Multimodal Conversational AI
A Survey of Datasets and Approaches

Anonymous ACL submission

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) 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 (Zhou et al., 2020; Adiwardana et al., 2020; Paranjape et al., 2020; Roller et al., 2020). 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 to convey specific meanings. 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.

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.

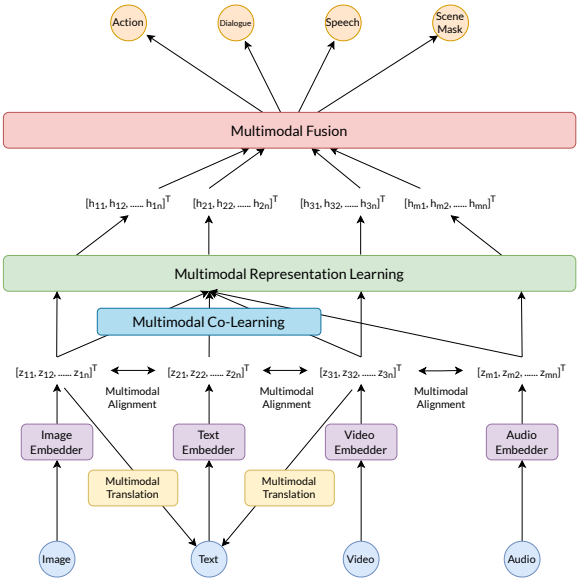


Figure 1: Taxonomy of research in multimodal Conversational AI

In this paper, we discuss the taxonomy of research challenges in Conversational AI as illustrated in Figure 1. We first motivate the need for multimodal conversational AI. 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

Research in multimodal conversational AI originated from the use of visual information to improve automatic speech recognition (ASR) (Yuhas et al., 1989; Meier et al., 1996; Duchnowski et al., 1994; Bregler and Konig, 1994; Ngiam et al., 2011).

Advances in client-side capabilities expanded ASR systems to more comprehensive interactive products such as ESPRIT's MASK (Lamel et al., 1998), Microsoft's MiPad (Huang et al., 2001), and AT&T's MATCH (Johnston et al., 2002) that parsed tactile, oral, and textual modalities for personalized tasks.

The need for systems with ocular skills motivated research in visual understanding. Deng et al. (2011) used semantic attributes to improve image retrieval. Ramanathan et al. (2014) grounded faces in video tracks to cast members. Kong et al. (2014) leveraged natural language descrptions 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 processing.

While modalities most commonly considered in the literature are text, visual signals like image/video, tactile information, and vocal signals, new sources of information such as eye-gaze, 3D scans, emotion, action and dialogue history, and virtual reality are gaining popularity. 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) present 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 $\mathbf{S}$ to a message m given underlying context c. Statistically, $\mathbf{S}$ is estimated as:

$$\mathbf{S} = \underset{r}{\operatorname{argmax}}\, p(r|c, m). \quad (1)$$

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).

$$p(r|c,m) = \prod_{t=1}^{T} p(r_t|r_1, \ldots r_{t-1}, c, m) \quad (2)$$

Conversational systems must evolve with each dialogue turn as the conversation progresses. To capture this dependence, we model the response as a function of completed dialogue turns $t-1$.

$$p(r|c,m,t-1) = \prod_{t=1}^{T} p(r_t|r_1, \ldots r_{t-1}, c, m, t-1) \quad (3)$$

For notational convenience, we consider the message m, turns $t-1$, and previous responses $r_1, r_2 \ldots, r_{t-1}$ as modalities $\{\mathbf{c}_i\}_{i=1}^{m}$.

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

$$\mathbf{S} = \underset{R}{\operatorname{argmax}}\, p(R|c,m) \quad (4)$$

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.

$$\mathbf{S} = \underset{R}{\operatorname{argmax}}\, p(R|f(\mathbf{c}_1, \mathbf{c}_2 \ldots \mathbf{c}_m)) \quad (5)$$

In the following sections, we describe each subtask 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 multimodal data 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 machine learning addresses. However, 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 Salakhutdinov, 2014), and sequential networks (Nicolaou et al., 2011). The success of the Transformer (Vaswani et al., 2017) when representing text and BERT (Devlin et al., 2019) when modeling language have inspired a variety of multimodal architectures for vision-and-language (Sun et al., 2019b; Lu et al., 2019; Gabeur et al., 2020; Chen et al., 2020b; Tan and Bansal, 2019; Singh et al., 2021a) and ASR (Baevski et al., 2020; Hsu et al., 2021; Chan et al., 2021).

