

# Beyond Text Unlocking Multimodal Magic: Harnessing New Reddit Thread Datasets for Multimodal Summarization

## 1 Relatedness between Text and Images

In addition to the editing process, we evaluate the CLIPScore, which checks the correlation between the text data and the image. This helps us assess how effective our summaries are in correlation to the images that are present in each thread. A high CLIPScore indicates strong alignment between the text and image, suggesting that our summaries, particularly the post summaries, accurately reflect the content depicted in the images. This further validates the comprehensiveness and relevance of our summaries, ensuring they effectively capture the essence of the visual content within each thread.

## 2 Experimental Setup

All experiments were conducted on a machine with an NVIDIA A100-PCIE-40GB GPU, a 64-core Intel Xeon processor, and 256 GB DDR4 RAM. PyTorch was used as the backend, along with libraries such as NumPy, Pandas and SpaCy. The models included BART for abstractive summarization, RoBERTa for semantic embeddings and clustering, CLIP for text-image alignment, DALL-E for image generation, and Agglomerative Clustering for clustering tasks. The models were trained for 40 epochs with a learning rate of  $5 \times 10^{-5}$ , a batch size of 16, and the Adam optimizer with an epsilon of  $1 \times 10^{-8}$ .

Performance was evaluated using ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore for text summarization, and CLIPScore for text-image alignment. The dataset was split into 80% training, 10% validation, and 10% testing, with a nested 10-fold cross-validation for robust evaluation.

## 3 Baselines

**Extractive Baselines :** In our baseline models, Lead-1 summarizes the document by selecting its first sentence, while Lead-Comment summarizes using the top five prioritized comments within the thread. An extractive model, achieves

the maximum possible ROUGE score by extracting passages from the document, representing the highest performance achievable within an extractive framework.

**Text-only Baselines :** We evaluated three fine-tuned models: BART-base, T5-base, and LongT5-base. BART-base and T5-base are strong in summarization, while LongT5-base is designed for handling longer input sequences. All models were trained on the CNN/DailyMail/MREDDITSUM dataset before being fine-tuned for multimodal summarization.

**Extensions with Image Captioning :** For BART-ImgCap, LongT5-ImgCap, and T5-ImgCap, visual information is incorporated by adding an image caption to the text input. Captions are generated using the BLIP2 model and evaluated by the CLIP model to select the most relevant one. The selected caption is then added to the input before fine-tuning the models.

**Extensions with Vision-Guidance :** Vision-Guided BART and T5 models were adapted for multimodal summarization using 768-D ViT-base image embeddings. VG-BART was pre-trained on COCO captions, while VG-T5 was initialized from scratch. Both models were fine-tuned on pre-trained BART and T5 for this task, incorporating cross-modal attention and image transformers.

## 4 Data Annotation and Preprocessing for MMR-SUM

**Annotation Framework** The MREDDITSUM dataset employs a robust annotation framework designed to create high-quality multimodal summaries of Reddit threads. Each thread contains a combination of textual posts, manually linked comment images, and associated user discussions. The dataset comprises 5,033 threads, each with a polished, human-written summary. The annotation process is structured into three distinct phases:

**1. Original Post Summarization** The first step involves summarizing the intent of the original post and its associated image. Annotators are presented with the textual content of the post alongside the attached image and are tasked with constructing a single-sentence summary that captures: - The key intent of the original poster. - Relevant visual information from the image.

The resulting summaries are self-contained, allowing comprehension without requiring the original image. For example, a post stating “Blue or black?” accompanied by an image of blue and black shoes is summarized as: “The original poster asked whether blue or black shoes would better match a knee-length blue dress.”

**2. Comment Cluster Summarization** The second step involves organizing and summarizing the comments within the thread: **Clustering:** Comments are grouped based on semantic similarity using a RoBERTa-based sentence embedding model. Agglomerative clustering is applied with cosine similarity as the distance metric and average linkage. A threshold of 0.5 ensures clusters are both compact and semantically coherent. **Image Integration:** Annotators manually incorporate relevant images into comments based on the textual content. For



Figure 1: An illustration from the HMMRSUM dataset showcases a comprehensive summary encompassing the reddit post, diverse opinions from the comments. Furthermore, our model successfully prioritizes user rankings to the comments using ClipScore based on the post similarity with the comments.

