Multimodal Persona Based Generation of Comic Dialogs

Harsh Agrawal
IIT Delhi
harsh.ag14901@gmail.com

Aditya M. Mishra
IIT Delhi
mishramohanaditya@gmail.com

Manish Gupta
Microsoft
gmanish@microsoft.com

Mausam
IIT Delhi
mausam@cse.iitd.ac.in

Abstract

We focus on the novel problem of persona based dialogue generation for comic strips. Dialogs in comic strips is a unique and unexplored area where every strip contains utterances from various characters with each one building upon the previous utterances and the associated visual scene. Previous works like DialoGPT, PersonaGPT and other dialog generation models encode two-party dialogues and do not account for the visual information. To the best of our knowledge we are the first to propose the paradigm of multimodal persona based dialogue generation.

We contribute a novel dataset, COMSET, consisting of 54K strips, harvested from 13 popular comics available online. Further, we propose a multimodal persona-based architecture, MPDIALOG, to generate dialogues for the next panel in the strip which decreases the perplexity score by ∼10 points over strong dialogue generation baseline models. We demonstrate that there is still ample opportunity for improvement, highlighting the importance of building stronger dialogue systems that are able to generate persona-consistent dialogues and understand the context through various modalities.

1 Introduction

Multimodal conversational agents build dialog systems that engage with modalities beyond text, in constructing next responses. They open up a novel direction of text-vision multimodality, where the agent is part of the scene, rather than being a distant observer. This facilitates research and creation of support based multimodal agents. These agents could be critical for various applications such as assistants for visually impaired, conversations with robots in physical settings, instruction following by a digital agent that is manipulating images, clarification discussions during a presentation and so on. Such agents can help to promote literacy and language skills, as users engage with the generated dialogue to create their own stories. In all such cases, a natural conversational experience will be emulated better if visual or other modal elements get incorporated in the AI models.

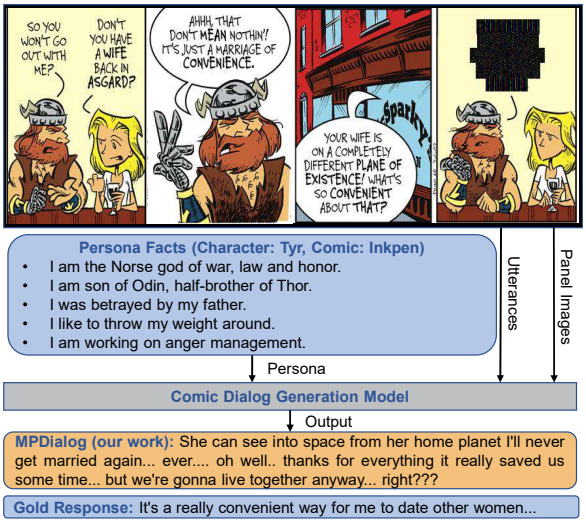


Figure 1: Comic Dialogue Generation: Input is a comic strip, with its text utterances, segmented visual panels and persona of target character; output is an utterance.

There is substantial recent research in building neural conversational AI systems for *text-only* task-oriented dialogues (Eric et al., 2017; Madotto et al., 2018; Wu et al., 2018, 2021; Hosseini-Asl et al., 2020; He et al., 2022) as well as open domain conversations (Gao et al., 2020; Zhang et al., 2020; Santra et al., 2021; Shuster et al., 2022). On the other hand, research on multimodal conversation is still in its early stages. A key exception is Visual Dialog (Das et al., 2017), where an agent answers multi-turn questions about a single static image. However, to the best of our knowledge, there is little work that builds dialog systems with multiple evolving images.

Our goal is to advance research in such multimodal dialog systems. A particular domain that enables us to study this is that of comic books. In contrast with Visual Dialog, a comic strip has *several* images with temporal progression and an aligned dialog. Building an effective comic dialog

system necessitates understanding the visual narrative, in addition to the textual context, making it a good testbed for multimodal dialog.

In addition to multimodality, comics have several other unique characteristics that make the domain challenging for AI systems. For instance, comic conversations are often multiparty, whereas most existing dialog agents assume dyadic (two party) conversations. Moreover, each character in a comic has a distinctive persona and style, and the dialog agent has to learn to follow the right style for each speaker. Finally, many comics are humorous, necessitating the model to be funny in its responses.

To study dialog systems in the comics domain, we first curate a novel dataset, COMSET, which consists of ∼54K strips from 13 comics. Each strip is associated with the visual panel (with text masked), along with the text transcript. We harvest strips from a publicly available online collection, GoComics.[1] Panel and dialogue segmentation on the visual scene data in these strips leads to a dataset with 200+ characters. To describe the distinctive persona of each lead character, we also curate a set of persona facts (inspired by Zhang et al. (2018)) from popular fandom websites.

We define the novel task of next utterance generation for comics conditioned on the textual dialog history, visual scene history, and the persona facts for the comic characters. Fig. 1 shows an example. Since existing dialogue generation models do not handle multi-image multimodal context along with persona, we implement a novel method (MPDIALOG) for the task, as illustrated in Fig. 4. Text utterances, persona facts, and visual scenes are passed into the MultiModal Embedding (MME) module which encodes them into tokens each of D=768 dimensions. These embeddings are then passed on to a language decoder to produce the output tokens. MME module (i) computes the text encodings using a text embedding (TE) layer, (ii) computes visual token embeddings of panel images using CLIP Vision encoder (VE), linearly projects (LP) each embedding of size D to $n \times D$ and reshaping it to n tokens each of size D, (iii) interleaves text and visual token embeddings. Interleaving occurs such that the dialogues of a panel are preceded by the respective panel embedding. Extensive comparisons show that MPDIALOG outperforms multiple text-only dialogue generation systems as well as those systems that do not use persona facts.

