Are Multimodal Transformers Robust to Missing Modality?

Mengmeng Ma[1] Jian Ren[2] Long Zhao[3] Davide Testuggine[2] Xi Peng[1]
[1]University of Delaware [2]Snap Inc. [3]Google Research
{mengma,xipeng}@udel.edu, jren@snap.com, longzh@google.com, davide.testuggine@gmail.com

Abstract

Multimodal data collected from the real world are often imperfect due to missing modalities. Therefore multimodal models that are robust against modal-incomplete data are highly preferred. Recently, Transformer models have shown great success in processing multimodal data. However, existing work has been limited to either architecture designs or pre-training strategies; whether Transformer models are naturally robust against missing-modal data has rarely been investigated. In this paper, we present the first-of-its-kind work to comprehensively investigate the behavior of Transformers in the presence of modal-incomplete data. Unsurprising, we find Transformer models are sensitive to missing modalities while different modal fusion strategies will significantly affect the robustness. What surprised us is that the optimal fusion strategy is dataset dependent even for the same Transformer model; there does not exist a universal strategy that works in general cases. Based on these findings, we propose a principle method to improve the robustness of Transformer models by automatically searching for an optimal fusion strategy regarding input data. Experimental validations on three benchmarks support the superior performance of the proposed method.

1. Introduction

Multimodal Transformers are emerging as the dominant choice in multimodal learning across various tasks [7], including classification [21,29], segmentation [34], and cross-model retrieval [18]. They have become the driving force in obtaining better performance on these tasks through a pre-train-and-transfer [3] paradigm.

Although Transformers have demonstrated remarkable success in processing multimodal data, they generally require modal-complete data. The completeness of modality may not always hold in the real world due to privacy or security constraints. For example, a social network might be unable to access location information if users decline to share their private location [20]; a healthcare application might not have all the records available when patients are unwilling to undergo risky or invasive examinations [37]. For this reason, it is important that Transformer models are *robust against missing-modal data*, *i.e.*, the model performance does not degrade dramatically.

Table 1. Evaluation of the Transformer robustness against missing-modal data on MM-IMDb, UPMC Food-101, and Hateful Memes. We use ViLT [18] as the backbone. Note that the multimodal performance is *even worse* than the unimodal one, when modality is missing severely (results are highlighted in shaded gray). *The reported evaluation scores are F1-Macro (MM-IMDb), Accuracy (UPMC Food-101), and AUROC (Hateful Memes). Higher scores indicate better results.

Dataset	Training		Testing		Evaluation*	Δ ↓
	Image	Text	Image	Text		
MM-IMDb [2]	100%	100%	100%	100%	55.3	0%
	100%	100%	100%	30%	31.2	**43.6%**
	100%	0%	100%	0%	35.0	36.7%
UPMC Food-101 [43]	100%	100%	100%	100%	91.9	0%
	100%	100%	100%	30%	65.9	**28.3%**
	100%	0%	100%	0%	71.5	22.2%
Hateful Memes [17]	100%	100%	100%	100%	70.2	0%
	100%	100%	100%	30%	60.2	14.2%
	100%	0%	100%	0%	56.3	**19.8%**

Despite its real-world importance, the robustness against missing modalities in multimodal Transformers is seldom investigated in the literature. So far, research on Transformer models has been limited to developing new architectures for fusion [29, 35, 38] or exploring better self-supervised learning tasks [1, 6, 8, 47, 48]. Recent work on Transformer robustness has primarily focused on noisy inputs rather than missing modalities [23].

A question naturally arises: *Are Transformer models robust against missing-modal data?* We empirically evaluate this problem across multiple datasets in Table 1. Unsurprisingly, we find that *Transformer models degrade dramatically with missing-modal data.* As shown, the multimodal performance drops when tested with modal-incomplete data, and, surprisingly, the multimodal performance is even worse than the unimodal when text are missing severely, *i.e.,* only 30% of text are available.

Table 2. Evaluation of the Transformer models under different fusion strategies on MM-IMDb and Hateful Memes. *Early* fusion refers to fusion at the first layer; *Late* fusion refers to fusion at the last layer. Different fusion strategies affect model robustness against the missing-modal data.

