Generating Need-Adapted Multimodal Fragments

Gaurav Verma
Adobe Research
Bangalore, India
gaverma@adobe.com

Suryateja BV
IIT Kanpur
Kanpur, India
suryab@iitk.ac.in

Samagra Sharma
IIT Roorkee
Roorkee, India
ssharma5@cs.iitr.ac.in

Balaji Vasan Srinivasan
Adobe Research
Bangalore, India
balsrini@adobe.com

ABSTRACT

Multimodal content is central to digital communications and has been shown to increase user engagement – making them indispensable in today's digital economy. Image-text combination is a common multimodal manifestation seen in several digital forums, e.g., banners, online ads, social posts. The choice of a specific image-text combination is dictated by *(a)* the information to be represented, *(b)* the strength of the image and text modalities in representing the information, and *(c)* the need of the reader consuming the content. Given an input content, representing the information to be represented in a multimodal fragment, creating variants accounting for these factors is a non-trivial and tedious task; calling for a need to automate. In this paper, we propose a holistic approach to automatically create multimodal image-text fragments derived from an unstructured input content tailored towards a target need. The proposed approach aligns the fragment to the target need both in terms of content as well as style. With the help of metric-based and human evaluations, we show the effectiveness of the proposed approach in generating multimodal fragments aligned to target needs while also capturing the information to be presented.

CCS CONCEPTS

• **Information systems** → *Multimedia and multimodal retrieval*; • **Computing methodologies** → *Causal reasoning and diagnostics*; *Image representations.*

KEYWORDS

Symbolic Concepts, Causal Models, Style Transfer, User Needs.

ACM Reference Format:
Gaurav Verma, Suryateja BV, Samagra Sharma, and Balaji Vasan Srinivasan. 2020. Generating Need-Adapted Multimodal Fragments. In *25th International Conference on Intelligent User Interfaces (IUI '20), March 17–20, 2020, Cagliari, Italy*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3377325.3377487

1 INTRODUCTION

Multimodal communication lies at the heart of human interactions. While the multiple modalities in communication can refer to both the medium of communication (e.g., touch, voice) and the nature of the content (e.g., visual, textual), this paper primarily deals with the latter. Such multimodal content manifests itself in a myriad of complicated combinations of image, text, videos, etc. While interacting with the Web, we come across several instances of such multimodal fragments that are image-text combinations acting as pointers to web blogs, news articles, product landing pages, etc. Such fragments feature in home pages, social media posts (Facebook, Twitter, Instagram, etc.) or online advertisements. Studies [12, 43] have shown the affinity of humans towards such multimodal content, which has also been exploited to create compelling and engaging experiences.

The popularity of multimodal content can be attributed to the effectiveness in communication achieved from individual content modalities. Take for instance the advertisement in Figure 1 from [33]. The statement "*Not everyone who drives drunk dies*" on its own might sound like a promotion for drunk driving. However, in the context of the whole advertisement, the meaning is more impactful; bringing out the horrible consequences of drunk driving. This is an example of **meaning multiplication with multimodal content** [5] – referring to the creation of new meaning by integrating the meaning from image and text modalities, that is not clear in the absence of either modality. Figure 1 demonstrates the impact and power of coordinated messages conveyed via multiple content modalities enabling a deeper meaning.

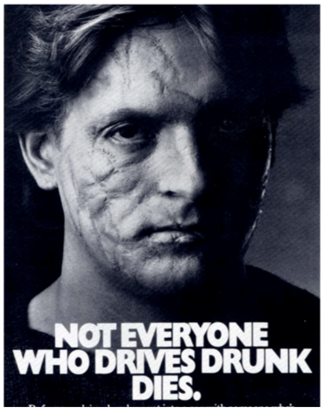
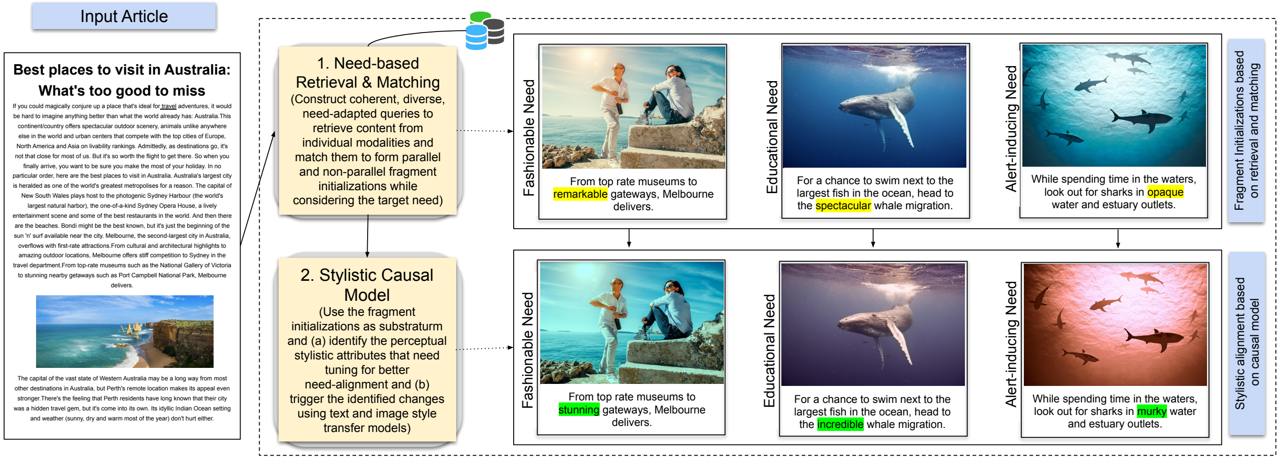


Figure 1: *Meaning Multiplication* with multimodal content: the caption *"Not everyone who drives drunk dies"* along with the image conveys an impactful message about drunk driving. [33].

IUI '20, March 17–20, 2020, Cagliari, Italy

ACM ISBN 978-1-4503-7118-6/20/03…$15.00
https://doi.org/10.1145/3377325.3377487


Input Article
1. Need-based Retrieval & Matching
2. Stylistic Causal Model
Fashionable Need
Educational Need
Alert-inducing Need
Fragment initializations based on retrieval and matching
Stylistic alignment based on causal model

Best places to visit in Australia: What's too good to miss

If you could magically conjure up a place that's ideal for travel adventures, it would be hard to imagine anything better than what the world already has: Australia.This continent/country offers spectacular outdoor scenery, animals unlike anywhere else in the world and urban centers that compete with the top cities of Europe, North America and Asia on livability rankings. Admittedly, as destinations go, it's not that close for most of us. But it's so worth the flight to get there. So when you finally arrive, you want to be sure you make the most of your holiday. In no particular order, here are the best places to visit in Australia. Australia's largest city is heralded as one of the world's greatest metropolises for a reason. The capital of New South Wales plays host to the photogenic Sydney Harbour (the world's largest natural harbor), the one-of-a-kind Sydney Opera House, a lively entertainment scene and some of the best restaurants in the world. And then there are the beaches. Bondi might be the best known, but it's just the beginning of the sun 'n' surf available near the city. Melbourne, the second-largest city in Australia, overflows with first-rate attractions.From cultural and architectural highlights to amazing outdoor locations, Melbourne offers stiff competition to Sydney in the travel department.From top-rate museums such as the National Gallery of Victoria to stunning nearby getaways such as Port Campbell National Park, Melbourne delivers.

The capital of the vast state of Western Australia may be a long way from most other destinations in Australia, but Perth's remote location makes its appeal even stronger.There's the feeling that Perth residents have long known that their city was a hidden travel gem, but it's come into its own. Its idyllic Indian Ocean setting and weather (sunny, dry and warm most of the year) don't hurt either.

(Construct coherent, diverse, need-adapted queries to retrieve content from individual modalities and match them to form parallel and non-parallel fragment initializations while considering the target need)

(Use the fragment initializations as substratum and (a) identify the perceptual stylistic attributes that need tuning for better need-alignment and (b) trigger the identified changes using text and image style transfer models)

From top rate museums to remarkable gateways, Melbourne delivers.

For a chance to swim next to the largest fish in the ocean, head to the spectacular whale migration.

While spending time in the waters, look out for sharks in opaque water and estuary outlets.

From top rate museums to stunning gateways, Melbourne delivers.

For a chance to swim next to the largest fish in the ocean, head to the incredible whale migration.

While spending time in the waters, look out for sharks in murky water and estuary outlets.

Figure 2: An overview of our proposed approach: our method takes an unstructured article as input and generates need-adapted multimodal fragments using a *Need-based Retrieval & Matching (NRM)* and *Stylistic Causal Model (SCM) based alignment.* While *NRM* caters to content-based aspects in the fragments, *SCM* handles finer need-alignment by triggering changes in perceptual stylistic attributes.

Existing literature in multimodal content [4, 14, 28, 41] focuses on understanding the literal connections between different content modalities such as describing objects and their spatial relationships. However, as seen from Figure 1, understanding the indirect relationships (irony, metaphor, symbolism, etc.) is key to leverage the power of multiple modalities and is a nascent area of exploration in multimodal content understanding. To this end, we take a step towards understanding the non-literal relations between text and image modalities, one of the common multimodal content manifestation, with an aim to generate multimodal fragments starting from an unstructured collection of content.