Overall, we make the following main contributions in this work. (1) We contribute a novel multimodal comics dataset, COMSET, containing ∼54K strips and persona facts for 200+ characters. (2) We propose a multimodal persona-based dialog generation baseline, MPDIALOG, which incorporates both the modalities and generates the next utterances effectively. (3) We demonstrate empirically that multimodality and persona orientation leads to better dialogues. This paper adds interesting questions around multimodal persona-based dialogue generation modeling and we hope that our study motivates more work in this area. We make code and dataset publicly available.[2]

2 Related Work

Our work is related to the following three areas: dialogue generation, multimodal models, and multimodal datasets for dialogue generation.

Dialogue Generation: Recently, several neural dialog generation models have been proposed (Gao et al., 2018; Ni et al., 2022); we discuss a few here. DialoGPT (Zhang et al., 2020) uses a GPT-2 (Radford et al., 2019) decoder pretrained on Reddit conversations and can effectively capture the contextual information in dialogues, thereby generating interesting and human-like responses. However DialoGPT does not allow explicit style control over the generated responses. EDGE (Gupta et al., 2021) allows for controlled response generation by conditioning on semantic frames of exemplar responses. A particular kind of style control models are persona-based models which use "persona" information for personalized dialog generation. Bert-over-Bert (Song et al., 2021) disentangles persona-based dialogue generation into two tasks: dialogue generation and consistency understanding; the model uses a shared BERT encoder but has two task-specific decoders. PersonaGPT (Tang et al., 2021) uses GPT-2 finetuned on PersonaChat (Zhang et al., 2018) dataset, with added persona fact tokens for personalized generation and question control codes for controlled generation. None of these models capture the multimodal multi-party context which is the setting for comic dialogues.

Multimodal Datasets for Dialogue Generation: The COMICS (Iyyer et al., 2017) dataset contains scanned images of comic strips but it does not contain manually extracted transcript information or

[1] https://gocomics.com/

[2] https://github.com/dair-iitd/MPdialog

information about comic characters. Further, as the authors mention, the dataset is unsuitable for generation tasks due to OCR detection inaccuracies. PersonaChat (Zhang et al., 2018) has conversations between two agents and their corresponding persona facts but it has no images. Other multimodal datasets include ImageChat (Shuster et al., 2020), PhotoChat (Zang et al., 2021) and VisualDialog (Das et al., 2017) which have a conversation between speakers about a single reference image. They differ from our setting, where the speakers are themselves a part of the image, and we have multiple panels (images).

Multimodal Models: Recently, several types of multimodal encoders and generators have been proposed for a variety of applications. Models like CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) are based on alignment of visual and textual embedding spaces. Frozen (Tsimpoukelli et al., 2021) and ClipCap (Mokady et al., 2021) also align text and visual embedding by projecting visual embeddings onto the textual embedding space. Text-image cross attention is used in VisualGPT (Chen et al., 2021), VC-GPT (Luo et al., 2022), CoCa (Yu et al., 2022). Perceiver-IO (Jaegle et al., 2021) is a fully attentional read-process-write architecture with variants like Uni-Perciever-MOE (Zhu et al., 2022) which use Mixture of Experts for response selection. In SimVLM (Wang et al., 2021) and VisualBERT (Li et al., 2019) the Visual and Textual Models are jointly trained on the task itself. Given its significant zero-shot image classification capabilities, we use CLIP as the image encoder for our MPDIALOG architecture.

3 The COMSET Dataset

We contribute a novel comics dataset, COMSET, containing 13 popular English comic strips, obtained from GoComics.[1] Each comic strip contains transcription and an image. We remove duplicate strips (re-broadcasts with minor modifications) based on Levenshtein distance between transcripts. For each comic, we also obtained persona facts (representative personality traits) for each character by manually curating such information from websites like Fandom,[3] Wikipedia,[4] and TV Tropes,[5] and paraphrasing all collected persona facts into first person English sentences. We describe data pre-processing and analysis in this section.

3.1 Dataset Pre-processing

The raw dataset was pre-processed as follows.

Parsing Transcripts: Parsing transcripts involves parsing speaker (character) and utterance pairs from unstructured conversation transcripts. We first obtained a list of comic characters (for our 13 comics) from same websites that were used to gather character personas. We also added character aliases to this list. Further, we mined frequent proper nouns with `PERSON` entity tag from all transcripts to search for all potential speaker candidates. We reduced infrequent characters into a catch-all character `OTHER`. Around 17% utterances in our corpus are attributed to `OTHER`. Further, there were some frequent speakers which were not named entities, for example, *Man, Woman, Stranger, Voice, Noise, Sound*. We conflated *Voice, Noise, Sound* into a single speaker (*Voice*) and added all such characters to list of characters. Finally, for all comics except *Doonesbury* and *Cleats*, we used list lookup for extraction of mention spans for character named entities. Using basic heuristics like word followed by colon or quotation characters, we could also do a fuzzy character name match to handle spelling errors in transcripts.

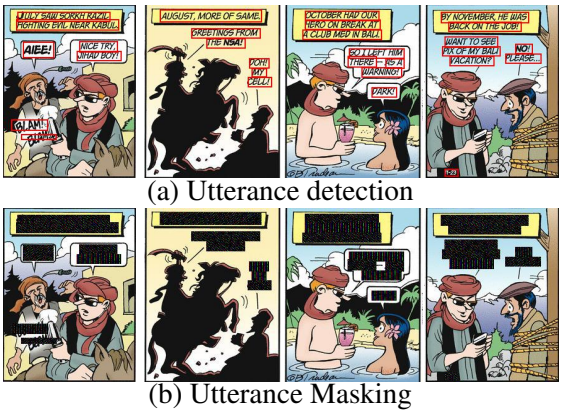
Transcripts for *Doonesbury* and *Cleats* contain free-form text like *Bucky is holding Smacky and says ...*. Typically each sentence contains four parts: character/speaker name (*Bucky*), action or attribute phrase (*is holding Smacky and*), speaking verb (*says, replies, asks, proclaims, etc.*[6]), and utterance. To obtain these parts from transcripts, we first perform part-of-speech tagging, named entity recognition, and dependency parsing using spaCy (Honnibal et al., 2020). Then we use heuristics like (a) speaker name should have the POS tag `PROPN`, must be the nominal subject (`nsubj`) and have the NER tag as `PERSON`, (b) The speaker should have a direct/indirect relation to the speaking verb.

