Abstract

As humans, we experience the world with all our senses or modalities (sound, sight, touch, smell, and taste). We use these modalities, particularly sight and touch, to convey and interpret specific meanings. Multimodal expressions are central to conversations; a rich set of modalities amplify and often compensate for each other. A multimodal conversational AI system answers questions, fulfills tasks, and emulates human conversations by understanding and expressing itself via multiple modalities. This paper motivates, defines, and mathematically formulates the multimodal conversational research objective. We provide a taxonomy of research required to solve the objective: multimodal representation, fusion, alignment, translation, and co-learning. We survey state-of-the-art datasets and approaches for each research area and highlight their limiting assumptions. Finally, we identify multimodal co-learning as a promising direction for multimodal conversational AI research.

1 Introduction

The proliferation of smartphones has dramatically increased the frequency of interactions that humans have with digital content. These interactions have expanded over the past decade to include conversations with smartphones and in-home smart speakers. Conversational AI systems (e.g., Alexa, Siri, Google Assistant) answer questions, fulfill specific tasks, and emulate natural human conversation (Hakkani-Tür et al., 2011; Gao et al., 2019).

Early examples of conversational AI include those based on primitive rule-based methods such as ELIZA (Weizenbaum, 1966). More recently, conversational systems were driven by statistical machine translation systems: translating input queries to responses (Ritter et al., 2011; Hakkani-Tür et al., 2012). Orders of magnitude more data led to unprecedented advances in conversational technology in the mid-part of the last decade. Techniques were developed to mine conversational training data from the web search query-click stream (Hakkani-Tür et al., 2011; Heck, 2012; Hakkani-Tür et al., 2013) and web-based knowledge graphs (Heck and Hakkani-Tür, 2012; El-Kahky et al., 2014). With this increase in data, deep neural networks gained momentum in conversational systems (Mesnil et al., 2014; Heck and Huang, 2014; Sordoni et al., 2015; Vinyals and Le, 2015; Shang et al., 2015; Serban et al., 2016; Li et al., 2016a,b).

Most recently, specialized deep learning-based conversational agents were developed primarily for three tasks: (1) goal-directed tasks in research systems (Shah et al., 2016; Eric et al., 2017; Liu et al., 2017, 2018; Li et al., 2019; Hosseini-Asl et al., 2020; Wu et al., 2020; Peng et al., 2021; Xu et al., 2021b) and commercial products (Siri, Cortana, Alexa, and Google Assistant), (2) question-answering (Yi et al., 2019; Raffel et al., 2020; Zaheer et al., 2021), and (3) open-domain conversations (Wolf et al., 2019; Zhou et al., 2020; Adiwardana et al., 2020; Paranjape et al., 2020; Roller et al., 2020; Bao et al., 2020; Henderson et al., 2020; Zhang et al., 2020a). However, developing a single system with a unified approach that achieves human-level performance on all three tasks has proven elusive and is still an open problem in conversational AI.

One limitation of existing agents is that they often rely exclusively on language to communicate with users. This contrasts with humans, who converse with each other through a multitude of senses. These senses or modalities complement each other, resolving ambiguities and emphasizing ideas to make conversations meaningful. Prosody, auditory expressions of emotion, and backchannel agreement supplement speech, lip-reading disambiguates unclear words, gesticulation makes spatial references, and high-fives signify celebration.

Alleviating this unimodal limitation of conversational AI systems requires developing methods...
to extract, combine, and understand information streams from multiple modalities and generate multimodal responses while simultaneously maintaining an intelligent conversation.

Similar to the taxonomy of multimodal machine learning research (Baltrušaitis et al., 2017), the research required to extend conversational AI systems to multiple modalities can be grouped into five areas: Representation, Fusion, Translation, Alignment, and Co-Learning. Representation and fusion involve learning mathematical constructs to mimic sensory modalities. Translation maps relationships between modalities for cross-modal reasoning. Alignment identifies regions of relevance across modalities to identify correspondences between them. Co-learning exploits the synergies across modalities by leveraging resource-rich modalities to train resource-poor modalities.

