

A Ablation Study on Different Sketch Choices

Table 5 demonstrates our early efforts at decomposing the semantic information of sketches. We compare different sketch representations, including pitch, pitch + chromagram, and semantic embeddings. As shown in the table, simple pitch-based sketches offer limited alignment capability, resulting in high PER and suboptimal performance in other metrics. The inclusion of chroma features slightly improves alignment, but still falls short of fully capturing the rich semantics needed for coherent vocal generation. These results validate the importance of incorporating abstract, semantically meaningful information into the sketch stage, and lay the groundwork for future exploration of interpretable yet powerful sketch formats.

Type	PER(%)↓	FAD↓	CE↑	CU↑	PC↑	PQ↑
no sketch	109.76	8.86	7.06	7.40	5.92	7.71
pitch	103.22	5.43	7.34	7.58	6.03	8.04
pitch + chroma	68.47	5.47	7.37	7.57	5.98	8.06
SSL embedding	9.44	5.60	7.55	7.66	6.00	8.25

Table 5: Impact of sketch types on SongBloom’s performance.

B Time Complexity of SongBloom Inference

We analyze the inference time complexity of SongBloom compared to a decoupled two-stage model, assuming both have the same number of layers.

Let L_1 denote the number of layers in the language model stage, and L_2 the number of layers in the diffusion stage. Let T be the total number of frames for both semantic and acoustic sequences (eg. $30 \text{ s} \times 25 \text{ fps} = 600$), P the patch size, $N = T/P$ the number of patches, and S the number of diffusion steps.

Assuming key-value caching is used during inference, we analyze the leading-order time complexity:

Decoupled 2-stage model:

$$O_0 = L_1 \cdot \frac{(1+T)T}{2} + L_2 \cdot (2T)^2 \cdot S$$

SongBloom:

$$O_1 = L_1 \cdot \frac{(1+T+N)(T+N)}{2} + L_2 \cdot (2P)^2 \cdot N \cdot S$$

Substituting $N = T/P$, we compute the difference:

$$O_1 - O_0 = \left(\frac{L_1}{2P^2} + \frac{L_1}{P} - 4L_2S \right) T^2 + \left(\frac{L_1}{2P} + 4L_2PS \right) T$$

When T is sufficiently large, the T^2 term dominates. In most practical cases, where:

$$L_1 < 4SP \cdot \frac{2P}{2P+1} \cdot L_2$$

the coefficient of T^2 is negative. Therefore, we conclude that SongBloom is asymptotically more efficient than the decoupled two-stage models, owing to its patch-wise diffusion mechanism and reduced per-step input length during inference.

C Criteria of the Subjective Listening Test

1. **Musicality of vocal:** (1–5 points) Does the main melody of the generated vocal match the subjective expectation?

- **5 points:** The melody is pleasant and emotionally expressive, with strong musical phrasing. It aligns well with expectations.
 - **4 points:** The melody generally meets expectations and conveys the song's theme and emotion, but lacks standout features.
 - **3 points:** The melody mostly aligns with expectations and conveys the theme and emotion, though some notes feel abrupt.
 - **2 points:** Only parts of the melody are coherent; most notes are scattered, and the theme and emotion are vaguely presented.
 - **1 point:** The melody significantly deviates from expectations, lacks coherent musical phrasing, and fails to convey the song's theme and emotion.
2. **Musicality of accompaniment:** (1–5 points) Does the accompaniment of the generated song sound harmonious?
- **5 points:** The accompaniment is richly colored and features diverse instrumentation. The melody is beautiful and complements the main melody harmoniously.
 - **4 points:** The accompaniment supports the main melody, but uses limited instrumentation or has a generally average melodic performance.
 - **3 points:** The accompaniment mostly supports the main melody, with only minor discord. However, it sometimes clashes with the main melody and lacks variety and color in instrumentation.
 - **2 points:** Some segments show disorganized instrumentation and monotonous melody, barely supporting the main melody.
 - **1 point:** The instrumentation is chaotic and the melody is discordant. There is a clear conflict with the main melody, failing to provide support.
3. **Quality of vocal:** (1–5 points) Is the vocal in the generated music clear and bright, with a full high-frequency range? Are there any noises or distortions present?
- **5 points:** The vocal quality is rich and clear, with no noise, approaching studio-recording quality.
 - **4 points:** The vocal quality is relatively clear, with slight noise that is either imperceptible or barely noticeable.
 - **3 points:** The vocal quality contains some noise or distortion, but it does not significantly affect the listening experience.
 - **2 points:** The vocal quality is unclear and unstable, resulting in a poor listening experience. Noticeable noise or distortion is present.
 - **1 point:** The vocal quality is extremely poor, with an unpleasant listening experience, and the vocal characteristics are barely recognizable.
4. **Quality of accompaniment:** (1–5 points) Is the high-frequency range of the generated music's accompaniment full? Are there any noises or instrumental distortions?
- **5 points:** The accompaniment has a full and clear sound quality with no flaws. The characteristics and melodies of different instruments are clearly distinguishable.
 - **4 points:** The accompaniment has good sound quality with slight noise. Only a few instruments in certain segments are hard to distinguish or slightly distorted, but this does not affect the overall listening experience.
 - **3 points:** The accompaniment has average sound quality. Some instruments are unclear or unidentifiable in certain segments. There is noticeable noise, distortion, or a lack of clarity.
 - **2 points:** The accompaniment has poor sound quality. In most parts of the piece, most instruments are unrecognizable. There is clear noise, distortion, or lack of clarity.
 - **1 point:** The accompaniment has extremely poor sound quality, with severe distortion, making it nearly impossible to identify any instrumental characteristics.
5. **Correctness of lyrics:** (1–4 points) Does the song content match the lyrics? Are there any errors such as extra words, missing words, or mechanical repetition?
- **4 points:** The song content fully matches the lyrics, with no missing or extra words, and no mechanical repetition of musical segments.

- 529
- 530
- 531
- 532
- 533
- 534
- 535
- 536
- 537
- 538
- 539
- 540
- 541
- 542
- 543
- 544
- 545
- **3 points:** The generated song contains a small number (within 5 words) of unclear, repeated, or missing lyrics.
 - **2 points:** The generated song contains multiple segments with unclear, repeated, or missing lyrics.
 - **1 point:** The generated song does not match the lyrics at all.
6. **Consistency of prompt:** (1–5 points) Does the musical style of the generated song match the style of the reference audio prompt?
- **5 points:** The musical style of the generated song fully matches the style specified in the prompt.
 - **4 points:** The musical style of the generated song is similar to the specified prompt, with only slight differences in some segments.
 - **3 points:** The musical style is somewhat similar to the specified style, but only vaguely reflects its characteristics.
 - **2 points:** The musical style does not resemble the specified style, with only faint traces of the intended musical elements.
 - **1 point:** The musical style has no relation to the specified style, making it difficult to connect the prompt with the resulting music.