MUGEN: A Playground for Video-Audio-Text Multimodal Understanding and GENeration
******** Supplementary Materials ********

A Dataset Collection

A.1 Full list of modifications

Here is a full list of the modifications we made to the game environment:

- Added game audio with two layers: sound effects and background music. Background music consists of 2 songs corresponding to the space and snow themes. There are 8 sound effects corresponding to Mugen's core actions: walk, jump, collect coin, kill monster, power-up, climb ladder, bump head, die. Each sound effect is triggered by these actions, and one sound effect plays at a time. Background music is layered with the sound effect audio to produce the full audio track.
- Selected a subset of assets for Mugen, the background, and ground to simplify the visual world to a single protagonist and two themes (space and snow).
- Slowed game physics to reduce the pace of gameplay.
- Increased the camera zoom to enlarge the relative size of characters.
- Enabled Mugen to jump on and kill some monsters (snail, worm, face).
- Added animations for when Mugen dies or kills a monster.
- Added a gem and power-up mode for Mugen. When Mugen collects a gem, a bubble shield is placed around Mugen which protects it from being killed by any monster. The shield disappears when Mugen collects a coin.
- Removed the camera tracking of Mugen in the y-axis. Now, the camera only follows Mugen's movements in the x-axis. We also reduced the map height because now the camera is stable in the y-axis.
- Added the barnacle and frog monsters to the game.
- Added hopping animations for the ladybug and frog monsters.
- Adjusted the move speeds and abilities of monsters so none are identical in motion and ability.

Before and after videos highlighting these modifications are in the supplementary video.

A.2 Manual Text

We split the recorded gameplay videos into 3.2 second clips and collected text descriptions by asking annotators to refer to each character with their specific names. Annotators can also adjust the video playback and volume. Figure 1 shows the annotation interface.

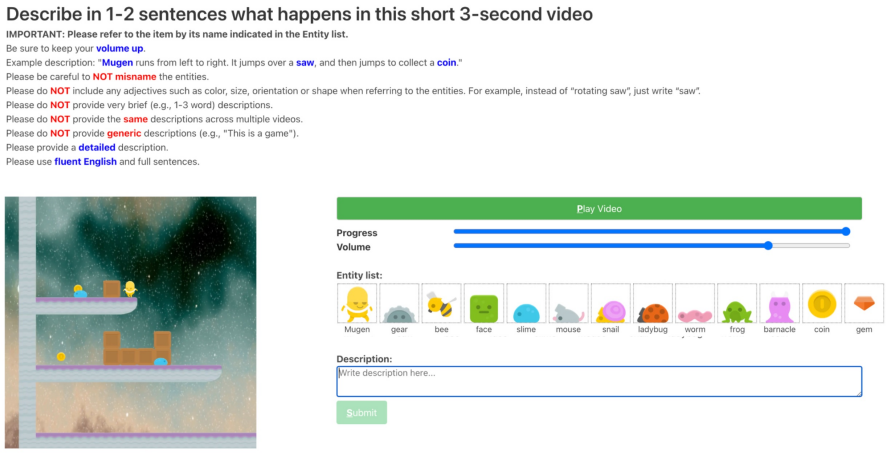


Fig. 1: Annotation interface to collect text descriptions for each video clip.

To ensure the annotation quality, we enforce several rules on the UI before any description can be submitted, such as: 1) Mugen must be mentioned in the description; 2) the video must be played; 3) the description length must be more than 3 words. In addition, we performed several post-processing steps to remove low quality text annotations. Specifically, we randomly sampled collected descriptions and manually inspected annotation quality to identify annotators who write poor quality annotations. We blocked these annotators from writing more descriptions and removed all existing annotations. We also removed text descriptions that are no more than 20 characters. In total, we filtered out 18,449 descriptions from blocked annotators and 4,778 short descriptions. After filtering, MUGEN consists of 378,902 text descriptions for 375,368 video clips, where a very small portion of the clips have more than one description.

A.3 Auto-Text

Since we have accurate metadata saved for all game elements and events, we can apply a template-based algorithm to automatically generate textual descriptions for each video in the following steps:

- Based on Mugen's pose (e.g. jump, walk, climb) saved in metadata for each frame, we merge frames with the same pose into a segment.
- We further merge segments with the same pose that are only interrupted by a few frames of other poses. This is to deal with the cases like Mugen jumping repeatedly. In this process, we also count the number of segments merged (to generate a phrase later).
- For each segment, we generate one short phrase based on a pre-defined template. For example, for a "jump" segment, we consider the following:

- The start and end height to decide whether Mugen jumps higher, lower, or on the same level.
- How many times Mugen jumped (saved from the previous merging step).
- Whether the jump has horizontal movement (left, right, or no move).
- Whether Mugen jumped over any enemy character.
- What kind of surface Mugen lands on.
- Whether the jump killed any enemy character in the end.