Dataset	Train		Test		Fusion Strategy	
	Image	Text	Image	Text	Early	Late
MM-IMDb	100%	100%	100%	100%	**55.3**	54.9
UPMC Food-101	100%	100%	100%	100%	**91.9**	91.8
Hateful Memes	100%	100%	100%	100%	**70.2**	64.5
MM-IMDb	100%	100%	100%	30%	**31.2**	31.0
UPMC Food-101	100%	100%	100%	30%	65.9	**69.1**
Hateful Memes	100%	100%	100%	30%	**60.2**	57.8

Prior work on Transformer models has shown that fusion strategies affect computation complexity and performances [3, 21, 29]. Another question arises: *Will the fusion strategy affect Transformer robustness against modal-incomplete data?* Unsurprisingly, we observe that different fusion strategies will significantly affect the robustness. What surprised us is that *the optimal fusion strategy is dataset-dependent; there does not exist a universal strategy that works in general cases in the presence of modal-incomplete data.* As shown in Table 2, when tested with missing-modal data, early fusion is preferred on MM-IMDb and Hateful Memes, while late fusion is preferred on the UPMC Food-101. This motivates us to improve the robustness of Transformers by automatically attain the optimal fusion strategy regarding different datasets.

We propose a new method to achieve this goal. Our main idea is to jointly optimize Transformer models with modal-complete and modal-incomplete data via multi-task optimization. On top of that, we propose a searching algorithm to attain the best fusion strategy regarding different datasets. Overall, the main contributions are as follows:

- To the best of our knowledge, this paper is first-of-its-kind study to investigate the Transformer robustness against modal-incomplete data.
- We observe that Transformer models degrade dramatically with missing-modal data. And surprisingly, the optimal fusion strategy is dataset dependent; there does not exist a universal strategy that works in the presence of modal-incomplete data.
- We improve the robustness of Transformer models via multi-task optimization. To further improve robustness, we develop an differentiable algorithm to attain the optimal fusion strategy.
- We conduct extensive experiments and ablation study on MM-IMDb [2], UPMC Food-101 [43], and Hateful Memes [17] to support our findings and validate the robustness of our method against missing modality.

2. Related Work

Multimodal learning. Different modalities, *e.g.,* natural language, visual signals, or vocal signals, are often complementary in content while overlapping for a common concept. Multimodal learning aims to utilize the complementary information of each modality to improve the performance of various computer vision tasks. A key aspect of multimodal learning is exploring efficient methods for multimodal fusion. Simple methods like concatenation have been widely studied in [32,42]. For efficient cross-modality interaction, a tensor fusion [46] mechanism is proposed by Zadeh *et al.* Following this effort, efficient low-rank fusion [25] is proposed to address the exponential dimension explosion of tensor fusion.

The aforementioned fusion mechanisms heavily depend on the completeness of modality, making multimodal fusion impossible with modality-incomplete data. Therefore, another important direction in multimodal learning is to build models that are robust against the modality-incomplete data [27,39]. For example, Ma *et al.* [27] propose a method based on Bayesian Meta-Learning to estimate the latent feature of the modality-incomplete data. However, the existing endeavors usually adopt modality-specific models for each modality, such as ResNet [12] for images and LSTM [13] for texts, which may lead to a larger set of architectural decisions and training parameters. Instead, we use Transformers as general architectures to jointly model each modality, leading to a simple design and reduced training parameters.

Multimodal transformer. Multimodal Transformers have been used in various tasks such as cross-model retrieval [18, 22], action recognition [29], and image segmentation [34, 45]. They provide several advantages over conventional backbones, *e.g.,* ResNet [12], regarding to flexibility and training load.

The *flexibility* to accommodate modality-incomplete samples is crucial for multimodal backbones, as the real-world multimodal data are often imperfect due to missing modality. Conventional backbones [31,39] are generally not flexible. These backbones output the joint multimodal representation by explicitly fusing the features of each modality via concatenation [32], tensor fusion [46], and others mechanisms. However, explicit fusion requires the presence of all modalities. Missing any modality will break the training pipeline. In contrast, multimodal Transformers use the self-attention mechanism [40] to generate a holistic representation of all modalities, allowing the absence of any modalities. When dealing with modality-incomplete samples, it can ignore the absent modalities by applying a mask on the attention matrix. Therefore, multimodal Transformers are more flexible in dealing with missing modalities. Besides, an *easy-to-train* model is vital for multimodal learning. The training load of conventional multimodal backbone grows as the number of modalities in-

Table 3. Multi-label classification scores (%) on the *MM-IMDb* [2] under different settings: train and test with full modality (100% Image + 100% Text); train and test with single modality (100% Image or 100% Text). † indicate our implementation.

