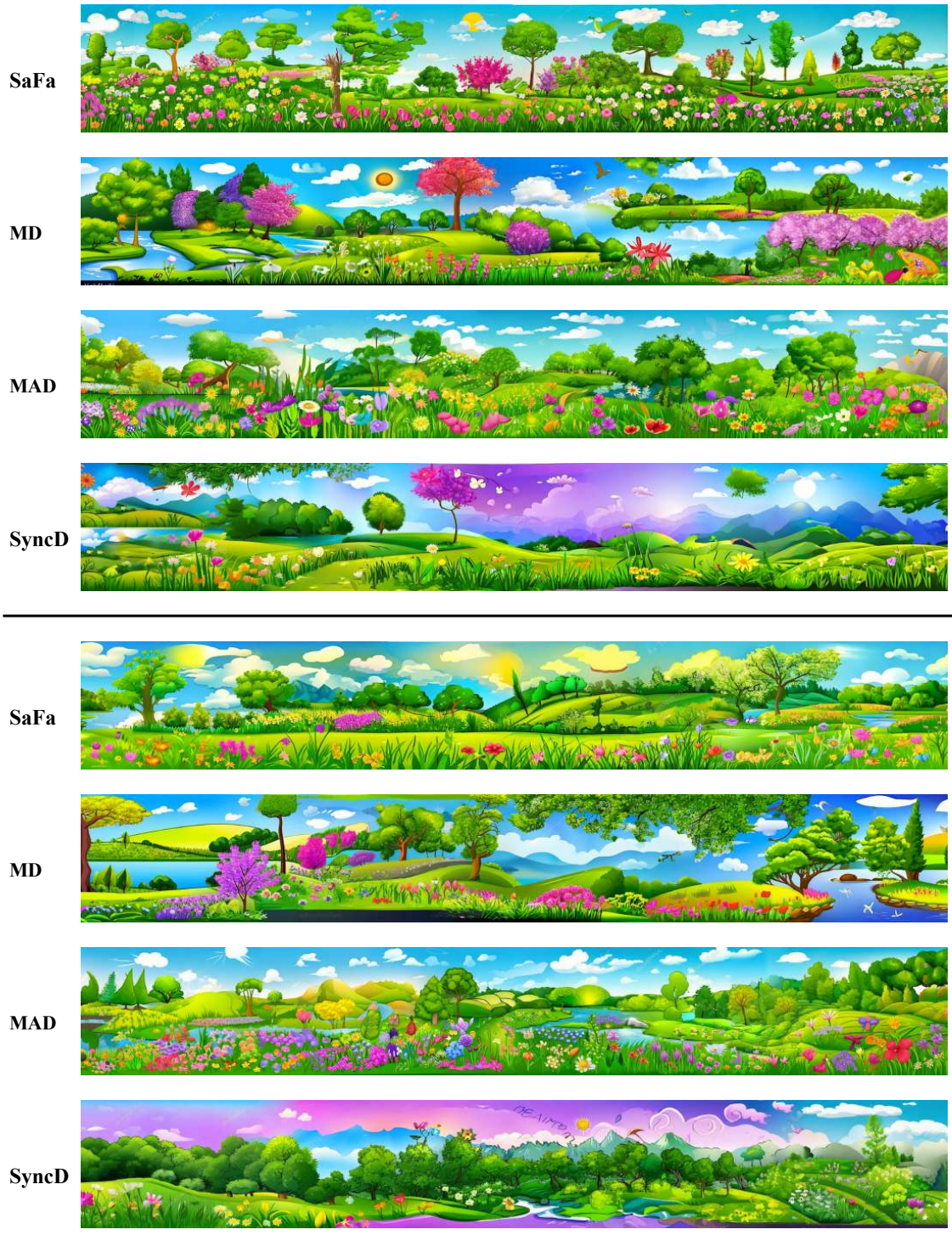


Figure 24. Qualitative comparison on panorama image generation. MD* represent an enhanced MD method with triangular windows.