Transformer-based models used as joint multimodal representations can be described as illustrated in the taxonomy of Figure 1. Modality specific encoders $\{j_i(\cdot)\}_{i=1}^{m}$ embed unimodal tokens $\{c_{i_k}\}_{k=1}^{n}$ to create latent features $\{z_{i_k}\}_{k=1}^{n}$ (Equation 6). Decoder networks use latent features to produce output symbols. A transformer $\Psi(\cdot)$ consists of stacked encoders and decoders with intra-modality attention. Attention heads compute relationships within elements of a modality, producing multimodal representations $\{h_{i_k}\}_{k=1}^{n}$ (Equation 7).

$$z_{i_1}, z_{i_2}, \ldots z_{i_n} = j_i(c_{i_1}, c_{i_2}, \ldots c_{i_n}) \quad (6)$$

$$h_{1_1} \ldots h_{m_n} = \Psi(z_{1_1}, z_{1_2}, \ldots z_{m_n}) \quad (7)$$

4.1.2 Coordinated Representations

In contrast, coordinated representations model each modality separately. Constraints coordinate representations of separate modalities by enforcing cross-modality 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 8). A notion of distance d between modalities in the coordinated space enables cross-modal retrieval.

$$d(g_a(\text{dog}), g_i(\text{dog})) < d(g_a(\text{dog}), g_i(\text{car})) \quad (8)$$

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., 2019; Shuster et al., 2020; Chen et al., 2020b; Geng et al., 2021; Xu et al., 2021b)

4.3 State-of-the-art Representation+Fusion Models for Conversational AI

Having introduced the multimodal representation and fusion challenges, we present the state-of-the-art in these sub-tasks for conversational AI.

4.3.1 Factor Graph Attention

Schwartz et al. (2020) develop 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 textual modalities using LSTMs. Nodes in the factor graph represent attention distributions over elements of each modality, and factors capture relationships between nodes.

There are two types of factors – local and joint. Local factors capture interactions between nodes of a single modality (e.g., words in the same sentence), while joint factors capture interactions between different modalities (e.g., a word in a sentence and an object in an image).

Representations from all modalities are concatenated via hybrid fusion and passed through a multi-layer perceptron network to retrieve the best candidate answer.

Table 1 compares the Recall-at-k (R@k) of discriminative models on VisDial v1.0 test-std. The F-RCNN version of FGA is the state-of-the-art.

Model	R@1	R@5	R@10
LF (Das et al., 2017)	40.95	72.45	82.83
HRE (Das et al., 2017)	39.93	70.45	81.50
Memory Network (Das et al., 2017)	40.98	72.30	83.30
CorefNMN (ResNet-152) (Kottur et al., 2018)	47.55	78.10	88.80
NMN (ResNet-152) (Hu et al., 2017)	44.15	76.88	86.88
FGA (F-RCNNx101) (Schwartz et al., 2020)	**52.75**	**82.92**	**91.07**

Table 1: Comparison of models on VisDial v1.0 test-std (Recall@k)(Schwartz et al., 2020)

4.3.2 TRANSRESNET

Shuster et al. (2020) present TRANSRESNET for image-based dialogue. Image-based dialogue is the task of choosing the optimal response on a dialogue turn given an image, an agent personality, and dialogue history. TRANSRESNET consists of separately learned sub-networks to represent input modalities. Images are encoded using ResNeXt 32×48d trained on 3.5 billion Instagram images (Xie et al., 2017), personalities are embedded using a linear layer, and dialogue is encoded by a transformer pretrained on Reddit (Mazaré et al., 2018) to create a joint representation.

TRANSRESNET compares model-agnostic and model-based fusion by using either concatenation or attention networks to combine representations. Like FGA, the chosen dialogue response is the candidate closest to the fused representation.

On the first turn, TRANSRESNET uses only style and image information to produce responses. Dialogue history serves as an additional modality on subsequent rounds. Ablation of one or more modalities diminishes the ability of the model to retrieve the correct response. Optimal performance on Image-Chat (Shuster et al., 2020) is achieved using multimodal concatenation of jointly represented modalities (Table 2).

Modalities	Turn 1	Turn 2	Turn 3	All
Image Only	37.6	28.1	20.7	28.7
Style Only	18.3	15.3	17.0	16.9
Dialogue History Only	1.0	33.7	32.3	22.3
Style + Dialogue	18.3	45.4	43.1	35.4
Image + Dialogue	37.6	39.4	32.6	36.5
Image + Style	**54.0**	41.1	35.2	43.4
Style + Dialogue + Image	**54.0**	**51.9**	**44.8**	**50.3**

Table 2: Recall@1 (%) on Image-Chat using TRANSRESNET$_{\text{RET}}$ (ResNeXt-IG-3.5B, MM-Sum)

4.3.3 MultiModal Versatile Networks (MMV)

Alayrac et al. (2020) present 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).

There are 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-modality reasoning enables multimodal conversational AI to hold meaningful conversations 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.