Panel Segmentation: Each strip image had several panels and utterances across panels. Classical vision methods like Hough Transform (Duda and Hart, 1972), polygon detection (Li et al., 2014), recursive cuts (Pang et al., 2014) and density gradients (Tanaka et al., 2007) led to poor panel segmentation due to their assumptions about uniform white background and clean gutters. Inspired

[3] https://www.fandom.com/topics/comics
[4] https://www.wikipedia.org/
[5] https://tvtropes.org/

[6] https://archiewahwah.wordpress.com/speech-verbs-list/

Figure 2: Few examples of Panel Segmentation



(a) Utterance detection

(b) Utterance Masking

Figure 3: Examples of utterance detection and masking

by Iyyer et al. (2017), we model the panel segmentation as an object detection problem. We used the 500 manually annotated panel bounding boxes out of comic strips provided by them to train a Faster-RCNN (Ren et al., 2015) architecture with a ResNet-50 (He et al., 2016) backbone, and used it to segment panels from our comic strips. Some segmentation results are shown in Fig. 2.

Dialogue Text Detection and Masking: While predicting the next utterance for a character in the current panel, the ground truth utterance in the panel image could lead to a label leak. Hence, to eliminate redundancies and to avoid possibilities of label leak, we mask the utterance text from panel images. Iyyer et al. (2017) detect utterance text on images by training a Faster-RCNN model on 1500 manually annotated panels to detect text boxes. This approach led to poor results for our dataset since text box structure is not consistent across comics, and often there is no explicit text box or bubble to encapsulate the dialogue, also evident from Figs. 2 and 3. Hence, we used off-the-shelf OCR, specifically EasyOCR,[7] to extract the text and bounding boxes from each segmented panel. We filled bounding boxes with random noise so as to not bias the model towards any color at utterance positions, as shown in Fig. 3.

Multimodal Alignment: For each comic strip c, panel segmentation yields a sequence of n_c panel images along with OCR text $\{P_j\}_{j=1}^{n_c}$ for each panel j, and transcript parsing yields a sequence of m_c utterances $\{D_i\}_{i=1}^{m_c}$ along with speaker labels. For next utterance prediction, the model needs both text and visual context *aligned* with each other. For each (D_i, P_j) pair, we calculate a string fuzzy Levenshtein distance-based similarity score S_{ij} which determines the extent to which D_i matches with text P_j. The panel index for the i^{th} utterance is then calculated as $\sigma_i = \arg\max_j S_{ij}$. The matched panel sequence can be written as $\Sigma = \{\sigma_1, \sigma_2, \ldots, \sigma_m\}$. Due to inaccurate OCR, Σ may not be monotonically increasing. We handle this inconsistency by transforming Σ to a sorted sequence $\overline{\Sigma} = \text{DP}(\Sigma)$ where DP is a dynamic programming method to sort an input sequence with minimum number of edits. We found that the DP filter was needed for only 2% of all the utterances.

3.2 Dataset Statistics and Quality

Across 13 comics, COMSET contains 53,903 strips covering a total of 159,610 panels and 238,484 utterances. Thus, there are 2.96 images per strip. On average, a dialogue contains 16.09 tokens. Each strip has 2.98 characters on average. The dataset contains 6.66 persona facts per character on average across 202 characters. Each persona fact contains 12.23 tokens on average. Table 1 shows key statistics for COMSET. Table 2 shows distribution of number of strips, panels, utterances and characters across the 13 comics. We split the 13 comics into a seen set of 8 comics and unseen set of 5 comics. Seen set was further split randomly 70:10:20 into train:val:test stratified by comic name.

We manually inspected our dataset quality using 50 randomly chosen examples. We found that our scripts for parsing speaker from transcripts had an accuracy of ∼98%. Some comics had bad transcripts, and speaker information was completely missing (<1%). In ∼2% of utterances, there were some parts of the speaker overflowing into the previous utterance due to whitespace in speaker names

[7] https://github.com/JaidedAI/EasyOCR

Metric	Value
Avg Unique characters (per strip)	2.98
Avg dialogue length (tokens)	16.09
Avg persona facts (per comic)	57.39
Avg persona facts (per character)	6.66
Avg persona fact length (tokens)	12.23
Image per Utterance	0.671
Image per dialog	2.96
Number of Strips/dialogs	53,903
Number of Panels/Images	159,610
Number of Utterances	238,484
Number of Characters(Personas)	202

Table 1: Key statistics of the proposed COMSET dataset.

Comic	# strips	# panels	# utterances	# characters
cleats	2588	5580	10064	33
bigtop	1752	5457	8977	11
heartofthecity	6544	14117	23499	14
garfield	10295	24578	27731	21
peanuts	2623	10612	7069	15
riphaywire	2730	7815	14638	18
bignate	5446	21339	32380	11
inkpen	2205	6722	9736	12
getfuzzy	2383	6630	12080	11
familytree	362	1119	2112	11
calvinandhobbes	2557	8120	10120	11
doonesbury	13821	45401	77329	21
cathy	597	2120	2749	13
Total	53903	159610	238484	202

Table 2: Comic wise distribution of dataset statistics.