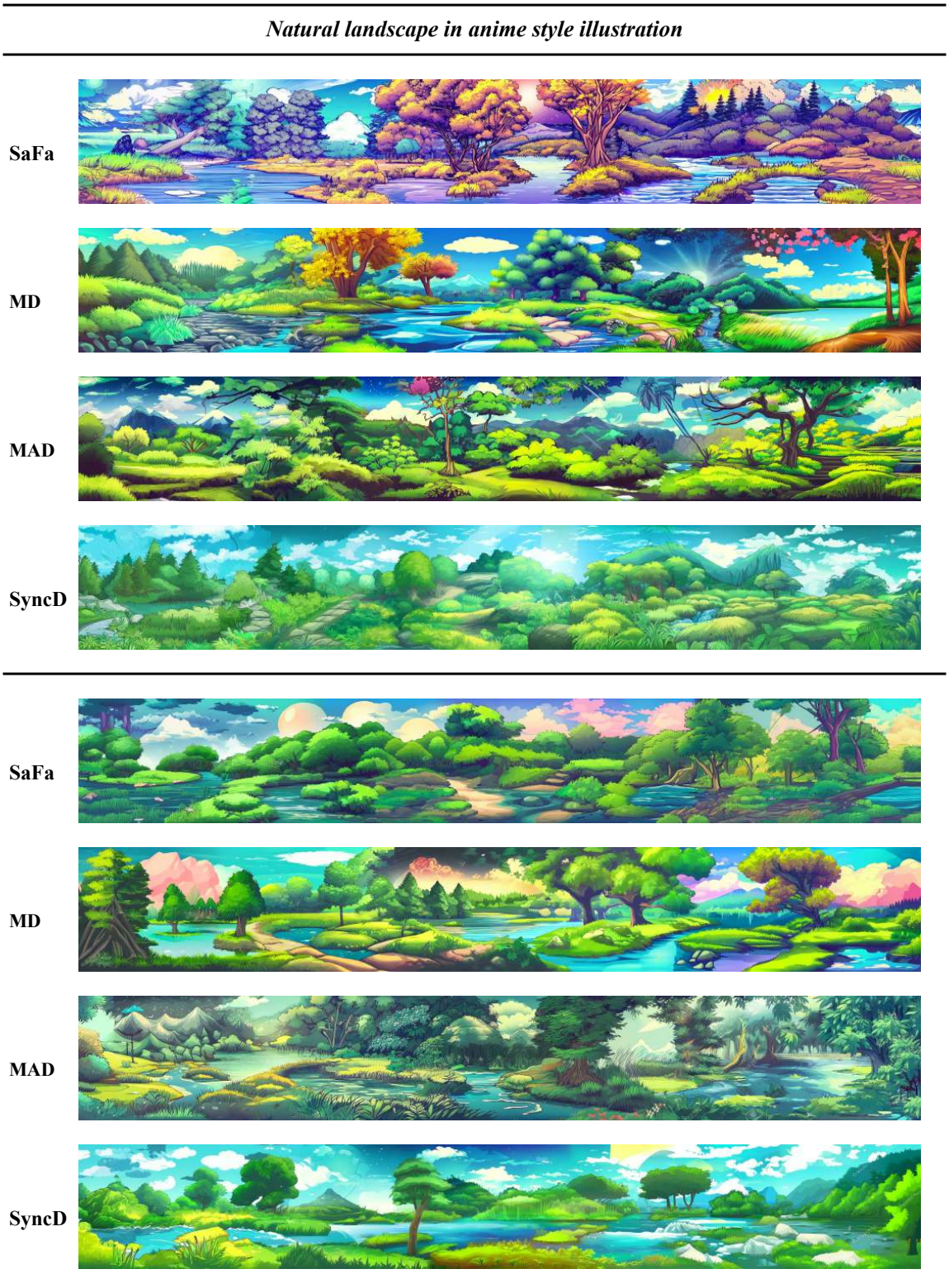


Figure 25. Qualitative comparison on panorama image generation. MD* represent an enhanced MD method with triangular windows.