The objective of MQA is a simplification of Equation 5 to a single-turn, single-timestep scenario ($t = 1$), producing a response to a question c_q given multimodal information $\{c_i\}_{i=1}^{m}$:

$$\mathbf{S}_{MQA} = \underset{R}{\operatorname{argmax}}\, p(R|f(c_1, \ldots c_m, c_q, 1)) \quad (9)$$

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 10) is an extension of MQA to include QA history $\mathbf{c}_{qa} = \{c_{q_1}, c_{a_1}, \ldots, c_{q_{t-1}}, c_{a_{t-1}}\}$. Note that

MTQA is a turn-agnostic objective, which we denote using the single-timestep notation $t = 1$.

$$\mathbf{S}_{MTQA} = \underset{R}{\operatorname{argmax}}\, p(R|f(c_1, \ldots, \mathbf{c}_{qa}, 1)) \quad (10)$$

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). SIMMC2.0 (Kottur et al., 2021) presents goal-oriented dialogue for shopping, leveraging dialogue and a state of the world to resolve references, track dialogue state, and recommend the correct object. IGC, Image-Chat, and SIMMC2.0 solve the MD objective:

$$\mathbf{S}_{MD} = \underset{R}{\operatorname{argmax}}\, p(R|f(c_1, \ldots c_m, c_d, t)) \quad (11)$$

In contrast to MTQA, MD depends on previous dialogue responses (c_d) and represents systems that evolve with dialogue turns t.

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 CHARADES (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-language 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 (Zhou 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 $\mathbf{y} \in \mathcal{Y}$ to a query video $\mathbf{x}$ (Equation 12). Video and text representation functions $g(\cdot)_{\text{video}}$ and $g(\cdot)_{\text{text}}$ embed modalities into a coordinated space where they are compared using a distance measure d.

$$\underset{y \in \mathcal{Y}}{\operatorname{argmin}}\, d(g_{\text{video}}(\mathbf{x}), g_{\text{text}}(\mathbf{y}_j)) \quad (12)$$

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.

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) present a self-supervised task to train modality embedding graphs for multimodal alignment. Sampling temporally aligned audio, visual clips,

Model	UCF101 (FT)	HMDB51 (FT)	ESC-50 (Linear)	AS	K600	YC2	MSR-VTT
MIL-NCE (S3D-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	/	/	/	/
AA+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-50x2) (Alayrac et al., 2020)	**95.2**	**75.0**	88.9	**30.9**	**70.5**	45.4	31.1

Table 3: Comparison of learnt representations on UCF101, HMDB51, ESC-50, AudioSet, 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).

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 and 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 5. 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:

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 humans 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

Dataset	Modalities	Task	Data Collection	POV
VQA (Antol et al., 2015)	I,Q	Question Answering	Human-Human	Third Person
Visual7W (Zhu et al., 2016)	I,Q	Question Answering	Human-Human	Third Person
TVQA (Lei et al., 2019)	V,Q,	Question Answering	Human-Human	Third Person
MovieQA (Tapaswi et al., 2016)	V,C,Q,T	Question Answering	Human	Third Person
MANYMODALQA (Hannan et al., 2020)	I,C,Q,T,Tables	Question Answering	Human-Human	Third Person
MIMOQA (Singh et al., 2021b)	I,Q,T	Question Answering	Machine	X
VisDial (Das et al., 2017)	I,H_Q,H_A,C,Q	Question Answering	Human-Human	Third person
Guesswhat (de Vries et al., 2017)	I, H_Q,H_A	Question Answering	Human-Human	Third Person
AVSD (Alamri et al., 2019)	V,A,H_Q,H_A,C,Q	Question Answering	Human-Human	Third Person
DVD (Le et al., 2021)	V,Q,H_Q,H_A	Question Answering	Machine	X
SIMMC2.0 (Kottur et al., 2021)	H_D, Q, VR	Shopping	Machine Self-play	First Person
IGC (Mostafazadeh et al., 2017)	I,Q,D	Chit-chat + Question Answering	Human-Human	Third-Person
Image-Chat (Shuster et al., 2020)	I,D,Personality	Chit-chat	Human-Human	Third Person
UCF101 (Soomro et al., 2012)	V,A (partial)	Action Recognition	YouTube	Third Person
HMDB51 (Kuehne et al., 2011)	V	Action Recognition	YouTube+Movies	Third Person
Kinetics 700 (Carreira et al., 2019)	V	Action Recognition	YouTube	Third Person
HowTo100M (Miech et al., 2019)	V,T	Text-Video Embeddings	YouTube	First+Third Person
YouCook2 (Zhou et al., 2017)	V,T	Text-video retrieval, activity recognition	YouTube	Third Person
MSRVTT (Xu et al., 2016)	V,A,T	Video-to-text	Web videos	First+Third Person
CVDN (Thomason et al., 2019)	VR, Actions, H_Q,H_A	Navigation from Dialog History	Human-Human	First Person
TEACh (Padmakumar et al., 2021)	Scene, Actions, D	Action prediction, Task from language	AI2-THOR	First+Third Person
Ego4D (Grauman et al., 2021)	V,T,A,Gaze,3D Scan	Spatial Reasoning	Human	First Person

Table 4: Datasets for multimodal representations. I=Image, V=Video, 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

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 5. 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$_{\text{RET}}$ 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 problems 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 train bilingual models to accelerate spoken language understanding in low resource languages. Selvaraju et al. (2019) use 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

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a Human-like Open-Domain Chatbot. *arXiv:2001.09977 [cs, stat]*. ArXiv: 2001.09977.

Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. 2021. VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text. *arXiv:2104.11178 [cs, eess]*. ArXiv: 2104.11178.

Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K. Marks, Chiori Hori, Peter Anderson, Stefan Lee, and Devi Parikh. 2019. Audio-Visual Scene-Aware Dialog. *arXiv:1901.09107 [cs]*. ArXiv: 1901.09107.

Jean-Baptiste Alayrac, Adrià Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. 2020. Self-Supervised MultiModal Versatile Networks. *arXiv:2006.16228 [cs]*. ArXiv: 2006.16228.

Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. 2020. Self-Supervised Learning by Cross-Modal Audio-Video Clustering. *arXiv:1911.12667 [cs]*. ArXiv: 1911.12667.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433, Santiago, Chile. IEEE.

Kent Bach and Robert M. Harnish. 1979. *Linguistic Communication and Speech Acts*. Cambridge: MIT Press.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *arXiv:2006.11477 [cs, eess]*. ArXiv: 2006.11477.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2017. Multimodal Machine Learning: A Survey and Taxonomy. *arXiv:1705.09406 [cs]*. ArXiv: 1705.09406.

Suzanna Becker and Geoffrey E. Hinton. 1992. Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355:161–163.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory - COLT' 98*, pages 92–100, Madison, Wisconsin, United States. ACM Press.

C. Bregler and Y. Konig. 1994. "Eigenlips" for robust speech recognition. In *Proceedings of ICASSP '94. IEEE International Conference on Acoustics, Speech and Signal Processing*, volume ii, pages II/669–II/672 vol.2.

Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. 2019. A Short Note on the Kinetics-700 Human Action Dataset. *arXiv:1907.06987 [cs]*. ArXiv: 1907.06987.

David M. Chan, Shalini Ghosh, Debmalya Chakrabarty, and Björn Hoffmeister. 2021. Multi-Modal Pre-Training for Automated Speech Recognition. *arXiv:2110.09890 [cs, eess]*. ArXiv: 2110.09890.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020a. A Simple Framework for Contrastive Learning of Visual Representations. *arXiv:2002.05709 [cs, stat]*. ArXiv: 2002.05709.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020b. UNITER: UNiversal Image-TExt Representation Learning. *arXiv:1909.11740 [cs]*. ArXiv: 1909.11740.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual Dialog. *arXiv:1611.08669 [cs]*. ArXiv: 1611.08669.

Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. GuessWhat?! Visual object discovery through multi-modal dialogue. *arXiv:1611.08481 [cs]*. ArXiv: 1611.08481.

Jia Deng, Alexander C. Berg, and Li Fei-Fei. 2011. Hierarchical semantic indexing for large scale image retrieval. In *CVPR 2011*, pages 785–792. ISSN: 1063-6919.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*. ArXiv: 1810.04805.

Kun Duan, David J. Crandall, and Dhruv Batra. 2014. Multimodal Learning in Loosely-Organized Web Images. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2465–2472, Columbus, OH, USA. IEEE.

P. Duchnowski, Uwe Meier, and Alex Waibel. 1994. See Me, Hear Me: Integrating Automatic Speech Recognition and Lip-reading. In *Proceedings of 3rd International Conference on Spoken Language Processing (ICSLP '94)*, pages 547 – 550.

Ali El-Kahky, Xiaohu Liu, Ruhi Sarikaya, Gokhan Tur, Dilek Hakkani-Tur, and Larry Heck. 2014. Extending domain coverage of language understanding systems via intent transfer between domains using knowledge graphs and search query click logs. In

2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 4067–4071. IEEE.

Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D. Manning. 2017. Key-Value Retrieval Networks for Task-Oriented Dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 37–49, Saarbrücken, Germany. Association for Computational Linguistics.

Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. 2020. Multi-modal Transformer for Video Retrieval. *arXiv:2007.10639 [cs]*. ArXiv: 2007.10639.

Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. 2015. Are You Talking to a Machine? Dataset and Methods for Multilingual Image Question Answering. *arXiv:1505.05612 [cs]*. ArXiv: 1505.05612.

Jianfeng Gao, Michel Galley, and Lihong Li. 2019. Neural Approaches to Conversational AI. *arXiv:1809.08267 [cs]*. ArXiv: 1809.08267.

Ruohan Gao, Yen-Yu Chang, Shivani Mall, Li Fei-Fei, and Jiajun Wu. 2021. ObjectFolder: A Dataset of Objects with Implicit Visual, Auditory, and Tactile Representations. *arXiv:2109.07991 [cs]*. ArXiv: 2109.07991.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio Set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, New Orleans, LA.

Shijie Geng, Peng Gao, Moitreya Chatterjee, Chiori Hori, Jonathan Le Roux, Yongfeng Zhang, Hongsheng Li, and Anoop Cherian. 2021. Dynamic Graph Representation Learning for Video Dialog via Multi-Modal Shuffled Transformers. *arXiv:2007.03848 [cs]*. ArXiv: 2007.03848.

Rohit Girdhar and Deva Ramanan. 2020. Cater: A diagnostic dataset for compositional actions and temporal reasoning. In *International Conference on Learning Representations*.