Concurrently, it is necessary for the research areas outlined above to address four main challenges in multimodal conversational reasoning – disambiguation, response generation, coreference resolution, and dialogue state tracking (Kottur et al., 2021). Multimodal disambiguation and response generation are challenges associated with fusion that determine whether available multimodal inputs are sufficient for a direct response or if follow-up queries are required. Multimodal coreference resolution is a challenge in both translation and alignment, where the conversational agent must resolve referential mentions in dialogue to corresponding objects in other modalities. Multimodal dialogue state tracking is a holistic challenge across research areas typically associated with task-oriented systems. The goal is to parse multimodal signals to infer and update values for slots in user utterances.

In this paper, we discuss the taxonomy of research challenges in multimodal Conversational AI as illustrated in Figure 1. Section 2 provides a history of research in multimodal conversations. In Section 3, we mathematically formulate multimodal conversational AI as an optimization problem. Sections 4, 5, and 6 survey existing datasets and state-of-the-art approaches for multimodal representation and fusion, translation, and alignment. Section 7 highlights limitations of existing research in multimodal conversational AI and explores multimodal co-learning as a promising direction for research.

2 Background

Early work in multimodal conversational AI focused on the use of visual information to improve automatic speech recognition (ASR). One of the earliest papers along these lines is by Yuhas et al. (1989) followed by many papers including work by Meier et al. (1996), Duchnowski et al. (1994), Bregler and Konig (1994), and Ngiam et al. (2011).

Advances in client-side capabilities enabled ASR systems to utilize other modalities such as tactile, voice, and text inputs. These systems supported more comprehensive interactions and facilitated a higher degree of personalization. Examples include ESPRIT’s MASK (Lamel et al., 1998), Microsoft’s MiPad (Huang et al., 2001), and AT&T’s MATCH (Johnston et al., 2002).

Vision-driven tasks motivated research in adding visual understanding technology into conversational AI systems. Early work in reasoning over text+video include work by Ramanathan et al. (2014) where they leveraged these combined modalities to address the problem of assigning names of people in the cast to tracks in TV videos. Kong et al. (2014) leveraged natural language descriptions of RGB-D videos for 3D semantic parsing. Srivastava and Salakhutdinov (2014) developed a multimodal Deep Boltzmann Machine for image-text retrieval and ASR using videos. Antol et al. (2015) introduced a dataset and baselines for multimodal question-answering, a challenge combining computer vision and natural language pro-
cessing. More recent work by Zhang et al. (2019b) and Selvaraju et al. (2019) leveraged conversational explanations to make vision and language models more grounded, resulting in improved visual question answering.

While modalities most commonly considered in the conversational AI literature are text, vision, tactile, and speech, other sources of information are gaining popularity within the research community. These include eye-gaze, 3D scans, emotion, action and dialogue history, and virtual reality. Heck et al. (2013) and Hakkani-Tür et al. (2014) use gesture, speech, and eye-gaze to resolve and infer intent in conversational web-browsing systems. Grauman et al. (2021) present ego-centric video understanding, Padmakumar et al. (2021), and Shridhar et al. (2020) present task completion from 3D simulations, and Gao et al. (2021) presents multisensory object recognition.

Processing conventional and new modalities brings forth numerous challenges for multimodal conversations. To answer these challenges, we will first mathematically formulate the multimodal conversational AI problem, then detail fundamental research sub-tasks required to solve it.

3 Mathematical Formulation

We formulate multimodal conversational AI as an optimization problem. The objective is to find the optimal response $S$ to a message $m$ given underlying multimodal context $c$. Based on the sufficiency of the context, the optimal response could be a statement of fact or a follow-up question to resolve ambiguities. Statistically, $S$ is estimated as:

$$S = \arg\max_R p(R|c, m).$$  \hspace{1cm} (1)

The probability of an arbitrary response $r$ can be expressed as a product of the probabilities of responses $\{r_i\}_{i=1}^T$ over $T$ turns of conversation (Sordoni et al., 2015).

$$p(r|c, m) = \prod_{i=1}^T p(r_i|r_1, \ldots, r_{i-1}, c, m)$$  \hspace{1cm} (2)

It is also possible for conversational AI to respond through multiple modalities. We represent the multimodality of output responses by a matrix $R := \{r_1^1, r_1^2, \ldots, r_1^l\}$ over $l$ permissible output modalities.

$$S = \arg\max_R p(R|c, m)$$  \hspace{1cm} (3)

Learning from multimodal data requires manipulating information from all modalities using a function $f(\cdot)$ consisting of five sub-tasks: representation, fusion, translation, alignment, and co-learning. We include these modifications and present the final multimodal conversational objective below.

$$S = \arg\max_{R} p(R|f(c, m))$$  \hspace{1cm} (4)

In the following sections, we describe each sub-task contained in $f(\cdot)$.

