System Introduction of SingVisio

System Overview

Figure 1 shows the overview of the explainer system which consists of a **Control Panel, Step View, Comparison View, Projection View**, and **Metric View.**

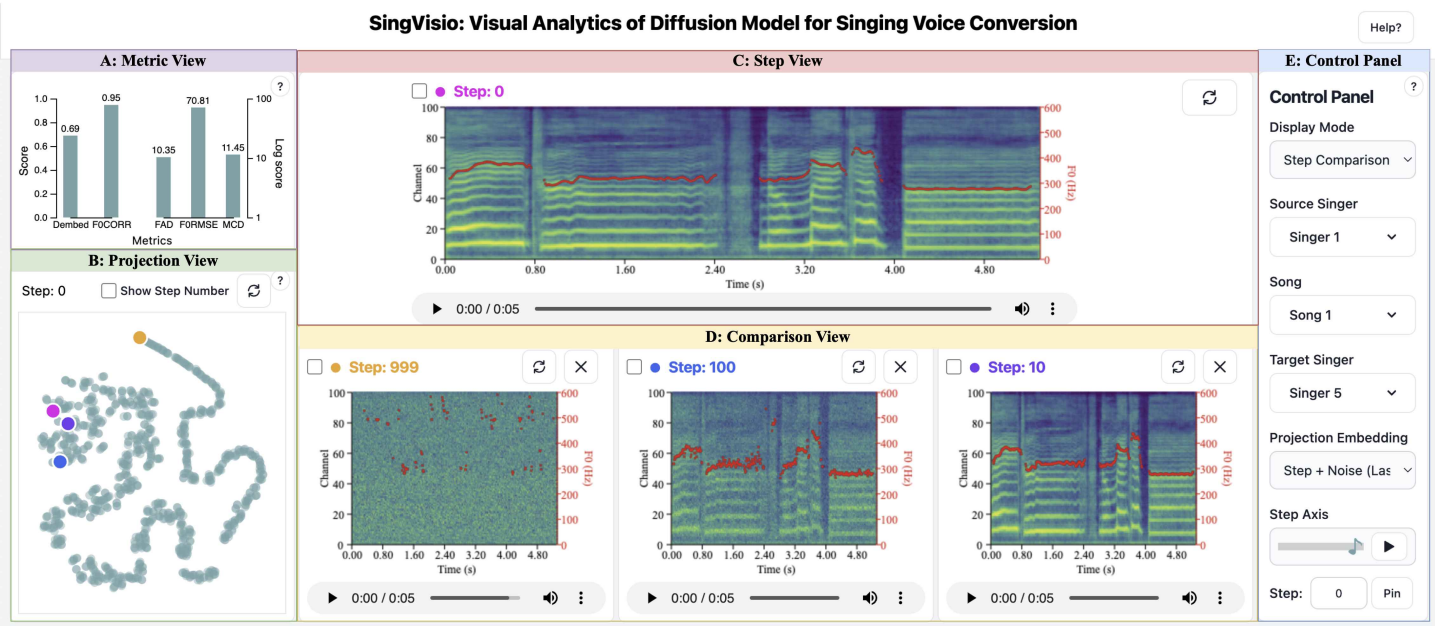

Figure 1 Visual analysis system for diffusion-based singing voice conversion. (A) ***Control Panel*** enables users to change the display mode and select data to visually understand and analyze the model behavior. (B) ***Step View*** gives users an overview of the whole diffusion generation process. (C) ***Comparison View*** facilitates users to pair-wisely compare generated results from different steps or different singers. (D) ***Projection View*** assists users in visually and interactively understanding the progression and dynamics of a generation process. (E) ***Metric View*** shows users the performance of the model on five objective metrics, allowing users to interactively analyze the trends of these metrics as they change with the diffusion steps.

A: Control Panel

The ***Control Panel***, depicted in Figure 2, comprises six elements, featuring five drop-down menus (Display Mode, Source Singer, Song, Target Singer, Projection Embedding) and a step controller. Users are allowed to choose a display mode and projection embedding, one or two source singers, source songs, and a target singer (two-singer selection available when *Display Mode* is set to one of three Comparison modes: Source Singer Comparison, Song Comparison,

Target Singer Comparison). The inclusion of a step controller facilitates user control of the diffusion step.