Palash Goyal, Saurabh Sahu, Shalini Ghosh, and Chul Lee. 2020. Cross-modal Learning for Multi-modal Video Categorization. *arXiv:2003.03501 [cs]*. ArXiv: 2003.03501.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Christian Fuegen, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. 2021. Ego4D: Around the World in 3,000 Hours of Egocentric Video. *arXiv:2110.07058 [cs]*. ArXiv: 2110.07058.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. 2020. Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284. Curran Associates, Inc.

Mihai Gurban, Jean-Philippe Thiran, Thomas Drugman, and Thierry Dutoit. 2008. Dynamic modality weighting for multi-stream hmms inaudio-visual speech recognition. In *Proceedings of the 10th international conference on Multimodal interfaces - IMCI '08*, page 237, Chania, Crete, Greece. ACM Press.

Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 297–304, Chia Laguna Resort, Sardinia, Italy. PMLR.

Dilek Hakkani-Tür, Larry Heck, and Gokhan Tur. 2011. Exploiting query click logs for utterance domain detection in spoken language understanding. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5636–5639. IEEE.

Dilek Hakkani-Tür, Larry Heck, and Gokhan Tur. 2013. Using a knowledge graph and query click logs for unsupervised learning of relation detection. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8327–8331. IEEE.

Dilek Hakkani-Tür, Gokhan Tur, Rukmini Iyer, and Larry Heck. 2012. Translating natural language utterances to search queries for slu domain detection using

query click logs. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4953–4956. IEEE.

Dilek Hakkani-Tür, Malcolm Slaney, Asli Celikyilmaz, and Larry Heck. 2014. Eye Gaze for Spoken Language Understanding in Multi-Modal Conversational Interactions. In *Proceedings of the 16th International Conference on Multimodal Interaction*, ICMI '14, pages 263–266, New York, NY, USA. Association for Computing Machinery. Event-place: Istanbul, Turkey.

Dilek Hakkani-Tür, Gokhan Tur, and Larry Heck. 2011. Research Challenges and Opportunities in Mobile Applications [DSP Education]. *IEEE Signal Processing Magazine*, 28(4):108–110. Conference Name: IEEE Signal Processing Magazine.

Darryl Hannan, Akshay Jain, and Mohit Bansal. 2020. ManyModalQA: Modality Disambiguation and QA over Diverse Inputs. *arXiv:2001.08034 [cs]*. ArXiv: 2001.08034.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, Seattle, WA, USA. IEEE.

Larry Heck. 2012. The conversational web. In *Keynote, IEEE Workshop on Spoken Language Technology*, Miami, FL, USA.

Larry Heck and Dilek Hakkani-Tür. 2012. Exploiting the semantic web for unsupervised spoken language understanding. In *2012 IEEE Spoken Language Technology Workshop (SLT)*, pages 228–233. IEEE.

Larry Heck, Dilek Hakkani-Tür, Madhu Chinthakunta, Gokhan Tur, Rukmini Iyer, Partha Parthasacarthy, Lisa Stifelman, Elizabeth Shriberg, and Ashley Fidler. 2013. Multimodal Conversational Search and Browse. IEEE Workshop on Speech, Language and Audio in Multimedia.

Larry Heck and Hongzhao Huang. 2014. Deep learning of knowledge graph embeddings for semantic parsing of twitter dialogs. In *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 597–601. IEEE.

Chiori Hori, Takaaki Hori, Teng-Yok Lee, Kazuhiro Sumi, John R. Hershey, and Tim K. Marks. 2017. Attention-Based Multimodal Fusion for Video Description. *arXiv:1701.03126 [cs]*. ArXiv: 1701.03126.

Wei-Ning Hsu, Yao-Hung Hubert Tsai, Benjamin Bolte, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: How Much Can a Bad Teacher Benefit ASR Pre-Training? In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6533–6537. ISSN: 2379-190X.

Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. 2017. Learning to Reason: End-to-End Module Networks for Visual Question Answering. _eprint: 1704.05526.

X. Huang, A. Acero, C. Chelba, L. Deng, J. Droppo, D. Duchene, J. Goodman, H. Hon, D. Jacoby, L. Jiang, R. Loynd, M. Mahajan, P. Mau, S. Meredith, S. Mughal, S. Neto, M. Plumpe, K. Steury, G. Venolia, K. Wang, and Y. Wang. 2001. MiPad: a multimodal interaction prototype. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, volume 1, pages 9–12 vol.1.

Aren Jansen, Daniel P. W. Ellis, Shawn Hershey, R. Channing Moore, Manoj Plakal, Ashok C. Popat, and Rif A. Saurous. 2019. Coincidence, Categorization, and Consolidation: Learning to Recognize Sounds with Minimal Supervision. *arXiv:1911.05894 [cs, eess, stat]*. ArXiv: 1911.05894.

Michael Johnston, Srinivas Bangalore, Gunaranjan Vasireddy, Amanda Stent, Patrick Ehlen, Marilyn Walker, Steve Whittaker, and Preetam Maloor. 2002. MATCH: An Architecture for Multimodal Dialogue Systems. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 376–383, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Sebastian Kalkowski, Christian Schulze, Andreas Dengel, and Damian Borth. 2015. Real-time Analysis and Visualization of the YFCC100m Dataset. In *Proceedings of the 2015 Workshop on Community-Organized Multimodal Mining: Opportunities for Novel Solutions - MMCommons'15*, pages 25–30, Brisbane, Australia. ACM Press.