(ex. 'Voice from television', 'Person on TV'). Text masking was evaluated on 1000 examples and we found ∼4% of all comics had italicized text, or font size too small, low character spacing, that made it difficult to detect and mask bounding boxes. In ∼3% comics, panel segmentation was challenging due to no clear demarcation between several frames, as depicted in Fig. 7 in the Appendix. In ∼5% of all utterances, the dialogue did not map correctly to its panel primarily due to OCR detection errors. We had also assumed that a dialogue can be mapped one-to-one to a panel which is not always true as a dialogue can sometimes overflow into multiple panels, in which case a panel with the most matching words was chosen. Overall, we find error percentages in each part of the dataset curation pipeline to be low. The end to end accuracy over 200 random datapoints from the test set came out to be 91.5% indicating that the resulting dataset is of high enough quality to study the comic generation task.

4 Methodology

In this section we formalize the next utterance prediction task in the multi-modal persona-based dialogue setting for benchmarking COMSET and propose a novel baseline architecture MPDIALOG.

4.1 Next Utterance Prediction Task

For a comic strip, consider a conversation history with utterances $\{C_i\}_{i=1}^{n}$ and an aligned sequence of images $\{I_j\}_{j=1}^{m}$. At any time step t, the objective is to generate C_t given the textual conversation history $\{C_i\}_{i=1}^{t-1}$ and the corresponding image history sequence $\{I_j\}_{j=1}^{k}$ where C_t is aligned with I_k, $t \leq n$, and $k \leq m$.

In practice, it may be useful to limit historical context to a history size h of past utterances and their corresponding panel images. While this problem formulation is generally applicable to any setting with multimodal conversation history, we propose a model for next utterance prediction for comics in this work.

4.2 Baseline Methods

We first describe our adaptation to the existing language model (LM) only methods, as well as LM+persona based methods.

LM only: LM only methods use only the text part of the conversations. We experiment with DialoGPT (Zhang et al., 2020) and EDGE (Gupta et al., 2021). DialoGPT is trained on a 147M multi-turn dialogue dataset from Reddit, and conditions response generation on the previous conversation context. EDGE (Gupta et al., 2021) allows controlling dialogue response generation based on semantic structure of exemplar responses. During inference, EDGE retrieves the exemplar responses of the test set context with train set dialogues as the candidate set using a ParlAI Polyencoder (Humeau et al., 2019) model[8] pretrained on the ConvAI2 dataset. EDGE then uses the opensesame (Swayamdipta et al., 2017) frame extraction model, which is a frame-semantic parser for automatically detecting FrameNet (Baker et al., 1998) frames and their frame-elements from sentences. We adapt these models to COMSET by extracting the conversation history $C_{t':t-1}$ and finetune the model to predict C_t, where $t' = \max(0, t-h)$. We set the maximum history size $h = 5$.

LM+Persona: These baselines utilize the conversation context along with persona facts for each character to generate persona consistent responses. Models evaluated include PersonaGPT (Tang et al., 2021) and BoB (Song et al., 2021). These models assume a dyadic conversation and require persona facts of both the speakers as input to generate responses. PersonaGPT is finetuned on the PersonaChat (Zhang et al., 2018) dataset, with added spe-

[8] https://parl.ai/projects/polyencoder

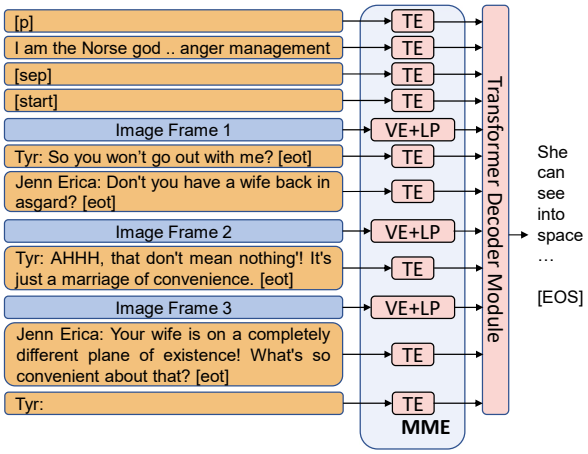


Figure 4: Overall Architecture of MPDIALOG

cial tokens (`[p1], [p2]`) to mark the persona facts. Since our problem is different and includes multi party conversations, we provide the persona facts of the speaker of utterance C_t as a prefix to the input (marked by a single special persona token `[p]`) with the objective of predicting C_t. Thus, we pre-train the models using the target character persona prefix on their respective datasets and later finetune it on COMSET.

4.3 MPDIALOG Architecture

The architecture for MPDIALOG is inspired by Frozen (Tsimpoukelli et al., 2021). It consists of a vision encoder and a language decoder with a linear projection module in between. The vision encoder encodes the visual comic panels into a single vector, which is then projected into visual tokens using the linear projection module and is fed into the language decoder along with text tokens for generation as shown in Fig. 4. We use a CLIP vision encoder (Radford et al., 2021), as it has been shown to be effective in aligning visual embeddings into the same semantic space as text embeddings. Similar to *Frozen*, we project the panel embedding of D dimensions into an $n \times D$ dimensional vector using a linear layer and reshape it into n visual tokens each with embedding size D. We use n=2 in our experiments as suggested in Tsimpoukelli et al. (2021).

As shown in Fig. 4, in the MME (MultiModal Embedding) module, we concatenate the dialog history tokens separated by the end of text separator token (`[eot]`) and insert visual token embeddings wherever the panel in the conversation changes (including the first) and feed the resulting sequence into the language decoder. When persona information is available we prepend the persona tokens to the sequence input of the language decoder, along with the persona start token `[p]`. The multimodal MME output is fed as input to the GPT-2 language decoder. We train the model using causal language modelling, i.e., auto regressive loss over the target prediction tokens. We use PersonaGPT-base and a pre-trained CLIP vision encoder as the textual and visual components of MPDIALOG respectively, and finetune it on COMSET in an end-to-end fashion. The projection module is simply a linear layer. Unlike Frozen, we do not freeze any component and train the entire architecture end-to-end. Once trained we generate responses using nucleus sampling (Holtzman et al., 2019) and set top-p=0.95, top-k=50. Other generation parameters are as follows; temperature=0.05, repetition penalty=1.2 (Keskar et al., 2019).