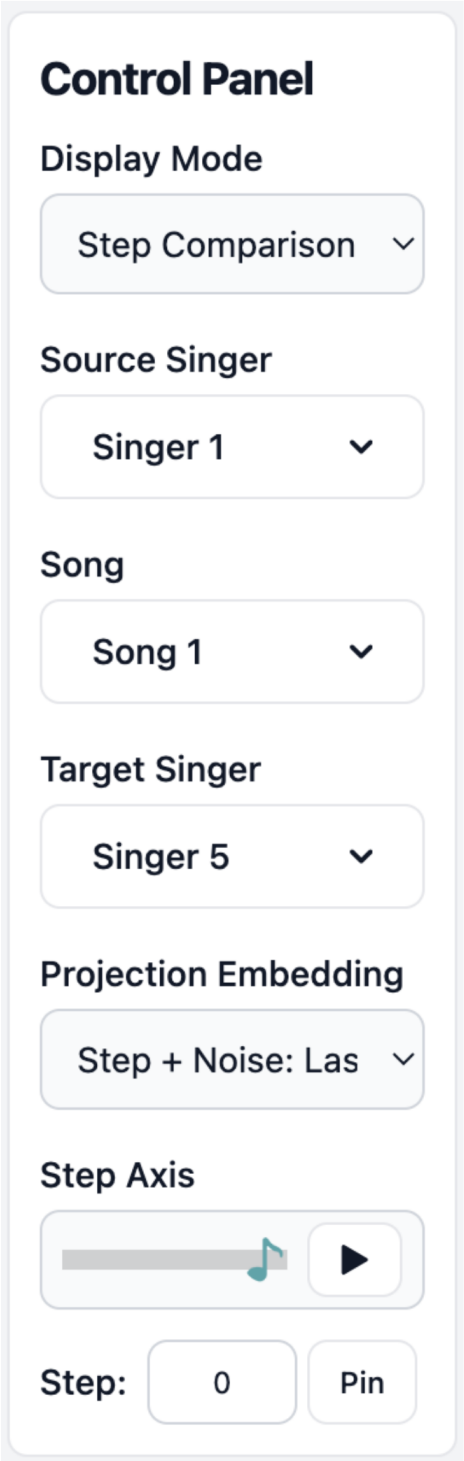

Figure 2 Overview of the ***Control Panel***.

Display Mode

There are five types of display modes, as shown in Figure 3, including *Step Comparison, Target Singer Comparison, Source Singer Comparison, Song Condition Comparison*, and *Metric Comparison*. Users can click the drop-down menu of "Display Mode" to choose a specific mode.

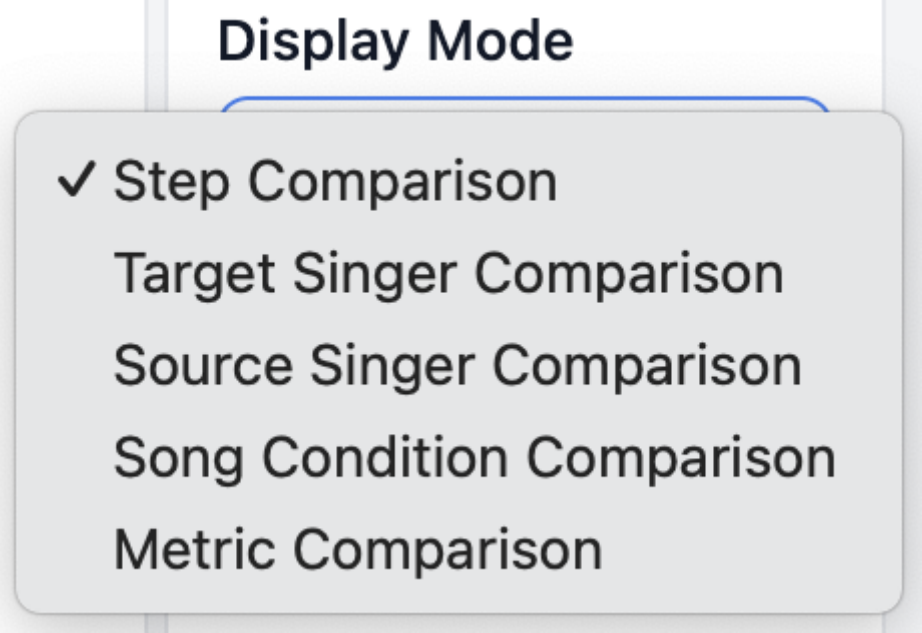

Figure 3 The drop-down menu of Display Mode

- *Step Comparison*

This mode primarily focuses on step-wisely comparing the diffusion steps in the generation process. It (1) provides an animation of random noise gradually refined for users to have an overview of the whole denoising process in ***Step View***, and (2) enables users to flexibly select and compare the generated results from different diffusion steps in ***Comparison View***.

- *Target Singer Comparison*

This mode focuses on the pair-wise comparison of converting the same source singing voice (also means the same song) to two different target singers. It (1) enables users to select a source singer and a source song, but two target singers in the Control Panel, (2) provides details (Mel spectrogram, F0 contour and audible audio) of the source singer's audio and two target singers' audio in ***Comparison View***, and (3) presents two conversion animations wherein random noise undergoes gradual refinement to transform into the singing voice of the target speaker in ***Step View***.

- *Source Singer Comparison*

This mode focuses on the pair-wise comparison of converting two different source singers' audio with the same song to the same target singer. It (1) allows users to select two different source singers, a source song and a target singer, (2) provides details (Mel spectrogram, F0 contour and audible audio) of the two source audio and the target audio in ***Comparison View***, (3) presents two conversion animations wherein random noise undergoes gradual refinement to transform into the singing voice of the target speaker in ***Step View***.

- *Song Condition Comparison*

This mode focuses on the pair-wise comparison of converting two different source audios that are derived from the same singer but contain different songs to the same target singer. It (1) allows users to select a source singer, a target singer but two songs, (2) provides details

(Mel spectrogram, F0 contour and audible audio) of the two source singers' audio and the target singer's audio in ***Comparison View***, (3) presents two conversion animations illustrating the progressive refinement of random noise into the singing voice of the target singer in ***Step View***.

- *Metric Comparison*

This mode allows users to (1) click on a specific metric bar, prompting the system to filter out an example that performs best on the corresponding metric, displaying metric curves along diffusion steps in the ***Comparison View***. Additionally, (2) users can hover over and slide their pointer along the step axis of the metric curve, prompting the system to display the values of that metric at different steps in the ***Comparison View***, while simultaneously showing the generated results at different steps in the ***Step View***.

Source Singer/Source Song/Target Singer

These three drop-down menus are designed to offer users options for source singer, source song, and target singer selection to compare different conditions in the singing voice conversion, as shown in Figure 4.

