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

Thomas Hayes*[1], Songyang Zhang*[2], Xi Yin[1], Guan Pang[1], Sasha Sheng[1], Harry Yang[1], Songwei Ge[3], Qiyuan Hu[1], and Devi Parikh[1]

[1]Meta AI [2]University of Rochester [3]University of Maryland
{thayes427,yinxi,gpang,sash,harryyang,isabellehu,dparikh}@fb.com,
szhang83@ur.rochester.edu,songweig@umd.com
https://mugen-org.github.io/

Abstract. Multimodal video-audio-text understanding and generation can benefit from datasets that are narrow but rich. The narrowness allows bite-sized challenges that the research community can make progress on. The richness ensures we are making progress along the core challenges. To this end, we present a large-scale video-audio-text dataset MUGEN, collected using the open-sourced platform game CoinRun. We made substantial modifications to make the game richer by introducing audio and enabling new interactions. We trained RL agents with different objectives to navigate the game and interact with 13 objects and characters. This allows us to automatically extract a large collection of diverse videos and associated audio. We sample 375K video clips (3.2s each) and collect text descriptions from human annotators. Each video has additional annotations that are extracted automatically from the game engine, such as accurate semantic maps for each frame and templated textual descriptions. Altogether, MUGEN can help progress research in many tasks in multimodal understanding and generation. We benchmark representative approaches on tasks involving video-audio-text retrieval and generation. Our dataset and code are released at: https://mugen-org.github.io/.

Keywords: video, audio, language, multimodal, retrieval, generation

1 Introduction

Research in multimodal understanding and generation brings together the subfields of vision and language in AI. Significant progress has been made on image-text understanding and generation tasks, such as CLIP [54] for image-text retrieval and DALL-E [55] for text-to-image generation. This progress has been made possible with large-scale image-text datasets [6,53,61,64,73] that are collected from the web. However, progress in the video-text domain lags due to challenges in data collection and modeling of spatiotemporal information.

* equal contribution, ordered alphabetically

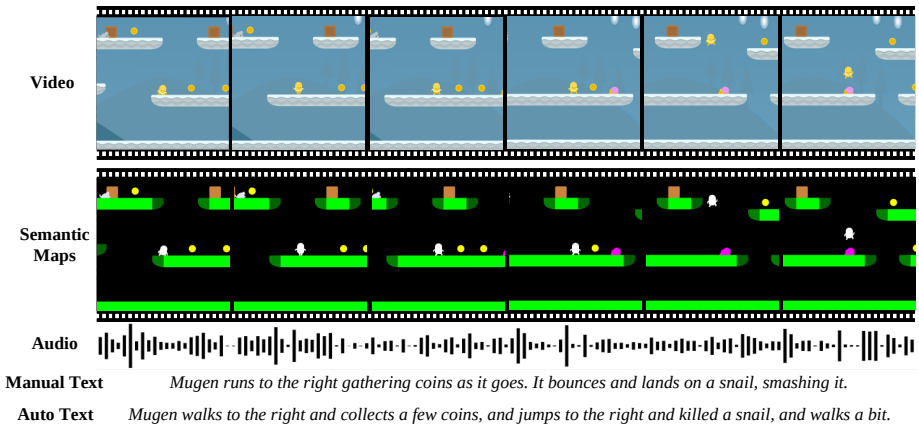


Fig. 1: An example from our dataset. For each 3.2s video clip, we have rich annotations including accurate semantic maps, synchronized audio, manual text collected from human annotators, and auto-text generated based on certain rules.

Many existing video-text datasets [68,40,79,50] are collected in the wild and are proposed for understanding tasks such as video-text retrieval [50], video question answering [40], and generation tasks like video captioning [68]. Yet the performance on these tasks is still far behind their image counterparts [10,42,76] due to the challenges in understanding the complex dynamics in these in-the-wild videos. Moreover, such video-text pairs are too challenging for text-to-video generation, where more constrained datasets are used instead, e.g., bouncing MNIST [33], KTH [51] and UCF-101 [48]. However, these are limited in actions and interactions between entities which are crucial to modeling real-world videos.

In this paper, we introduce MUGEN, a large-scale controllable video-audio-text dataset with rich annotations for multimodal understanding and generation. MUGEN is collected in a closed world based on the open-sourced platform game CoinRun [11]. We have made substantial modifications to the game engine to make the videos more diverse (and delightful) by introducing audio, adjusting camera zoom and stabilization, and enabling new interactions between characters. We name the protagonist "Mugen", and collected videos about Mugen's interactions with the other characters and objects.

To collect videos, we train reinforcement learning (RL) agents to navigate the world and record gameplay. To increase video diversity and reduce bias towards the actions of any single agent, we trained 14 RL agents with different objectives. We record 233K videos of gameplay where the game environment is procedurally generated, so there are no video duplicates. We then sample 375K 3.2s video clips from this video set to collect text descriptions from human annotators (which we call "manual text"). For each video clip, there are additional annotations that come for free: 1) audio is generated from a set of background music and foreground sound effects; 2) accurate semantic maps are generated for each frame using the game assets; 3) automatic text descriptions ("auto-text") are generated

Table 1: Comparison between MUGEN and other video-text datasets. Sent., Sem. and Cust. represent sentence, semantic and customizable. M, A, W and ASR represent descriptions that are manually annotated, alt-text collected from the web, and translated from speech. R and G represent sound recorded with the video and generated based on the video.

Dataset	Video Content	Sent. Source	Number of Sent.	Number of Videos	Number of Clips	Properties Sem.	Properties Audio	Properties Cust.
BMNIST [33]	digit	A	-	-	-	✗	✗	✓
KTH [51]	human action	A	2K	-	2K	✗	✗	✗
TVR [41]	TV show	M	11K	-	22K	✗	R	✗
YouCook2 [79]	cooking	M	14K	2K	14K	✗	R	✗
MSVD [7]	open	M	70K	-	2K	✗	R	✗
A2D [67]	human action	M	7K	-	4K	✓	R	✗
Charades [62]	daily life	M	16K	-	10K	✗	R	✗
FLINTSTONES [25]	cartoon	M	25K	-	25K	✓	R	✗
MSRVTT [68]	human activity	M	200K	7180	10K	✗	R	✗
WebVid-10M [5]	open	W	10.7M	-	10.7M	✗	R	✗
HowTo100M [50]	instructional	ASR	136M	1.2M	136M	✗	R	✗
HD-VILA-100M [69]	open	ASR	100M	3.3M	100M	✗	R	✗
MUGEN (ours)	platform game	M+A	379K	233K	375K	✓	G	✓

based on Mugen's actions and language templates. This results in 375K video-audio-text samples in the MUGEN dataset. One example is shown in Figure 1.