4 Multimodal Representation + Fusion

Multimodal representation learning and fusion are primary challenges in multimodal conversations. Multimodal representation is the encoding of multidimensional signals in a format amenable to computational processing. Multimodal fusion concerns joining features from multiple modalities to make predictions.

4.1 Multimodal Representations

Using multimodal information of varying granularity for conversations necessitates techniques to represent high-dimensional signals in a latent space. These latent multimodal representations encode human senses to improve a conversational AI’s perception of the real-world. Success in multimodal tasks requires that representations satisfy three desiderata (Srivastava and Salakhutdinov, 2014):

1. Similarity in the representation space implies similarity of the corresponding concepts
2. The representation is easy to obtain in the absence of some modalities
3. It is possible to infer missing information from observed modalities

There exist numerous representation methods for the range of problems multimodal conversational AI addresses. Multimodal representations are broadly classified as either joint representations or coordinated representations (Baltrušaitis et al., 2017).

4.1.1 Joint Representations

Joint representations combine unimodal signals into the same representation space. Traditional techniques to learn joint representations include multimodal autoencoders (Ngiam et al., 2011), multimodal deep belief networks (Srivastava and
A notion of distance would be closer to the dog’s image representation (Vaswani et al., 2017) and BERT when modeling (Nicolaou et al., 2011).

4.1.2 Coordinated Representations

In contrast, coordinated representations model each modality separately. Constraints coordinate representations of separate modalities by enforcing cross-modal similarity over concepts. For example, the audio representation $g_a(\cdot)$ of a dog’s bark would be closer to the dog’s image representation $g_i(\cdot)$ and further away from a car’s (Equation 7). A notion of distance $d$ between modalities in the coordinated space enables cross-modal retrieval.

$$d(g_a(\text{dog}), g_i(\text{dog})) < d(g_a(\text{dog}), g_i(\text{car}))$$

In practice, contrastive objectives are used to coordinate representations between pairs of modalities. Contrastive learning has been successful in relating separate views of the same image (Becker and Hinton, 1992; Chen et al., 2020a; He et al., 2020; Grill et al., 2020; Radford et al., 2021), images and their natural language descriptions (Weston et al., 2010; Kiros et al., 2014; Zhang et al., 2020; Li et al., 2020), and videos with their corresponding audio and natural language descriptions (Owens et al., 2016; Korbar et al., 2018; Sun et al., 2019a; Miech et al., 2020; Alayrac et al., 2020; Akbari et al., 2021; Xu et al., 2021a; Qian et al., 2021; Morgado et al., 2021).

4.2 Multimodal Fusion

Multimodal fusion combines features from multiple modalities to make decisions, denoted by the final block before the outputs in Figure 1. Fusion approaches are broadly classified into model-agnostic and model-based methods.

Model-agnostic methods are independent of specific algorithms and are split into early, late, and hybrid fusion. Early fusion integrates features following extraction, projecting features into a shared space (Potamianos et al., 2003; Ngiam et al., 2011; Nicolaou et al., 2011; Jansen et al., 2019). In contrast, late fusion integrates decisions from unimodal predictors (Becker and Hinton, 1992; Korbar et al., 2018; Shuster et al., 2020; Alayrac et al., 2020; Akbari et al., 2021). Early fusion is predominantly used to combine features extracted in joint representations while late fusion combines decisions made in coordinated representations. Hybrid fusion exploits both low and high level modality interactions (Wu et al., 2005; Schwartz et al., 2020; Piergiovanni et al., 2020; Goyal et al., 2020).

Model-based methods consist of graphical techniques like Hidden Markov Models (Nefian et al., 2002; Gurban et al., 2008), neural networks (Nicolaou et al., 2011; Antol et al., 2015; Gao et al., 2015; Malinowski et al., 2015; Kottur et al., 2018; Qian et al., 2021), and transformers (Xu and Saenko, 2016; Hori et al., 2017; Peng et al., 2019; Zhang et al., 2019a; Shuster et al., 2020; Chen et al., 2020b; Geng et al., 2021; Xu et al., 2021b).