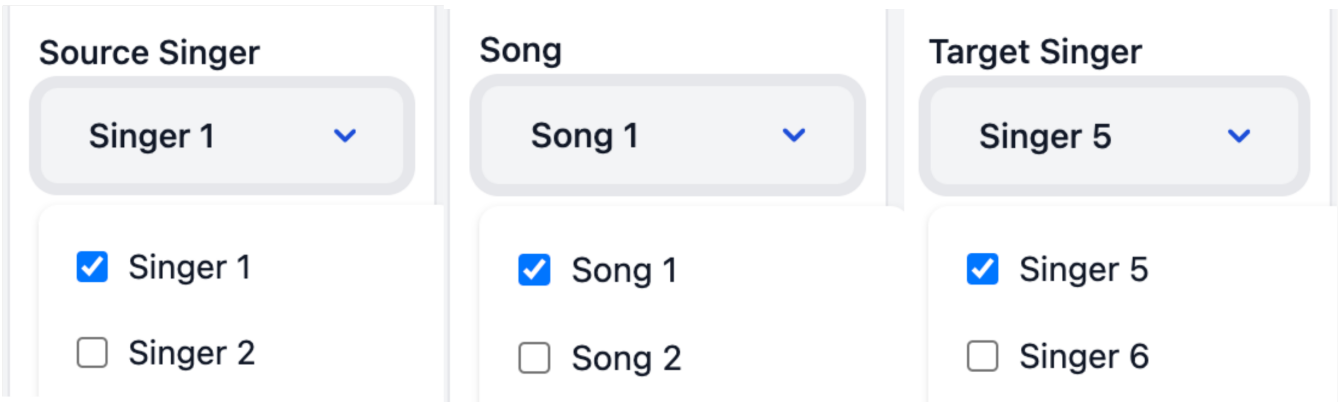

Figure 4 Three drop-down menus for source singer, source song and target singer selection, respectively.

Projection Embedding

This drop-down menu enables users to select projection embeddings from different layers, as shown in Figure 5. After a specific projection embedding is selected, the system displays 2D t-SNE visualization results of the high-dimensional diffusion steps in the ***Projection View*** (which will be introduced in Section D: Projection View). Specifically, the projection embedding can be the vanilla diffusion steps, or the combined results involving steps and noise, step, noise and diffusion conditions at the first, middle, or final layers of the diffusion model.

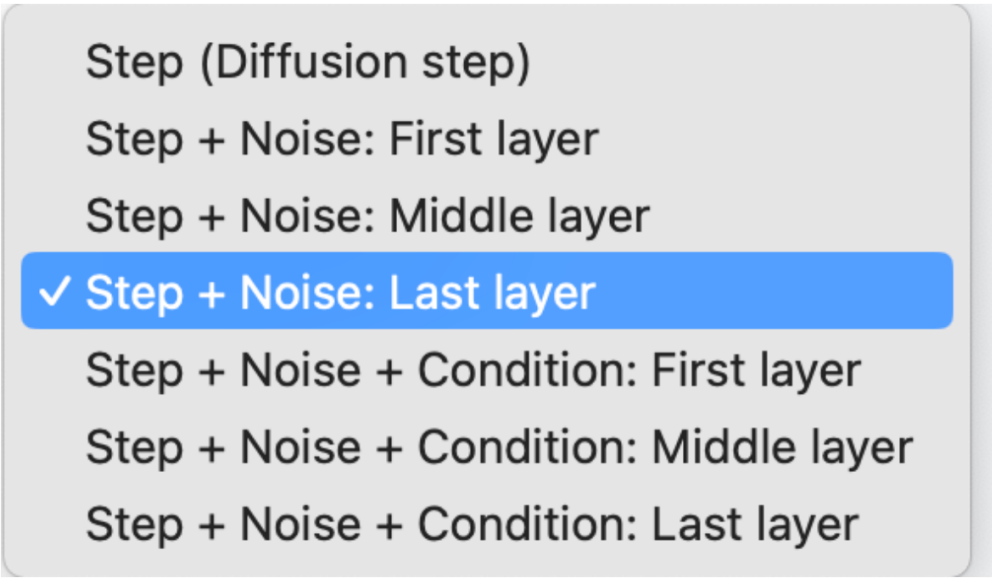

Figure 5 The drop-down menu for projection embedding

Step Controller

The Step controller includes (1) a step slider to smoothly control the diffusion step, (2) a tooltip to display or input a specific step number, and (3) a button named 'Pin' that enables users to add a specific step's generated result in the ***Comparison View***.

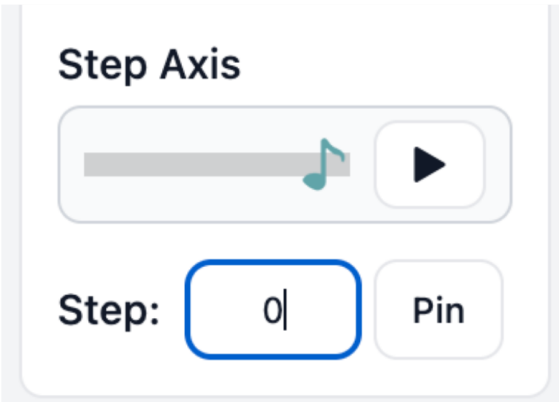

Figure 6 The illustration of the step controller

B: Step View

Step View enables users to visualize the whole generation process of diffusion in the context of singing voice conversion tasks, which means users can observe how the spectral characteristics change over time as noise is subsequently removed, leading to the desired singing voice conversion. Specifically, it can be observed that the Mel spectrogram transitions from being completely noisy to gradually becoming clearer, and the fundamental frequency curve also

transforms from scattered points into a smooth curve. The audio undergoes a process of gradual optimization from being pure noise to having improved sound quality and intelligibility.