Table 1 shows a comparison between MUGEN and other multimodal datasets. There are several advantages of MUGEN. First, the videos in MUGEN are collected in a closed world with a limited set of visually simple objects and scenes (i.e., simpler than in-the-wild datasets) but with diverse motions and interactions between entities that capture some of the core challenges in video understanding and generation (i.e., richer than other closed world datasets). Not only does the narrowness allow for bite-sized challenges to make progress on, it also alleviates the need for web-scale data and correspondingly massive compute resources in studying multimodal understanding and generation. Second, there are rich annotations for each video including accurate semantic maps, synchronized audio, and auto-text and manual text descriptions that can enable a wide variety of tasks in multimodal research. Third, the game engine setup allows us to render videos at different resolutions on the fly, which is more flexible and storage efficient. Fourth, the game engine is modifiable, and once released, will allow the research community to collect more data to study a diverse set of problems.

MUGEN enables study of many multimodal video-audio-text tasks. In this paper, we focus on several tasks including retrieval and generation between all pairs of modalities. For the research community to make progress, it is vital to have consistent evaluation protocols. While many automatic metrics have been proposed for evaluation of generative models, human judgement is still the gold standard. Prior works have compared to ground-truth for text [12]

or audio generation [9], but video generation evaluation is usually conducted by comparing to baselines because ground-truth is too challenging. This makes it difficult to compare methods and calibrate progress over time. Given that MUGEN is a closed world dataset with simplified visual elements, it is possible to compare to ground-truth videos for evaluation. In this paper, we conduct a comprehensive human evaluation for various cross-modal generation baselines. We evaluate both the generation quality as well as faithfulness to input modality. We hope this evaluation protocol will be adopted by the community. We will make our evaluation interfaces publicly available.

We summarize the contributions of this paper as follows:

- We propose MUGEN, a large-scale dataset of 375K video-audio-text samples with additional annotations of semantic maps and auto-text to facilitate research in multimodal understanding and generation.
- We benchmark the performance of video-audio-text retrieval and generation between every pair of modalities in a unified framework. To our knowledge, this is the first work that benchmarks all these tasks on one dataset.
- We formulate a standard protocol for human evaluation of quality and faithfulness for four generation tasks.
- We will release the dataset and the game platform so the community can generate more data for a variety of tasks to push the field forward.

2 Related Work

Multimodal Datasets. Existing multimodal datasets belong to two categories based on the visual content: open world (in-the-wild environments) and closed world (constrained environments). Open world datasets such as MSCOCO [45], ConceptualCaptions [6], and WIT [64] are widely used for image-text research. CLEVR [30] is a closed world dataset collected by arranging different 3D shapes on a clean background, which enables systematic progress in visual reasoning by reducing the complexity and bias from the real world.

Most video-text datasets are open world. MSRVTT [68], ANetCap [37], MSVD [7], and DiDeMo [3] contain videos of sports and human actions collected from the web. YouCook2 [79] and HowTo100M [50] contain instructional videos collected from YouTube. TVR [41], TVQA [40], and LSMDC [57] are collected from TV series and movies. Ego4D [24] is collected by people wearing an egocentric camera recording everyday activities around the world. Videos in these datasets contain complex backgrounds and diverse events, which makes them very challenging. Datasets from constrained environments, *e.g.* Bouncing MNIST (BMNIST) [33] and KTH [51], have been proposed. These datasets don't capture some of the core challenges in videos such as multiple entities interacting with each other in meaningful ways. FLINTSTONES [25] is created from an animated series, but the scenes are too diverse for the size of the dataset. In contrast, MUGEN simplifies the visual complexities of the scenes and objects, but captures complex motion and interactions between multiple entities.

The text in existing datasets are either collected from humans [25,68,79] or extracted from speech [50]. Besides human descriptions, MUGEN also allows generating templated auto-text descriptions for videos of arbitrary lengths.

Most open world video-text datasets are associated with audio recorded from human speech and/or events. AudioSet [22] and VGGSound [8] are collected with video-audio pairs for audio event recognition. However, the video and audio are often not well-aligned. (E.g., the speech may describe things not related to or aligned with the video and background noise is common.) In contrast, the video and audio in MUGEN are synchronized based on Mugen's actions, making it feasible to study less explored tasks like audio generation from video or text.

Multimodal Understanding and Generation. Multimodal research typically involves four modalities: image, video, audio and text. Image-text tasks are widely studied, such as VQA [4], image captioning [1,31,72], image-text retrieval [32], visual storytelling [27], text-to-image generation [56], etc. Earlier methods aimed to design effective models for specific tasks [23,43,52,71]. Later work [10,42,78] focused on large-scale pre-training to learn cross-modal representations that can be transferred to various downstream tasks. More recently, CLIP [54], CogView [14], and DALL-E [55] leverage even larger-scale training to improve model generalization and zero-shot learning. FLAVA [63] and Florence [74] were proposed as foundation models for both vision and language.

Many video-text tasks have been proposed, such as video QA [40], video-text retrieval [68], video grounding [19], video captioning [68], text-to-video generation [25], etc. Similar to image-text research, early approaches focused on a single task [20,17,38,77]. Some recent work proposed novel architectures to learn task-agnostic video-text embeddings, such as MIL-NCE [49] and ClipBERT [39]. Compared to video-text retrieval and video captioning, text-to-video generation is relatively understudied, largely due to a lack of feasible datasets. Early methods [51,44,46] were evaluated on simple datasets like BMNIST [33] and KTH [51]. Mazaheri and Shah [48] annotated 10 action classes from UCF-101 [48]. However, the limited motion in these datasets restricts the diversity of the collected text descriptions, making them sub-optimal for studying text-to-video generation.

There are also efforts on audio, such as audio-text retrieval [36], audio captioning [35], audio-to-video generation [47], video-to-audio generation [28], etc. We explore video-audio-text retrieval and generation between all pairs of modalities on MUGEN, and conduct extensive evaluations including human evaluation.