Douwe Kiela, Luana Bulat, and Stephen Clark. 2015. Grounding Semantics in Olfactory Perception. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 231–236, Beijing, China. Association for Computational Linguistics.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2014. Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. *arXiv:1411.2539 [cs]*. ArXiv: 1411.2539.

Chen Kong, Dahua Lin, Mohit Bansal, Raquel Urtasun, and Sanja Fidler. 2014. What Are You Talking About? Text-to-Image Coreference. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3558–3565. ISSN: 1063-6919.

Bruno Korbar, Du Tran, and Lorenzo Torresani. 2018. Cooperative Learning of Audio and Video Models from Self-Supervised Synchronization. *arXiv:1807.00230 [cs]*. ArXiv: 1807.00230.

Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. 2021. SIMMC 2.0: A Task-oriented Dialog Dataset for Immersive Multimodal Conversations. *arXiv:2104.08667 [cs]*. ArXiv: 2104.08667.

Satwik Kottur, José M. F. Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. 2018. Visual Coreference Resolution in Visual Dialog using Neural Module Networks. *arXiv:1809.01816 [cs]*. ArXiv: 1809.01816.

H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. 2011. HMDB: A large video database for human motion recognition. In *2011 International Conference on Computer Vision*, pages 2556–2563. ISSN: 2380-7504.

Lori Lamel, Samir Bennacef, Jean-Luc Gauvain, Hervé Dartigues, and Jean-Noel Temem. 1998. User Evaluation Of The Mask Kiosk.

Hung Le, Chinnadhurai Sankar, Seungwhan Moon, Ahmad Beirami, Alborz Geramifard, and Satwik Kottur. 2021. DVD: A Diagnostic Dataset for Multi-step Reasoning in Video Grounded Dialogue. *arXiv:2101.00151 [cs]*. ArXiv: 2101.00151.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L. Berg. 2019. TVQA: Localized, Compositional Video Question Answering. *arXiv:1809.01696 [cs]*. ArXiv: 1809.01696.

Jie Lei, Licheng Yu, Tamara L. Berg, and Mohit Bansal. 2020. TVQA+: Spatio-Temporal Grounding for Video Question Answering. *arXiv:1904.11574 [cs]*. ArXiv: 1904.11574.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A Diversity-Promoting Objective Function for Neural Conversation Models. *arXiv:1510.03055 [cs]*. ArXiv: 1510.03055.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P. Spithourakis, Jianfeng Gao, and Bill Dolan. 2016b. A Persona-Based Neural Conversation Model. *arXiv:1603.06155 [cs]*. ArXiv: 1603.06155.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In *Computer Vision – ECCV 2020*, pages 121–137, Cham. Springer International Publishing.

Ji Lin, Chuang Gan, and Song Han. 2019. TSM: Temporal Shift Module for Efficient Video Understanding. *arXiv:1811.08383 [cs]*. ArXiv: 1811.08383.

Bing Liu, Gokhan Tur, Dilek Hakkani-Tur, Pararth Shah, and Larry Heck. 2017. End-to-end optimization of task-oriented dialogue model with deep reinforcement learning. *arXiv preprint arXiv:1711.10712*.

Bing Liu, Gokhan Tur, Dilek Hakkani-Tur, Pararth Shah, and Larry Heck. 2018. Dialogue learning with human teaching and feedback in end-to-end trainable task-oriented dialogue systems.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. *arXiv:1908.02265 [cs]*. ArXiv: 1908.02265.

Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. 2015. Ask Your Neurons: A Neural-based Approach to Answering Questions about Images. *arXiv:1505.01121 [cs]*. ArXiv: 1505.01121.

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training Millions of Personalized Dialogue Agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2775–2779, Brussels, Belgium. Association for Computational Linguistics.

U. Meier, W. Hurst, and P. Duchnowski. 1996. Adaptive bimodal sensor fusion for automatic speechreading. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, volume 2, pages 833–836 vol. 2.

Grégoire Mesnil, Yann Dauphin, Kaisheng Yao, Yoshua Bengio, Li Deng, Dilek Hakkani-Tur, Xiaodong He, Larry Heck, Gokhan Tur, Dong Yu, et al. 2014. Using recurrent neural networks for slot filling in spoken language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(3):530–539.

Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2020. End-to-End Learning of Visual Representations from Uncurated Instructional Videos. *arXiv:1912.06430 [cs]*. ArXiv: 1912.06430.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. *arXiv:1906.03327 [cs]*. ArXiv: 1906.03327.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*. ArXiv: 1301.3781.

Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. 2021. Audio-Visual Instance Discrimination with Cross-Modal Agreement. *arXiv:2004.12943 [cs]*. ArXiv: 2004.12943.

Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios P. Spithourakis, and Lucy Vanderwende. 2017. Image-Grounded Conversations: Multimodal Context for Natural Question and Response Generation. *arXiv:1701.08251 [cs]*. ArXiv: 1701.08251.