5 Experiments and Results

5.1 Hyper-parameters for Reproducibility

All results are computed on a GeForce GTX 1080 Ti (12 GB) cluster with 64 cores each of Intel(R) Xeon(R) Gold 6142 CPU @ 2.60GHz. There are 12 layers each in both the vision encoder and the text decoder, with 8 multi attention heads in each transformer layer. The vision encoder weights are initialized from `openai/clip-vit-base-patch32` available on HuggingFace[9] and the text decoder weights are initialized from a PersonaGPT-base model trained on PersonaChat dataset. The model was trained for 3 epochs on 2 GPUs with a learning rate of 5e-5 and a linear decay schedule with an initial warmup of 500 steps using the AdamW (ϵ=1e-8) optimizer on a batch size of 12. Further details can be found in our repository.[2]

5.2 Evaluation Metrics

We report the results of various baselines and our proposed method on several natural language generation metrics. We report the perplexity score for each model which measures the uncertainty of a model to output the target sequence given the context words. Further, we evaluate the models on lexical metrics like unigram precision, recall and F1 scores as well as neural metrics like BLEURT (Sellam et al., 2020) and MaUde (Sinha et al., 2020). MaUde is a particularly relevant metric as it is curated specifically for dialogue generation and mea-

[9] https://huggingface.co/

	Model	Params	PPL	BLEURT	MaUde	Prec.	Rec.	F1
Seen	DialoGPT	117M	30.12	0.221	0.807	0.040	0.110	0.050
	EDGE	124M	30.17	0.256	0.897	**0.107**	0.087	0.083
	PersonaGPT	117M	19.40	0.233	0.894	0.040	0.130	0.054
	BoB	330M	40.00	0.224	0.896	**0.107**	0.114	**0.097**
	MPDIALOG	213M	**19.02**	**0.266**	**0.898**	0.064	**0.254**	0.093
Unseen	DialoGPT	117M	37.57	0.219	0.792	0.041	0.109	0.051
	EDGE	124M	36.86	0.254	0.862	0.130	0.081	0.083
	PersonaGPT	117M	**24.79**	0.230	0.896	0.045	0.126	0.058
	BoB	330M	52.69	0.240	0.872	**0.133**	0.085	0.090
	MPDIALOG	213M	25.75	**0.257**	**0.904**	0.066	**0.227**	**0.093**

Table 3: Comparison across various models for next utterance prediction on both seen and unseen comics.

	Comic	DialoGPT	EDGE	PersonaGPT	BoB	MPDIALOG
Seen	familytree	0.224	0.250	0.230	0.207	**0.265**
	doonesbury	0.221	**0.256**	0.202	0.227	0.240
	getfuzzy	0.230	0.252	0.242	0.226	**0.281**
	bigtop	0.230	0.266	0.246	0.220	**0.269**
	garfield	0.207	0.255	0.227	0.233	**0.270**
	inkpen	0.237	0.259	0.243	0.233	**0.277**
	cathy	0.205	0.246	0.227	0.217	**0.259**
	calvin and hobbes	0.221	0.263	0.251	0.234	**0.274**
Unseen	rip haywire	0.224	0.243	0.231	0.234	**0.255**
	cleats	0.220	**0.263**	0.228	0.239	0.254
	peanuts	0.200	0.249	0.232	0.235	**0.257**
	bignate	0.221	0.259	0.225	0.247	**0.261**
	heart of the city	0.234	0.258	0.237	0.247	**0.260**

Table 4: Comic wise BLEURT for various models.

sures the coherence of responses with the previous conversation context.

5.3 Main Results

We design extensive experiments to answer the following questions: (1) Can existing dialogue generation language models adapt their knowledge to the comic setting? (2) To what extent does persona orientation of language models help in generating comics? (3) Does adding multimodality help the language model in better understanding the context and thereby generating coherent responses? (4) How does the generalizability to unseen comics (zero-shot setting) vary across architectures. To answer these questions we finetune each of the baseline language-only models and those with persona alignment on only the textual component and later train MPDIALOG on the multimodal dataset. We generate response for each of the trained models using nucleus sampling. This was done for both the seen (finetuned) and unseen (zero-shot) splits of our dataset. The results for these experiments are shown in Table 3.

Performance on Seen Dataset: We observe that the proposed model, MPDIALOG, outperforms both the language model only as well as persona-based baselines. Language only models (like DialoGPT and EDGE) cannot generate coherent responses (high perplexity and low MaUde) in the comic setting. This is expected as it is very hard to understand the context of a comic without any information about the characters or the visual scene. We observe that adding persona information of the characters significantly boosts performance as is evident from the perplexity scores, BLEURT and MaUde, of PersonaGPT-base. We conducted a Welch-T (Welch, 1947) test on results of MPDialog with other baselines for precision, recall, F1, MaUde, BLEURT and we got $\max(p) < 0.025$ indicating statistical significance. Persona information delivers meaningful insights into the context and helps the model in understanding the conversation better. Moreover, adding visual scene information along with the persona also boosts performance as the model has now access to the actual scene of the comic in which the conversation is happening. As an ablation we also trained a model with language and visual components but without the persona information. Its perplexity came out to be 23.76 on the seen set, which is better than just the language part (DialoGPT) but worse than MPDialog (which also incorporates persona). For illustration, we show a cherry-picked example in Fig. 5 where PersonaGPT-base and MPDIALOG are able to align their responses with the character persona, whereas responses for other language only models are either too banal (EDGE and BoB) or completely nonsensical (DialoGPT).