3 MUGEN Dataset

Environment. In-the-wild video understanding and generation poses several challenges, including understanding motion of objects, interactions among objects, physics, camera vs. object and scene motion, 3D depth of scenes, diverse object appearances and semantics, lighting conditions, etc. Our goal was to develop a dataset that is rich along some of these dimensions, but narrower along others, to enable focused advances in some of the core challenges of multimodal video research. Specifically, we desired a closed world dataset where physics

are simplified, the camera angle is fixed, the number of objects is limited, and lighting is consistent. Yet, we sought diverse motions and interactions between entities (dataset statistics are shown in Figure 2). For these purposes, we chose an open source video game which (1) enables training RL agents to collect video data at scale and (2) gives access to the game engine that provides additional high quality annotations for free, such as precise frame-level semantic maps and automatic text descriptions. Amongst open source games, we chose OpenAI's CoinRun [11] because of its ease of modification.

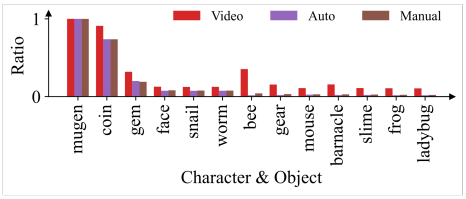
OpenAI's CoinRun is a platform game developed for quantifying generalization of RL agents [11,15,29]. The game has a single main character (who we call Mugen) with the objective to collect coins without being killed by monsters. Each level has a number of coins and monsters, and the level ends when Mugen collects all coins, Mugen is killed by a monster, or the level times out after 21 seconds. The environment is procedurally generated, with each level having a unique configuration of platforms, coins, and monsters.

We made a number of modifications to increase the diversity of game events and enhance richness, such as adding audio, slowing game physics, adjusting camera zoom and stabilization, and enabling new interactions between characters. Altogether, our updated version of CoinRun features Mugen, 10 monsters, coin and gem objects, and 2 world themes, space and snow. Mugen can take 16 different actions (see Figure 2c for the most frequent actions). Monsters differ in their action vocabulary; some walk, others hop, and one flies. A full list of modifications, before and after videos highlighting these changes, and images of these characters, objects, and themes can be found in the appendix.

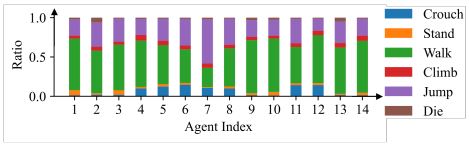
Audio. The audio consists of two layers, sound effects and background music. We chose 8 sound effects corresponding to Mugen's core actions: walk, jump, collect coin, kill monster, power-up, climb ladder, bump head, die. Each sound effect is triggered by these actions, and one sound effect plays at a time. Background music features 2 themes for the space and snow game themes. Background music is layered with the sound effect audio to produce the full audio track.

Video Collection. We train RL agents to navigate the environment and collect gameplay videos. We use an IMPALA-CNN architecture [16] and train agents with Proximal Policy Optimization [59]. Inputs to the agent include the current game frame and the agent's velocity. The agent's performance in the game is immaterial to us; we care about maximizing the diversity of video data. To this end, we trained 14 agents with modified reward functions to achieve different behaviors. For example, decreasing the reward discount factor makes the agent more myopic and risk-tolerant, so the agent dies frequently. Figure 2b shows the distribution of Mugen's poses for each agent where the variation in time spent in different poses indicates differing actions. To further increase diversity, we ensured that the seed for map procedural generation is always unique. We have verified there are no duplicate videos in MUGEN.

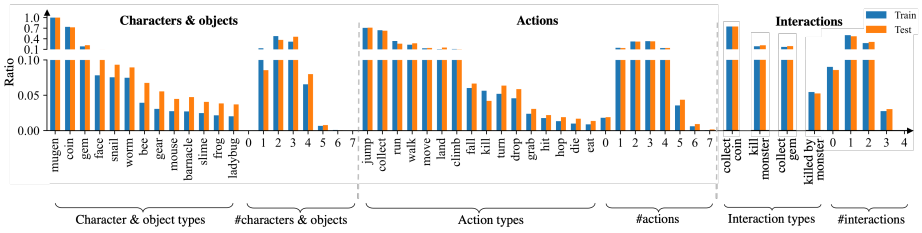
To efficiently handle large-scale game video data and enable easy data customization (e.g., swapping characters or objects, changing background), we do not save the rendered videos. Instead, we save all metadata such as world lay-



(a) Characters and objects in video and text.



(b) Mugen's poses across the 14 RL agents.



(c) Characters, objects, actions and interactions in three splits and their occurrence per video.

Fig. 2: Distribution of characters, objects, actions and interactions.

out and character movements in a json format, from which we can render RGB frames and pixel-accurate segmentation maps at any resolution up to 1400×1400 on-the-fly, resulting in more efficient data storage. [1]

We recorded 233K videos of gameplay ranging from 3.2s to 21s (level timeout) at 30 frames per second. Each video corresponds to a whole level of gameplay. We will release the game engine so others can customize the data environment or agents for their own purposes.

Manual Text. We split the 233K videos into 3.2s (96 frames) clips and ask annotators to describe in 1-2 sentences what happens in the short video. After filtering low quality annotations, MUGEN consists of $378,902$ text descriptions for $375,368$ video clips. [2] Refer to the appendix for the annotation interface and details on annotation quality control.

Auto-Text. In addition to collecting human annotation, we also developed a template-based algorithm to automatically generate textual descriptions for videos based on game engine metadata. See the appendix for details.

Note that both video and auto-text can be generated automatically. We can generate arbitrary amounts of video-text data with arbitrary lengths. This makes it feasible to study more tasks where manual annotations are expensive to acquire, such as text-conditioned long video generation [21], video grounding [75], and dense video captioning [37]. Auto-text is also highly structured in nature. This simplifies the text and improves model explainability since each action and interaction in the video has a unique description in the text.

Dataset Statistics. In total, MUGEN consists of 375K 3.2s video clips paired with 379K manual text descriptions, as well as 233K longer (3.2s to 21s) videos.

[1] Storage is $> 100\times$ smaller than 1024×1024 videos stored with lossless encoding.
[2] A very small portion of the clips have more than one description.

Each video clip or long video also comes with semantic maps, auto-text, and audio. There are 11 characters, 2 objects, 16 different actions for Mugen, and 4 classes of interactions with other objects and characters: collect coin, collect gem (power up), kill monster, killed by monster.