– An example templated phrase generated with the above information: "jumps up to the right over a snail to a platform". Note the character "Mugen" is not in the phrase here; it's only added once we merge all phrases in the next step.
– Finally, we merge all short phrases for each segment into the full auto-text description, connect them with "and", and put "Mugen" at the beginning. To avoid excessively long descriptions, segments with less than 5 frames are filtered out at this step. An example of merged auto-text: "Mugen walks to the right, and jumps up to the right to a ladder, and killed by a frog".

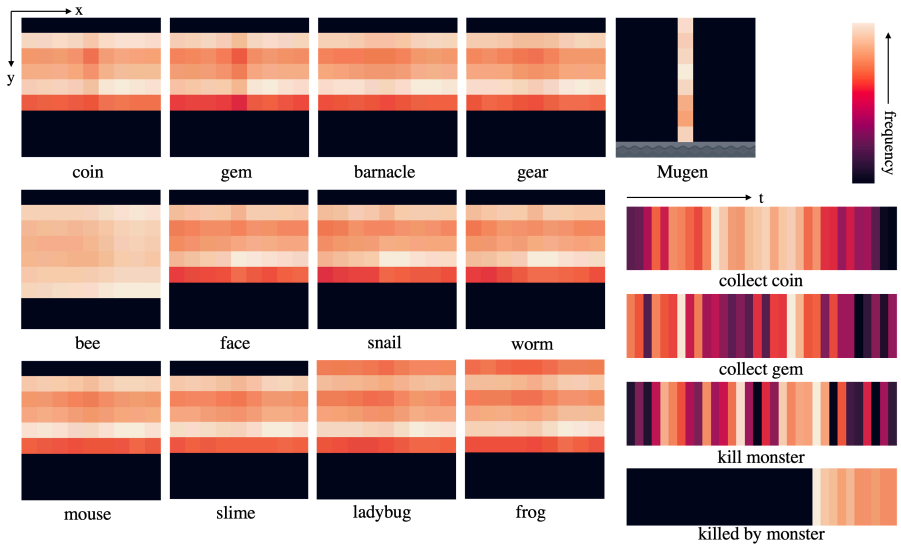


Fig. 2: Location heat maps for all characters and objects and temporal heat maps for the 4 classes of interactions. All heat maps are shown on log scale to make tail events more visible.

B Dataset Statistics

To better understand the distribution of entities and interactions in our dataset, we construct position and temporal heat maps. Specifically, for each character

and object, we compute the distribution of positions in each frame in our 3.2s video dataset, where the position is the center of the entity asset. For each of the 4 classes of interactions, we compute the distribution of interaction time over the 96 frames of each video.

Figure 2 shows these location and temporal heat maps. We see that all of the entities primarily lie on horizontal bands that correspond to the various y positions of platforms (all entities besides Mugen and bee only appear on platforms). Mugen is always centered along the x-axis. Also, we observe that monsters that are able to be killed by Mugen (snail, worm, face) tend to appear closer to the center of the frame. This occurs because Mugen has incentive to approach these monsters and jump on them. Other monsters less frequently appear in the center because Mugen avoids them. Gem and coin also appear less often in the center because they disappear when Mugen collects them. Another observation is that all entities besides Mugen occur slightly more frequently on the right side of the frame than the left, which is due to map procedural generation being slightly biased towards placing platforms on the right side of the map. As for temporal heat maps of interactions, we see that all interactions are more or less uniformly distributed except "killed by monster", which only occurs at the end of videos because the level ends once Mugen dies.

C Diversity of Generated Samples

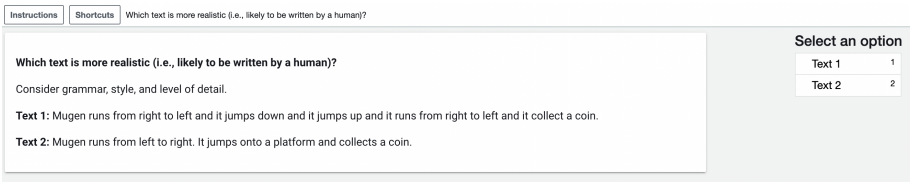
We measure the relative diversity of generated samples with respect to ground truth. Specifically, we use the three encoders jointly trained in the retrieval task to calculate the embeddings and compute the similarity (range in $[-1, 1]$) between every pair of samples within each modality. We report the average similarity in Table 1. Smaller similarity indicates larger diversity within each modality. Our generated samples achieve a similar level of diversity as the ground-truth samples across all modalities. Note that our platform is also customizable to increase the diversity.

