

Supplementary Materials for TANGO 2: Aligning Diffusion-based Text-to-Audio Generations through Direct Preference Optimization

Anonymous Authors

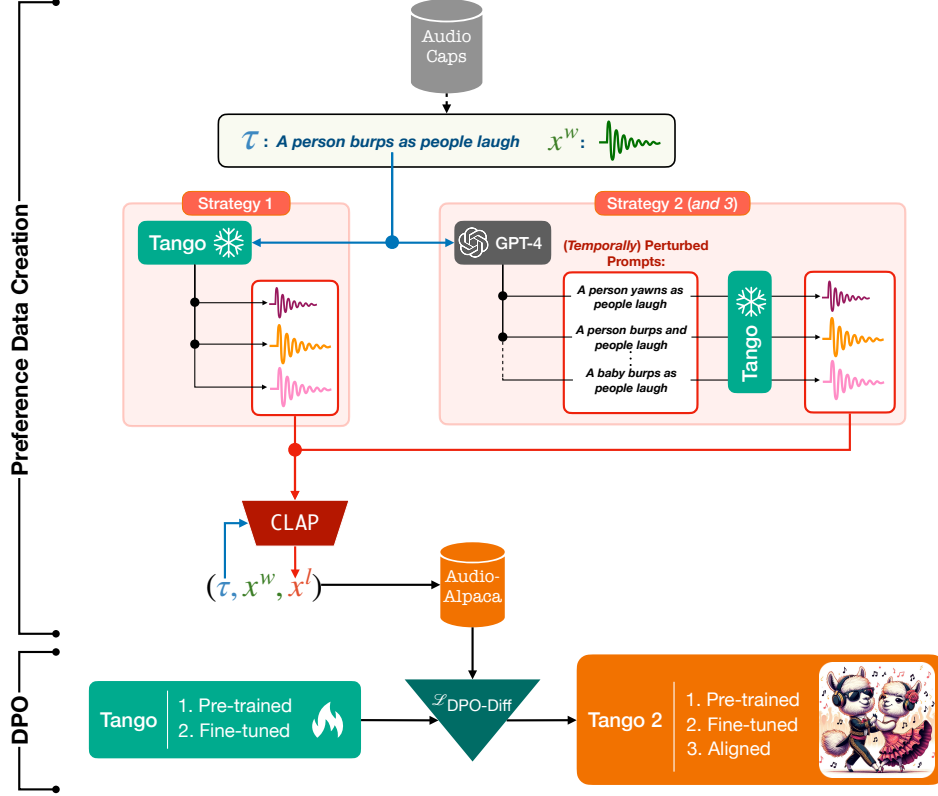


Figure 1: An updated illustration of our pipeline for text-to-audio alignment. The top part depicts the preference dataset creation where three strategies are deployed to generate the undesirable audio outputs to the input prompts. These samples are further filtered to form Audio-alpaca. This preference dataset is finally used to align TANGO using DPO-diffusion loss, resulting in TANGO 2.

1 CODE, DEMO, AND DATASET

We plan to release the code and dataset on GitHub and Hugging Face, respectively, upon acceptance. Meanwhile, for review purposes we share the code anonymously at <https://github.com/339wef0493/tango2>. The dataset is available at <https://huggingface.co/datasets/339wef0493/audio-alpaca>. Some comparative audio samples are also presented at <https://339wef0493.github.io>.

2 ABLATIONS ON AUDIO-ALPACA

We conducted an ablation study on Audio-alpaca to gauge the impact of different negative data construction strategies. As shown in Table 1, excluding the data samples from by *strategies 2 and 3* notably diminishes the performance of TANGO 2. This underscores the significance of event and temporal prompt perturbations.

3 IMPACT OF CONCEPT OR EVENT COUNT

We categorize prompts based on the presence of multiple concepts or events, exemplified by phrases like "A woman speaks while cooking". As underlined, this prompt contains two distinct events i.e., "sound of a woman speaking" and "sound of cooking". Through manual scrutiny, we discovered that pinpointing prompts with such multi-concepts is challenging using basic parts-of-speech or named entity-based rules. Consequently, we task GPT-4 with extracting the various concepts or events from the prompts using in-context exemplars. The specific prompt is displayed in Table 2. To evaluate GPT-4’s performance on this task, we randomly selected 30 unique prompts and manually verified their annotations from GPT-4’s. No errors attributable to GPT-4 were found. In general, TANGO 2 outperforms AUDIO-LDM2 and TANGO across most objective and subjective metrics, following Table 3. We proceed to visualize the CLAP scores

Table 1: Text-to-audio generation results on AudioCaps evaluation set. Due to time and budget constraints, we could only subjectively evaluate AudioLDM 2-Full-Large and Tango-full-FT. Notably these two models are considered open-sourced SOTA models for text-to-audio generation as reported in AudioBox.