We first analyze the occurrences of characters and objects in video, manual text, and auto-text in Figure 2a. [3] We observe that not all characters and objects appearing in the video are mentioned in the text. This is because annotators are more likely to describe characters that interact with Mugen than those in the background. Given the unbalanced distribution of characters and objects, when splitting our dataset, we sample fewer videos featuring only Mugen or Mugen's interaction with coins for the validation and test sets. Both validation and test sets contain only one manual text per video. This results in $349,666$, $12,851$, $12,851$ video clips in training, validation, and test sets, respectively.

The distributions of characters and objects, actions, and interactions are shown in Figure 2c. "Jump" and "collect" are the top 2 most frequent actions, consistent with "collect coin" being the most frequent class of interaction (this is CoinRun after all!). The rarest interaction type is Mugen being killed by a monster. We also show the distribution of the number of characters and objects, actions, and interactions in each video. Most videos contain 2-4 characters and objects, 2-4 actions, and 2-3 interactions. This is more diverse than other closed world datasets with one or two digits moving [33] or a single person moving in a scene [51]. We also show the location heatmap of each charchater/object and temporal heatmap of each action/interaction in the appendix.

As shown in Table 1, MUGEN is several orders of magnitude larger than existing closed world datasets such as BMNIST [33], KTH [51] and FLINTSTONES [25]. While it is smaller than some open world datasets including HowTo100M [50] and WebVid-2M [5], it is also visually less diverse, making it feasible to train effective models without having to work with web-scale data. Moreover, our dataset provides audio aligned with video, accurate frame-level semantic maps, and automatically generated text descriptions which enable studying a variety of tasks. Finally, this dataset is customizable with the released game engine, so the community can generate more data of different distributions.

4 Video-Audio-Text Retrieval and Generation

While MUGEN can enable many tasks, we focus on retrieval and generation between every pair of modalities. We first present the cross-modal retrieval framework and then a unified pipeline for cross-modal generation.

4.1 Video-Audio-Text Retrieval

Cross-modal retrieval, which retrieves samples from one modality given a query from another, is a fundamental task with many real-world applications. For example, text-to-video retrieval is widely used for video search.

[3] The occurrence of one character is counted at most once in each video/text.

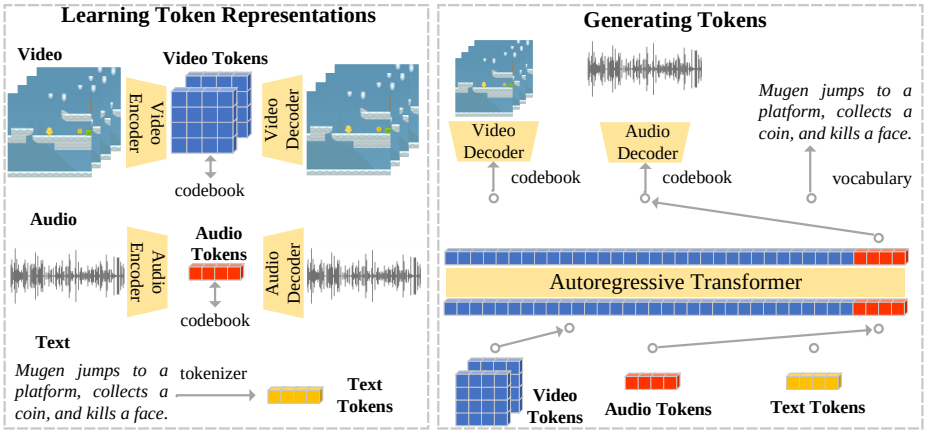


Fig. 3: A unified framework for generation between every pairs of modalities. The right part shows an example of video-to-audio generation.

We use an encoder F_x to map input x of each modality to a feature vector $\mathbf{f}_x = F_x(x)$. It is projected into a joint embedding space $\mathbf{e}_x = \mathbf{f}_x \cdot \mathbf{W}_x$, where $\mathbf{W}_x$ are the learnable parameters. Given inputs p and q from two modalities P and Q, the similarity can be computed by a scaled cosine function, $s(p,q) = cos(\mathbf{e}_p, \mathbf{e}_q) \cdot e^{\tau_{PQ}}$, where τ_{PQ} is a learnable temperature parameter. The matching loss L_{PQ} is computed as:

$$L_{PQ} = -\frac{1}{2N}\sum_{i=1}^{N}(\log(\frac{e^{s(p_i,q_i)}}{\sum_{j=1}^{N} e^{s(p_i,q_j)}}) + \log(\frac{e^{s(p_i,q_i)}}{\sum_{k=1}^{N} e^{s(p_k,q_i)}})), \tag{1}$$

where N is the number of samples in a batch, p_i and q_i represent the ith sample from P and Q modalities within the batch.

We train three models with L_{VA}, L_{VT} and L_{AT} separately for video(V)-audio(A), video-text(T), and audio-text retrieval. For comparison, we also sum three losses to learn a joint model.

During inference, to retrieve samples from modality P given a query from modality Q, we rank the samples based on the similarities $s(p,q)$. To retrieve modality P based on queries from two modalities Q and R, we sum the similarity from two modalities, $s(p,q)+s(p,r)$. $s(p,q)$ and $s(p,r)$ can either come from two models independently trained by L_{PQ} and L_{PR} or the joint model.

4.2 Video-Audio-Text Generation

Cross-modal generation has gained increasing interest in recent years. Amongst video-audio-text cross-modal generation tasks, video-to-text generation (video captioning) is most studied, while other tasks (video-to-audio, text-to-video, etc.) are relatively under-explored.

Table 2: Performance comparison on video(V)-audio(A)-text(T) retrieval. For retrieval in modality P, $Q+R$ denotes the ensemble of two models independently trained by L_{PQ} and L_{PR}; $Q+R*$ denotes the joint model. Recalls are shown in percentage(%).

Query	Video Retrieval			Query	Audio Retrieval			Query	Text Retrieval		
Type	R1	R5	R10	Type	R1	R5	R10	Type	R1	R5	R10
A	58.59	88.83	94.41	V	61.14	88.99	94.59	V	10.61	25.72	34.70
T	8.54	22.50	31.71	T	2.40	8.35	13.38	A	2.95	9.36	14.80
A+T	**81.50**	**96.10**	**98.26**	V+T	**69.59**	**92.48**	**96.42**	V+A	**11.68**	**27.13**	**36.60**
A+T*	62.54	87.33	92.62	V+T*	41.83	73.04	82.68	V+A*	10.95	26.24	35.33

Inspired by the success of using a VQ-VAE [65] and transformer for image [55] and video [70] generation, we adopt a similar and unified framework for cross-modal generation, as shown in Figure 3. For each modality, we first learn a discrete codebook to encode the data. Then an decoder-only transformer is used to learn token generation from one modality to another.