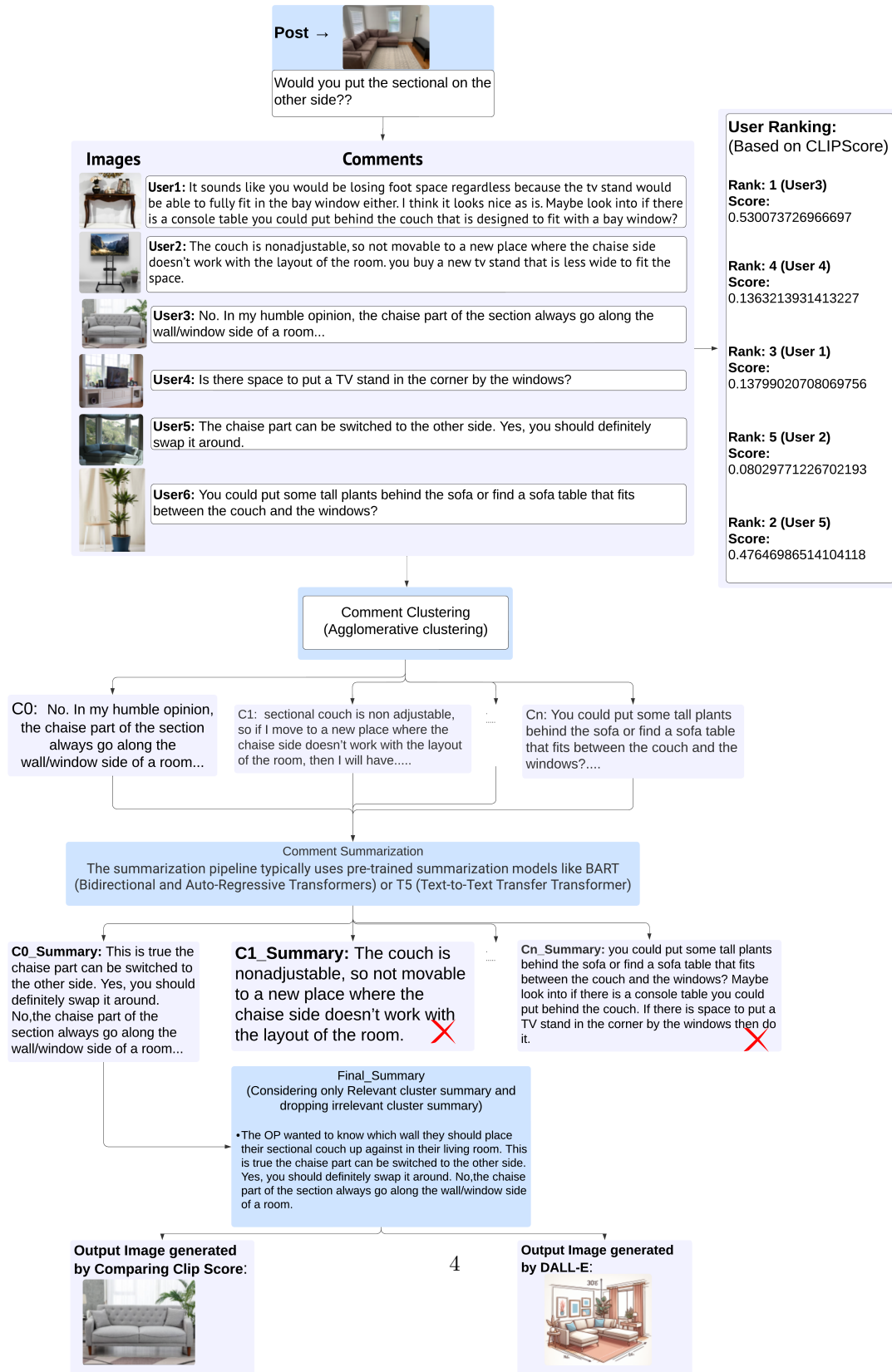


Figure 2: A Case Study of Relevant-cluster-based multi-stage summarization(RCMS)

example, if a commenter suggests a specific furniture design or accessory, an appropriate image is added to provide visual context. Saliency Ranking: Clusters are ranked by the aggregated Reddit scores of their constituent comments, prioritizing the top clusters for annotation to maximize relevance. Summarization: Annotators generate concise 1–2 sentence summaries for each cluster, ensuring that references to the accompanying image are explicitly incorporated when relevant. Consistency in terminology is enforced, referring to users as "Commenters."