In the *Step Comparison Mode* and *Condition Comparison Mode*, the content presented in the ***Step View*** is slightly different. In the *Step Comparison Mode*, we focus on comparing and analyzing the converted results from different steps, so only one diffusion process animation is displayed in the view. While, in the *Condition Comparison Mode*, the main objective is to compare the conversion results under different conditions, e.g., source singer, source song, target singer, at this time, the ***Step View*** shows pair-wise diffusion process animations for two different conditions. In Figure 7 and 8 below, animations illustrating the whole diffusion generation process of singing voice conversion are presented for the *Step Comparison Mode* and *Condition Comparison Mode*, respectively.

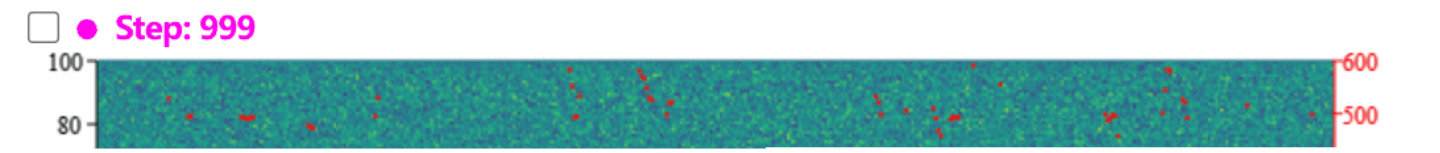

Figure 7 The animation of the whole diffusion generation process for singing voice conversion in ***Step View*** for *Step Comparison Mode*.

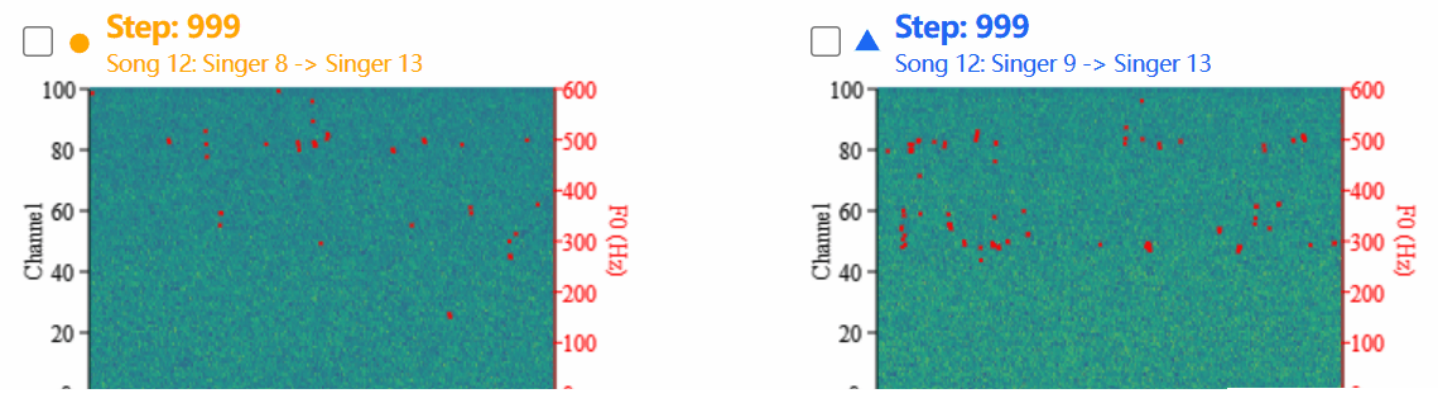

Figure 8 The animation of the whole diffusion generation process for singing voice conversion in ***Step View*** for Condition *Comparison Mode*.

The ***Control Panel***, mentioned earlier, allows users to interact with the diffusion process through the Step Controller. Users can adjust the diffusion time step to observe the intermediate results of any step in the generation process, enlarge the Mel spectrogram to observe detailed information through a brush operation, and restore it to the original Mel spectrogram using the refresh button in the top right corner in the ***Step View***.

C: Comparison View

To facilitate a more convenient and detailed observation of the intermediate results generated by the diffusion model, we introduce the ***Comparison View***. This view provides basic information, including the audible audio, as well as its corresponding Mel spectrogram and fundamental frequency contour. All information related to the clip of audio forms a basic block, which we refer to as the 'basic display unit', as shown in Figure 9. The specific content displayed in ***Comparison View*** differs among the *Step Comparison Mode, Condition Comparison Mode, and Metric Comparison Mode.*