Ara V. Nefian, Luhong Liang, Xiaobo Pi, Liu Xiaoxiang, Crusoe Mao, and Kevin Murphy. 2002. A coupled HMM for audio-visual speech recognition. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages II–2013–II–2016. ISSN: 1520-6149.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. 2011. Multimodal Deep Learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pages 689–696, Madison, WI, USA. Omnipress. Event-place: Bellevue, Washington, USA.

Mihalis A. Nicolaou, Hatice Gunes, and Maja Pantic. 2011. Continuous Prediction of Spontaneous Affect from Multiple Cues and Modalities in Valence-Arousal Space. *IEEE Transactions on Affective Computing*, 2(2):92–105. Conference Name: IEEE Transactions on Affective Computing.

Andrew Owens, Jiajun Wu, Josh H. McDermott, William T. Freeman, and Antonio Torralba. 2016. Ambient Sound Provides Supervision for Visual Learning. *arXiv:1608.07017 [cs]*. ArXiv: 1608.07017.

Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramithu, Gokhan Tur, and Dilek Hakkani-Tur. 2021. TEACh: Task-driven Embodied Agents that Chat. *arXiv:2110.00534 [cs]*. ArXiv: 2110.00534.

Ashwin Paranjape, Abigail See, Kathleen Kenealy, Haojun Li, Amelia Hardy, Peng Qi, Kaushik Ram Sadagopan, Nguyet Minh Phu, Dilara Soylu, and Christopher D. Manning. 2020. Neural Generation Meets Real People: Towards Emotionally Engaging Mixed-Initiative Conversations. *arXiv:2008.12348 [cs]*. ArXiv: 2008.12348.

Mandela Patrick, Yuki Markus Asano, Ruth Fong, João F. Henriques, Geoffrey Zweig, and Andrea Vedaldi. 2020. Multi-modal self-supervision from generalized data transformations. *CoRR*, abs/2003.04298.

Gao Peng, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven Hoi, Xiaogang Wang, and Hongsheng Li. 2019. Dynamic Fusion with Intra- and Inter- Modality Attention Flow for Visual Question Answering. *arXiv:1812.05252 [cs, eess]*. ArXiv: 1812.05252.

A. J. Piergiovanni, Anelia Angelova, and Michael S. Ryoo. 2020. Evolving Losses for Unsupervised Video Representation Learning. *arXiv:2002.12177 [cs]*. ArXiv: 2002.12177.

G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.W. Senior. 2003. Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the IEEE*, 91(9):1306–1326. Conference Name: Proceedings of the IEEE.

Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. 2021. Spatiotemporal Contrastive Video Representation Learning. *arXiv:2008.03800 [cs]*. ArXiv: 2008.03800.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. *arXiv:2103.00020 [cs]*. ArXiv: 2103.00020.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv:1910.10683 [cs, stat]*. ArXiv: 1910.10683.

Vignesh Ramanathan, Armand Joulin, Percy Liang, and Li Fei-Fei. 2014. Linking People in Videos with "Their" Names Using Coreference Resolution. In *Computer Vision – ECCV 2014*, pages 95–110, Cham. Springer International Publishing.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *arXiv:1506.01497 [cs]*. ArXiv: 1506.01497.

Alan Ritter, Colin Cherry, and William B. Dolan. 2011. Data-Driven Response Generation in Social Media. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 583–593, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Stephen Roller, Y.-Lan Boureau, Jason Weston, Antoine Bordes, Emily Dinan, Angela Fan, David Gunning, Da Ju, Margaret Li, Spencer Poff, Pratik Ringshia, Kurt Shuster, Eric Michael Smith, Arthur Szlam, Jack Urbanek, and Mary Williamson. 2020. Open-Domain Conversational Agents: Current Progress, Open Problems, and Future Directions. *arXiv:2006.12442 [cs]*. ArXiv: 2006.12442.

Idan Schwartz, Seunghak Yu, Tamir Hazan, and Alexander Schwing. 2020. Factor Graph Attention. *arXiv:1904.05880 [cs]*. ArXiv: 1904.05880.

Ramprasaath Ramasamy Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Shalini Ghosh, Larry Heck, Dhruv Batra, and Devi Parikh. 2019. Taking a HINT: Leveraging Explanations to Make Vision and Language Models More Grounded. In *2019*

IEEE/CVF International Conference on Computer Vision (ICCV), pages 2591–2600, Seoul, Korea (South). IEEE.

Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models. *arXiv:1507.04808 [cs]*. ArXiv: 1507.04808.

Pararth Shah, Dilek Hakkani-Tur, and Larry Heck. 2016. Interactive reinforcement learning for task-oriented dialogue management.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural Responding Machine for Short-Text Conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1577–1586, Beijing, China. Association for Computational Linguistics.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2020. ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. *arXiv:1912.01734 [cs]*. ArXiv: 1912.01734.

Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. 2020. Image Chat: Engaging Grounded Conversations. *arXiv:1811.00945 [cs]*. ArXiv: 1811.00945.

Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. *arXiv:1604.01753 [cs]*. ArXiv: 1604.01753.

Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556 [cs]*. ArXiv: 1409.1556.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2021a. FLAVA: A Foundational Language And Vision Alignment Model. *arXiv:2112.04482 [cs]*. ArXiv: 2112.04482.

Hrituraj Singh, Anshul Nasery, Denil Mehta, Aishwarya Agarwal, Jatin Lamba, and Balaji Vasan Srinivasan. 2021b. MIMOQA: Multimodal Input Multimodal Output Question Answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5317–5332, Online. Association for Computational Linguistics.

Richard Socher and Li Fei-Fei. 2010. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 966–973. ISSN: 1063-6919.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. *arXiv:1212.0402 [cs]*. ArXiv: 1212.0402.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A Neural Network Approach to Context-Sensitive Generation of Conversational Responses. *arXiv:1506.06714 [cs]*. ArXiv: 1506.06714.

Nitish Srivastava and Ruslan Salakhutdinov. 2014. Multimodal Learning with Deep Boltzmann Machines. *Journal of Machine Learning Research*, 15(84):2949–2980.

Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. 2019a. Learning Video Representations using Contrastive Bidirectional Transformer. *arXiv:1906.05743 [cs, stat]*. ArXiv: 1906.05743.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019b. VideoBERT: A Joint Model for Video and Language Representation Learning. *arXiv:1904.01766 [cs]*. ArXiv: 1904.01766.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. MovieQA: Understanding Stories in Movies through Question-Answering. *arXiv:1512.02902 [cs]*. ArXiv: 1512.02902.

Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. 2019. Vision-and-Dialog Navigation. *arXiv:1907.04957 [cs]*. ArXiv: 1907.04957.

Shyam Upadhyay, Manaal Faruqui, Gokhan Tür, Hakkani-Tür Dilek, and Larry Heck. 2018. (Almost) Zero-Shot Cross-Lingual Spoken Language Understanding. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6034–6038. ISSN: 2379-190X.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Oriol Vinyals and Quoc Le. 2015. A Neural Conversational Model. *arXiv:1506.05869 [cs]*. ArXiv: 1506.05869.

Joseph Weizenbaum. 1966. ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Jason Weston, Samy Bengio, and Nicolas Usunier. 2010. Large Scale Image Annotation: Learning to Rank with Joint Word-Image Embeddings. In *European Conference on Machine Learning*.

Zhiyong Wu, Lianhong Cai, and Helen Meng. 2005. Multi-level Fusion of Audio and Visual Features for Speaker Identification. In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Dough Tygar, Moshe Y. Vardi, Gerhard Weikum, David Zhang, and Anil K. Jain, editors, *Advances in Biometrics*, volume 3832, pages 493–499. Springer Berlin Heidelberg, Berlin, Heidelberg. Series Title: Lecture Notes in Computer Science.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated Residual Transformations for Deep Neural Networks. *arXiv:1611.05431 [cs]*. ArXiv: 1611.05431.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. 2021a. VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding. *arXiv:2109.14084 [cs]*. ArXiv: 2109.14084.

Huijuan Xu and Kate Saenko. 2016. Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering. *arXiv:1511.05234 [cs]*. ArXiv: 1511.05234.

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.

Nancy Xu, Sam Masling, Michael Du, Giovanni Campagna, Larry Heck, James Landay, and Monica S. Lam. 2021b. Grounding Open-Domain Instructions to Automate Web Support Tasks. *arXiv:2103.16057 [cs]*. ArXiv: 2103.16057.

Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Joshua B. Tenenbaum. 2019. Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding. *arXiv:1810.02338 [cs]*. ArXiv: 1810.02338.

B.P. Yuhas, M.H. Goldstein, and T.J. Sejnowski. 1989. Integration of acoustic and visual speech signals using neural networks. *IEEE Communications Magazine*, 27(11):65–71. Conference Name: IEEE Communications Magazine.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2021. Big Bird: Transformers for Longer Sequences. *arXiv:2007.14062 [cs, stat]*. ArXiv: 2007.14062.

Heming Zhang, Shalini Ghosh, Larry Heck, Stephen Walsh, Junting Zhang, Jie Zhang, and C.-C. Jay Kuo. 2019. Generative Visual Dialogue System via Adaptive Reasoning and Weighted Likelihood Estimation. *arXiv:1902.09818 [cs]*. ArXiv: 1902.09818.

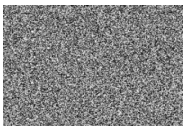
Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. 2020. Contrastive Learning of Medical Visual Representations from Paired Images and Text. *arXiv:2010.00747 [cs]*. ArXiv: 2010.00747.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The Design and Implementation of XiaoIce, an Empathetic Social Chatbot. *Computational Linguistics*, 46(1):53–93.

Luowei Zhou, Chenliang Xu, and Jason J. Corso. 2017. Towards Automatic Learning of Procedures from Web Instructional Videos. *arXiv:1703.09788 [cs]*. ArXiv: 1703.09788.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7W: Grounded Question Answering in Images. *arXiv:1511.03416 [cs]*. ArXiv: 1511.03416.

A Conversations with TransResNet

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

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.