Performance on Unseen Dataset: We also show results on unseen comics set, for various models in Table 3. Again, MPDIALOG outperforms baselines across most metrics. Perplexity scores of language only models (both DialoGPT and EDGE) are quite high. Adding persona information and evaluating PersonaGPT we find that it significantly boosts the models' ability to generate responses with more certainty, as the perplexity score drops by a factor of $\sim$12 points. It is worthwhile to note that PersonaGPT performs much better on neural metrics like BertScore and MauDe as compared to language only baselines. Our multimodal model (MPDIALOG) further improves these neural metrics and shows that it is able to produce much more coherent responses by taking into account the associated visual information. Thus the trend (LM + persona + images > LM + persona > LM) can still be observed as in the "seen comics" setting.

Persona Facts (Character: Arlene, Comic: Garfield)
- I am Garfield's on-and-off girlfriend
- I can unsuccessfully crack jokes at the expense of Garfield on a regular basis.
- I love to make smart remarks about Garfield, most likely due to his enormous ego.
- I am often interested in a deeper relationship with Garfield

Context: I adore you, I worship you! I love you more than life itself. I can't live without you.
Gold Response: Uh.. Garfield?

MPDialog (our work): You're not alone...you've been there for years and your whole life has just come full circle around me!! And now that we have each other again..we'll be together foreverrrr!!! We will rule this world!!

PersonaGPT (LM + persona): You're not alone??!...you know what this is like?? It must be so much better when we get back together and have some sorta conversation over it all....but that doesn't amaze me too hard for no reason.

EDGE (LM + frames): I can't live without you

BoB (LM + persona): I'm going to have to eat.

DialoGPT (LM only): I'm sobsomn'ing to the world is a great place, and it's not just for me?!???!!! what do we have in common with each other??? oh my god....what are they doing ' 'out here again

Figure 5: Comparison of predictions from various models for a test instance from Garfield. Left: The comic strip, context and the persona facts of the speaker (Arlene). Right: Predictions of various models.

	Comic	DialoGPT	EDGE	PersonaGPT	BoB	MPDIALOG
Seen	familytree	0.829	0.894	0.884	0.861	**0.901**
	doonesbury	0.835	0.908	0.877	0.915	**0.926**
	getfuzzy	0.829	0.904	**0.917**	0.915	0.903
	bigtop	0.813	0.898	0.899	**0.904**	0.901
	garfield	0.755	**0.867**	0.858	0.864	0.819
	inkpen	0.832	0.922	0.918	**0.935**	0.923
	cathy	0.763	0.883	0.894	**0.903**	0.889
	calvin and hobbes	0.804	0.908	0.911	0.906	**0.923**
Unseen	rip haywire	0.836	0.869	0.927	0.898	**0.932**
	cleats	0.807	0.862	0.885	**0.895**	0.891
	peanuts	0.695	0.845	0.892	0.842	**0.897**
	bignate	0.796	0.868	0.886	0.860	**0.908**
	heart of the city	0.828	0.868	0.890	0.867	**0.892**

Table 5: Comic wise MaUde for various models.

	Comic	DialoGPT	EDGE	PersonaGPT	BoB	MPDIALOG
Seen	familytree	43.50	38.10	**22.30**	59.38	**22.30**
	doonesbury	27.29	28.29	19.92	36.39	**19.35**
	getfuzzy	32.22	33.91	22.08	41.39	**21.53**
	bigtop	28.07	29.18	18.40	36.75	**18.28**
	garfield	21.87	21.36	12.70	29.08	**12.38**
	inkpen	27.57	30.16	17.99	32.87	**17.62**
	cathy	35.08	33.40	23.95	49.69	**23.07**
	calvin and hobbes	25.40	27.02	17.92	34.44	**17.65**
Unseen	rip haywire	52.92	51.24	**33.50**	72.06	36.22
	cleats	31.26	31.04	**20.23**	47.79	21.32
	peanuts	35.20	33.17	26.19	54.40	**25.80**
	bignate	31.12	30.42	**19.67**	41.22	20.02
	heart of the city	37.34	38.41	**24.35**	48.00	25.40

Table 6: Comic wise PPL for various models.

5.4 Comic-wise Quantitative Analysis

Table 4 shows comic-level BLEURT scores for both the seen as well as unseen test sets. We also show MaUde and perplexity scores in Tables 5 and 6 respectively. For most comics across all the three metrics, MPDIALOG performs better than other models. Unlike most comic strips, Cleats comic focuses on the relationships between the characters, their sportsmanship and the challenges of being part of a team. We believe that images in Cleats do not contain much additional information and hence multi-modality of MPDIALOG does not lead to improved results.

5.5 Qualitative Analysis

The proposed method, MPDIALOG, is persona-based. How well does it capture the persona style in the generations, compared to other persona-based baselines? To answer this question, we perform the following experiment. For every character c in the train set, we obtain its unigram vocabulary distribution Train_c. Given a model, over the entire test set, we also compute unigram vocabulary distribution Outputs_c from combined text of all generations for character c. If the model has captured persona for character c well, the symmetric KL-divergence between Train_c and Outputs_c should be small. Hence, we compare MPDIALOG with other persona-based baseline models (PersonaGPT and BoB) using the symmetric KL divergence metric. We observe that symmetric KL divergence is 4.41, 4.56 and 3.36 for PersonaGPT, BoB and MPDIALOG respectively. Thus, we infer that MPDIALOG is the best at capturing the persona information.