Table 1: Diversity comparison between generated samples and ground truth (GT). V, A, T represent video, audio and text. The right arrow indicates the direction of generation.

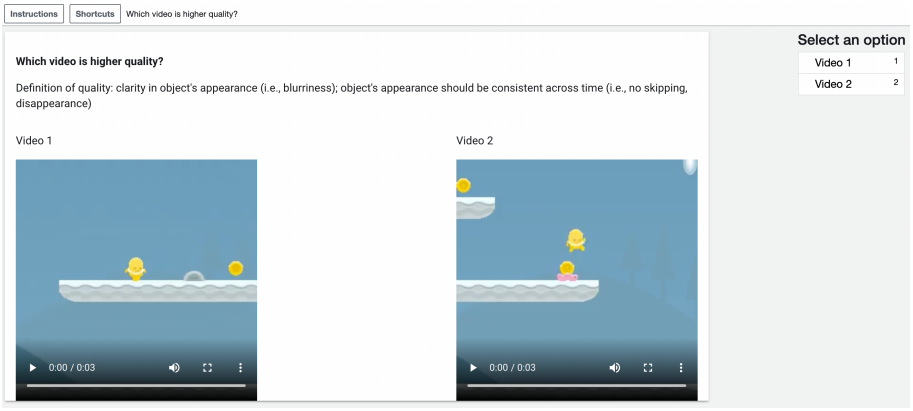
Video		Audio		Text	
T/A → V	GT	V/T → A	GT	V/A → T	GT
0.55/0.55	0.53	0.61/0.60	0.61	0.39/0.43	0.42

D Human Evaluation User Interface

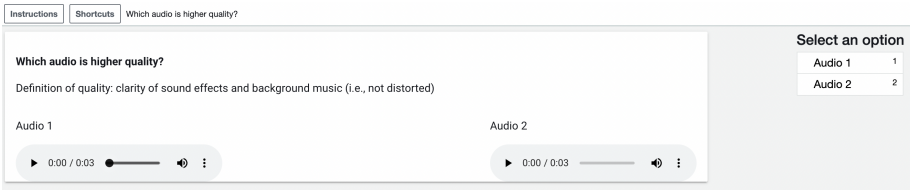
We demonstrate the user interface used in the human evaluation for multimodal generation, as shown in Figure 3. Figure 3a-3c are used for quality evaluation, while Figure 3d-3g are used for faithfulness evaluation.



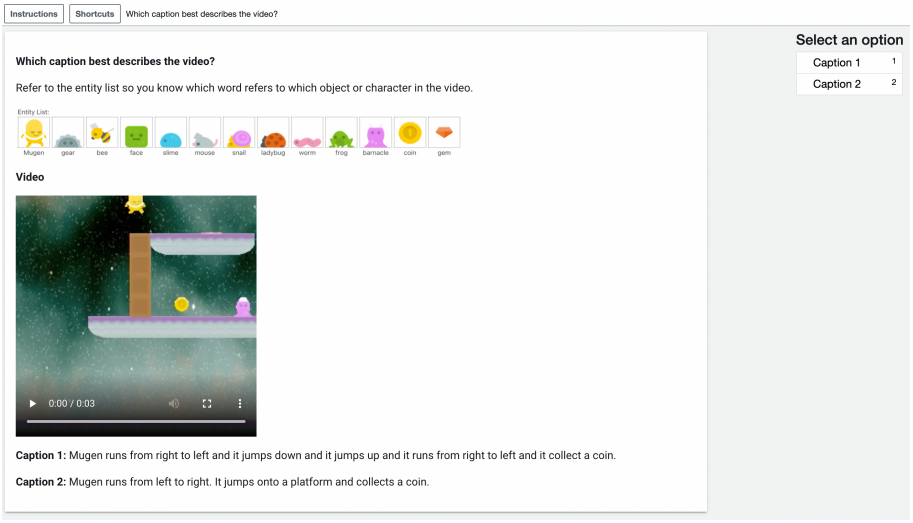
(a) Text quality evaluation.