There are several scenarios requiring literal and non-literal understanding towards constructing a multimodal fragment. For example, consider a travel blogger who has a set of notes and images from her recent travel. Such an unstructured content can be used to create a "card" for her home page with a fun fact about the place to kindle interest among users looking for more *information* about it. She can use it to *alert* her Twitter followers about precautions while travelling. A hotel she stayed during her trip might want to use a *fashionable* variant of the information around their brand for their endorsements/advertisements. While all these combinations can be constructed from her notes, each of these is aligned to a unique '*need*' that the traveller wants to cater to with the content.

Such needs in multimodal content can come from the objective of the writer or the requirements of the reader and dictate both the choice of content (concepts) featuring in the output fragment and the style with which the content is expressed. Figure 2 illustrates the same idea where a common article is represented by different multimodal fragments catering to different needs. Creating multimodal content variants catered to such target needs puts a heavy creative load on the author in finding appropriate images and texts that go along well *and* satisfy various target needs. Such a process is both time consuming and exhausting if it has to be scaled for different content, scenarios and needs. This calls for automating the process of creating multimodal fragments from an unstructured image and text content (e.g., travel notes).

Our algorithm takes an unstructured content (like the traveler's notes and images) along with a target need and creates several fragment variants that are tailored to the target need with respect to both the information presented as well as the style of the content presented. As shown in Figure 2, our proposed system consists of two key components: *(a)* ***Need-based Retrieval & Matching*** (*NRM*), *(b)* ***Stylistic Causal Model*** (*SCM*). While the former component (*NRM*) is responsible for synthesizing several possible initializations of the multimodal fragments, the latter (*SCM*) is responsible to identify and improve upon the perceptual attributes of the synthesized fragments by inducing stylistic changes with the help of state-of-the-art style transfer models. Figure 2 shows the output of the need-based retrieval and stylistic causal models for a given input. We briefly describe the two parts, namely *NRM* and *SCM*, below and elaborate in Section 3.1 and 3.2.

(1) Need-based Retrieval and Matching (*NRM*): A need tailored retrieval framework that creates seed fragment variants that simultaneously account for *(a)* the relevance to the input content, *(b)* the relationship between content modalities (parallel and divergent – as defined in marketing science [5]), and *(c)* their suitability for a target need. The retrieval creates several possible initializations of the multimodal fragments accounting for the content relevance to the input article and symbolic connections of different concepts in the fragment to the target user need. Different concepts symbolically enhance the relationship to a target need, e.g. 'wreck' or 'danger' often symbolizes the advertising need of inducing 'alert'. The presence of such concepts symbolizing the target need enhances the alignment of the fragment to the need. Parallelism and non-parallelism are key to the impact of multimodal content – since the former amplifies the message conveyed and the latter deepens the impact of the fragment.

(2) Stylistic Causal Model (*SCM*): A need-adapted causal framework that connects different target needs to stylistic perceptual attributes expressed across different content modalities yielding the stylistic aspects of the content (along with their extent of contribution) towards meeting a target need. This model allows for an adaptation of the stylistic attributes of the fragments to ensure better alignment with the target need leveraging existing unimodal style-transfer models.

2 RELATED WORK

While there is a large volume of work in generating variants of a specific content modality (i.e., in our context, either only text or image) – explorations are scant in coordinated multimodal content generation, which is our core contribution in this paper.

Single Modality Content Generation: There has been a surge in image generation and style transfer techniques with the advent of Generative Adversarial Networks (GANs) [20] – photo-realistic image generation [6], photo realistic style transfer [26, 51], 3D image generation [13, 35, 49, 64], etc. Perceptual attributes in images such as virality, aesthetics, memorability are being modelled [2, 25, 27] and utilized for style-transfer as well. Inspired from style transfer in images, several techniques [9, 19, 30, 45, 63, 65] have been proposed to achieve generation of particular style in textual modality such as generating humorous puns [48], satirical news headings [55], and romantic caption variation [17]. These methods typically rely on learning the properties of certain lexicons that invoke particular styles to streamline content towards the target styles [23, 57] while retaining the core information in the content.

A few explorations [8, 10, 16, 18, 21, 22] aim to understand relationships between various modalities and utilize them towards cross-modal translation – i.e., generating content in a given modality from seed-content in a different modality. Models such as Attention GAN [59] and GILT [14] generate images based on textual descriptions. Image captioning models such as DenseCap [28] – an entity-aware caption generation model, generate text from images. Multimodal VAEs [58] have been explored to learn joint embeddings on image-label pairs to generate an image given label.

A key shortcoming of all these approaches, given our problem setting, is that they generate content in a single modality and cannot be used for generating coherent multimodal content – which is crucial for creating a multimodal fragment. Moreover, it is unclear how each of the stylistic attributes yield towards achieving the target need. To alleviate this, we propose a stylistic causal model in our approach that determines the changes that should be made to perceptual attributes of the fragment (i.e., style of various content modalities) in order to ensure better alignment with the target-need and utilize these unimodal engines to enhance the variants.

Multimodal Generation: While work on the simultaneous generation of multiple modalities is limited, there have been a few explorations on multimodal summarization [50, 53], which generate a multimodal summary of a multimodal input. Zhu et al. [66] introduced a multimodal corpus for this task and have developed an attention-based framework to produce image-text pairs as a summary. While such a summary can be one of the variants for an input text, it cannot cater across all the needs of a consumer/author. Our proposed technology includes a need-based retrieval and stylistic-adaptation to ensure informational and stylistic alignment to the target need – which is different from the task of summarization.

Modeling Cross-modal Relationships: Recent efforts towards learning common multimodal embeddings for image and text, for instance VSE++ [15] and the work by Sikka et al. [47], have been found to be effective for modeling content similarities across modalities. In the advertising literature, there exist a large volume of studies that aim to model the effect and interaction of multiple content modalities. Bateman et al. [5] classify multimodal fragments as parallel and non-parallel based on the interplay between the semantics of the different modalities. They identify parallel fragments as those where different modalities contribute towards the same meaning. Non-parallel fragments are either additive or divergent; in additive fragments content modalities amplify the meaning while divergent fragments have modalities conveying different meanings combining towards the larger fragment purpose. Taking inspiration from this prevalent idea in advertising literature, we use the common embeddings along with visual and textual concepts to initialize different parallel and non-parallel variants of the fragments.

Multimodal datasets: Given the recent interest in problems around multimodal content, researchers have curated several interesting datasets with multiple content modalities. RecipeQA [60] and Recipe1M+ [40] contain a sequential recipe description along with corresponding images; Zalando Fashion corpus [34] contains an image of a cloth along with its title and description (in German); Kruk et al. [32] introduce a dataset of Instagram posts that can be used to determine authorial intent from multimodal content. The Pitts Ads dataset [24] introduced ads along with their metadata like slogans, sentiments, topics. These varied datasets have propelled studies into how images and texts interact – such as Alikhani et al.'s [3] work on discourse relations based off RecipeQA dataset [60]. Zhang et al. [62] have modelled notions of parallelism between images and text and Ye et al. [61] have built common visual semantic embeddings from the Pitts Ad dataset [24]. We use the Pitts Ads dataset to establish symbolic mapping between content and target-needs and learn stylistic causal modeling, as it contains highly semantic relationships between images and texts *while* serving a *need.*

3 MULTIMODAL FRAGMENT GENERATION

A multimodal fragment comprises both the content and style elements. We use a conceptual map and a need-based ranker to account for the content preferences in the variant and ensure informational alignment of the fragment to the need. We use a stylistic causal model coupled with the generation engines to ensure stylistic alignment of the fragments to the need.

Given the input content to create the multimodal fragment variants, we start with retrieving text segments and images that are relevant to the input content based on the information overlap between the candidates and the input. The retrieval can be on the entire corpus or from a limited subset. For every candidate text or image retrieved, we use a combination of relevance to the input and appropriateness to the target need to identify the top text or image candidates. To model the content appropriateness to a target need, we need a corpus of content and their alignment to a target need. We use the Pitts Ads dataset [24] to learn a symbolic map of different concepts in the Pitts Ads dataset and various target user needs to achieve an understanding of a concept beyond its semantics; for example, 'flower' is a concept that symbolizes 'beauty' and

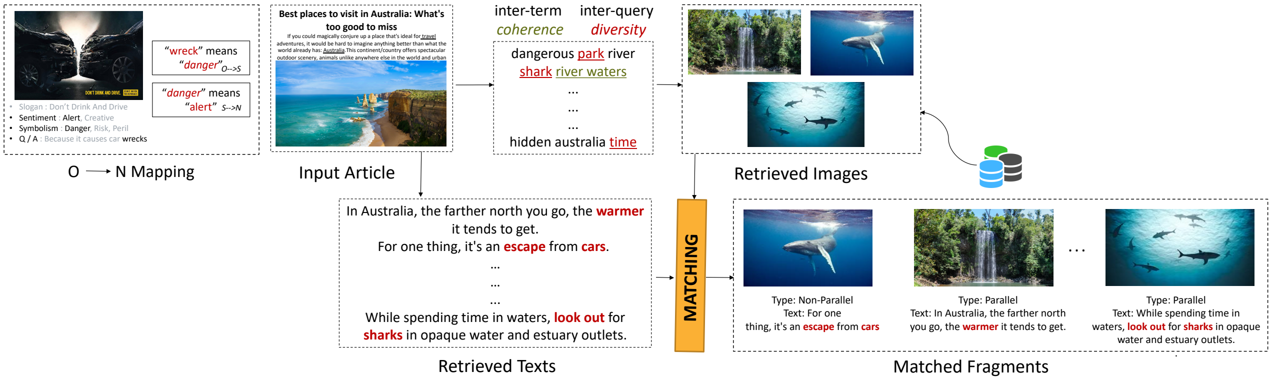

DON'T DRINK AND DRIVE
"wreck" means "danger"O-->S
"danger" means "alert" S-->N
Slogan : Don't Drink And Drive
Sentiment : Alert, Creative
Symbolism : Danger, Risk, Peril
Q / A : Because it causes car wrecks
O → N Mapping
Best places to visit in Australia: What's too good to miss
If you could magically conjure up a place that's ideal for travel adventures, it would be hard to imagine anything better than what the world already has: Australia.This continent/country offers spectacular outdoor scenery, animals unlike anywhere else in the world and urban
Input Article
inter-term coherence
inter-query diversity
dangerous park river
shark river waters
...
hidden australia time
Retrieved Images
In Australia, the farther north you go, the warmer it tends to get.
For one thing, it's an escape from cars.
...
While spending time in waters, look out for sharks in opaque water and estuary outlets.
Retrieved Texts
MATCHING
Type: Non-Parallel
Text: For one thing, it's an escape from cars
Type: Parallel
Text: In Australia, the farther north you go, the warmer it tends to get.
Type: Parallel
Text: While spending time in waters, look out for sharks in opaque water and estuary outlets.
Matched Fragments

Figure 3: An overview of the need-based retrieval and matching (NRM) method to obtain several need-adapted initializations.