We also attempt to understand the image patch attribution for a generated dialogue by our model as applied on Fig. 5. We conducted a Grad-CAM (Selvaraju et al., 2017) analysis to check where the model "looks" while generating its utterances. Since generation is stochastic and dependent on nucleus sampling, we cannot attribute the model's output to a particular attention map over the image. As a surrogate, we calculate the attention map over the visual panels when the model generates the last `[eot]` token. In Fig. 5, we were able to observe that the model does indeed look at Arlene's face and Garfield's face (as indicated in Fig. 6) and gives less relative importance to the background and the bubble above it. It helped us confirm that our model is able to contextualize within the images as well and generates tokens

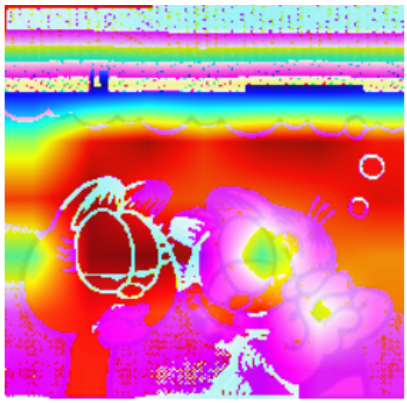


Figure 6: Grad-CAM output heatmap overlayed on image in Fig. 5.

based on meaningful and interpretable image features.

5.6 Human Evaluation Results

We obtain manual annotations for the utterances generated by various models on fluency, engagingness, dialog-consistency, scene-consistency and persona-detection. Four annotators performed judgments on a set of 65 examples, randomly sampled from the test set. We compute inter-annotator agreement as pairwise correlation between method rankings, averaged across the five criteria. It was found to be 0.318 (Kendall's Tau 'B') which is considered as strong agreement. Detailed annotation guidelines are mentioned in the appendix. Specifically, we measure persona detection as follows. Given persona facts of two characters, and a response, the annotator is asked to guess which of the twi persona the response matches to. Table 7 shows that MPDIALOG performs best on all measures except for dialog consistency where EDGE performs the best. EDGE uses semantic frame examplars to guide a structure for the uttterance leading to better consistency. All the other models do not make use of this extra structural input, and amongst them, MPDIALOG performs best. On persona detection, MPDIALOG performs comparably to PersonaGPT. Overall, MPDIALOG performs quite well on human perceived quality of generated comic dialogues.

As an additional qualitative analysis for the proposed model, we performed the following experiment. We considered examples, where in the multimodal input context, we changed the last character prompt to some other character from the same or other comic. The goal was to check how would another character (say "Tyr") respond in a situation in a comic (say "Garfield"). We found that the

Model	Engagingness	Fluency	Dialog consistency	Scene consistency	Persona detection
DialoGPT	3.06	2.46	1.60	1.93	0.43
EDGE	1.77	3.46	**2.57**	2.39	0.48
PersonaGPT	3.17	2.53	1.83	1.89	**0.63**
BoB	2.13	2.82	2.18	2.25	0.60
MPDIALOG	**3.52**	**3.50**	2.19	**2.89**	0.62

Table 7: Human Evaluation Results

generated responses often reflect the persona of the injected character. For example, in Garfield, we found for the same situation: (1) Mom's response showing her down-to-earth, exasperated and sensitive nature who loves her son dearly, and (2) Susie's response to be teasing Calvin, thereby showing her love-hate relationship with Calvin. Thus, our model seems to be capturing the persona behavior somewhat, but we feel there is much more work to be done to generate responses that are contextually more coherent, and at the level of human skill.

6 Conclusions and Future Work

We propose a novel problem of next utterance prediction for comics given historical multimodal context consisting of previous utterances and panel images. We contribute a novel dataset, COMSET, which contains 53,903 strips, 159,610 panels and 238,484 utterances from 13 comics. We also propose a multimodal persona-based baseline model, MPDIALOG, which performs better compared to strong language-only and persona-based dialogue generation models, both in the seen comic and the unseen comic settings. We make our code and dataset publicly available[2]. In the future we plan to (1) focus on generation of humor-focused text, and (2) explore generation of next utterances and panel images together.

Acknowledgements

This work is supported by grants by Google, Verisk, and 1MG, an IBM SUR award, and the Jai Gupta chair fellowship by IIT Delhi. We also acknowledge travel support from Google and Yardi School of AI travel grants. We thank the IIT Delhi HPC facility for its computational resources. We also thank Rocktim Jyoti Das for his help with the code for MPDialog.

Limitations

In this paper, we focused on English comics only because of their ease of availability. Although we have not experimented with non-English text, we

expect the proposed model to work well in multilingual settings if we replace GPT-2 decoder with other decoders like BLOOM (Scao et al., 2022).

Ethics Statement

Most of our dataset has been obtained from GoComics (https://gocomics.com/). The website allows downloads of comic images for research purposes. However, they do not allow redistribution of images. Hence, in our dataset release, we have only provided links to images on GoComics website. Providing links to images or webpages is a common trend (e.g., Google Landmarks, GoogleConceptualCaptions, WIT datasets). That said, our code base provides all the scripts needed to (1) do pre-processing and modeling based on this images (2) gather transcripts and align with the panels in comic strips. Thus, overall, all steps in the paper are reproducable. Further, we have also provided character identification annotations that we perform on these images as part of the dataset.

Natural language generation is in general prone to issues like biased, offensive, harmful, misinformative text generation. Fortunately, in this work, we finetune our models using relatively clean comics dataset. Also, given that these generations are meant to be consumed in a humorous form, we do not foresee the bias (if at all) generated by our model to be hurtful. To the extent we browsed over the generations produced by our model, we did not observe any biased, offensive, harmful, misinformative text getting generated.

References

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. 2021. Visualgpt: Data-efficient image captioning by balancing visual input and linguistic knowledge from pretraining. *arXiv preprint arXiv:2102.10407*.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335.

Richard O Duda and Peter E Hart. 1972. Use of the hough transformation to detect lines and curves in pictures. *Communications of the ACM*, 15(1):11–15.

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.

Jianfeng Gao, Michel Galley, and Lihong Li. 2018. Neural approaches to conversational ai. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1371–1374.

Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brockett, and William B Dolan. 2020. Dialogue response ranking training with large-scale human feedback data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 386–395.