Learning Token Representations. For video representation, we train a 3D VQ-VAE to learn a codebook following the training losses in [70]. The encoder is used to encode videos as inputs for the transformer during training, and the decoder is used for video generation during inference. Similarly, we train a 1D VQ-VAE to learn audio compression following the training losses in [13]. For text representation, we learn a tokenizer from manual text in the training set.

Generating Tokens. We use a decoder-only transformer to do auto-regressive token generation. During training, the input to the transformer is a sequence of tokens concatenated from modality P and Q. Video tokens are flattened from 3D latent codes into 1D. Text tokens are truncated or padded to have the same length. Causal attention is used where each token can only attend to prior tokens. The transformer learns to predict the token ids at every location. The loss functions for the two modalities are summed, similar to DALL-E [55]. During inference, given the tokens from P, we auto-regressively generate all tokens for Q. For audio or video, we use the predicted tokens to look up the codebook embeddings and feed them into VQ-VAE's decoder to reconstruct the video or audio. For text, the vocabulary is used to reconstruct the sentences.

5 Experiments

5.1 Video-Audio-Text Retrieval

The retrieval task is to find the true match from the test set in one modality, given queries of either one of the other modalities or both modalities. For retrieval based on queries from two modalities, we compare the ensemble of two separately trained models and the joint model. We report recall at rank 1, 5 and 10. For all experiments, if not specified, the video dimension is 256×256×32, where the 32 frames are evenly sampled from the 96 frames to save computation.

Implementation Details. Pre-trained models are used as initialization including ResNet-18 [26] pre-trained on VGGSound [8] for the audio encoder, S3D [66] pre-trained on Kinetics 400 [34] for the video encoder, and DistilBERT [58] for the text encoder. The parameters in the text encoder are fixed [4] and the other two encoders are learnable. The temperature τ_{PQ} is initialized as 0.07 and the maximum is 100, the learning rate is 0.001, and batch size is 16. All models are trained for 400K steps and checkpoints are selected based on the validation set.
Results. The results are shown in Table 2. We have the following observations: 1) across all single modality retrieval tasks, video-to-audio and audio-to-video retrieval perform the best and text-to-audio and audio-to-text perform the worst. This is because video and audio are synchronized, while audio and text are only sparsely aligned on Mugen's actions and interactions; different text descriptions can map to similar audio samples. 2) retrieval based on two modalities with an ensemble of models ($P + Q$) consistently outperforms single modality retrieval. This is because the other modality can provide complementary information. For example, text contains information of Mugen's moving direction that is available in video but not audio. 3) the performance of the joint model ($P + Q^*$) typically falls between the separately trained models, which indicates that it is challenging to learn a joint embedding space for all modalities.

5.2 Video-Audio-Text Generation

We evaluate cross-modal generation for all pairs of modalities. We use V, A, T to denote video, audio, text, and P2Q to represent the task (e.g., T2V for text-to-video generation). We focus on quantitative evaluations of the quality of the output and faithfulness to the input.
Implementation Details. The 3D VQ-VAE is similar to [70] except that we use a kernel size of 3, which significantly sped up training compared to the original kernel size of 4. We use a down-sample ratio of 32×32×4 for video compression and a vocabulary of size 2048. The 3D VQ-VAE is trained for 600K steps with a learning rate of 0.003 and a batch size of 8. The 1D VQ-VAE for audio features non-causal, dilated 1D convolutions where the dilation is grown by a factor of 3. The vocabulary size is 1024. Audio sample rate is 22kHz. The 1D VQ-VAE is trained for 1M steps with a learning rate of 0.0003 and a batch size of 4. We use Byte-Pair Encoding (BPE) [60,18] for text tokenization and train a tokenizer from the manual text annotations in the training set. All P2Q generation models are trained with the same transformer architecture and optimization hyper-parameters. For the transformer, we use 12 layers with a hidden dimension of 768 and 8 attention heads. All models are trained for 600K steps with a learning rate of 0.0003 and batch size of 4. Model checkpoints are selected based on the performance on the validation set.
Inference. During inference, we perform token sampling from the estimated distribution with filtering. For video or audio generation, we use top-k= 100 and

[4] Our initial experiments show unstable training with learnable text encoder.

Table 3: Performance comparison on all generation tasks. D.S. denotes the down-sampled training set. F(V/A)D denotes FVD for video quality and FAD for audio quality. R.Sim. denotes the Relative Similarity to evaluate the faithfulness to the input. "B4", "M.", "R." and "C." denote BLEU4, METEOR, ROUGE and CIDEr. "Q." and "F." stand for quality and faithfulness. Video frame lengths map to different frame rates (8/16/32 represent 2.5/5/10 frames per second). Audio token lengths map to different compression ratios (68/137/275/551 represent 1024/512/256/128 compression ratios in VQ-VAE). All metrics except F(V/A)D are shown in percentage (%). All metrics are better when higher except F(V/A)D, which is the lower the better.