4.3.1 Factor Graph Attention

Schwartz et al. (2020) develops Factor Graph Attention (FGA), a joint representation for multi-turn question answering grounded in images. FGA embeds images using VGG-16 (Simonyan and Zisserman, 2015) or F-RCNN (Ren et al., 2016) and...
using multimodal concatenation of jointly represented modalities (Table 2).

<table>
<thead>
<tr>
<th>Modalities</th>
<th>Turn 1</th>
<th>Turn 2</th>
<th>Turn 3</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image Only</td>
<td>37.6</td>
<td>28.1</td>
<td>20.7</td>
<td>28.7</td>
</tr>
<tr>
<td>Style Only</td>
<td>18.3</td>
<td>15.3</td>
<td>17.0</td>
<td>16.9</td>
</tr>
<tr>
<td>Dialogue History Only</td>
<td>1.0</td>
<td>33.7</td>
<td>32.3</td>
<td>22.3</td>
</tr>
<tr>
<td>Style + Dialogue</td>
<td>18.3</td>
<td>45.4</td>
<td>43.1</td>
<td>35.4</td>
</tr>
<tr>
<td>Image + Dialogue</td>
<td>37.6</td>
<td>39.4</td>
<td>32.6</td>
<td>36.5</td>
</tr>
<tr>
<td>Image + Style</td>
<td><strong>54.0</strong></td>
<td>41.1</td>
<td>35.2</td>
<td>43.4</td>
</tr>
<tr>
<td>Style + Dialogue + Image</td>
<td><strong>54.0</strong></td>
<td><strong>51.9</strong></td>
<td><strong>44.8</strong></td>
<td><strong>50.3</strong></td>
</tr>
</tbody>
</table>

Table 2: Recall@1 (%) on Image-Chat using TRANSRESNET_RET (ResNetx-IG-3.5B, MM-Sum)

4.3.3 MultiModal Versatile Networks (MMV)

Alayrac et al. (2020) presents a training strategy to learn coordinated representations using self-supervised contrastive learning from instructional videos. Videos are encoded using TSM with a ResNet50 backbone (Lin et al., 2019), audio is encoded using log MEL spectrograms from ResNet50, and text is encoded using Google News pre-trained word2vec (Mikolov et al., 2013).

Alayrac et al. (2020) defines three types of coordinated spaces: shared, disjoint, and ‘fine+coarse’. The shared space enables direct comparison and navigation between modalities, by assuming equal granularity. The disjoint space sidesteps navigation to solve the granularity problem by creating a space for each pair of modalities. The ‘fine+coarse’ space solves both issues by learning two spaces. A fine-grained space compares audio and video, while a lower-dimensional coarse-grained space compares fine-grained embeddings with text. We further discuss the MMV model in Section 6.3.

5 Multimodal Translation

Multimodal translation maps embeddings from one modality to signals from another for cross-modality reasoning (Figure 1). Cross-modal reasoning enables multimodal conversational AI to hold meaningful conversations and resolve references across multiple senses, specifically language and vision. To this end, we survey existing work addressing the translation of images and videos to text. We discuss multimodal question-answering and multimodal dialogue, translation tasks that extend to multimodal conversations.

5.1 Image

Antol et al. (2015) and Zhu et al. (2016) present Visual Question-Answering (VQA) and Visual7W for...
multimodal question answering (MQA). The MQA challenge requires responding to textual queries about an image. Both datasets collect questions and answers using crowd workers, encouraging trained models to learn natural responses. Heck and Heck (2022) presents the Visual Slot dataset, where trained models learn answers to questions grounded in Uls.

The objective of MQA is a simplification of Equation 4 to a single-turn, single-timestep scenario ($T = 1$), producing a response to a question $m_q$ given multimodal context $\{c_i\}_{i=1}^n$:

$$S_{MQA} = \arg\max_R p(R|f(c_1, \ldots, c_n, m_q))$$

Multi-turn question-answering (MTQA) is the next step towards multimodal conversational AI. VisDial (Das et al., 2017) extends VQA to multiple turns, translating over QA history in addition to images. GuessWhat?! (de Vries et al., 2017) is a guessing game, discovering objects in a scene through dialogue. MANYMODALQA (Hannan et al., 2020) requires reasoning over prior knowledge, images, and databases. MIMOQA (Singh et al., 2021b) is an example of multimodal responses, where answers are image-text pairs.