'delicateness'. As shown in Figure 3, we extract the concepts from the images and text in a content and compute their frequencies in fragments that have historically satisfied the target need. This mapping captures the concepts that symbolize the target need as per the underlying historical data, for a target need. We use this mapping along with the relevance to retrieve relevant images/text that are also informationally aligned to the target need.

We match the retrieved images and texts in a common visual semantic embedding space [15] using the cosine similarity between the embedding to find the content overlap between the different modalities. For every such pair – we evaluate the parallel-ness and non-parallel-ness of the content to yield several initializations for the fragment variants. From the Pitts Ads dataset [24], we have considered three needs that are popular within advertising domain – educational, alert-inducing, and fashionable [24, 62].

We further use the Pitts Ads dataset to capture the relationships between stylistic perceptual attributes of the multimodal fragment and the target user needs using a causal deconfounding model [54]. For instance, our model indicates that the 'happiness' of an image makes it more fashionable and is a causal attribute. We extract several such attributes from each of the content modalities – e.g., aesthetics, memorability, happiness, gloominess, scariness of an image; emotions from texts. For every target need, we deconfound the effect of these attributes on the target need to identify top stylistic attributes that causally impact the target need.

As shown in Figure 5, for every fragment initialized as a parallel or non-parallel variant based on the retrieved content, we compute the attribute vector and use the stylistic causal model to arrive at an expected attribute vector. We compute the difference between these two attributes vectors and trigger appropriate generations to improve the identified style attribute using state-of-the-art style transfer models. Finally, the generated fragments are ranked using our conceptual map and the causal model to identify the most appropriate fragment for a target need. To summarize, our solution consists of 2 major parts:

(1) **Need-based Retrieval and Matching (*NRM*) (Fig. 3)**: A conceptual map-based model indicating the content preferences of a target need and a need-based ranker that utilizes this map to retrieve need and content specific content to initialize multimodal combinations accounting for symbolic relevance to the target need.

(2) **Stylistic Causal Modeling (*SCM*) (Fig. 5)**: A causal model capturing the stylistic preferences of a target need and a generation engine that uses the causal model to improve specific stylistic attributes of the fragment to better align with the target need.

In the subsequent subsections, we elaborate on the details of the two aforementioned components of our proposed method.

3.1 Need-based Retrieval and Matching (*NRM*)

Figure 3 shows an overview of our approach for Need-based Retrieval and Matching (*NRM*). The Pitts Ads Dataset [24] contains 13, 938 image-text pairs across different needs – with 53 different symbols. We use this to learn symbolic preferences between concrete concepts in an advertisement comprising of image and text , and the target needs. For example, the concept 'wreck' symbolizes 'danger', which in turn relates to 'Alert'. We use this mapping to understand the symbolic need association of various concepts. In our fragment, we prefer content related to such concepts that are better associated to the target need. For example, choosing images and text that have concepts related to *'wreck'* will aid in getting better initializations for a fragment catering to *'alert-inducing'* need.

We first extract various objects referred in the ad images (using [28]) and text (using entities in QA pairs and slogans in the meta data). We then learn two mappings – object-to-symbol and symbol-to-need. For this, we compute average similarity between 53 symbols and various objects extracted from the ad images using the embeddings from [61] resulting in an object-symbol mapping $O \longrightarrow S$. Symbol-need mapping $S \longrightarrow N$ is computed in the form of a frequency table over the 13, 938 ads. Multiplying the 2 mappings yields an $O \longrightarrow N$ mapping from different objects to multiple needs with appropriate weights. We use this concept preference in tandem with the relevance to the input content in our retrieval stage to extract need and input relevant content to initialize the fragments.

3.1.1 *Need-aware Retrieval.*

Given an input collection of content, the first step in our solution is to retrieve relevant image/text units that can be part of the target fragments. We independently extract the text units and images relevant to the input content.

For the text, we encode the sentences using a sentence encoder [1] and compute the similarity of the sentences to the average sentence embedding of the input content. Additionally, we introduce a notion of need vectors to semantically represent the user's need.

These vectors are calculated as the average of the embeddings of the symbols and objects for each need from the Pitts Ads Dataset. For every sentence, the average similarity of a sentence to the input text and the target need is used to rank the most appropriate representative sentences for each need.

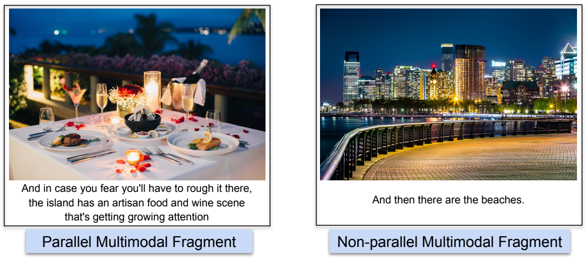
For the images, we extract keywords based on their frequency in the input content and extract the top-n keywords to construct the query. The choice of top-n keywords is biased by giving relatively higher importance to words that symbolize the need (based on the mapping learned above). We construct a query using a combination of k ($k < n$) words from the extracted words – such that the inter-term coherence and inter-query diversity of the query is high. Inter-term coherence ensures that the terms within a query are not semantically too different ("beautiful laptop award" is bad in terms of inter-term coherence), while inter-query diversity ensures that two queries are sufficiently different ("beautiful young woman" and "beautiful young lady" are not very diverse queries). We first choose a total of k words which are a combination of words in the input content that are often used in some symbolic sense and frequent words in the corpus (based on tf-scores). We then construct potential queries by combining these words into m word combinations for a total of ${}^{k}C_{m}$ combinations. Of these combinations, the top 20 queries that have high inter-word similarity [42] (i.e., query-coherence score) are retained. To ensure inter-query diversity we choose queries from the set of coherent queries that have highest Levenshtein score (a quantification of character-level edit distance between two strings). Explicitly incorporating measures for diversity ensures that retrieved images are diverse and there is a larger pool to choose from in the later stages of the pipeline.

3.1.2 Multimodal Matching. Once we have retrieved representative images and sentences, we combine them into different fragment initializations. Following the formulation of Zhang et al. [62], we classify fragments into two classes – parallel and non-parallel, while preserving relevance to input content and alignment to target needs. Parallel pairs are those pairs that 'convey the same meaning when presented in individual modalities', and non-parallel pairs are those that offer 'different unrelated meanings in individual modalities'.

We rank the image-text pairs, on a parallel to non-parallel spectrum, by sorting on the basis of similarity between VSE++ embeddings [15] of an image and a sentence. However, a mere ranking on parallelism does not account for user needs and relevance to input. Therefore, along with parallelism, we also incorporate a need-based relevance score for ranking of the fragments, to ensure that the fragments are well representative of the article and well-suited to the target need. We define a net relevance score, R_{net} as a combination of (a) parallel-ness and non-parallel-ness score ($R_{p/np}$), (b) relevance of the text and image to the input content (R_t and R_i respectively) and (c) how strongly the fragment demonstrates affiliation to a certain need (N_t and N_i respectively), given by,

$$R_{net} = \alpha |R_{p/np}| + (1 - \alpha) \frac{R_t \|N_t\| + R_i \|N_i\|}{2}$$

We rank all possible pairs using the above net relevance score, R_{net}. Here, $R_{p/np}$ is computed using cosine similarity between image and text embeddings in the Visual Semantic Embedding space [15], R_t is the relevance of retrieved text computed using cosine similarity between the current sentence embedding and the input document embedding, R_i is the relevance of retrieved image computed using cosine similarity between image tag embeddings and the input document embedding, N_t and N_i are need vectors of the text and image respectively, α is a hyperparameter with value in the range $[0, 1]$ which acts as the weight we give to our notions of parallelness or need-based relevance. In our experiments, we have chosen α to be 0.7, biasing towards need-based relevance.


And in case you fear you'll have to rough it there, the island has an artisan food and wine scene that's getting growing attention

And then there are the beaches.
Parallel Multimodal Fragment
Non-parallel Multimodal Fragment

Figure 4: Example of parallel (left) and non-parallel (right) multimodal fragments as synthesized by the multimodal matching part of our need-based retrieval method.