Model	#Parameters	Objective				Subjective	
		FAD ↓	KL ↓	IS ↑	CLAP ↑	OVL ↑	REL ↑
AudioLDM-M-FULL-FT	416M	2.57	1.26	8.34	0.43	—	—
AudioLDM-L-FULL	739M	4.18	1.76	7.76	0.43	—	—
AudioLDM 2-FULL	346M	2.18	1.62	6.92	0.43	—	—
AudioLDM 2-FULL-LARGE	712M	2.11	1.54	8.29	0.44	3.56	3.19
TANGO-FULL-FT	866M	2.51	1.15	7.87	0.54	3.81	3.77
TANGO 2	866M	2.69	1.12	9.09	0.57	3.99	4.07
w/o Strategy 2 & 3	866M	2.64	1.13	8.06	0.54	—	—
w/o Strategy 1	866M	2.47	1.13	8.58	0.56	—	—
w/o Strategy 2	866M	2.28	1.12	8.38	0.55	—	—
w/o Strategy 3	866M	2.46	1.13	8.63	0.56	—	—

Table 2: GPT-4 prompt used to extract events or concepts from prompts describing audios.

You are to extract all the indivisible events in the given text, labeled as input. Imagine experiencing the events in the input as you are reading it and write down the indivisible events one by one. After writing your experience, list all the events in the sequence you observed them as a python list. Think step-by-step. Do not directly give the answer. Please refer to these following examples as reference for input and output:
<i>Example 1 -</i>
Input: An aircraft engine runs and vibrates, metal spinning and grinding occur, and the engine accelerates and fades into the distance
Output: Firstly, an aircraft engine runs and vibrates. Then, I hear metal spinning and grinding. Then, the aircraft engine accelerates. Finally, the aircraft fades into the distance.
So, here is the list of events that I observed:
["aircraft engine runs", "aircraft engine vibrates", "metal spinning", "metal grinding", "aircraft engine accelerates", "aircraft fades into the distance"]
<i>Example 2 -</i>
Input: Bubbles gurgling and water spraying as a man speaks softly while crowd of people talk in the background
Output: Firstly, I hear bubble gurgling. Also, I hear water spraying. Simultaneously, a man is speaking softly. Also, a crowd of people are talking in the background.
So, here is the list of events that I observed:
["bubble gurgling", "water spraying", "a man is speaking softly", "crowd talking"]
<i>Example 3 -</i>
Input: A man talking then meowing and hissing
Output: Firstly, I hear a man talking. Subsequently, I hear meowing. I also hear hissing.
So, here is the list of events that I observed:
["a man talking", "meowing", "hissing"]
**** Examples end here
Now, given the input text below extract all the indivisible events one by one as explained above with examples. Also, remember to follow the exact format of the examples.
Input: {PROMPT}
Output:

of the models in Figure 2. This visualization illustrates that TANGO 2 consistently outperforms the baselines as the number of events or concepts per prompt increases. In particular, specifically, TANGO closely matches the performance of TANGO 2 only when the textual prompt contains a single concept. However, the disparity between these two models widens as the complexity of the prompt increases

with multiple concepts. Another interesting observation is the relatively low performance of TANGO 2 on single-concept prompts. We suppose this could be the ascribed to the relatively larger influence of noise over concept presence and order on the CLAP-score for single-concept prompts. This supposition is also backed by TANGO 2's poorer IS score—a reference-free measure of acoustic clarity

Table 3: Evaluation results for audio generation in the presence of multiple or single concept(s)/event(s) in the text prompts of the AudioCaps test set. Note: single event IS* has been amended from the main manuscript due to an error in our code.

Model	Multiple Events/Concepts						Single Event/Concept					
	FAD ↓	Objective			Subjective		FAD ↓	Objective			Subjective	
		KL ↓	IS ↑	CLAP ↑	OVL ↑	REL ↑		KL ↓	IS* ↑	CLAP ↑	OVL ↑	REL ↑
AudioLDM 2-Full	2.03	1.64	7.88	0.43	—	—	7.93	1.24	4.50	0.47	—	—
AudioLDM 2-Full-Large	2.33	1.58	8.14	0.44	3.54	3.16	5.82	1.09	6.60	0.49	3.65	3.41
Tango-full-FT	2.69	1.16	7.85	0.54	3.83	3.80	7.52	1.01	6.41	0.51	3.67	3.49
TANGO 2	2.60	1.11	8.98	0.57	3.99	4.07	5.48	1.00	4.95	0.52	3.95	4.10

and diversity across a set of samples—for single event prompts, as shown in Table 3. More on this is explored in Section 5.

4 HUMAN EVALUATION SETUP

We setup a *Gradio*¹ app (UI shown in Fig. 3) for human evaluation. Each annotator was presented with 20 prompts, each prompt having randomly ordered audios from AudioLDM2, TANGO, and TANGO 2. The following instructions were given to the annotators:

Welcome *username*

Instructions for evaluating audio clips

Please carefully read the instructions below.

Task

You are to evaluate three 10 sec long audio outputs to each of the 20 prompts below. These three outputs are from three different models. You are to judge each output with respect to two qualities:

- Overall Quality (OVL): Overall quality of the audio is to be judged based on five grades:
 - 1 : Completely Unnatural
 - 2 : Mostly Unnatural
 - 3 : Somewhat Natural
 - 4 : Mostly Natural
 - 5 : Completely Natural

Overall fidelity, clarity, and noisiness of the audio is important here.