The objective of MTQA (Equation 9) is an extension of MQA to include QA history $h_{qa} = \{m_{q_1}, r_{a_1}, m_{q_2}, r_{a_2}, \ldots, m_{q_{n-1}}, r_{a_{n-1}}\}$:

$$S_{MTQA} = \arg\max_R p(R|f(c_1, \ldots, c_n, h_{qa}))$$

Image-Grounded Conversations (IGC) (Mostafazadeh et al., 2017) builds on MTQA by presenting a dataset for multimodal dialogue (MD): machine perception and conversation through language. Image-Chat (Shuster et al., 2020) extends IGC to agents with personalities. Crowd workers hold three-turn conversations about an image with one of 215 emotions (e.g., peaceful, erratic, skeptical). Motivated by the popularity of visual content in instant-messaging, Meme incorporated Open-domain Dialogue (MOD) (Fei et al., 2021) contains natural language conversations interspersed with behavioral stickers. SIMMC (Moon et al., 2020) and SIMMC2.0 (Kottur et al., 2021) present goal-oriented dialogue for shopping. The challenge requires leveraging dialogue and a state of the world to resolve references, track dialogue state, and recommend the correct object. IGC, Image-Chat, MOD, SIMMC, and SIMMC2.0 solve the MD objective that depends on previous dialogue responses $h_d = \{m_{d_1}, r_{d_1}, m_{d_2}, r_{d_2}, \ldots, m_{d_{n-1}}, r_{d_{n-1}}\}$:

$$S_{MD} = \arg\max_R p(R|f(c_1, \ldots, c_n, h_d))$$

5.2 Video

An extension of VQA to the video domain includes TVQA, TVQA+ (Lei et al., 2020) built on TV shows, MovieQA (Tapaswi et al., 2016) based on movies, and Audio Visual Scene-Aware Dialog (AVSD) (Alamri et al., 2019) based on CHA-RADIES (Sigurdsson et al., 2016). DVD (Le et al., 2021) presents video-QA over videos synthesized from the CATER dataset (Girdhar and Ramanan, 2020). Besides visual reasoning, video-QA requires temporal reasoning, a challenge addressed by multimodal alignment that we discuss in the following section.

6 Multimodal Alignment

While image-based dialogue revolves around objects (e.g., cats and dogs), video-based dialogue revolves around objects and associated actions (e.g., jumping cats and barking dogs) where spatial and temporal features serve as building blocks for conversations. Extracting these spatiotemporal features requires multimodal alignment – aligning sub-components of different modalities to find correspondences. We identify action recognition and action from modalities as alignment challenges relevant to multimodal conversations.

6.1 Action Recognition

Action recognition is the task of extracting natural language descriptions from videos. UCF101 (Soomro et al., 2012), HMDB51 (Kuehne et al., 2011), and Kinetics-700 (Carreira et al., 2019) involve extracting actions from short YouTube and Hollywood movie clips. HowTo100M (Miech et al., 2019), MSR-VTT (Xu et al., 2016), and YouCook2 (Zhu et al., 2017) are datasets containing instructional videos on the internet and require learning text-video embeddings. YouCook2 and MSR-VTT are annotated by hand while HowTo100M uses existing video subtitles or ASR.

Mathematically, the goal is to retrieve the correct natural language description $y \in Y$ to a query video $x$ (Equation 11). Video and text representation functions $g(\cdot)_\text{video}$ and $g(\cdot)_\text{text}$ embed modalities into a coordinated space where they are com-
pared using a distance measure $d$.  

$$\arg\min_{y \in Y} d(g_{\text{video}}(x), g_{\text{text}}(y_j))$$ \hfill (11)

### 6.2 Action from Modalities

Equipping multimodal conversational agents with the ability to perform actions from multiple modalities provides them with an understanding of the real world, improving their conversational utility.