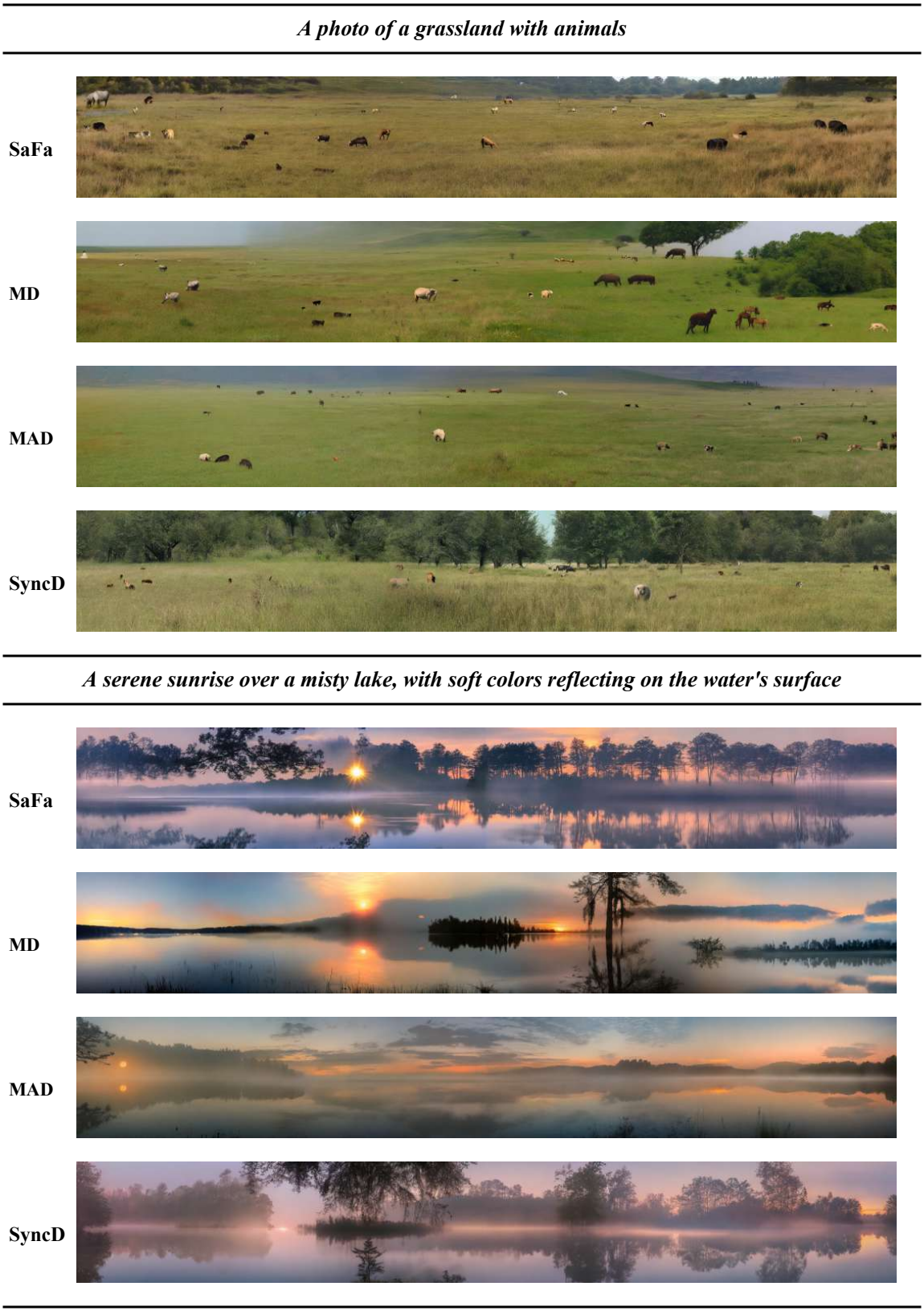


Figure 26. Qualitative comparison on panorama image generation. MD* represent an enhanced MD method with triangular windows.