Prakhar Gupta, Jeffrey P Bigham, Yulia Tsvetkov, and Amy Pavel. 2021. Controlling dialogue generation with semantic exemplars. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3018–3029.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Wanwei He, Yinpei Dai, Yinhe Zheng, Yuchuan Wu, Zheng Cao, Dermot Liu, Peng Jiang, Min Yang, Fei Huang, Luo Si, et al. 2022. Galaxy: A generative pre-trained model for task-oriented dialog with semi-supervised learning and explicit policy injection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10749–10757.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python. To appear.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. *Advances in Neural Information Processing Systems*, 33:20179–20191.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2019. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *International Conference on Learning Representations*.

Mohit Iyyer, Varun Manjunatha, Anupam Guha, Yogarshi Vyas, Jordan Boyd-Graber, Hal Daume, and Larry S Davis. 2017. The amazing mysteries of the gutter: Drawing inferences between panels in comic

book narratives. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, pages 7186–7195.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. 2021. Perceiver io: A general architecture for structured inputs & outputs. In *International Conference on Learning Representations*.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Luyuan Li, Yongtao Wang, Zhi Tang, and Liangcai Gao. 2014. Automatic comic page segmentation based on polygon detection. *Multimedia Tools and Applications*, 69(1):171–197.

Ziyang Luo, Yadong Xi, Rongsheng Zhang, and Jing Ma. 2022. Vc-gpt: Visual conditioned gpt for end-to-end generative vision-and-language pre-training. *arXiv preprint arXiv:2201.12723*.

Andrea Madotto, Chien-sheng Wu, and Pascale Ngan Fung. 2018. Mem2seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems. In *ACL 2018-56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, volume 1, page 1468.

Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, and Erik Cambria. 2022. Recent advances in deep learning based dialogue systems: A systematic survey. *Artificial intelligence review*, pages 1–101.

Xufang Pang, Ying Cao, Rynson WH Lau, and Antoni B Chan. 2014. A robust panel extraction method for manga. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 1125–1128.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *https://github.com/openai/gpt-2*.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Bishal Santra, Sumegh Roychowdhury, Aishik Mandal, Vasu Gurram, Atharva Naik, Manish Gupta, and Pawan Goyal. 2021. Representation learning for conversational data using discourse mutual information maximization. *arXiv preprint arXiv:2112.05787*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. 2020. Image-chat: Engaging grounded conversations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2414–2429.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, et al. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *arXiv preprint arXiv:2208.03188*.

Koustuv Sinha, Prasanna Parthasarathi, Jasmine Wang, Ryan Lowe, William L Hamilton, and Joelle Pineau. 2020. Learning an unreferenced metric for online dialogue evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2430–2441.

Haoyu Song, Yan Wang, Kaiyan Zhang, Weinan Zhang, and Ting Liu. 2021. Bob: Bert over bert for training persona-based dialogue models from limited personalized data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 167–177.

Swabha Swayamdipta, Sam Thomson, Chris Dyer, and Noah A Smith. 2017. Frame-semantic parsing with softmax-margin segmental rnns and a syntactic scaffold. *arXiv preprint arXiv:1706.09528*.

Takamasa Tanaka, Kenji Shoji, Fubito Toyama, and Juichi Miyamichi. 2007. Layout analysis of tree-structured scene frames in comic images. In *IJCAI*, volume 7, pages 2885–2890.

Fengyi Tang, Lifan Zeng, Fei Wang, and Jiayu Zhou. 2021. Persona authentication through generative dialogue. *arXiv preprint arXiv:2110.12949*.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2021. Simvlm: Simple visual language model pretraining with weak supervision. In *International Conference on Learning Representations*.

B. L. Welch. 1947. The Generalization of 'Student'S' problem when several different population variances are involved. *Biometrika*, 34(1-2):28–35.

Chien-Sheng Wu, Richard Socher, and Caiming Xiong. 2018. Global-to-local memory pointer networks for task-oriented dialogue. In *International Conference on Learning Representations*.

Qingyang Wu, Yichi Zhang, Yu Li, and Zhou Yu. 2021. Alternating recurrent dialog model with large-scale pre-trained language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1292–1301.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Xiaoxue Zang, Lijuan Liu, Maria Wang, Yang Song, Hao Zhang, and Jindong Chen. 2021. Photochat: A human-human dialogue dataset with photo sharing behavior for joint image-text modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6142–6152.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. 2020. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278.

Jinguo Zhu, Xizhou Zhu, Wenhai Wang, Xiaohua Wang, Hongsheng Li, Xiaogang Wang, and Jifeng Dai. 2022. Uni-perceiver-moe: Learning sparse generalist models with conditional moes. In *Advances in Neural Information Processing Systems*.

A Panel Segmentation Errors

Our method produced errors where the demarcation between frames was not very clear as shown in a few examples in Fig. 7.

Figure 7: Panel Segmentation Error Analysis: The erroneous segmentations are colored separately from red

B Annotation details

Human annotations were done by four undergraduate Computer Science students (3 male, 1 female) with an interest in comics in the age group 21-22 years. They were paid as per the rules of our institute for the task. The annotators were informed that this data will be used for research on dialogue generation for comics.

The following guidelines were provided to the annotators for evaluation.

- Fluency: How fluent is the response on it's own? (1-5), where 1 is "not fluent at all", 5 is "extremely fluent". Fluency encompasses how easy to understand the response is.
- Engagingness: How much engaging is the response on its own? (1-5), where 1 is "not engaging at all" or "generic", 5 is "extremely engaging" or "unique". Engagingness is defined as how interesting and unique the response is. Repetition and generic responses are scored low and highly detailed and attention grabbing responses are scored high.
- Dialog Consistency: How consistent is the response to the dialogue history? (1-5) 1 is "totally unrelated" and 5 is "Fully consistent".
- Scene Consistency: How much consistent is the response to the image history? (1-5) 1 is "totally unrelated" and 5 is "Fully consistent" and 3 is "OK".
- Persona Detection: Given persona facts of two characters, which persona does the response match to?