Talk the Walk (de Vries et al., 2018) presents the task of navigation conditioned on partial information. A “tourist” provides descriptions of a photo-realistic environment to a “guide” who determines actions. Vision-and-Dialog Navigation (Thomason et al., 2019) contains natural dialogues grounded in a simulated environment. The task is to predict a sequence of actions to a goal state given the world scene, dialogue, and previous actions. TEACh (Padmakumar et al., 2021) extends Vision-and-Dialog Navigation to complete tasks in an AI2-THOR simulation. The challenge involves aligning information from language, video, as well as action and dialogue history to solve daily tasks. Ego4D (Grauman et al., 2021) contains text annotated egocentric (first person) videos in real-world scenarios. Ego4D includes 3D scans, multiple camera views, and eye gaze, presenting new representation, fusion, translation, and alignment challenges. It is associated with five benchmarks: Video QA, object state tracking, audio-visual diarization, social cue detection, and camera trajectory forecasting.

### 6.3 Multimodal Versatile Networks (MMV)

In addition to a representation, Alayrac et al. (2020) presents a self-supervised task to train modality embedding graphs for multimodal alignment. Sampling temporally aligned audio, visual clips, and narrations from the same video creates positive training examples, while those from different videos comprise negative training examples. A Noise-Contrastive Estimation (NCE) loss (Gutmann and Hyvärinen, 2010) is minimized to ensure similarity between embeddings of positive training examples while forcing negative pairs further apart. A Multiple Instance Learning (MIL) (Miech et al., 2020) variant of NCE measures loss on pairs of modalities of different granularity. MIL accounts for misalignment between audio/video and text by measuring the loss of fine-grained information with multiple temporally close narrations.

The network is trained on HowTo100M (Miech et al., 2019) and AudioSet (Gemmeke et al., 2017). Table 3 compares the performance of MMV on action classification, audio classification, and zero-shot text-to-video retrieval.

### 7 Discussion

The current datasets used for research in multimodal conversational AI are summarized in Table 4. While MQA and MTQA are promising starting points for multimodal natural language tasks, extending QA to conversations is not straightforward. Inherently, MQA limits itself to direct questions targeting visible content, whereas multimodal conversations require understanding information that is often implied (Mostafazadeh et al., 2017). Utterances in dialogue represent speech acts and are classified as constatives, directives, commissives, or acknowledgments (Bach and Harnish, 1979). Answers belong to a single speech act (constatives) and represent a subset of natural conversations.

Similarly, the work to-date on action recognition is incomplete and insufficient for conversational systems. Conversational AI must represent and understand spatiotemporal interactions. However, current research in action recognition attempts to learn relationships between videos and their natural language descriptions. These descriptions are not speech acts themselves. Therefore, they do not adequately represent dialogue but rather only serve as anchor points in the interaction.

In contrast, Image-Chat (Shuster et al., 2020) presents a learning challenge directly aligned with the multimodal dialogue objective in Equation 4. Image-Chat treats dialogue as an open-ended discussion grounded in the visual modality. Succeeding in the task requires jointly optimizing visual and conversational performance. The use of crowd workers that adopt personalities during data collection encourages natural dialogue and captures conversational intricacies and implicatures.

MQA answers explicit questions about an image (Is this at a farm?), and action recognition describes videos (Mountain biking). On the other hand, Image-Chat requires both implicit knowledge (Halloween, Exercise) and multi-turn reasoning (Halloween → Holiday, Exercise → Fitness).

Despite its advantages over other datasets, Image-Chat makes three assumptions about multimodal conversations limiting its extension to the multimodal conversational objective:
Table 3: Comparison of learnt representations on UCF101, HMDB51, ESC-50, Kinetics600, YouCook2, and MSR-VTT. Top-1 Accuracy for UCF101, HMDB51, ESC-50, Kinetics600, mean Average Precision (mAP) for AudioSet, Recall@10 for YouCook2 and MSR-VTT (Alayrac et al., 2020).