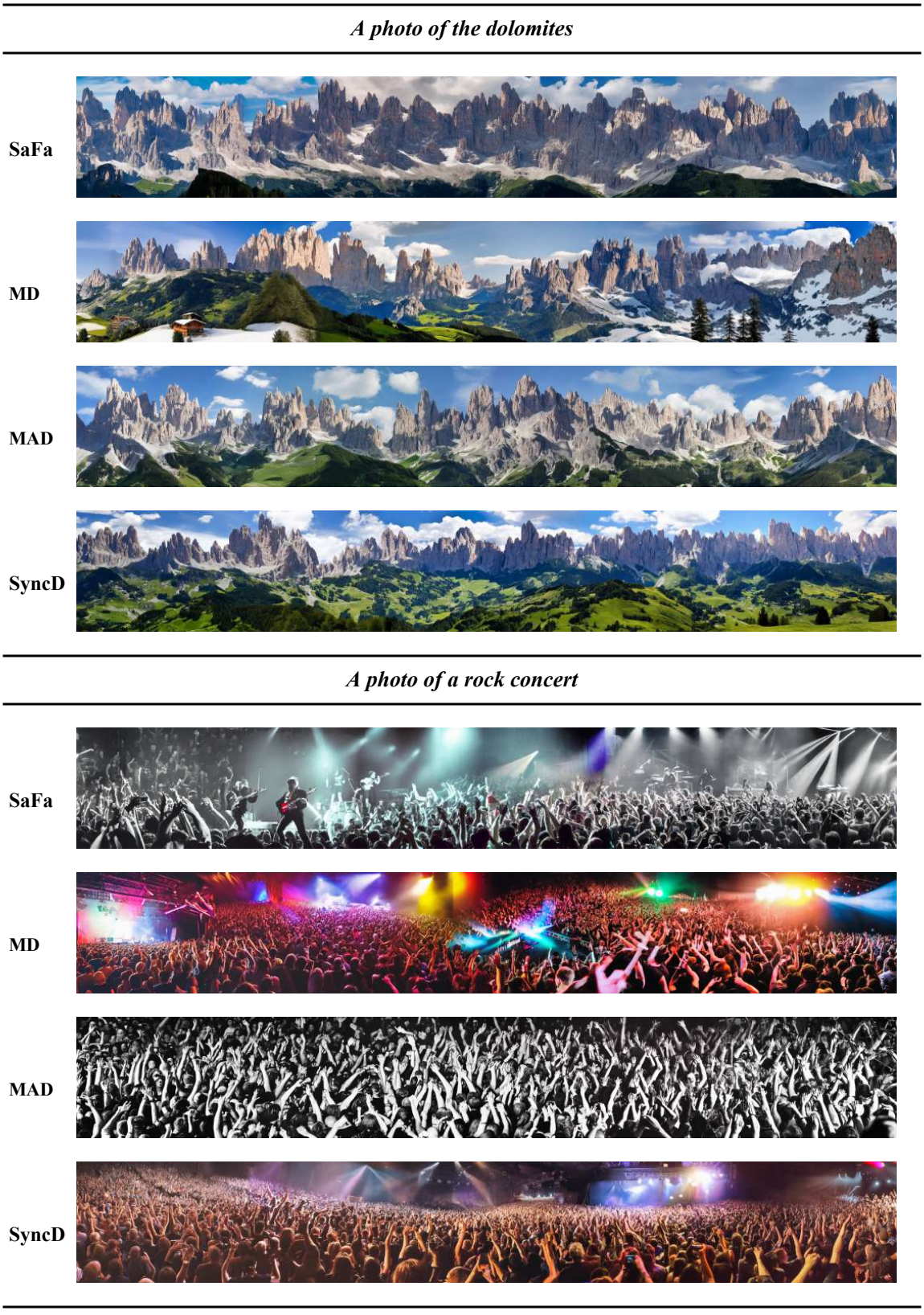


Figure 27. Qualitative comparison on panorama image generation. MD* represent an enhanced MD method with triangular windows.