**3. Summary Synthesis** The final step synthesizes the original post and comment cluster summaries into a cohesive thread-level summary: - Original post summaries and top comment cluster summaries are concatenated in descending order of cluster saliency. - Annotators refine the concatenated text by: - Removing redundancy. - Enhancing fluency and readability. - Ensuring logical sentence flow. - All summaries are written in the past tense for consistency and improved readability.

**Preprocessing Pipeline** The preprocessing pipeline ensures high-quality threads are curated while maintaining ethical standards.

**Thread Selection Criteria** Threads were selected based on the following criteria: 1. Image Inclusion: Each thread must contain an image in the original post, as Reddit does not allow images in comments so manually images were added in the comments by annotators as per their textual relevancy. 2. Image-Centric Discussion: Threads were chosen where the discussion relied heavily on the image, such as evaluating furniture arrangements or outfit styling. 3. Summarizability: Threads with a clear advice-seeking or opinion-sharing purpose were prioritized, while those eliciting simple reactions or humor were excluded.

**Source Subreddits** Threads were sourced from nine subreddits, focusing on: - Fashion and Accessories (e.g., r/fashionadvice, r/handbags). - Interior Design (e.g., r/designmyroom, r/malelivingspace).

Posts and comments were collected using a modified RedCaps tool, covering content from MREDDITSUM dataset expansion and new 2000 threads content from reddit from 2022 to November, 2024. Only threads with at least five comments were included.

**Filtering and Cleaning** - Content Filtering: - NSFW content and images containing faces were removed to ensure privacy. - Comments generated by bots and their replies were excluded. - Text Preprocessing: - URLs in the text were replaced with a placeholder token ('[URL]'). - Annotators manually added images to comments necessarily, ensuring visual references matched textual content.

**Quality Assurance** Annotation quality was prioritized through rigorous standards: - Annotator Selection: - Annotators were recruited from English-speaking regions with their expertise in their respective domains like fashion designer, interior designer, etc with a 98% approval rate and over 5,000 completed tasks on Amazon Turk. - A qualification task was administered, with results manually reviewed for quality assurance. - Compensation: - Annotators were paid an average rate of \$3/hour, exceeding minimum wage across all recruitment regions. - Guidelines: - Detailed instructions, including examples of

acceptable and unacceptable outputs, were provided to annotators.

**Dataset Characteristics** The MMRSUM dataset contains 5,033 annotated threads split into training, validation, and test sets of 4,529, 252, and 252 threads, respectively. The threads characterization details are mentioned in paper in detail.

This comprehensive annotation and preprocessing pipeline enables MMRSUM to serve as a benchmark dataset for multimodal summarization tasks. The inclusion of manually annotated images enhances its utility for research involving the interplay of textual and visual modalities in online discussions.

**For more detailed information please refer to the research paper itself**

**Post:**



help me find a tie?

**Comments:**


User 1: The beautiful thing about this colour is you could pair pretty much any tie with it (depending on what pants you want to wear) if you were going to wear this with dark navy pants or blazer's look for a creamy yellow tie.

User 2 : yellow with a light grey suit??

User 3 : yellow but go with a navy blue.

**MREDDITSUM**


**Post:**




help me find a tie?

**Comments:**


User 1: The beautiful thing about this colour is you could pair pretty much any tie with it (depending on what pants you want to wear) if you were going to wear this with dark navy pants or blazer's look for a creamy yellow tie.



User 2 : yellow with a light grey suit??



User 3 : yellow but go with a navy blue.



**MMRSUM**

Figure 3: Representation of MREDDITSUM and MMRSUM datasets.