We select fragments that have a high relevance score and also ensure diversity by picking only unique pairs. In particular, we pick 3 parallel ($R_{p/np} > 0$) and 3 non-parallel ($R_{p/np} < 0$) pairs with high net relevance score while enforcing diversity. Figure 4 illustrates two fragment initializations obtained using *NRM*, and their classification as parallel and non-parallel. We take an average of the need vectors N_t and N_i and pick the need that is maximally satisfied. At the end of this stage, we have multimodal fragments that are not only relevant to the article but also informationally aligned to target needs.

3.2 Stylistic Causal Model (*SCM*)

Figure 5 shows an overview of our causal models to achieve stylistic alignment of the fragments towards the target needs. The obtained matched pairs might lack in certain stylistic attributes that can better cater to the target need. That is, while the content remains the same, the style (or the perceptual attributes) of a fragment can be modified to better suit the target need under consideration – a bright image is more elating than a dull one, while the content of the image remains the same; formal text establishes more credulity for a communication than informal text. We extract six image attributes, namely - Aesthetics, Memorability, Happiness, Scariness, Gloominess, Peacefulness; and use ad slogans annotated by humans to compute 14 text attributes, namely Anger, Anticipation, Disgust, Fear, Joy, Sadness, Surprise, Trust, Specificity, Sentiment, Formality, Colloquial-ness, Objectivity, and Concreteness using state-of-the-art scorers [1, 7, 29, 31, 46, 56]. To capture the stylistic preferences, we model the relations between the extracted attributes and the needs using a causal deconfounding model [54].

3.2.1 Causal Modeling of Stylistic Attributes. There has been an abundance of research that aims to leverage supervised learning approaches that range from simple logistic regression models to more sophisticated deep neural networks in order to establish correlations between the observed features of the content ($\mathbf{x}$) and the target variables (y) that quantify target goals. It is well known that causations cannot be inferred from correlations – the strongly correlated features may not necessarily be the ones that *influence* the

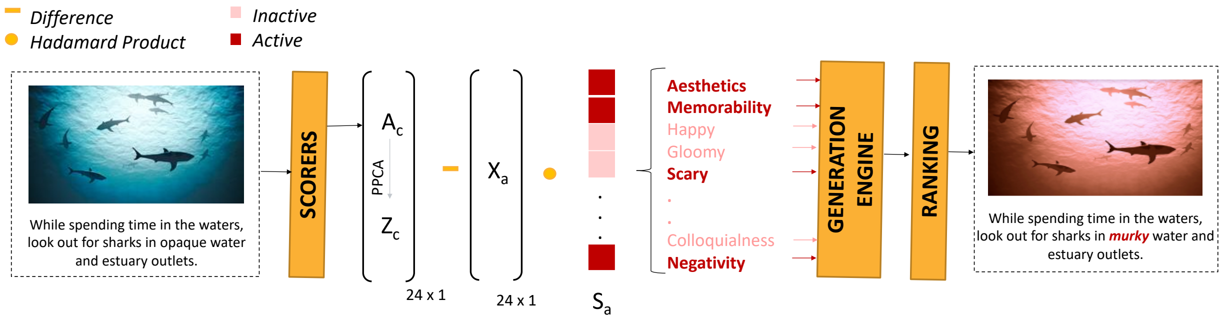

Difference
Hadamard Product
Inactive
Active
While spending time in the waters, look out for sharks in opaque water and estuary outlets.
SCORERS
A_c
PPCA
Z_c
24 x 1
X_a
24 x 1
S_a
Aesthetics
Memorability
Happy
Gloomy
Scary
Colloquialness
Negativity
GENERATION ENGINE
RANKING
While spending time in the waters, look out for sharks in murky water and estuary outlets.

Figure 5: An overview of the stylistic causal model that takes the output of need-based retrieval and matching, and triggers style transfer models to make changes in stylistic perceptual attributes of individual modalities.

target variables [54]; they may indirectly do so by being jointly influenced by an unobserved confounder $\mathbf{z}$ – a variable that influences both $\mathbf{x}$ and y. Given this, there have been recent efforts that enable causal inference in a multiple-cause setting, by inferring confounders $\mathbf{z}$ [54]. We leverage this progress to align the process of multimodal fragment creation with need to identify causal stylistic elements that influence a target need.

Let $\mathbf{X}$ be the matrix of stylistic attributes extracted from the entire dataset and Y denote the matrix of the need that the corresponding ad caters to. Given $\mathbf{X}$ and Y we can build three logistic regression models for each need capturing $p\left(Y_{n,i} = 1\right) = \sigma\left(W_n^T \mathbf{X_i}\right)$, where n indicates the need, $\sigma(.)$ is the sigmoid function, and $p(Y_{n,i} = 1)$ represents the probability that the ith input attribute is causally significant for the target need. Once such models have been trained, we can now use the learned weight matrices W_a, W_e, W_f to determine correlations between the various attributes and the target need – alert, educational and fashionable respectively. However, the effect of these attributes might be confounded by an unobserved variable that jointly effects the overall need.

To overcome the issue of confounded attributes, we use the Causal Deconfounder model [54] and learn a latent matrix $\mathbf{Z}$ from $\mathbf{X}$ using the standard PPCA model [52] and apply logistic regression on the concatenated matrix $[\mathbf{X} \oplus \mathbf{Z}]$. The latent matrix $\mathbf{Z}$ acts as a substitute confounder, thus eliminating any other implicit confounding effects. We found experimentally that using a latent dimension of 4 gave the best results. Since the confounding effects are eliminated, all the remaining relations between attributes and target need are not mere correlations but causations.

The Causal Deconfounding model helps in establishing causal relations between the stylistic attributes towards a certain target need as well as their extent. Following [54], we consider that there is a causal influence if $p - value$ is less than 0.05 in the learned models. We prepare a causal signal matrix $\mathbf{S} = [S_a S_e S_f]$ between needs and style attributes, with each entry being 0 if there is no causal signal, else taking the causal mean value. We leverage these causal relations to modify the image or text attributes in order to attain better stylistic alignment with the target need.

3.2.2 *Stylistic Attribute Identification.*

For every matched fragment, which could be either classified as parallel or non-parallel, we compute the attribute vector $\mathbf{A}_c$ using various scorers [1, 7, 29, 31, 46, 56]. We concatenate the learned substitute confounder $\mathbf{Z}_c$ with the attribute vector $\mathbf{A}_c$ to form $\mathbf{X}_c$,

$$\mathbf{X}_c = [\mathbf{A}_c \oplus \mathbf{Z}_c]$$

We then use the causal signal matrix $\mathbf{S}$ from our learned causal model to arrive at what attributes to improve upon for the top need that a pair satisfies. The difference between desired attribute vector, which is $\mathbf{X_n} = (\mathbf{W_n}^+)^T Y_n$ (where $Y_a = [1, 0, 0]^T$, for example), and the actual attribute vector $\mathbf{X}_c$, is multiplied point-wise with S_n, and is used to identify the attributes that should be tuned in order to achieve better alignment with the target need Y_n. Let this set of attributes be $A = \{a_1, a_2, \ldots, a_k\}$. Based on the attributes determined to be improved by our causal model, we apply style-transfer for the particular attribute in the corresponding image or text component to improve the attribute in the final variant. If there are multiple attributes identified by our causal model, we take the top attributes of each content modality.

For example, suppose the causal model indicates the improvement of aesthetics, memorability and happiness in image, and positivity in text to better tailor to a target need. Since simultaneous style attribute transfer is not feasible with current state-of-the-art techniques [46], we transfer <aesthetics of image and positivity of text>, <memorability of image and positivity of text> and <happiness of image and positivity of text> to create 3 variants. Let $X_{c,a}$, $X_{c,m}$ and $X_{c,h}$ be the values for the aesthetics, memorability, and happiness improved fragments respectively along with text positivity. We then use the stylistic causal predictor W_n learned in Subsection 3.2.1, and apply it on these attribute vectors to rank which variation satisfies target need n the most. For each of these variants, the predictor score for the target need is computed from the trained model which can be used to identify the best combination among these or rank them. Effectively, we carry out the following for a target need n:

$$y_n = \sigma\left\{(\mathbf{W_n}^+)^T \mathbf{X}_c\right\}$$

$$\left\{y_{n,a}, y_{n,m}, y_{n,h}\right\} = \sigma\left((\mathbf{W_n}^+)^T \left\{X_{c,a}, X_{c,m}, X_{c,h}\right\}\right)$$

$$\theta = argmax\left(y_{n,a} - y_n, y_{n,m} - y_n, y_{n,h} - y_n\right)$$

and pick the attribute θ which maximally improves on the need score y_n of our given fragment. More generally, for every variant improved on the identified attribute, we extract their attributes $\mathbf{X}_c$ and compute their need score $(\mathbf{W_n}^+)^T X_c$ based on the trained models $\mathbf{W_n}$ for the target need n, and pick the variant that maximally improves on the target need or use it to rank the variant.

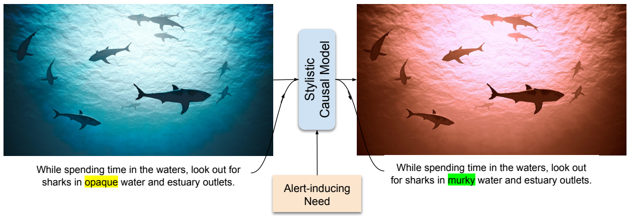

Stylistic Causal Model
While spending time in the waters, look out for sharks in opaque water and estuary outlets.
Alert-inducing Need
While spending time in the waters, look out for sharks in murky water and estuary outlets.