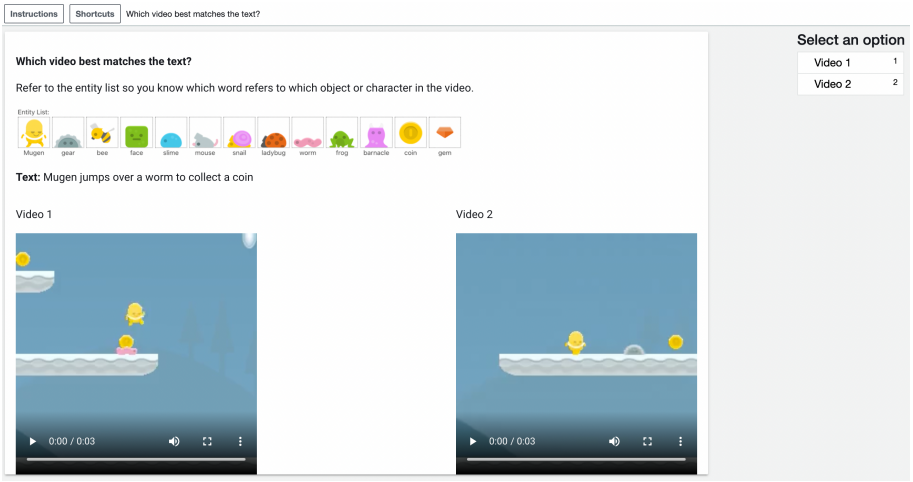
(b) Video quality evaluation.



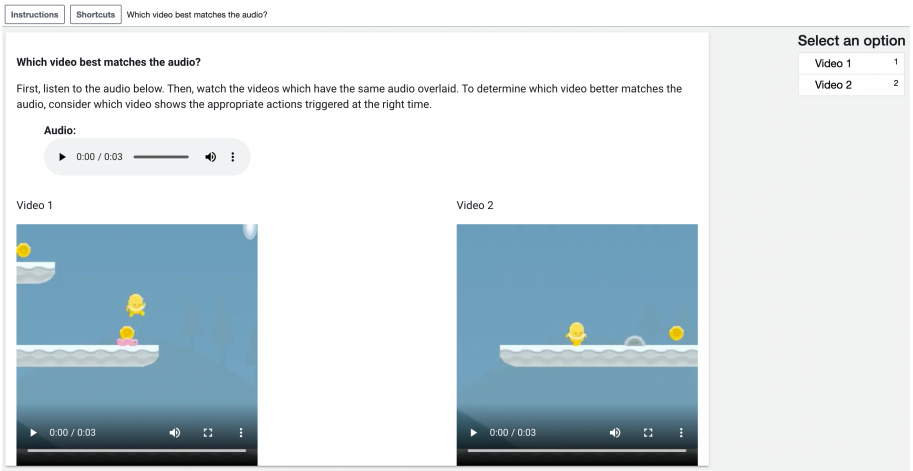
(c) Audio quality evaluation.



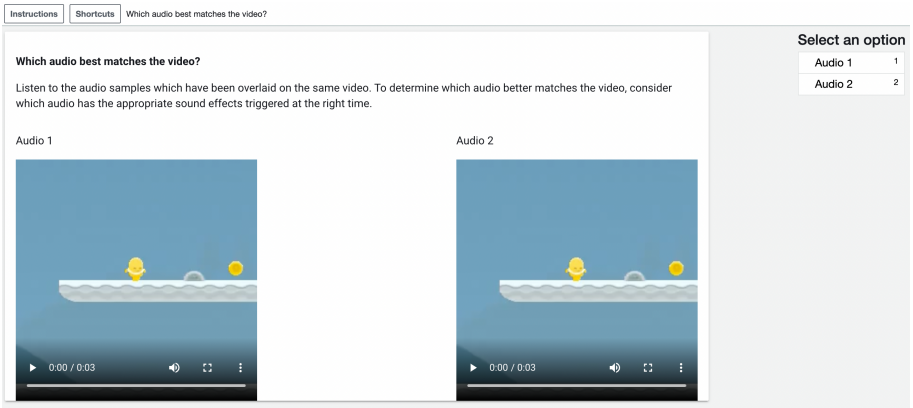
(d) Video-to-text faithfulness evaluation.



(e) Text-to-video faithfulness evaluation.



(f) Audio-to-video faithfulness evaluation.



(g) Video-to-audio faithfulness evaluation.

Fig. 3: Human evaluation user interface.