Method	Modality Image	Modality Text	F1 Micro	F1 Macro	F1 Weighted	F1 Samples
MFAS [30]	✓		47.8	25.6	42.1	48.4
		✓	60.2	48.9	58.5	60.6
CentralNet [41]	✓		—	33.5	49.2	—
		✓	—	45.9	57.5	—
ViLT [18]†	✓		51.8	35.0	48.0	51.1
		✓	63.3	52.5	62.0	62.9
MFAS [30]	✓	✓	—	55.7	62.5	—
CentralNet [41]	✓	✓	63.9	56.1	63.1	63.9
ViLT [18]†	✓	✓	64.7	55.3	64.4	64.6

creases since the backbone usually consists of modality-specific sub-models that need to be trained independently for each modality [27, 41]. Instead, Transformer models process all modalities simultaneously using a single model [18, 24], which greatly reduces the training load.

Dynamic neural networks. Our work is also related to dynamic neural networks, which adapt the network structure to different inputs, resulting in noticeable gains in accuracy, computation efficiency, or flexibility [11]. We follow the spirit of the Dynamic Depth [11] method. Numerous methods have been proposed to dynamically select layers for inference to reduce computation costs. Our idea is inspired by AdaShare [36], which focuses on learning a policy that selects layers for sharing in multi-task learning. Its main idea is to use Gumbel Softmax Sampling [14,28] to learn the policy and network parameters without replying on Reinforcement Learning [44] or extra policy network [10]. However, the direct application of Gumbel Softmax Sampling to our problem results in a large search space with many invalid policies. As a result, we develop an efficient method without using Gumbel Softmax Sampling.

3. Analysis of Multimodal Transformer

3.1. Background

In this paper, we focus on the multimodal Transformer that adopts Vision Transformer (ViT) [7] as the backbone. ViT consists of a sequence of L Transformer layers, each of which contains a Multi-Head Attention (MHA) layer, Multilayer Perceptron (MLP), and Layer Normalization (LN). The MHA computes the dot-product attention [40] on the input sequence, resulting in an attention matrix indicating the similarity between each token.

We follow the vision-language Transformer [18] to preprocess the data. The input text is mapped into the word embedding through a word embedding codebook and a position embedding codebook. The input image is first partitioned into patches and then flattened into vectors. These vectors are then transformed into latent embedding using linear projection and position embedding. Finally, the image and text embedding are integrated with their corresponding modality-type embedding [15, 18]. The final multimodal input sequence is the concatenation of vision and text embedding.

Table 4. Classification accuracy (%) on the *UPMC Food-101* [43]. † indicate our implementation.

Method	Modality Image	Modality Text	Accuracy
BERT+LSTM [9]	✓		71.7
		✓	84.4
ViLT [18]†	✓		71.5
		✓	84.4
BERT+LSTM [9]	✓	✓	92.5
MMBT [15]	✓	✓	92.1
ViLT [18]†	✓	✓	92.0

3.2. Robustness Against Missing Modality

Question: Are Transformer models robust against modal-incomplete data?

Observation: Unsurprisingly, Transformer models degrade dramatically with modal-incomplete data.

We begin by defining how Transformer robustness is measured. Specifically, we adopt two different evaluation settings: a "full" test set with full-modal data and a "missing" test set with missing-modal data. We evaluate Transformer robustness by comparing model performance on the "missing" test set to the "full" test set: the smaller the difference, the better the robustness.

First, we empirically verify that model performance degrades dramatically in the presence of missing-modal data. Table 1 shows the evaluation results on three widely-used multimodal datasets. As shown, when only 30% text modality is observed, the multimodal performance drops by 43.6%, 28.3%, and 14.2%, respectively. Moreover, when the modality is severely missing, the multimodal performance is even worse than the unimodal one on MM-IMDb and UPMC Food-101.

Second, we observe that the modality importance varies on different datasets. We use unimodal performance to indicate the importance of each modality. Results of unimodal performance are shown in Tables 3, 4, and 5. As shown, the text modality is more important than the image modality on MM-IMDb and Food-101, while text and image are equally important on Hateful Memes. In specific, in the first two datasets, text has higher performance than the image. Moreover, the performance gap between unimodal (text) and full modalities is smaller then that between unimodal (image) and full modalities (10% vs. 22%), indicating that text is the dominant modality. In contrast, in the Hateful Memes dataset, the performances of text and image are comparable,

Table 5. AUROC (%) on the unseen test set of *Hateful Memes* [17]. *denotes the results from hateful memes challenge [16]. † indicate our implementation.