Figure 6: Stylistic Causal Model identifies the perceptual stylistic attributes, in both the image and the text, that need to be changed for better target-need alignment and trigger style transfer models.

3.2.3 Image & Text Style Transfer. We leverage the Fast Photo Style Transfer framework [36] and Flair Contextual Word Embeddings [1] to perform the image and text style transfers respectively. The style images are learned using Deviant Art Dataset [44, 46]. Note that the selection of the stylistic attributes is based on the availability of models to transfer the style of an identified attribute. Since that is not the core of our innovation here, we have limited our model to the attributes above. For an identified attribute θ we apply the corresponding style transfer model. Since the scorer is independent of the transfer models, we compute the attribute scores of the transferred and the original content to address any discrepancies of these transfer model. Fig 6 shows an example where a fragment from *NRM* is stylistically aligned to alert-inducing need via *SCM*.

4 METRIC BASED EVALUATION

We evaluate the proposed framework on the corpus of articles in the multimodal summarization dataset [66] – which contains articles from Daily Mail corpus and human curated summary units. We take the article (along with the images) to be input to our system.

As mentioned before, we use the Ads corpus in [24] as our corpus of multimodal fragments to learn the symbolic correlations and causal attributes for the target needs. The dataset categorizes the ads units into 10 classes – alert-inducing, fashionable, educational, eagerness, creative, activeness, amused, inspirational, youthful, empathetic – depending on their end-goal as perceived by human annotators. We use three of these classes as the target user needs in our experiments – educational, alert-inducing, and fashionable .

For modelling style relationship to target need, we extract 6 attributes in images – Aesthetics, Memorability, Happiness, Scariness, Gloominess, Peacefulness and 14 attributes in text – Anger, Anticipation, Disgust, Fear, Joy, Sadness, Surprise, Trust, Specificity, Sentiment, Formality, Colloquial-ness, Objectivity, Concreteness. We build the causal model of style attributes on these 20 aspects on the Ads dataset for every need – yielding a causal weighting of these attributes that cause/aid in achieving the target need.

Table 1 shows the performance of the proposed approach in generating the fragments. In the summarization literature [37, 38], Rouge scores are an indicator of how the summary is capturing the information in the article. Here we report the Rouge-1, Rouge-2 and Rouge-L scores to show how the generated fragments capture the original article. High Rouge scores of our approaches indicate that we do not compromise on the information representation. While achieving this, we are able to get a diverse set of fragment variants (measured by the diversity in concepts across different fragments based on the score in [11, 39]) and also satisfying the target need. We use the following formulation to evaluate inter-fragment diversity in a set of fragments generated using a particular method.

$$S_{div} = 1 - \sum \frac{sim(I_i, I_j) + sim(T_i, T_j)}{2n(n-1)}; \qquad (1)$$
$$(i, j) \in \{1, \ldots, n\} \times \{1, \ldots, n\} : i \neq j$$

For calculating the similarity between I_i and I_j we use the cosine similarity between their VSE++ image embeddings [15]. For calculating the similarity between T_i and T_j we use to cosine similarity between their sentence encoder embeddings [1].

For quantifying the content relevance between fragment-text and article, we use cosine similarity between sentence encoder embedding and the document embedding of the input article (as discussed earlier in Section 3.1.1). Similarly, for relevance between fragment-image and the input article, we evaluate the cosine similarity between VSE++ embeddings of images in article and the ones chosen by our method (as discussed in Section 3.1.2 earlier). We evaluate the quality of retrieved sentences in terms of their ability to capture information of the input content with respect to the ground truth summaries provided by human annotators [66]. As it can be inferred, the proposed technology retrieves and modifies sentences without a hefty loss of information with respect to ground truth summary; this is further substantiated by reasonable values of fragment-text's relevance with input article. While the images used in retrieved and final fragments are not very similar to those used in the content, qualitative analysis suggests that they are representative to a good extent – a part of this may be attributed to the fact that our generation explicitly aims to generate non-parallel fragments which are found to be crucial in marketing literature – this involves a mild compromise on the relevance.

It is interesting to note that the relevance to target need is significantly higher in comparison to ground truth summary units – this establishes the efficacy of need-alignment capabilities of our proposed model. It is also encouraging that the proposed technique also improves the content relevance of fragment-text to input article (from 0.74 to 0.76) while improving the need-alignment of the fragment. Finally, unlike the human-generated multimodal summaries, the proposed method generates a diverse set of fragments and hence can be useful in several scenarios which are discussed in Section 6 below.

5 HUMAN EVALUATION

In this section, we discuss the extensive human evaluations of the proposed methodology, as well as their results. While Table 2 serves as a reference, we present the associated analysis and deeper insights in the following subsections. A representative diagram that illustrates the two annotation experiments conducted to evaluate the proposed methodology is presented in Figure 7.

5.1 Need-Alignment of Multimodal Fragments

Ensuring that the generated fragments are aligned to the target needs is a crucial component of our proposed methodology. While the metrics in Table 1 present an objective measure for this, it is important to establish the same with human evaluations, given the subjective interpretation of target needs. To this end, we randomly selected a subset of 50 articles and ask the Amazon Mechanical Turk (MTurk) annotators to first read the article and then compare the generated multimodal fragments with respect to their alignment

	Similarity to Textual Summary			Relevance to article		Relevance to need	Diversity
Model	*(Rouge-1)*	*(Rouge-2)*	*(Rouge-L)*	*(image)*	*(text)*	*(fragment)*	*(fragment)*
Human	1.0^*	1.0^*	1.0^*	0.53	1.0^*	0.33	0.36
NRM	**0.55**	**0.28**	**0.33**	0.74	**0.45**	0.68	0.61
NRM + SCM	**0.55**	0.23	**0.33**	**0.76**	0.11	**0.77**	**0.68**

Table 1: Metric-based evaluation of the summary units curated by human annotators in [66], retrieval based fragments from *NRM* and the ones generated by our entire approach (*NRM+SCM*). * denotes that the text in human-generated summary was used as the ground truth, and hence the metric value is 1.

	Need-alignment			Relevance to article			Fluency	Naturalness	Share?	Click?
Model	*(fragment)*	*(image)*	*(text)*	*(fragment)*	*(image)*	*(text)*	*(text)*	*(image)*	*(fragment)*	*(fragment)*
Human	3.16 ± 1.28	3.20 ± 1.23	3.11 ± 1.25	**3.63** ± 1.35	**3.69** ± 1.39	**3.61** ± 1.35	3.54 ± 1.23	**3.50** ± 1.11	3.19 ± 0.93	3.14 ± 1.08
NRM	3.26 ± 1.30	3.31 ± 1.24	3.21 ± 1.26	3.59 ± 1.33	3.68 ± 1.34	3.59 ± 1.33	**3.56** ± 1.24	3.31 ± 1.03	3.37 ± 0.95	3.32 ± 1.13
NRM + SCM	**3.34** ± 1.32	**3.58** ± 1.28	**3.37** ± 1.37	3.61 ± 1.34	3.66 ± 1.37	3.58 ± 1.33	3.53 ± 1.24	3.23 ± 0.98	**3.54** ± 0.99	**3.50** ± 1.16

Table 2: Human evaluation: We report mean values on 5-level Likert scale, and the corresponding standard deviations ($\mu \pm \sigma$). '$Human$' denotes the fragments generated by combining human curated summary units from [66], NRM denotes the fragments based on need-based retrieval model only, $NRM + SCM$ denotes the fragments synthesized by combining need-based retrieval model and stylistic causal model.

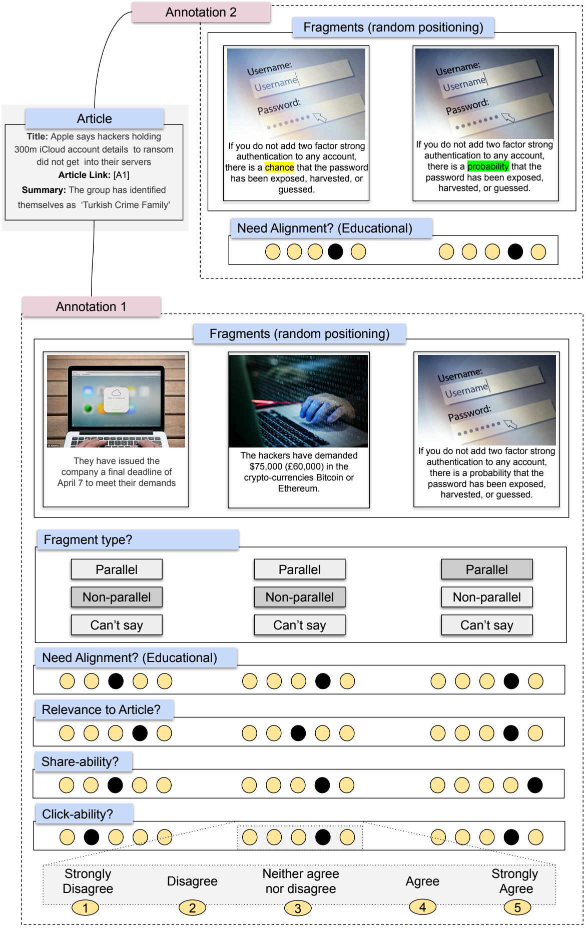
with a target need. To make sure that the annotators actually read and understood the article, we ask them to assess whether a given a statement is a valid summary of the article that they have just read. We only consider the responses of annotators who correctly identify the presented statement as a valid or invalid summary while reporting the results in Table 2.