| Dataset          | VQA (Antol et al., 2015) | Visual7W (Zhu et al., 2016) | Visual Slot (Heck and Heck, 2022) | TVQA (Lei et al., 2019) | MovQA (Tapaswi et al., 2016) | MANYMODALQA (Hamman et al., 2020) | MIMOQA (Singh et al., 2021b) | VisDial (Das et al., 2017) | Guesswhat (de Vries et al., 2017) | AVSD (Alamri et al., 2019) | DVD (Le et al., 2021) | SIMMC (Moon et al., 2020) | SIMMCC2.0 (Kotur et al., 2021) | IGC (Mostafazadeh et al., 2017) | Image-Chat (Shuster et al., 2020) | MOD (Fei et al., 2021) | UCF101 (Soomro et al., 2012) | HMDB51 (Kuehne et al., 2011) | Kinetics 700 (Carreira et al., 2019) | HowTo100M (Miech et al., 2019) | YouCook2 (Zhou et al., 2017) | MSR-VTT (Xu et al., 2016) | Talk the Walk (de Vries et al., 2018) | CViDN (Thomason et al., 2019) | TEACh (Padmakumar et al., 2021) | Ego4D (Grauman et al., 2021) |
|------------------|---------------------------|------------------------------|----------------------------------|--------------------------|-------------------------------|--------------------------------|--------------------------------|-----------------------------|--------------------------------|-------------------------|-----------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-------------------------------|-----------------------------|------------------------------|-------------------------------|-----------------------------|-----------------------------|-----------------------------|
| Model            | UCF101 (FT) | HMDB51 (FT) | ESC-50 (Linear) | AS | K600 | YC2 | MSR-VTT |
| MIL-NCE (3D-G)   | (Miech et al., 2020)  | 91.3 | 61.0 | / | / | / | 51.2 | 32.4 |
| AVTS (MC3)       | (Korbar et al., 2018)  | 89.0 | 61.6 | 80.6 | / | / | / | / |
| A+AV CC          | (Jansen et al., 2019)  | / | / | / | 28.5 | / | / | / |
| CVRL             | (Qian et al., 2021)    | / | / | / | / | 64.1 | / | / |
| XDC              | (Alwassel et al., 2020)| 91.2 | 61.0 | 84.8 | / | / | / | / |
| ELo              | (Piergiovanni et al., 2020) | 93.8 | 67.4 | / | / | / | / | / |
| AVID             | (Morgado et al., 2021) | 91.5 | 64.7 | 89.2 | / | / | / | / |
| GDT (IG65M)      | (Patrick et al., 2020) | 95.2 | 72.8 | 88.5 | / | / | / | / |
| MMV FAC (TSM-50k2) (Alayrac et al., 2020) | 95.2 | 75.0 | 88.9 | / | / | / | / |

Table 4: Datasets for multimodal representations. I=Image, V=Video, UI=User Interface, C=Caption, Q=Question, T=Text, H$_Q$=Question history, H$_A$=Answer history, H$_D$=Dialogue history, VR=Virtual Reality, D=Dialogue, A=Audio, S=Speech

1. Conversations are limited to three turns, devoid of long-term dialogue dependencies.
2. Language and images are the only modalities.
3. Personalities are independent of previous responses. This differs from natural human conversations where human tend to understand and reciprocate the personality of a dialogue partner (Rashkin et al., 2019).

The discussion above highlights the limitations of existing datasets for the multimodal conversational AI task. Datasets need to be improved to better capture and represent more natural, multi-turn dialogues over multiple modalities; dialogues that more closely resemble how humans converse with each other and their environment.

In addition, algorithmic improvements are required to advance the field of multimodal conversational AI - particularly with respect to the objective function. Current approaches such as MQA and action recognition models optimize a limited objective compared to Equation 4. We postulate that the degradation of these methods when applied to multimodal conversations is largely caused by this and, therefore, motivates investigation.

Another open research problem is to improve performance on Image-Chat. The current state-of-the-art TRANSRESNET$_{RET}$ is limited. The model often hallucinates, referring to content missing in the image and previous dialogue turns. The model also struggles when answering questions and holding extended conversations. We suspect these prob-
lems are a reflection of the limiting assumptions Image-Chat makes and the absence of multimodal co-learning to extract relationships between modalities. For further details, we refer readers to example conversations in Appendix A.

Different modalities often contain complementary information when grounded in the same concept. Multimodal co-learning exploits this cross-modality synergy to model resource-poor modalities using resource-rich modalities. An example of co-learning in context of Figure 1 is the use of visual information and audio to generate contextualized text representations.

