C3Net: Compound Conditioned ControlNet for Multimodal Content Generation

Supplementary Material

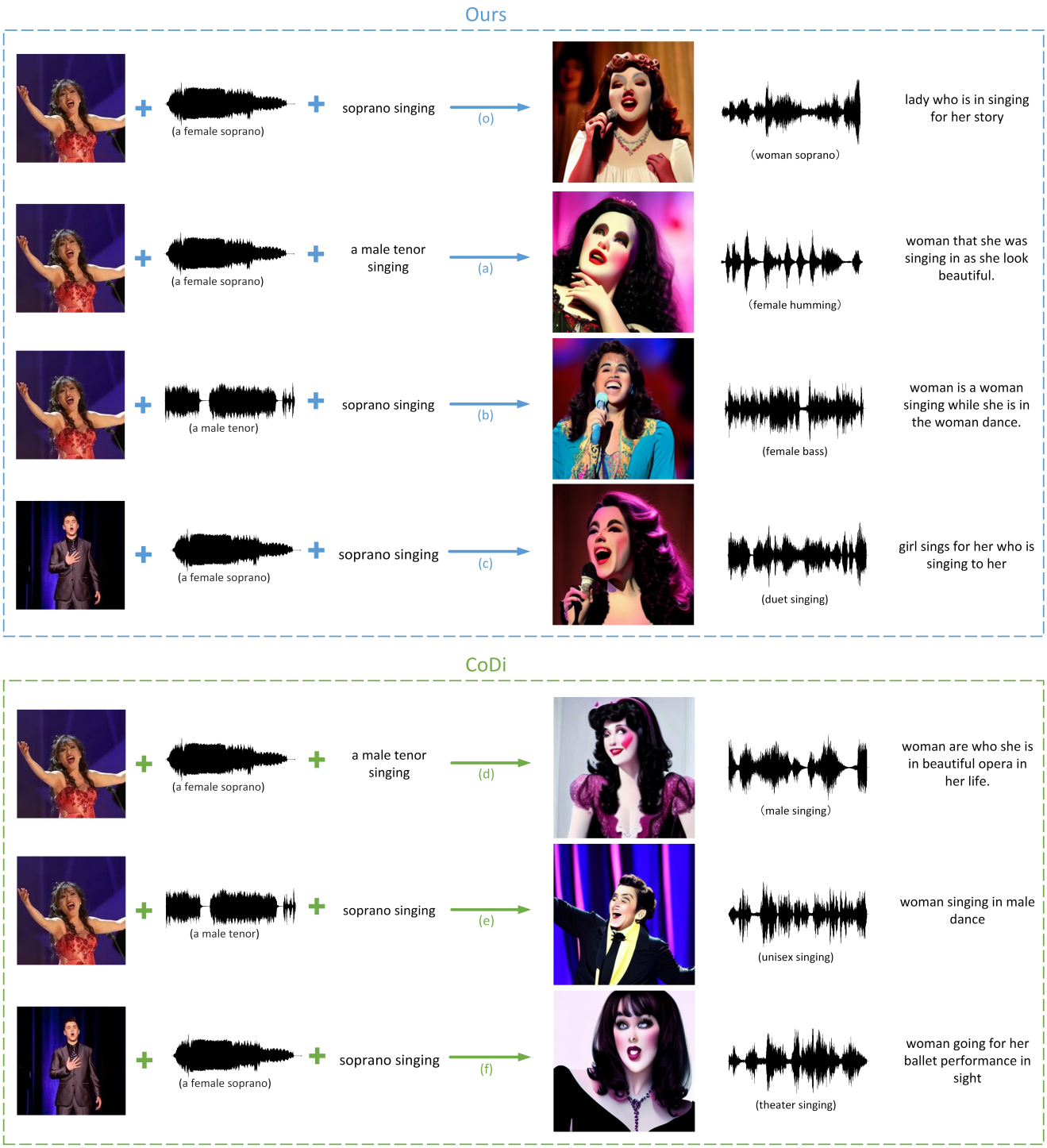

Figure 1. Further experiments in the following settings are conducted to investigate the effect of multimodal conditions which are in conflict or contradictory among each other. In *(a)*, the text condition relates to a male tenor, while the image and audio depict a female soprano. In *(b)*, male tenor audio is used as a condition, while the image and text respectively relate to a female soprano. In *(c)*, the image indicates a male singer, while the audio and text describe a female soprano. In *(o)*, which is a control, we use consistent conditions to compare the differences in the above experimental scenarios. We repeated the experiment with same settings and produce generations using the baseline [1].

1. Robustness to Contradictory Multimodal Conditions

We further explored how contradictory conditions can influence the generation of C3Net. Through extensive experiments, we found that rather than causing collapse and generating nonsense or dominated by a single condition, C3Net coordinates contradictory inputs innovatively. We elaborate on C3Net's robustness by an example in Figure 1, where all other conditions indicate a female soprano except one, which describes a male tenor.

The example shows four scenarios, including one control and three experimental scenarios, where two conditions describe a female soprano, while the remaining one relates to a male tensor. C3Net generates images, audio, and text conditioned on these contradictory inputs. The generated images and text all describe a female, the most frequent subject in three conditions. In scenarios *(a)* and *(b)*, the discrepancy changes the tone of generated audio from a female soprano to a lower frequency humming. In *(c)*, the generated audio and text indicate another singer, coordinating the condition of *female soprano* and *male tenor*. In *(o)*, a control, we use consistent conditions to compare the differences in the above experimental scenarios.

We repeated the experiment with the same settings to produce multi-modal generations using the baseline [1]. We found that without the Control C3-UNet structure, the generated contents tend to be intermediates of the contradictory conditions (e.g., scenario *(e)*) or shifted in meaning (e.g., text generated in scenario *(a)* and *(e)*). These defects are likely resulted from using simple interpolation to coordinate multiple conditions.

2. Audio Files

We put in our supplemental zip file all the relevant audio files depicted in each figure in the main paper and this supplementary document, including all condition audios and generated audios.

References

[1] Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion, 2023. 1, 2