The triplets of multimodal fragments – the curated summary units (*Human*), output of the need-based retrieval model (*NRM*), and output of the combined need-based retrieval and stylistic causal model (*NRM+SCM*); all representing the same article are compared against each other rated on a five-level Likert scale for their need alignment, by five annotators each for a given triplet of fragments. As presented in Table 2, the multimodal fragments generated by our proposed *NRM + SCM* has better need alignment than the ones synthesized by need-based retrieval model (*NRM*), as well as the summary units (*Human*). It is interesting to note that *NRM + SCM* improves on the output of *NRM* by making stylistic changes, which was a core component of our hypothesis. From Table 2, it can also be observed that both text as well as image – the individual modalities of our multimodal fragments, show better need alignment for *NRM + SCM* than for *NRM* and *Human*. To further substantiate the claim, we note that for a given article, 61.3% of the annotators prefer the fragments generated by *NRM + SCM* over those generated by *NRM* and *Human*, in terms of their need-alignment; while the corresponding percentage for individual modalities, i.e., image and text is 64.1% and 59.8%, respectively. All ratings showed moderate inter-annotator agreement – a Fleiss Kappa score of 0.58, 0.53, 0.55 for *fragment*, *image*, and *text*, respectively.

5.2 Relevance of Multimodal Fragments

While adapting to a target need, it is important to do so without compromising on the relevance of the generated multimodal fragments to the original article. To assess this, we ask the MTurk annotators to first read one article from the randomly chosen 50 articles, and then rate the relevance of three multimodal fragments on a five-level Likert scale. Like above, each triplet of multimodal fragments representing a common article was shown to five different annotators. We use the same summary-based filtering mechanism as above to filter out the responses of annotators who did not read or understand the article properly. From the results presented in Table 2, we can note that the relevance of fragments generated from summary units are only marginally better than those generated by *NRM* and *NRM + SCM*. In terms of fractions of annotators, 62.7% of the annotators preferred the fragments generated by either *NRM + SCM* or *NRM* over those generated by *Human* in terms of their relevance to the article; the corresponding share of annotators for individual modalities, i.e., image and text was 63.6% and 58.7%, respectively. This trend, when analyzed in conjunction with the trend pertaining to need-alignment discussed above, highlights that our proposed approach (*NRM + SCM*) generated need-adapted multimodal fragments without compromising extensively on the relevance to input article. On showing the text and image individually to the annotators, instead of the whole fragment, the inferred trends persist. The inter-annotator agreement for *fragment*, *image* and *text*, as quantified by Fleiss Kappa score, was 0.52, 0.49, 0.54, respectively; all signifying moderate agreement.

Apart from evaluating need-alignment and relevance, we also ask the annotators to rate the multimodal fragments generated by the three methods on aspects like fluency of the text, naturalness of the image, share-ability and click-ability of the multimodal fragment. While fluency of text and naturalness of images are subtle qualitative aspects of individual modalities, share-ability and click-ability are aspects that quantify how engaging a multimodal fragment is, as a whole. While share-ability relates to the tendency of annotators to share an article that is represented by given multimodal fragments, click-ability relates to their tendency of interacting with the fragments in order to access the original article. Together, these two cover the two aspects of interactions that are predominant on the Web – *encouraging* others to consume, and *consuming* itself. As we can note in Table 2, even though the fluency and naturalness of individual modalities for *NRM + SCM* is marginally lower than those for *NRM* and *Human*, the annotators have a considerably higher tendency to share and click on multimodal fragments generated by *NRM + SCM*, in comparison to that for *NRM* and *Human*. 67.2% of the annotators considered the outputs of *NRM + SCM* to be more share-able than the outputs of *NRM* and *Human*; while 64.8% considered them to be more click-able. It is again worth noting that *NRM + SCM* further *improves* the output of *NRM* on high-level engagement metrics. The Fleiss Kappa score for share-ability and



Title: Apple says hackers holding 300m iCloud account details to ransom did not get into their servers
Article Link: [A1]
Summary: The group has identified themselves as 'Turkish Crime Family'

If you do not add two factor strong authentication to any account, there is a chance that the password has been exposed, harvested, or guessed.

If you do not add two factor strong authentication to any account, there is a probability that the password has been exposed, harvested, or guessed.

They have issued the company a final deadline of April 7 to meet their demands

The hackers have demanded $75,000 (£60,000) in the crypto-currencies Bitcoin or Ethereum.

If you do not add two factor strong authentication to any account, there is a probability that the password has been exposed, harvested, or guessed.

Figure 7: Representative diagrams to illustrate the two human annotation experiments performed. The marked results are the actual actual responses of annotators. For the given instance of annotation experiment (AE) 1, the shown fragments are *Human*, *NRM*, and *NRM* + *SCM* from left to right; for AE 2, the shown fragments are *NRM* and *NRM* + *SCM* from left to right.

click-ability was found to be 0.62 and 0.67, respectively, signifying substantial agreement; for fluency of text and naturalness of images the same was 0.48 and 0.45, signifying moderate agreement.

5.3 Evaluating Parallel/Non-Parallel Matching

As described earlier, our proposed methodology comprises of two major components, wherein the stylistic causal model improves on the multimodal fragments synthesized by need-based retrieval by inducing stylistic changes that would ensure better alignment with target need. In Table 2, we note that even though *NRM* + *SCM* generates the most need-aligned multimodal fragments, *NRM* provides a concrete substratum by retrieving relevant and need-focused content in both the individual modalities. This observation is consistent not only in terms of need-alignment of individual modalities, but also across evaluation aspects like relevance to article and engagement metrics (i.e., share-ability and click-ability).

We have designed our need-based retrieval mechanism to address the advertising requirement of generating parallel and non-parallel multimodal fragments. To assess this capability of the proposed retrieval method, we ask the MTurk annotators to analyze 100 synthesized multimodal fragments, obtained from 50 randomly chosen distinct articles, and mark them as either a 'parallel fragment', a 'divergent fragment', or 'can't say'. Each fragment was analyzed and annotated by 5 different annotators. We note that the annotations of the human annotators were same as those provided by our need-based retrieval method 74.9% of the times. Of the annotations that did not agree with our labels (i.e., 25.1%), 64.8% were 'can't say' – implying that the annotators were themselves undecided. While the former observation reinforces the efficacy of the proposed method in correctly synthesizing parallel and divergent fragments, the latter affirms that a majority of the mis-labeled fragments are not absolutely incorrect, but only ambiguous.

5.4 Goal-directed Changes in Attributes

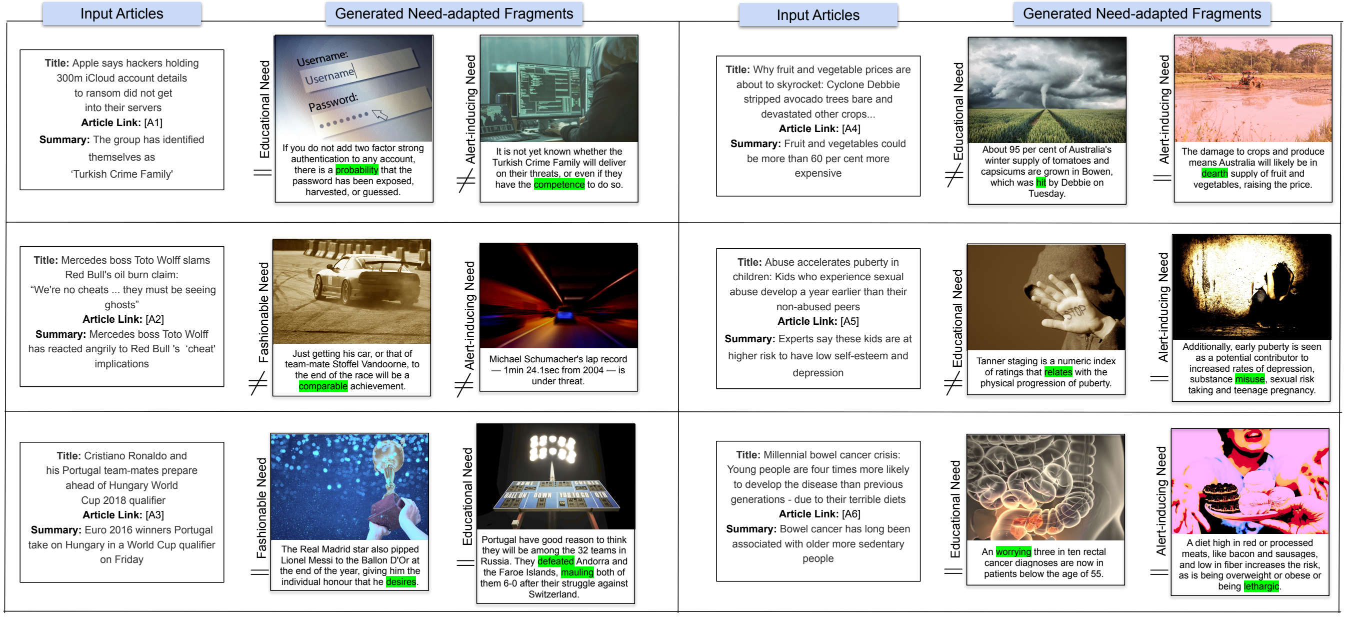
To assess the contribution of stylistic changes made to the initialized multimodal fragments, we asked the MTurk annotators to compare the need-alignment of the initialized fragment with that of the final output of the stylistic causal model. Specifically, the annotators were asked to provide a rating on the five-level Likert Scale, signifying their (dis)agreement with the following statement: the multimodal fragment towards the *left* shows *better* need alignment with a target need [1]. A total of 300 fragment pairs, obtained from 50 articles, were compared, each by 3 annotators. 44.3% of the annotators expressed moderate to strong agreement (i.e., a score of 4 or 5, respectively) indicating that they identified the stylistic changes, as facilitated by the causal model, to enable better need-alignment for the *Educational* need. The percentage of annotators who expressed moderate to strong agreement for *Alert* and *Fashionable* needs was 37.1% and 31.9%, respectively. Fleiss Kappa score was 0.51, which signifies moderate inter-annotator agreement.