- Relevance (REL): The extent of alignment of the audio with the prompt is to be judged based on five grades:
 - 1 : Completely Irrelevant
 - 2 : Mostly Irrelevant
 - 3 : Somewhat Aligned
 - 4 : Mostly Aligned
 - 5 : Completely Aligned

You are to judge if the concepts from the prompt appear in the audio in the correct temporal order.

Listening guide

- (1) Please use a head/earphone to listen to minimize exposure the external noise.
- (2) Please move to a quiet place as well, if possible.

UI guide

- (1) Each audio clip has two attributes OVL and REL below. You may select the appropriate option from the dropdown list.

(2) To save your judgements, please click on any of the *save* buttons. All the *save* buttons function identically. They are placed everywhere to avoid the need to scroll to save.

Hope the instructions were clear. Please feel free to reach out to us for any queries.

5 ERROR ANALYSIS

In Table 3, the IS of TANGO 2 appears to be worse than the other models. This indicates a more prominent presence of noise and less fidelity in the single concept outputs of TANGO 2. Notably, the subjective evaluation does not indicate this, as the among those 50 samples only six were with single event. However, upon further inspection of a portion of the single event prompts in the entire test set, the validity of the single-event IS score is further substantiated. We reckon this is caused by the lack of focus on single event generation in the preference dataset, as all three of the strategies were inherently focused on multi-event scenarios. We present a few of such single-event samples from TANGO 2 and TANGO in the *Single-Event Comparison* section of the anonymized demo website (<https://339wef0493.github.io>).

¹<https://www.gradio.app>

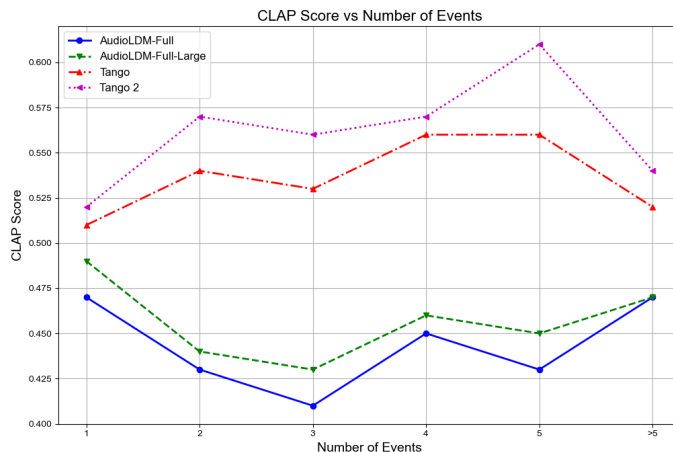


Figure 2: CLAP score of the models vs the number of events or concepts in the textual prompt.

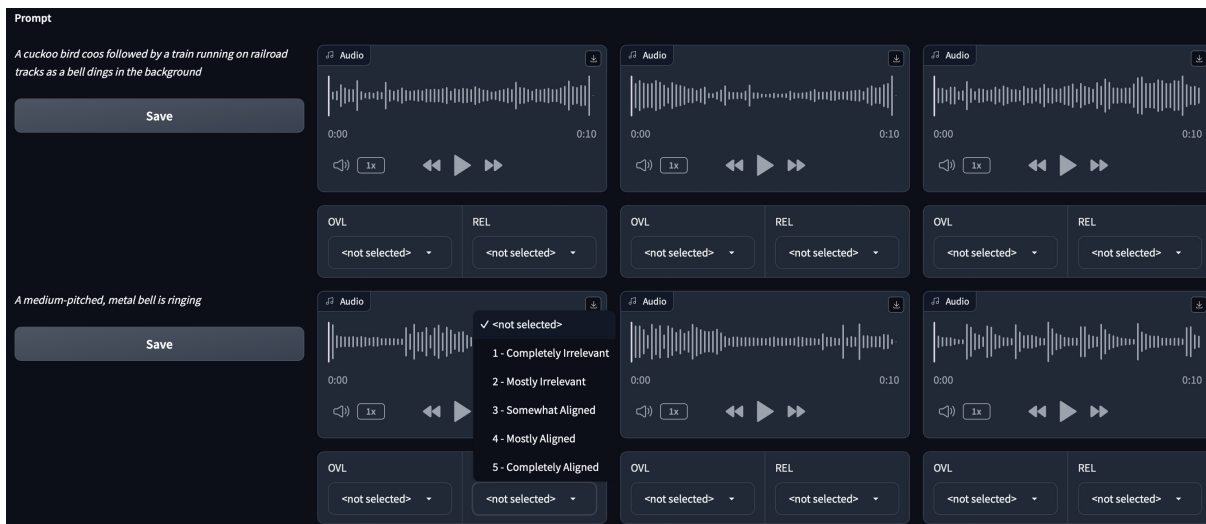


Figure 3: A sample of the UI for human evaluation.