Out	In	Train	Out	In	Text	Auto						Human	
Mod.	Mod.	Data	Len.	Len.	Type	F(V/A)D	R.Sim.	B4	M.	R.	C.	Q.	F.
Text	Video	Full	-	8	M	-	83.5	7.4	20.2	27.9	18.2	-	-
				16			101.5	**7.8**	20.8	28.7	**20.2**	-	-
				32			**108.0**	**7.8**	**21.3**	**29.1**	19.9	**31.3**	**42.6**
	Audio	Full	-	68	M	-	101.0	6.0	19.3	26.5	14.1	-	-
				137			103.9	6.3	19.3	26.6	14.4		
				275			106.7	6.5	19.3	26.8	14.7		
				551			**107.5**	**6.7**	**19.4**	**27.1**	**15.5**		
Video	Text	Full	8	-	M	$112.7_{\pm 0.2}$	39.5	5.1	15.2	21.7	11.1	-	-
			16		M	$72.7_{\pm 2.0}$	63.9	7.3	18.5	26.5	15.3	-	-
			32		M	$\mathbf{61.0_{\pm 0.6}}$	64.9	8.1	19.9	28.1	**19.2**	**17.0**	**31.6**
			32		A	$140.7_{\pm 3.1}$	14.3	6.4	17.8	25.3	14.9	9.2	11.7
			32		M+A	$61.4_{\pm 1.0}$	**66.5**	**8.2**	**20.0**	**28.2**	19.1	-	-
		D.S.	8		M	$112.7_{\pm 1.1}$	42.0	5.0	15.3	21.9	11.0	-	-
			16		M	$72.2_{\pm 1.7}$	69.5	7.3	18.6	26.7	15.8	-	-
			32		M	$62.0_{\pm 0.7}$	70.6	7.9	20.0	28.2	19.0	**18.8**	**35.7**
			32		A	$151.7_{\pm 3.4}$	20.5	6.2	17.7	24.9	14.2	12.1	13.1
			32		M+A	$\mathbf{61.0_{\pm 1.4}}$	**72.2**	**8.2**	**20.2**	**28.4**	**19.7**	-	-
	Audio	Full	32	68	-	$64.0_{\pm 0.9}$	79.9	-	-	-	-	-	-
				137		$66.4_{\pm 0.4}$	91.4					-	-
				275		$\mathbf{63.4_{\pm 0.4}}$	93.1					**17.6**	**37.1**
				551		$64.5_{\pm 1.0}$	**93.6**					-	-
Audio	Video	Full	68	32	-	$523.8_{\pm 1.0}$	86.5	-	-	-	-	-	-
			137			$128.5_{\pm 0.4}$	95.4					-	-
			275			$52.3_{\pm 0.5}$	**96.7**					**15.2**	**31.1**
			551			$\mathbf{50.0_{\pm 0.8}}$	92.7					-	-
	Text	Full	68	-	M	$574.0_{\pm 2.8}$	**91.3**	7.0	**18.5**	26.5	15.3	-	-
			137			$171.1_{\pm 2.2}$	88.8	6.9	**18.5**	26.5	15.0		
			275			$\mathbf{93.7_{\pm 1.6}}$	86.9	**7.1**	**18.5**	**26.9**	**16.2**		
			551			$109.4_{\pm 2.3}$	78.7	6.9	18.1	26.1	15.9		

top-p= 0.9 for filtering. For text generation, we use top-k= 1, which is the same as beam search [2] with size 1 in the common captioning setup.

Automatic Evaluation. For text generation, we use metrics that are widely used in captioning evaluation including BLEU4, METEOR, ROUGE, and CIDEr. For video quality, we follow prior practices and use I3D pre-trained on Kinetics 400 to calculate FVD. For audio quality, we use the pre-trained audio encoder on VGGSound to calculate FAD. To automatically evaluate faithfulness to input, we propose a new metric Relative Similarity (R.Sim.) that leverages the retrieval models. Specifically, we calculate the average similarity between the input and output divided by the average similarity between the input and the ground truth. For T2V and T2A generation, we use the V2T and A2T models applied on the generated video/audio to calculate the captioning metrics.

Human Evaluation. We establish a human evaluation protocol to calibrate towards the Ground Truth (GT). We randomly selected 512 samples from the test set and manually inspected the descriptions to ensure the samples were diverse and not too simple (e.g., to avoid multiple samples where Mugen simply jumps onto a platform). For each task, we evaluate both quality and faithfulness. We use "quality" [5] to measure the single modality quality and "faithfulness" to measure the alignment between the input and output modality. As it is not straightforward for humans to judge the alignment between audio and text, we do not evaluate T2A and A2T but focus on the other cross-modal generation tasks. For quality, we ask human judges to select the higher quality sample (video, audio, or text) between the generation and GT. For faithfulness, human judges are asked to select the media which better aligns with the input [6]. Each comparison is evaluated by 5 judges and the majority vote is taken. We report the percentage of samples that are chosen over the GT as the final metric. The upper bound for these evaluations is around 50% when a human judge cannot tell the difference between the generation and the GT.

We took several steps to mitigate bias and improve replicability in human evaluation. We shuffle sample order, shuffle the presentation order of models, anonymize model generations, and recruit diverse raters with Amazon Mechanical Turk. We also remove confounding factors. For instance, for video comparison, we render the GT video using the same theme (snow or space), frame rate, and resolution as the generated video. For generated text, several post-processing steps are used: capitalize the first letter of the first word in each sentence, use "Mugen" to replace "mugen", and remove duplicated spaces.

Text Generation from Video or Audio. As shown in Table 3, we vary the video frame rate and audio compression ratio (a higher compression ratio results in fewer tokens) for comparison. For V2T generation, higher frame rate leads to stronger performance. Human evaluation shows high faithfulness with 42.6% of samples chosen over GT, and relatively lower quality with 31.25% of samples considered more realistic than GT. For A2T generation, a smaller compression ratio (more tokens) is better. A2T performs worse than V2T as video-text are more densely aligned than audio-text.

[5] Even within "quality", there are different kinds of deficiencies and more fine-grained evaluation could be part of future work.

[6] We will release the annotation UIs for others to follow this protocol.

Video Generation from Audio or Text. For T2V generation, we experiment with training using manual text and auto-text. As mentioned earlier, we balanced the characters in the validation and test sets. Correspondingly, to study the effects of data balancing, we also generate a smaller training set with 233K samples by down-sampling videos with Mugen or Mugen and coins only. As shown in Table 3, for T2V generation, we have the following observations: 1) Larger frame rate leads to better performance in all automatic metrics. 2) Auto-text performs worse than manual text and cannot noticeably improve performance when combined with manual text. This is because we evaluate on manual text for all comparisons. We hypothesize that auto-text may be useful when manual text is not available or is limited. 3) Models trained on the down-sampled training set consistently outperform those on the full set. Future work can explore other sampling strategies to fully utilize the training set. 4) Human evaluation results show better faithfulness compared to quality. The trends between automatic metrics and human evaluation results are similar.

For A2V generation, a smaller compression ratio (longer token sequence) leads to better quality and faithfulness in the automatic metrics. Human evaluation shows higher faithfulness compared to quality, similar to the T2V task.

Audio Generation from Video or Text. For audio generation, generating a longer audio sequence (less compression) leads to better quality in FAD for both T2A and V2A. But the R.Sim. may not follow the same trend. Human evaluation also shows higher faithfulness than quality, similar to other tasks.

When comparing the human evaluation results for all tasks, we see V2T is the easiest with the highest quality and faithfulness. V2T is also the most studied task in literature. For the other three tasks, faithfulness is considerably higher than quality. Improving video and audio reconstruction in VQ-VAE can potentially lead to higher quality. This also suggests that humans can reasonably ignore generation quality in faithfulness evaluation. We also compare the diversity of generated samples in Appendix C.