There are two interesting aspects of evaluating the generated output of stylistic causal model against the synthesized output of the need-based retrieval model. Firstly, since the need-based retrieval explicitly retrieves content in individual modalities that takes the target need into account, it acts as a competitive baseline because the substratum for stylistic causal model to improve upon is already need-aligned to quite some extent (as it is also evident from Table 1). Given this, it is encouraging that a substantial number of annotators agree with the contribution of *SCM* towards target need-alignment. Secondly, the existing style transfer models which are used by our proposed stylistic causal model in an ad hoc manner have some well known shortcomings that might adversely affect the generated fragment. Given the purview of this work, we do not aim to address the shortcomings that pertain to style transfer models. Nonetheless, the human evaluations, along with the metric-based evaluations in Table 1 establish the role of stylistic causal model in further aligning the fragments towards a target need.

6 DISCUSSION

As evidenced by the metric-based results and the human evaluations, our method (i.e., *NRM* + *SCM*) generates multimodal fragments that are, to a great extent, *(a)* need-aligned, *(b)* related to the

[1] To ensure bias-free results, we randomly chose to populate the left and right fragments with retrieved initializations or outputs of causal models. While inferring the ratings of the annotators, Likert scale was reversed in accordance with the random choice.



Title: Apple says hackers holding 300m iCloud account details to ransom did not get into their servers
Article Link: [A1]
Summary: The group has identified themselves as 'Turkish Crime Family'

If you do not add two factor strong authentication to any account, there is a probability that the password has been exposed, harvested, or guessed.

It is not yet known whether the Turkish Crime Family will deliver on their threats, or even if they have the competence to do so.

Title: Mercedes boss Toto Wolff slams Red Bull's oil burn claim: "We're no cheats ... they must be seeing ghosts"
Article Link: [A2]
Summary: Mercedes boss Toto Wolff has reacted angrily to Red Bull 's 'cheat' implications

Just getting his car, or that of team-mate Stoffel Vandoorne, to the end of the race will be a comparable achievement.

Michael Schumacher's lap record — 1min 24.1sec from 2004 — is under threat.

Title: Cristiano Ronaldo and his Portugal team-mates prepare ahead of Hungary World Cup 2018 qualifier
Article Link: [A3]
Summary: Euro 2016 winners Portugal take on Hungary in a World Cup qualifier on Friday

The Real Madrid star also pipped Lionel Messi to the Ballon D'Or at the end of the year, giving him the individual honour that he desires.

Portugal have good reason to think they will be among the 32 teams in Russia. They defeated Andorra and the Faroe Islands, mauling both of them 6-0 after their struggle against Switzerland.

Title: Why fruit and vegetable prices are about to skyrocket: Cyclone Debbie stripped avocado trees bare and devastated other crops...
Article Link: [A4]
Summary: Fruit and vegetables could be more than 60 per cent more expensive

About 95 per cent of Australia's winter supply of tomatoes and capsicums are grown in Bowen, which was hit by Debbie on Tuesday.

The damage to crops and produce means Australia will likely be in dearth supply of fruit and vegetables, raising the price.

Title: Abuse accelerates puberty in children: Kids who experience sexual abuse develop a year earlier than their non-abused peers
Article Link: [A5]
Summary: Experts say these kids are at higher risk to have low self-esteem and depression

Tanner staging is a numeric index of ratings that relates with the physical progression of puberty.

Additionally, early puberty is seen as a potential contributor to increased rates of depression, substance misuse, sexual risk taking and teenage pregnancy.

Title: Millennial bowel cancer crisis: Young people are four times more likely to develop the disease than previous generations - due to their terrible diets
Article Link: [A6]
Summary: Bowel cancer has long been associated with older more sedentary people

An worrying three in ten rectal cancer diagnoses are now in patients below the age of 55.

A diet high in red or processed meats, like bacon and sausages, and low in fiber increases the risk, as is being overweight or obese or being lethargic.

Figure 8: Qualitative illustration of need-adapted multimodal fragments generated by our proposed method. The original articles can be accessed at [A1], [A2], [A3], [A4], [A5], [A6]. The associated target-need as well as the type of the multimodal fragment is shown alongside; = denotes a parallel fragment while ≠ denotes a non-parallel fragment.

content in the input article, *(c)* diverse, and (d) well-performing on high-level engagement metrics like share-ability and click-ability. On comparing the performance with human-curated summary units (i.e., *Human*), we note that while the generated multimodal fragments underperform marginally in terms of relevance to article, fluency of text, and naturalness of image, they are better in terms of need-alignment, share-ability and click-ability. The trade-off can be seen as an inherent side-effect of automating the process of fragment creation. Moreover, the diversity of multimodal fragments generated, as it can be seen from a few qualitative example in Figure 8, allows for quite a few interesting downstream interaction scenarios to emerge.

In the life-cycle of an article, the involved personas primarily includes the authors, the virtual distributors, and the consumers. The authors, who often act as distributors of their own articles, might want to adapt the pointers to their articles (i.e., the multimodal fragments themselves) to align well with the platform-level user-characteristics (for instance, Instagram has different user-characteristics than Twitter or Facebook). The same idea applies to end-user characteristics where the distributor might want to use fragments that are well-suited for an individual (for instance, some of us are attracted towards educational aspects of the content, while others are attracted towards fashionable aspects). Both the scenarios involve personalization of multimodal fragments that are acting as pointers to the original article – while the former is a platform-level personalization, the latter is an individual-level personalization. Although the aspects of personalization discussed here briefly, are not in the purview of current work, it is interesting to note how the notion of generating diverse multimodal fragments will aid in this process. Tying it to consumer-end of the spectrum, we present empirical indications that the generated multimodal fragments perform well on high-level engagement metrics. We plan to study these personalization and interaction scenarios in future.

7 CONCLUSION

Several manifestations of multimodal fragments exist on the Web as advertisement banners, social media posts, etc., often acting as pointers to product landing pages, web blogs, and news articles. In this work, we provide a method to generate such multimodal fragments from the source content. Since the multimodal fragments, which in our case comprise of images and text, are often adapted to meet the needs of the target audiences as well as those of the publishers, our proposed method is aimed towards generating need-adapted multimodal fragments. The two core components of our method, i.e., need-based retrieval and stylistic causal model, both take the target need into account while synthesizing need-adapted initializations and generating stylistic variants of those initializations, respectively. Our extensive metric-based and human evaluations present strong empirical evidence that the generated fragments are *(i)* good representations of the source content by being relevant, *(ii)* well-aligned with the target-need, *(iii)* diverse, and *(iv)* induce higher engagement tendencies (in terms of sharing and clicking) than relevant and competitive baselines. Although the experiments conducted here are in the context of news articles and directed towards three specific target needs, we believe that the proposed method can be extended to cover a larger variety of documents as well as target needs. As future work, we aim to enhance the quality of the generated fragments by ensuring better fluency in text and more naturalness of generated images. We also plan to conduct further studies on interaction scenarios that entail the use of such multimodal fragments in the life-cycle of articles, as hinted upon in the brief discussion above.

REFERENCES

[1] Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual String Embeddings for Sequence Labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*. 1638–1649.

[2] Xavier Alameda-Pineda, Andrea Pilzer, Dan Xu, Nicu Sebe, and Elisa Ricci. 2017. Viraliency: Pooling local virality. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6080–6088.

[3] Malihe Alikhani, Sreyasi Nag Chowdhury, Gerard de Melo, and Matthew Stone. 2019. CITE: A Corpus of Image–Text Discourse Relations. *arXiv preprint arXiv:1904.06286* (2019).

[4] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 2 (2018), 423–443.

[5] John Bateman. 2014. *Text and image: A critical introduction to the visual/verbal divide*. Routledge.

[6] Andrew Brock, Jeff Donahue, and Karen Simonyan. 2018. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096* (2018).

[7] Julian Brooke and Graeme Hirst. 2013. A multi-dimensional Bayesian approach to lexical style. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 673–679.

[8] Lluis Castrejon, Yusuf Aytar, Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. 2016. Learning Aligned Cross-Modal Representations from Weakly Aligned Data. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*. IEEE.

[9] Pablo Samuel Castro and Maria Attarian. 2018. Combining Learned Lyrical Structures and Vocabulary for Improved Lyric Generation. *arXiv preprint arXiv:1811.04651* (2018).

[10] Lele Chen, Sudhanshu Srivastava, Zhiyao Duan, and Chenliang Xu. 2017. Deep cross-modal audio-visual generation. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*. ACM, 349–357.