Blum and Mitchell (1998) introduced an early approach to multimodal co-training, using information from hyperlinked pages for web-page classification. Socher and Fei-Fei (2010) and Duan et al. (2014) presented weakly-supervised techniques to tag images given information from other modalities. Kiela et al. (2015) grounded natural language descriptions in olfactory data. More recently, Upadhyay et al. (2018) jointly trains bilingual models to accelerate spoken language understanding in low resource languages. Selvaraju et al. (2019) uses human attention maps to teach QA agents “where to look”. Despite the rich history of work in multimodal co-learning, extending these techniques to develop multimodal conversational AI that understands and leverages cross-modal relationships is still an open challenge.

8 Conclusions

We define multimodal conversational AI and outline the objective function required for its realization. Solving this objective requires multimodal representation and fusion, translation, and alignment. We survey existing datasets and state-of-the-art methods for each sub-task. We identify simplifying assumptions made by existing research preventing the realization of multimodal conversational AI. Finally, we outline the collection of a suitable dataset and an approach that utilizes multimodal co-learning as future steps.

References


Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 5288–5296, Las Vegas, NV, USA. IEEE.


### Conversations with TRANSRESNET

<table>
<thead>
<tr>
<th>Image</th>
<th>Personality</th>
<th>Conversations</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image" /></td>
<td>Glamorous</td>
<td>The star of the show is the man and his fabulous car</td>
</tr>
<tr>
<td></td>
<td></td>
<td>I don’t see a man</td>
</tr>
<tr>
<td></td>
<td>Glamorous</td>
<td>That tends to be the case with men that buy flashy cars and BIG houses! [-]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>I would like to drive that car</td>
</tr>
<tr>
<td></td>
<td>Glamorous</td>
<td>That tends to be the case with men that buy flashy cars and BIG houses! [-]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Would you drive that car?</td>
</tr>
<tr>
<td></td>
<td>Glamorous</td>
<td>And I would pay him to do that in my backyard as well</td>
</tr>
<tr>
<td></td>
<td>Wise</td>
<td>I love going to new restaurants</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Well, easy come easy go with your food then, raw is dangerous but it’s your life</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Do you like spicy food?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>It is most likely too spicy for you</td>
</tr>
<tr>
<td><img src="image2.png" alt="Image" /></td>
<td>Businesslike</td>
<td>This is Lucy, my labrador</td>
</tr>
<tr>
<td></td>
<td></td>
<td>I should start a dog walking business [-]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>That’s a good idea!</td>
</tr>
<tr>
<td></td>
<td>Suave</td>
<td>He would make a lovely companion for me. [-]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Do you have any pets?</td>
</tr>
<tr>
<td></td>
<td>Considerate</td>
<td>I think I will. Dogs are a nice animal for the regular person to have</td>
</tr>
<tr>
<td><img src="image3.png" alt="Image" /></td>
<td>Caring</td>
<td>I hope you’re injured too. If not, I can help you with that [-]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>I’m not injured</td>
</tr>
<tr>
<td></td>
<td>Extreme</td>
<td>At least if you crash you’ll beat the paramedics to the scene by at least 20 minutes! [-]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Haha that’s a nice joke</td>
</tr>
<tr>
<td></td>
<td>Silly</td>
<td>I would die laughing if you did fall</td>
</tr>
<tr>
<td></td>
<td></td>
<td>That doesn’t look great</td>
</tr>
<tr>
<td></td>
<td>Angry</td>
<td>It makes me mad. Why did they have to hang it before the paint was dry? [-]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>I didn’t say it wasn’t artistic</td>
</tr>
<tr>
<td></td>
<td>Angry</td>
<td>Oh come on, you’re so negative! How is that NOT artistic?! [-]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Oh come on, you’re so negative! How is that NOT artistic?!</td>
</tr>
<tr>
<td></td>
<td>Angry</td>
<td>Oh come on, you’re so negative! How is that NOT artistic?! [-]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Why are you repeating yourself?</td>
</tr>
<tr>
<td></td>
<td>Angry</td>
<td>Because you have ruined the artistic integrity of the image. If you had the talent that I have, you would understand this. [-]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>I agree with you</td>
</tr>
<tr>
<td></td>
<td>Angry</td>
<td>Why are people so judgemental about art? It’s infuriating!</td>
</tr>
</tbody>
</table>

Conversations from TRANSRESNET. Images are chosen from YFCC100M using the browser by Kalkowski et al. (2015) and the internet, and personalities are chosen at random. 

[-] indicates lines spoken by the author.