6 Conclusion

We introduce MUGEN – a closed world, large-scale multimodal dataset based on a significantly enhanced version of the platform game CoinRun [11]. MUGEN has videos, human-annotated text descriptions, automatically generated templated text descriptions, frame-level pixel-accurate semantic segmentation maps, as well as audio. The multiple modalities and rich annotations in MUGEN enable research progress in various tasks in multimodal understanding and generation without requiring web-scale data or massive compute. We explore retrieval and generation between every pair of modalities. To evaluate generative models, we establish a human evaluation protocol by calibrating towards the ground-truth samples, making it easier to compare performance and show progress. The MUGEN dataset, the modified game engine, our training code and models, and the human evaluation UIs can be found at: https://mugen-org.github.io/.

References

1. Agrawal, H., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., Lee, S., Anderson, P.: Nocaps: Novel object captioning at scale. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8948–8957 (2019) 5
2. Anderson, P., Fernando, B., Johnson, M., Gould, S.: Guided open vocabulary image captioning with constrained beam search. arXiv preprint arXiv:1612.00576 (2016) 12
3. Anne Hendricks, L., Wang, O., Shechtman, E., Sivic, J., Darrell, T., Russell, B.: Localizing moments in video with natural language. In: Proceedings of the IEEE international conference on computer vision. pp. 5803–5812 (2017) 4
4. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2425–2433 (2015) 5
5. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In: IEEE International Conference on Computer Vision (2021) 3, 8
6. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3558–3568 (2021) 1, 4
7. Chen, D., Dolan, W.B.: Collecting highly parallel data for paraphrase evaluation. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies. pp. 190–200 (2011) 3, 4
8. Chen, H., Xie, W., Vedaldi, A., Zisserman, A.: Vggsound: A large-scale audio-visual dataset. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 721–725. IEEE (2020) 5, 11
9. Chen, L., Srivastava, S., Duan, Z., Xu, C.: Deep cross-modal audio-visual generation. In: Proceedings of the on Thematic Workshops of ACM Multimedia 2017. pp. 349–357 (2017) 4
10. Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Universal image-text representation learning. In: European conference on computer vision. pp. 104–120. Springer (2020) 2, 5
11. Cobbe, K., Klimov, O., Hesse, C., Kim, T., Schulman, J.: Quantifying generalization in reinforcement learning. In: International Conference on Machine Learning. pp. 1282–1289. PMLR (2019) 2, 6, 14
12. Cui, Y., Yang, G., Veit, A., Huang, X., Belongie, S.: Learning to evaluate image captioning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5804–5812 (2018) 3
13. Dhariwal, P., Jun, H., Payne, C., Kim, J.W., Radford, A., Sutskever, I.: Jukebox: A generative model for music. arXiv preprint arXiv:2005.00341 (2020) 10
14. Ding, M., Yang, Z., Hong, W., Zheng, W., Zhou, C., Yin, D., Lin, J., Zou, X., Shao, Z., Yang, H., et al.: Cogview: Mastering text-to-image generation via transformers. Advances in Neural Information Processing Systems **34** (2021) 5
15. Edwards, A., Sahni, H., Schroecker, Y., Isbell, C.: Imitating latent policies from observation. In: International conference on machine learning. pp. 1755–1763. PMLR (2019) 6
16. Espeholt, L., Soyer, H., Munos, R., Simonyan, K., Mnih, V., Ward, T., Doron, Y., Firoiu, V., Harley, T., Dunning, I., et al.: Impala: Scalable distributed deep-rl with

importance weighted actor-learner architectures. In: International Conference on Machine Learning. pp. 1407–1416. PMLR (2018) 6
17. Gabeur, V., Sun, C., Alahari, K., Schmid, C.: Multi-modal Transformer for Video Retrieval. In: European Conference on Computer Vision (ECCV) (2020) 5
18. Gage, P.: A new algorithm for data compression. C Users Journal **12**(2), 23–38 (1994) 11
19. Gao, J., Sun, C., Yang, Z., Nevatia, R.: Tall: Temporal activity localization via language query. In: ICCV (2017) 5
20. Gao, L., Guo, Z., Zhang, H., Xu, X., Shen, H.T.: Video captioning with attention-based lstm and semantic consistency. IEEE Transactions on Multimedia **19**(9), 2045–2055 (2017) 5
21. Ge, S., Hayes, T., Yang, H., Yin, X., Pang, G., Jacobs, D., Huang, J.B., Parikh, D.: Long video generation with time-agnostic vqgan and time-sensitive transformer. arXiv preprint arXiv:2204.03638 (2022) 7
22. Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 776–780. IEEE (2017) 5
23. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6904–6913 (2017) 5
24. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4d: Around the world in 3,000 hours of egocentric video. arXiv preprint arXiv:2110.07058 (2021) 4
25. Gupta, T., Schwenk, D., Farhadi, A., Hoiem, D., Kembhavi, A.: Imagine this! scripts to compositions to videos. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 598–613 (2018) 3, 4, 5, 8
26. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 11
27. Huang, T.H., Ferraro, F., Mostafazadeh, N., Misra, I., Agrawal, A., Devlin, J., Girshick, R., He, X., Kohli, P., Batra, D., et al.: Visual storytelling. In: NAACL (2016) 5
28. Iashin, V., Rahtu, E.: Taming visually guided sound generation. arXiv preprint arXiv:2110.08791 (2021) 5
29. Igl, M., Ciosek, K., Li, Y., Tschiatschek, S., Zhang, C., Devlin, S., Hofmann, K.: Generalization in reinforcement learning with selective noise injection and information bottleneck. Advances in neural information processing systems **32** (2019) 6
30. Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., Girshick, R.: Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2901–2910 (2017) 4
31. Johnson, J., Karpathy, A., Fei-Fei, L.: Densecap: Fully convolutional localization networks for dense captioning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4565–4574 (2016) 5
32. Johnson, J., Krishna, R., Stark, M., Li, L.J., Shamma, D., Bernstein, M., Fei-Fei, L.: Image retrieval using scene graphs. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3668–3678 (2015) 5