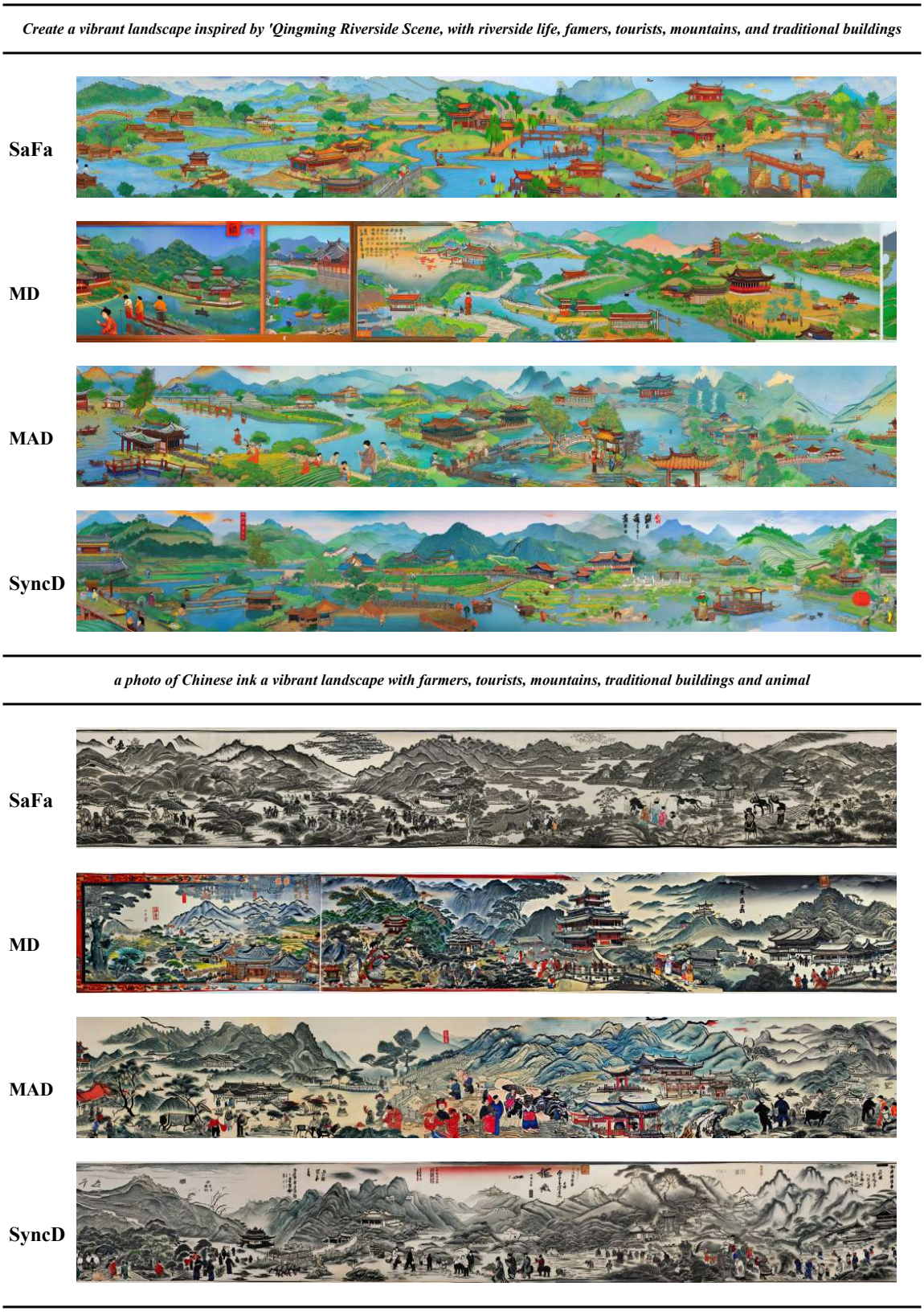


Figure 28. Qualitative comparison on panorama image generation. MD* represent an enhanced MD method with triangular windows.