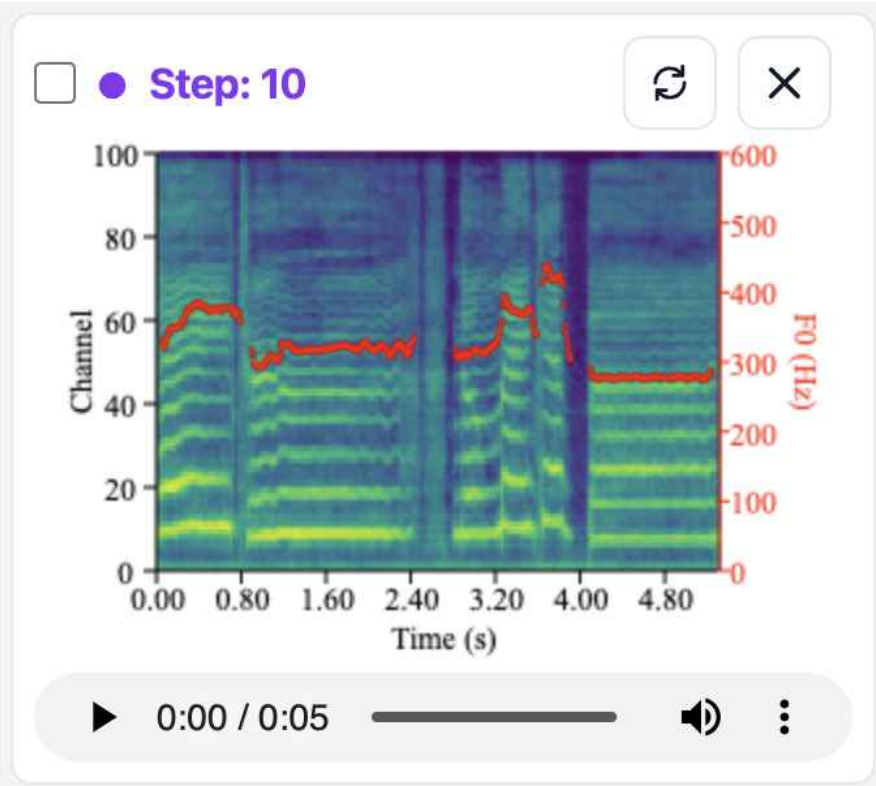

Figure 9 The basic display unit.

The *Step Comparison Mode* is primarily used to compare the results of different diffusion steps. In this mode, the ***Comparison View*** will display three basic display units from three different

steps, based on the steps selected by the user, as illustrated in Figure 10. The relative position of different basic display units (corresponding to different steps) can be directly adjusted by dragging.

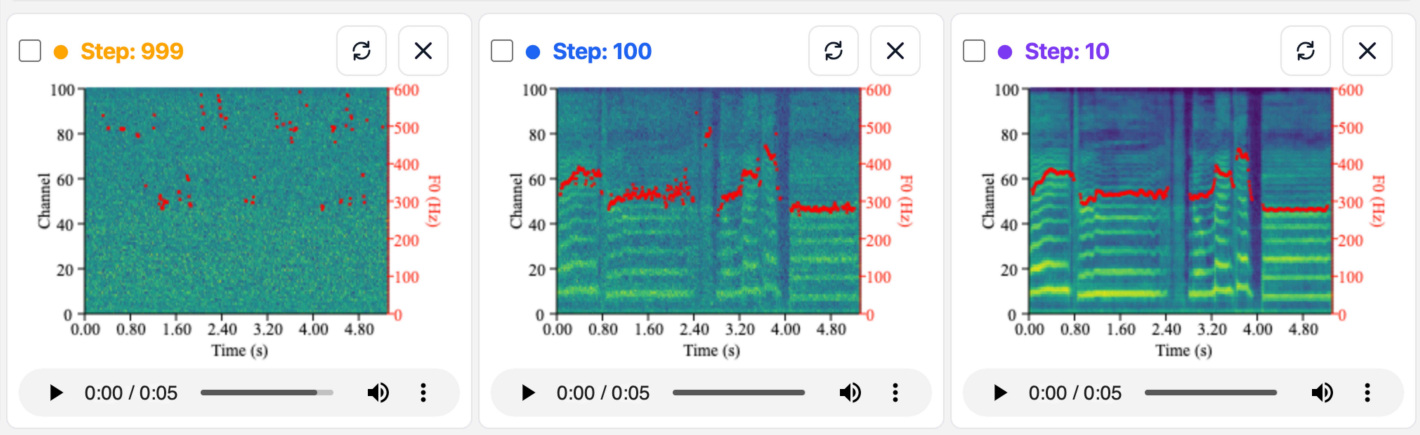

Figure 10 Comparison view under the step comparison mode.

The *Condition Comparison Mode* mainly compares the results of singing voice conversion under different conditions, so the ***Comparison View*** displays the basic display units corresponding to different source and target singers selected by users, as presented in Figure 11.

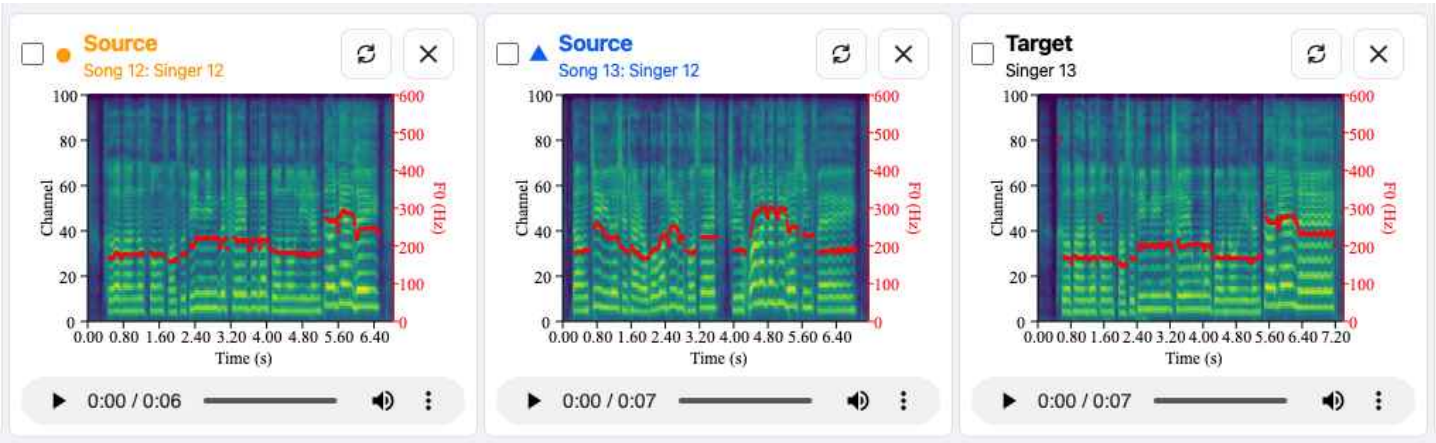

Figure 11 ***Comparison View*** under the condition comparison mode.