[11] Paul Clough, Mark Sanderson, Murad Abouammoh, Sergio Navarro, and Monica Paramita. 2009. Multiple approaches to analysing query diversity. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 734–735.

[12] Edgar Dale. 1969. Audiovisual methods in teaching. (1969).

[13] Nima Dehmamy, Luca Stornaiuolo, and Mauro Martino. 2018. Vox2Net: From 3D Shapes to Network Sculptures. In *NeurIPS Workshop on Machine Learning for Creativity and Design 2018*. 1–3.

[14] Ori Bar El, Ori Licht, and Netanel Yosephian. 2019. GILT: Generating images from long text. *arXiv preprint arXiv:1901.02404* (2019).

[15] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2017. Vse++: Improved visual-semantic embeddings. *arXiv preprint arXiv:1707.05612* 2, 7 (2017), 8.

[16] Rebecca Anne Fiebrink. 2011. *Real-time human interaction with supervised learning algorithms for music composition and performance*. Citeseer.

[17] Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. 2017. Stylenet: Generating attractive visual captions with styles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3137–3146.

[18] Spandana Gella, Mirella Lapata, and Frank Keller. 2016. Unsupervised visual sense disambiguation for verbs using multimodal embeddings. *arXiv preprint arXiv:1603.09188* (2016).

[19] Katy Ilonka Gero, Giannis Karamanolakis, and Lydia Chilton. [n.d.]. Transfer Learning for Style-Specific Text Generation. ([n. d.]).

[20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.

[21] Tanmay Gupta, Dustin Schwenk, Ali Farhadi, Derek Hoiem, and Aniruddha Kembhavi. 2018. Imagine this! scripts to compositions to videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 598–613.

[22] Jack Hessel, Lillian Lee, and David Mimno. 2019. Unsupervised Discovery of Multimodal Links in Multi-Image, Multi-Sentence Documents. *arXiv preprint arXiv:1904.07826* (2019).

[23] Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 1587–1596.

[24] Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. 2017. Automatic understanding of image and video advertisements. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1705–1715.

[25] Phillip Isola, Devi Parikh, Antonio Torralba, and Aude Oliva. 2011. Understanding the intrinsic memorability of images. In *Advances in Neural Information Processing Systems*. 2429–2437.

[26] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1125–1134.

[27] Xin Jin, Jingying Chi, Siwei Peng, Yulu Tian, Chaochen Ye, and Xiaodong Li. 2016. Deep image aesthetics classification using inception modules and fine-tuning connected layer. In *2016 8th International Conference on Wireless Communications & Signal Processing (WCSP)*. IEEE, 1–6.

[28] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4565–4574.

[29] Neel Kant, Raul Puri, Nikolai Yakovenko, and Bryan Catanzaro. 2018. Practical Text Classification With Large Pre-Trained Language Models. *arXiv preprint arXiv:1812.01207* (2018).

[30] Sanjeev Kumar Karn, Mark Buckley, Ulli Waltinger, and Hinrich Schütze. 2018. News Article Teaser Tweets and How to Generate Them. *arXiv preprint arXiv:1807.11535* (2018).

[31] Wei-Jen Ko, Greg Durrett, and Junyi Jessy Li. 2019. Domain agnostic real-valued specificity prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 6610–6617.

[32] Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. 2019. Integrating Text and Image: Determining Multimodal Document Intent in Instagram Posts. *arXiv preprint arXiv:1904.09073* (2019).

[33] Marie Anna Lee. [n.d.]. Relationship between words and images. ([n. d.]). http://marieannalee.com/arts091/lectures/text&image.pdf.

[34] Leonidas Lefakis, Alan Akbik, and Roland Vollgraf. 2018. FEIDEGGER: A Multimodal Corpus of Fashion Images and Descriptions in German. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

[35] Chun-Liang Li, Eunsu Kang, Songwei Ge, Lingyao Zhang, Austin Dill, Manzil Zaheer, and Barnabas Poczos. 2018. Hallucinating Point Cloud into 3D Sculptural Object. *arXiv preprint arXiv:1811.05389* (2018).

[36] Yijun Li, Ming-Yu Liu, Xueting Li, Ming-Hsuan Yang, and Jan Kautz. 2018. A closed-form solution to photorealistic image stylization. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 453–468.

[37] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[38] Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*. 150–157.

[39] Hao Ma, Michael R Lyu, and Irwin King. 2010. Diversifying query suggestion results. In *Twenty-fourth AAAI Conference on Artificial Intelligence*.

[40] Javier Marin, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, and Antonio Torralba. 2019. Recipe1M+: A Dataset for Learning Cross-Modal Embeddings for Cooking Recipes and Food Images. *IEEE transactions on pattern analysis and machine intelligence* (2019).

[41] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H Adelson, and William T Freeman. 2016. Visually indicated sounds. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2405–2413.

[42] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[43] Michael Sankey, Dawn Birch, and Michael Gardiner. 2010. Engaging students through multimodal learning environments: The journey continues. In *Proceedings ASCILITE 2010: 27th annual conference of the Australasian Society for Computers in Learning in Tertiary Education: curriculum, technology and transformation for an unknown future*. University of Queensland, 852–863.

[44] Andreza Sartori, Victoria Yanulevskaya, Almila Akdag Salah, Jasper Uijlings, Elia Bruni, and Nicu Sebe. 2015. Affective analysis of professional and amateur abstract paintings using statistical analysis and art theory. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 5, 2 (2015), 8.

[45] Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. 2017. Speaking the same language: Matching machine to human captions by adversarial training. In *Proceedings of the IEEE International Conference on Computer Vision*. 4135–4144.

[46] Aliaksandr Siarohin, Gloria Zen, Nicu Sebe, and Elisa Ricci. 2018. Enhancing perceptual attributes with bayesian style generation. In *Asian Conference on Computer Vision*. Springer, 483–498.

[47] Karan Sikka, Lucas Van Bramer, and Ajay Divakaran. 2019. Deep Unified Multimodal Embeddings for Understanding both Content and Users in Social Media Networks. *arXiv preprint arXiv:1905.07075* (2019).

[48] Jonathan A Simon. [n.d.]. Entendrepreneur: Generating Humorous Portmanteaus using Word-Embeddings. ([n. d.]).

[49] Xavier Snelgrove and Matthew Tesfaldet. [n.d.]. Interactive CPPNs in GLSL. ([n. d.]).

[50] Adam Summerville, Sam Snodgrass, Matthew Guzdial, Christoffer Holmgård, Amy K Hoover, Aaron Isaksen, Andy Nealen, and Julian Togelius. 2018. Procedural content generation via machine learning (PCGML). *IEEE Transactions on Games* 10, 3 (2018), 257–270.

[51] Matthew Tesfaldet, Marcus A. Brubaker, and Konstantinos G. Derpanis. 2018. Two-Stream Convolutional Networks for Dynamic Texture Synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[52] Michael E Tipping and Christopher M Bishop. 1999. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61, 3 (1999), 611–622.

[53] Quoc-Tuan Truong and Hady Lauw. 2019. Multimodal Review Generation for Recommender Systems. In *The World Wide Web Conference*. ACM, 1864–1874.

[54] Yixin Wang and David M Blei. 2018. The blessings of multiple causes. *arXiv preprint arXiv:1805.06826* (2018).

[55] Robert West and Eric Horvitz. 2019. Reverse-Engineering Satire, or" Paper on Computational Humor Accepted Despite Making Serious Advances". *arXiv preprint arXiv:1901.03253* (2019).

[56] Michael J Wilber, Chen Fang, Hailin Jin, Aaron Hertzmann, John Collomosse, and Serge Belongie. 2017. Bam! the behance artistic media dataset for recognition beyond photography. In *Proceedings of the IEEE International Conference on Computer Vision*. 1202–1211.

[57] Sam Wiseman, Stuart M Shieber, and Alexander M Rush. 2018. Learning neural templates for text generation. *arXiv preprint arXiv:1808.10122* (2018).

[58] Mike Wu and Noah Goodman. 2018. Multimodal generative models for scalable weakly-supervised learning. In *Advances in Neural Information Processing Systems*. 5575–5585.

[59] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1316–1324.

[60] Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. 2018. Recipeqa: A challenge dataset for multimodal comprehension of cooking recipes. *arXiv preprint arXiv:1809.00812* (2018).

[61] Keren Ye and Adriana Kovashka. 2018. Advise: Symbolism and external knowledge for decoding advertisements. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 837–855.

[62] Mingda Zhang, Rebecca Hwa, and Adriana Kovashka. 2018. Equal but not the same: Understanding the implicit relationship between persuasive images and text. *arXiv preprint arXiv:1807.08205* (2018).

[63] Xingxing Zhang, Mirella Lapata, Furu Wei, and Ming Zhou. 2018. Neural latent extractive document summarization. *arXiv preprint arXiv:1808.07187* (2018).

[64] Bo Zhao, Xiao Wu, Zhi-Qi Cheng, Hao Liu, Zequn Jie, and Jiashi Feng. 2018. Multi-view image generation from a single-view. In *2018 ACM Multimedia Conference on Multimedia Conference*. ACM, 383–391.

[65] Sanqiang Zhao, Rui Meng, Daqing He, Saptono Andi, and Parmanto Bambang. 2018. Integrating Transformer and Paraphrase Rules for Sentence Simplification. *arXiv preprint arXiv:1810.11193* (2018).

[66] Junnan Zhu, Haoran Li, Tianshang Liu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2018. MSMO: Multimodal Summarization with Multimodal Output. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 4154–4164.