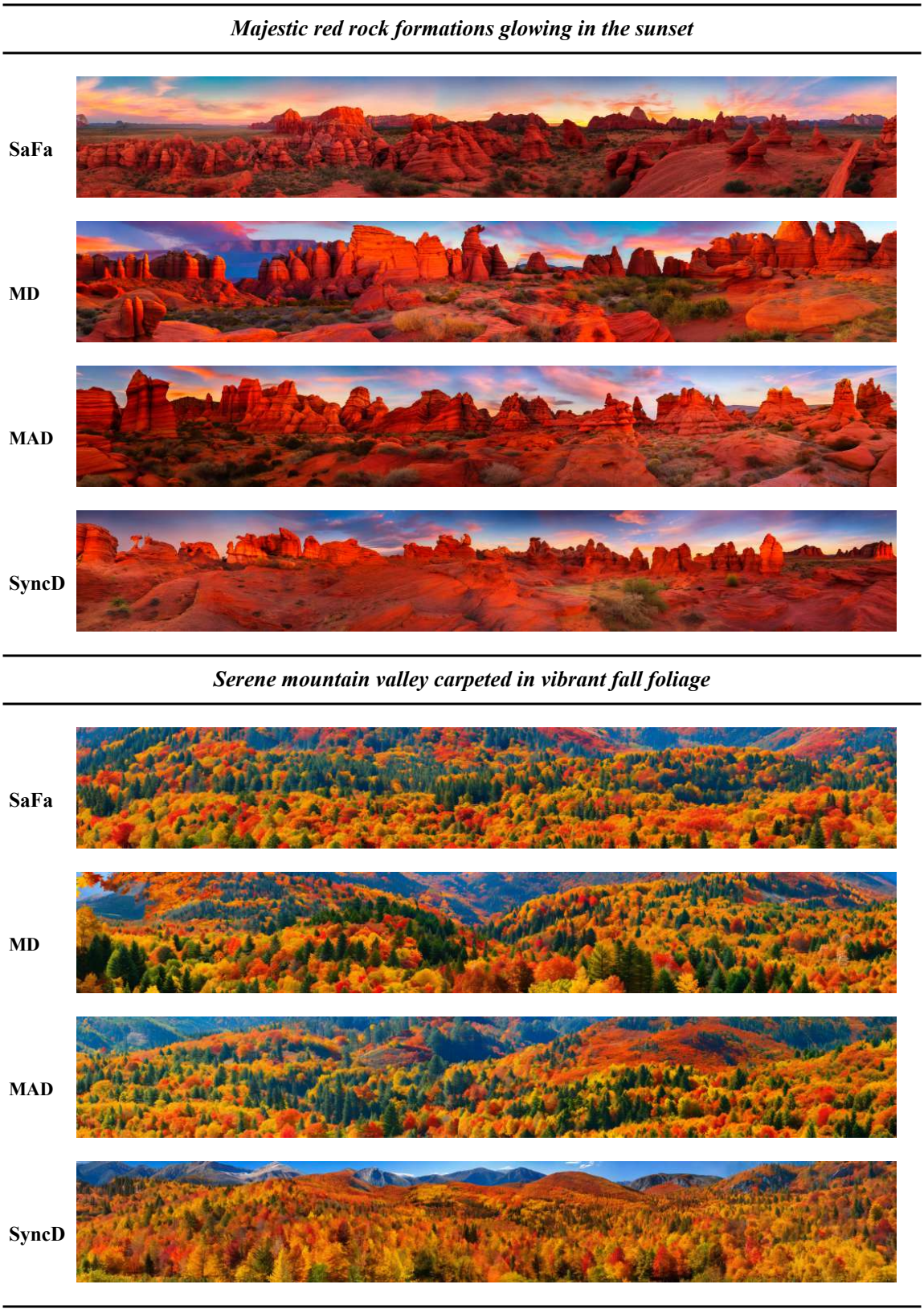


Figure 29. Qualitative comparison on panorama image generation. MD* represent an enhanced MD method with triangular windows.