Although there are different modes, there are some common operations in the ***Comparison View*** under both *Step Comparison Mode* and *Condition Comparison Mode*. Specifically, on this basic display unit, we can observe the range of the F0 and the pattern of the F0 contour. Through the brush operation, we can synchronously magnify all Mel spectrograms illustrated in this view, thus enabling a more detailed comparison and examination of the spectral differences. When there are multiple basic display units, users can select the checkboxes in the top left corner of any two basic display units. The page will then pop up the visualization of the difference between the two Mel spectrograms in the selected basic display units, allowing for a clearer and more convenient comparison. Specifically, the differences are represented by colors, allowing

the viewer to understand the data visually. Warmer colors like reds and oranges signify larger differences, while cooler colors like blues and greens represent smaller differences.

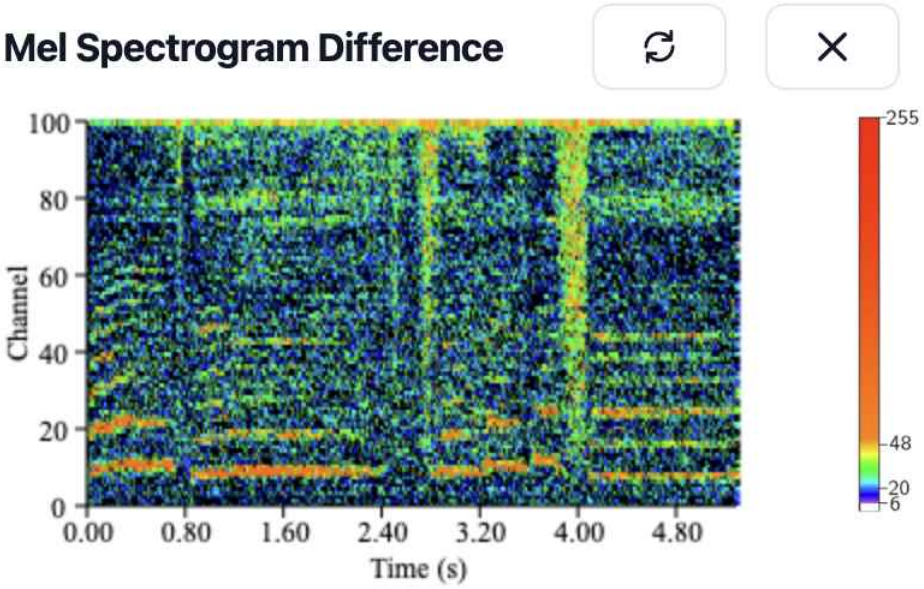

Figure 12 Mel spectrogram difference between two selected Mel spectrograms.

In *Metric Comparison Mode*, the ***Comparison View*** illustrates the metric curve over the diffusion step, which is described in Section E: Metric View.

D: Projection View

The diffusion step in the diffusion model plays a pivotal role in the generative process, representing a crucial stage where random noise is iteratively refined to produce the final output, such as the target speaker's singing voice. This step encapsulates the gradual transformation and enhancement of the input noise into meaningful and coherent content, influencing the overall quality and fidelity of the generated output.

High-dimensional diffusion step embeddings can be challenging to interpret directly. t-SNE reduces the dimensionality, projecting the step embeddings into a lower-dimensional space. It allows researchers to understand the intricate structure and relationships within the high-dimensional diffusion step space. Specifically, projected embedding reveals diffusion steps' patterns and trajectories, enabling a visual exploration of the dynamic evolution of the diffusion step in the generation process.

As mentioned earlier in Section A: Control Panel, the drop-down menu of projection embedding provides multiple projection embedding sources, including not only the vanilla diffusion step but also the combination of the diffusion step with noise and condition. By examining the

projection embedding results of combining diffusion step with noise and condition, users can compare the differences in diffusion step trajectories under different condition scenarios.

Consequently, we design a ***Projection View*** to present the step embeddings obtained by projecting high-dimensional diffusion step embeddings into two-dimensional embeddings, as shown in Figure 13. In the top right corner of this view, there is a button represented by a question mark. When users click this button, tips concerning the ***Projection View*** will be shown to aid understanding. Each point in the graph represents a diffusion step, and all 1000 diffusion steps together form a trajectory in space. Users can hover their mouse over the points and slide to inspect the step trajectory. While sliding the pointer, users can simultaneously observe the SVC results transitioning from a coarse state to a fine state in ***Step View***.