33. Kahou, S.E., Michalski, V., Memisevic, R., Pal, C., Vincent, P.: Ratm: recurrent attentive tracking model. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 1613–1622. IEEE (2017) 2, 3, 4, 5, 8
34. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017) 11
35. Kim, C.D., Kim, B., Lee, H., Kim, G.: Audiocaps: Generating captions for audios in the wild. In: NAACL-HLT (2019) 5
36. Koepke, A.S., Oncescu, A.M., Henriques, J., Akata, Z., Albanie, S.: Audio retrieval with natural language queries: A benchmark study. IEEE Transactions on Multimedia (2022) 5
37. Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Niebles, J.C.: Dense-captioning events in videos. In: International Conference on Computer Vision (ICCV) (2017) 4, 7
38. Le, T.M., Le, V., Venkatesh, S., Tran, T.: Hierarchical conditional relation networks for video question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9972–9981 (2020) 5
39. Lei, J., Li, L., Zhou, L., Gan, Z., Berg, T.L., Bansal, M., Liu, J.: Less is more: Clipbert for video-and-language learning via sparse sampling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7331–7341 (2021) 5
40. Lei, J., Yu, L., Bansal, M., Berg, T.L.: Tvqa: Localized, compositional video question answering. In: EMNLP (2018) 2, 4, 5
41. Lei, J., Yu, L., Berg, T.L., Bansal, M.: Tvr: A large-scale dataset for video-subtitle moment retrieval. In: European Conference on Computer Vision. pp. 447–463. Springer (2020) 3, 4
42. Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al.: Oscar: Object-semantics aligned pre-training for vision-language tasks. In: European Conference on Computer Vision. pp. 121–137. Springer (2020) 2, 5
43. Li, Y., Gan, Z., Shen, Y., Liu, J., Cheng, Y., Wu, Y., Carin, L., Carlson, D., Gao, J.: Storygan: A sequential conditional gan for story visualization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6329–6338 (2019) 5
44. Li, Y., Min, M., Shen, D., Carlson, D., Carin, L.: Video generation from text. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018) 5
45. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014) 4
46. Liu, Y., Wang, X., Yuan, Y., Zhu, W.: Cross-modal dual learning for sentence-to-video generation. In: Proceedings of the 27th ACM International Conference on Multimedia. pp. 1239–1247 (2019) 5
47. Mama, R., Tyndel, M.S., Kadhim, H., Clifford, C., Thurairatnam, R.: Nwt: Towards natural audio-to-video generation with representation learning. arXiv preprint arXiv:2106.04283 (2021) 5
48. Mazaheri, A., Shah, M.: Video generation from text employing latent path construction for temporal modeling. arXiv preprint arXiv:2107.13766 (2021) 2, 5
49. Miech, A., Alayrac, J.B., Smaira, L., Laptev, I., Sivic, J., Zisserman, A.: End-to-End Learning of Visual Representations from Uncurated Instructional Videos. In: CVPR (2020) 5

50. Miech, A., Zhukov, D., Alayrac, J.B., Tapaswi, M., Laptev, I., Sivic, J.: Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2630–2640 (2019) 2, 3, 4, 5, 8
51. Mittal, G., Marwah, T., Balasubramanian, V.N.: Sync-draw: Automatic video generation using deep recurrent attentive architectures. In: Proceedings of the 25th ACM international conference on Multimedia. pp. 1096–1104 (2017) 2, 3, 4, 5, 8
52. Niu, Y., Tang, K., Zhang, H., Lu, Z., Hua, X.S., Wen, J.R.: Counterfactual vqa: A cause-effect look at language bias. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12700–12710 (2021) 5
53. Ordonez, V., Kulkarni, G., Berg, T.: Im2text: Describing images using 1 million captioned photographs. Advances in neural information processing systems **24** (2011) 1
54. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021) 1, 5
55. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International Conference on Machine Learning. pp. 8821–8831. PMLR (2021) 1, 5, 10
56. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. In: International conference on machine learning. pp. 1060–1069. PMLR (2016) 5
57. Rohrbach, A., Rohrbach, M., Tandon, N., Schiele, B.: A dataset for movie description. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3202–3212 (2015) 4
58. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019) 11
59. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017) 6
60. Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1715–1725. Association for Computational Linguistics, Berlin, Germany (Aug 2016). https://doi.org/10.18653/v1/P16-1162, https://aclanthology.org/P16-1162 11
61. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2556–2565 (2018) 1
62. Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: Crowdsourcing data collection for activity understanding. In: European Conference on Computer Vision. pp. 510–526. Springer (2016) 3
63. Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., Kiela, D.: Flava: A foundational language and vision alignment model. arXiv preprint arXiv:2112.04482 (2021) 5
64. Srinivasan, K., Raman, K., Chen, J., Bendersky, M., Najork, M.: Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In: SIGIR (2021) 1, 4
65. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017) 10

66. Xie, S., Sun, C., Huang, J., Tu, Z., Murphy, K.: Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In: Proceedings of the European conference on computer vision (ECCV). pp. 305–321 (2018) 11
67. Xu, C., Hsieh, S.H., Xiong, C., Corso, J.J.: Can humans fly? action understanding with multiple classes of actors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2264–2273 (2015) 3
68. Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5288–5296 (2016) 2, 3, 4, 5
69. Xue, H., Hang, T., Zeng, Y., Sun, Y., Liu, B., Yang, H., Fu, J., Guo, B.: Advancing high-resolution video-language representation with large-scale video transcriptions. In: International Conference on Computer Vision and Pattern Recognition (CVPR) (2022) 3
70. Yan, W., Zhang, Y., Abbeel, P., Srinivas, A.: Videogpt: Video generation using vq-vae and transformers (2021) 10, 11
71. Yang, Z., Gong, B., Wang, L., Huang, W., Yu, D., Luo, J.: A fast and accurate one-stage approach to visual grounding. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019) 5
72. You, Q., Jin, H., Wang, Z., Fang, C., Luo, J.: Image captioning with semantic attention. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4651–4659 (2016) 5
73. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics **2**, 67–78 (2014) 1
74. Yuan, L., Chen, D., Chen, Y.L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., et al.: Florence: A new foundation model for computer vision. arXiv preprint arXiv:2111.11432 (2021) 5
75. Zeng, R., Xu, H., Huang, W., Chen, P., Tan, M., Gan, C.: Dense regression network for video grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10287–10296 (2020) 7
76. Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., Gao, J.: Vinvl: Revisiting visual representations in vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5579–5588 (2021) 2
77. Zhang, S., Peng, H., Fu, J., Lu, Y., Luo, J.: Multi-scale 2d temporal adjacent networks for moment localization with natural language. In: TPAMI (2021) 5
78. Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J., Gao, J.: Unified vision-language pre-training for image captioning and vqa. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 13041–13049 (2020) 5
79. Zhou, L., Xu, C., Corso, J.J.: Towards automatic learning of procedures from web instructional videos. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018) 2, 3, 4, 5