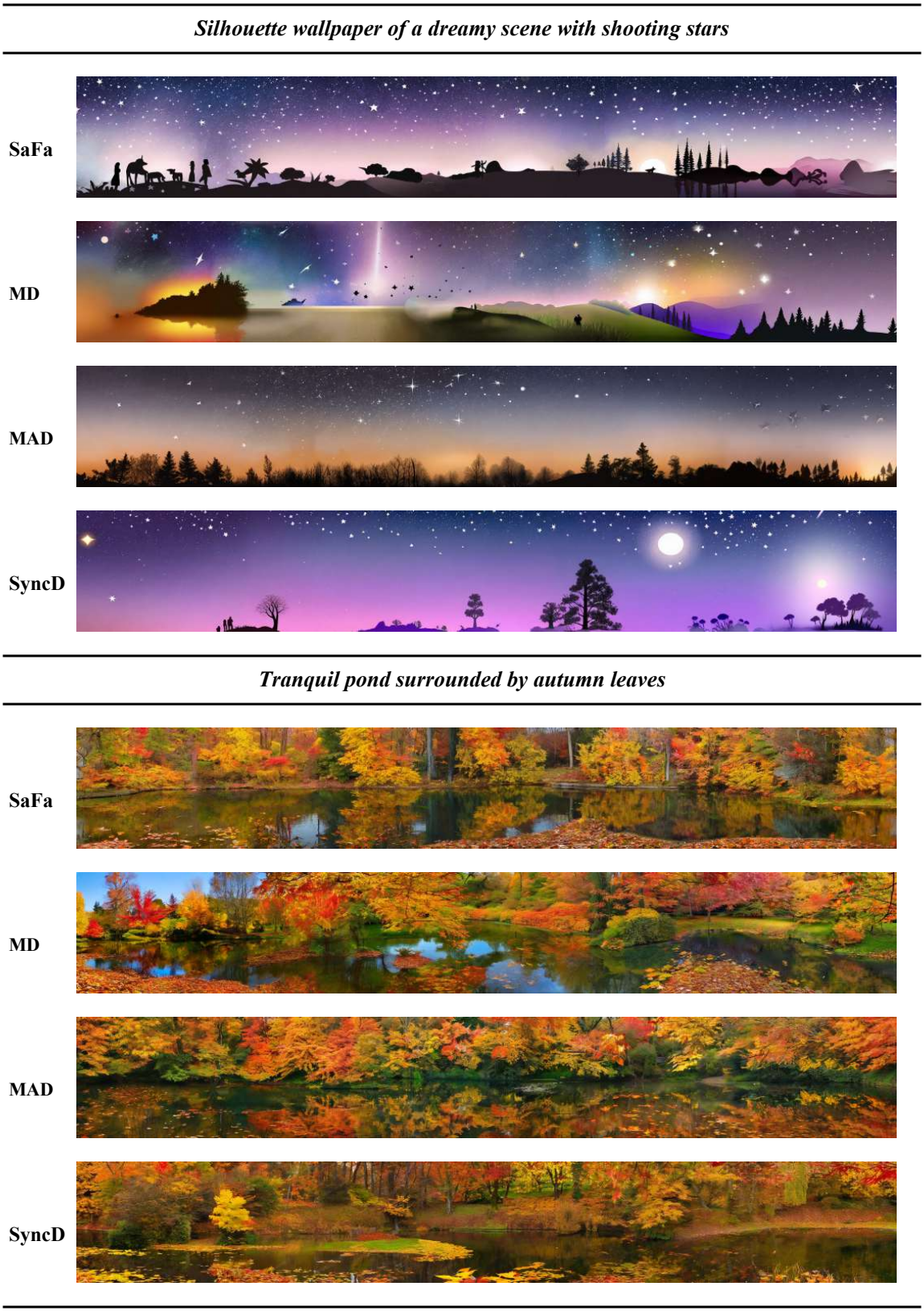


Figure 30. Qualitative comparison on panorama image generation. MD* represent an enhanced MD method with triangular windows.