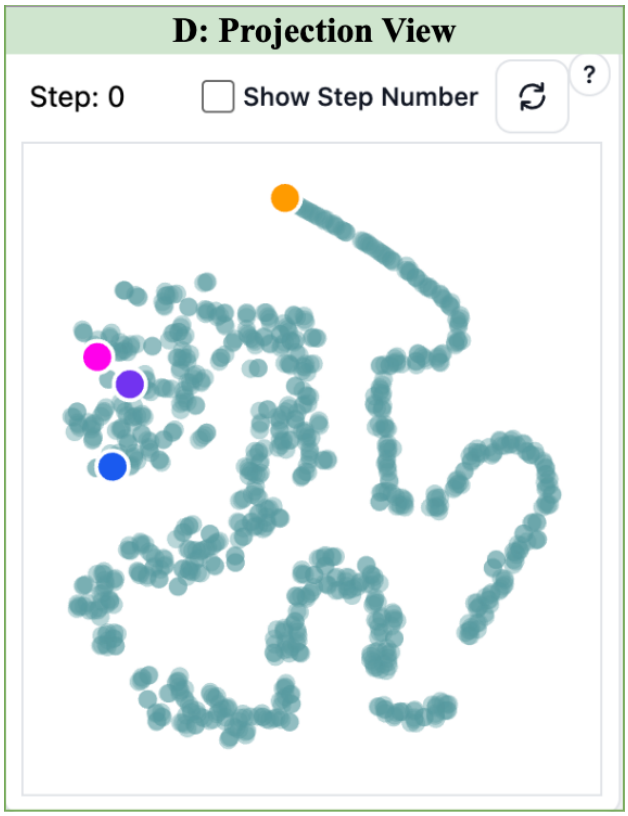

Figure 13 The illustration of ***Projection View.***

Additionally, users can enable the "Show Step Number" checkbox, which reveals the step number associated with each step. By scrolling the mouse wheel, they can zoom in or out on the points in space to examine the distribution of diffusion steps. By clicking on a specific step point, a basic display unit corresponding to the step will be added to the ***Comparison View***, and users can then observe the generated audio, its corresponding Mel spectrogram, and F0 contour for further analysis.

E: Metric View

Metric view is designed to show the performance of the model by showing five metrics and further explaining how different metrics vary across steps. This view aids the users in getting information on the performance of model. The five metrics, including Dembed, F0CORR, FAD, F0RMSE, and MCD, are divided into two groups based on their correlation with the model performance: positively-correlated metrics and negatively-correlated metrics. The metrics are then drawn as histograms, as presented in Figure 14. Here, the labels on the x-axis denote different metrics, and the labels on the y-axis are scores (the higher, the better) and log scores (the lower, the better). At the top of each rectangular bar, the corresponding metric value evaluated on the model is displayed. Additionally, there is a question mark button located at the top right corner of this view. When users click this button, a text box will appear providing the descriptions of each metric, as shown in Figure 15.

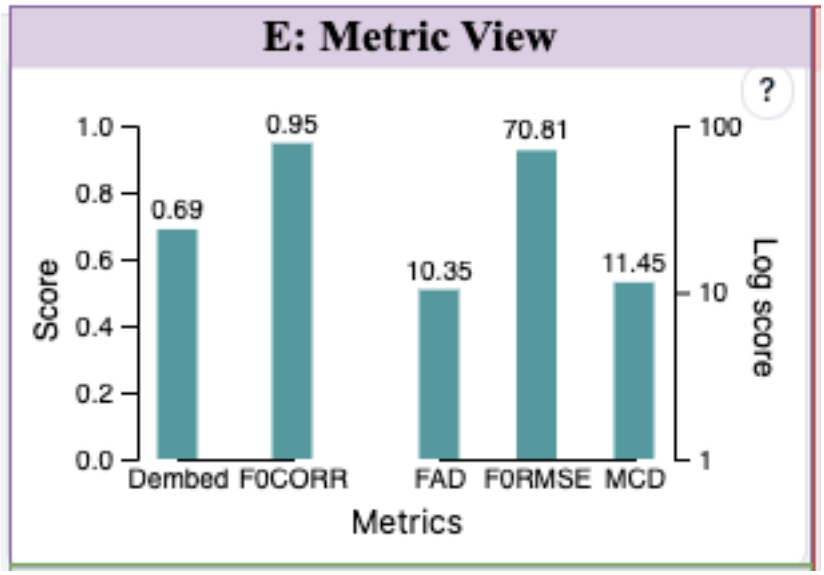

Figure 14 The illustration of ***Metric View***.

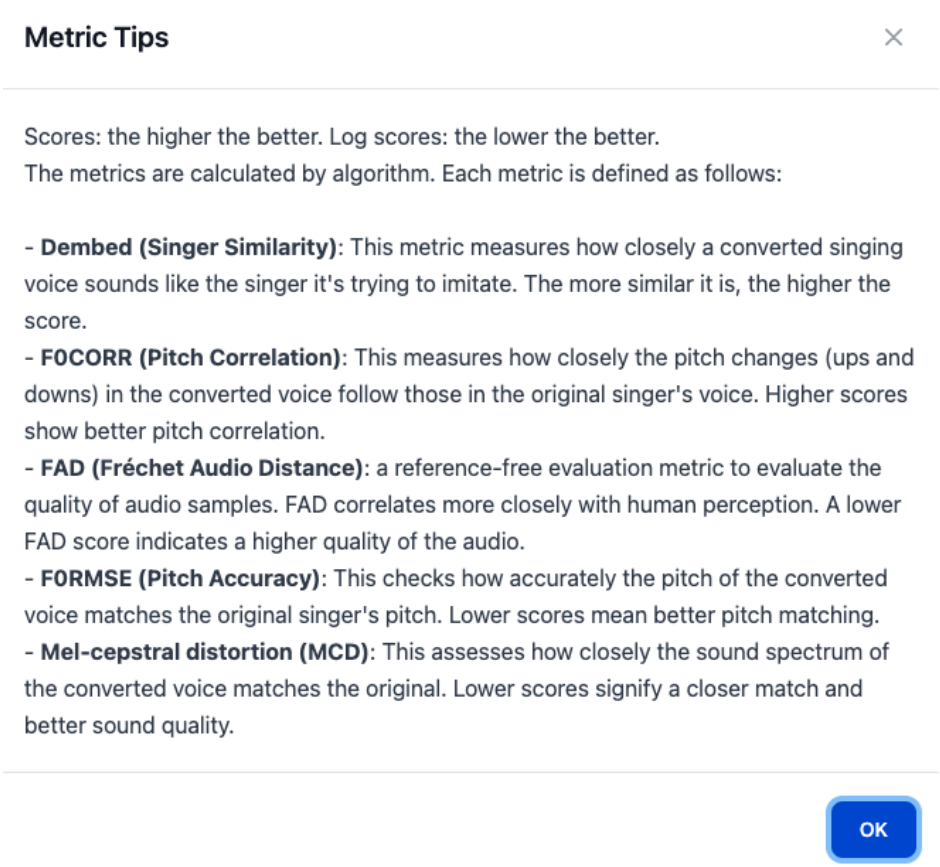

Figure 15 Metric descriptions in ***Metric View.***

The users can interact with each rectangular bar by hovering on it and clicking it to select a metric for detailed evaluation result display, which is represented by the utterance that obtains the best score in this metric. Specifically, after a metric is selected, the ***Comparison View*** will be replaced with a new panel showing the metric curve across the diffusion generation steps, as shown in Figure 16. The users can then observe the trend of different metrics from the curve. At the top of the chart, there are five legends representing five different metrics respectively in distinct colors. The x-axis shows different steps ranging from 999 to 0, and the y-axis displays the score for the evaluation metrics. The users can check the specific metric value for each step by hovering on the curve. Also, the step preview will update as the cursor moves along the curve.

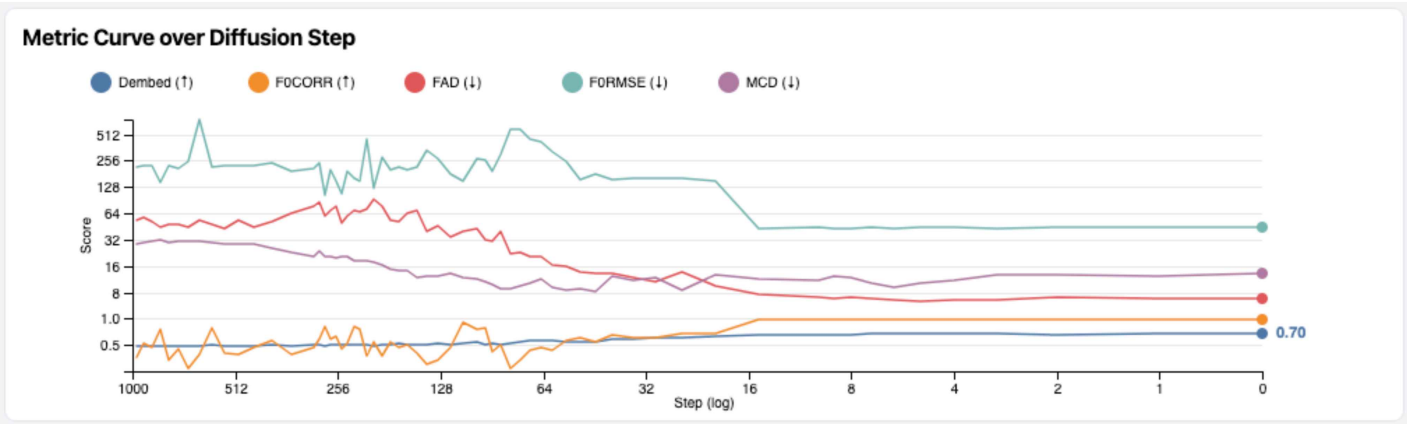

Figure 16 The illustration of the metric curve over the diffusion steps.

Appendix

The following is an explanation of the technical terms used in this system.

Mel spectrogram is extracted from the audio using a signal processing algorithm. It is a 2D representation with the dimensions of Time*Channel, where the time axis captures the progression of the audio signal over time, and the channel axis represents the frequency components or mel bins, providing a comprehensive view of the signal's spectral content. The Mel spectrogram is color-coded to indicate the intensity or magnitude of different frequencies over time. Bright colors, such as yellow and red, represent high energy or the presence of specific frequencies, while darker colors represent lower energy or the absence of those frequencies.

Fundamental Frequency Contour (F0 contour) tracks the fundamental frequency (F0) of the singing voice over time. This contour line may be drawn as a continuous curve that rises and falls to depict changes in F0. The F0 contour line is colored red, making it distinguishable from the Mel spectrogram.

Singer similarity (Dembed) quantitatively assesses the similarity between the timbre of the original singer's voice and the converted voice. It's calculated using the cosine similarity between feature vectors representing the timbre characteristics of the two voices. A higher similarity score indicates higher timbre similarity.

F0 Pearson Correlation Coefficient (F0CORR) measures the Pearson Correlation Coefficient between the F0 values of the converted singing voice and the target voice. It assesses the linear relationship between the F0 contours of the two voices. A higher F0CORR indicates a stronger correlation and higher F0 similarity.

Fréchet Audio Distance (FAD) is a reference-free evaluation metric to evaluate the quality of audio samples. FAD correlates more closely with human perception of audio quality. A lower FAD score indicates higher audio quality.

F0 Root Mean Square Error (F0RMSE) measures the Root Mean Square Error of the Fundamental Frequency (F0) values between the converted singing voice and the target voice. It quantifies how accurately the F0 of the converted voice matches that of the target voice. A lower F0RMSE indicates higher F0 accuracy.

Mel-cepstral distortion (MCD) assesses the quality of the generated speech by evaluating the discrepancy between the generated and ground-truth singing voices. It quantifies the differences between the two sequences of Mel-cepstra. A lower MCD value indicates a smaller deviation and hence higher quality in generated speech.