Method	Modality		AUROC
	Image	Text	
Unimodal*	✓		54.6
		✓	62.7
ViLT [18]†	✓		56.3
		✓	58.3
MMBT-Grid [15]*	✓	✓	67.3
MMBT-Region [15]*	✓	✓	72.2
ViLBERT [26]*	✓	✓	73.4
ViLBERT CC [26]*	✓	✓	72.8
Visual BERT [24]*	✓	✓	73.2
ViLT [18]†	✓	✓	70.2

and the performance gap between multimodal and unimodal is large (>20%), demonstrating that the two modalities are equally important.

Finally, we empirically observe that Transformer models tend to overfit to dominate modalities. Specifically, we first train our model with multimodal data and test with different unimodal data. Then we examine the performance gap between unimodal and multimodal testing – the larger the gap, the more severe the overfitting. Experimental results are shown in Table 6. As shown, for MM-IMDb dataset, text-only testing performs better than image-only testing, which means that text-only testing is closer to full-modal testing. Therefore text-only testing has a smaller gap than image-only testing, indicating that models trained on this dataset tend to overfit to text modality.

3.3. Optimal Fusion Strategy

Question: Will the fusion strategy affect Transformer robustness against modal-incomplete data?

Observation: Different fusion strategies do affect the robustness of Transformer models. Surprisingly, the optimal fusion strategy is dataset-dependent; there does not exist a universal strategy that works in general cases.

Typically, there exists two widely used fusion strategies: early and late fusion. For early fusion, cross-modal interaction happened in early layers, ensuring the model to have sufficient capacity to exploit multimodal information, but at the expense of larger computing costs. For late fusion, cross-modal interaction happened in later layers, which significantly lower computation costs, but the resulting model might have limited capacity to take the full advantage of multimodal information.

It remains an open question on how to determine the optimal layer for fusion [3]. Existing solutions in multimodal Transformer adopt a fixed fusion strategy [18,21,24,26,33]. However, the one-size-fits-all approach may not be optimal on all datasets. As discussed in Sec. 1, the optimal fusion

Table 6. Evaluation on the overfitting issue of Transformer models on MM-IMDb and Hateful Memes. Transformer models tend to overfit to dominate modality.

Dataset	Training		Testing		Evaluation
	Image	Text	Image	Text	
MM-IMDb	100%	100%	100%	100%	55.3
	100%	100%	0%	100%	47.4
	100%	100%	100%	0%	35.0
Hatful Memes	100%	100%	100%	100%	70.2
	100%	100%	0%	100%	55.7
	100%	100%	100%	0%	54.9

strategy is dataset-dependent.

4. Robust Multimodal Transformer

Without loss of generality, we consider a multimodal dataset with two modalities. Formally, let $\mathcal{D} = \left\{\boldsymbol{x}_i^1, \boldsymbol{x}_i^2, y_i\right\}_i$ denote the multimodal dataset, where $\boldsymbol{x}_i^1$ and $\boldsymbol{x}_i^2$ represent two different modalities and y_i is the corresponding label. Our target is to improve the Transformer robustness against modal-incomplete data, *i.e.,* model performance does not degrade dramatically. To this end, we propose to leverage multi-task optimization and the optimal fusion strategy to improve the robustness.

Multi-task learning. We intend to improve the performance of Transformer models in dealing with modal-incomplete data. In the missing-modal scenario, training data are modal-complete, while testing samples are modal-incomplete. This discrepancy motivates us to incorporate missing-modal data in the training process. By doing so, the Transformer model will be more confident in its prediction on modal-incomplete data, resulting in a robust Transformer. The key idea is to leverage the masking mechanism to "generate" the modal-incomplete data during training and jointly optimize the Transformer model with modal-complete and modal-incomplete data via multi-task optimization. Our method is simple to implement with minimal modification to the Transformer.

Optimal fusion strategy. The goal is to automatically search for the optimal fusion strategy on different datasets. Manually finding the optimal strategy is not practical, especially for lager-scale models [5, 7, 47], due to the heavy training load. However, designing such an algorithm is non-trivial in light of the non-differentiable nature of the discrete searching space [36]. Existing methods, such as Reinforcement Learning (RL) [44] and policy network (PN) [10], are either inefficient in training or adding additional parameters to the model. We propose a differentiable method to obtain the fusion strategy through standard back-propagation. The key idea is to learn a policy to obtain optimal layers for fusion. Specifically, each layer is assigned with a policy parameter[1] to decide fusion or not. The fusion strategy is

[1]Policy parameters are negligible compared to model parameters.

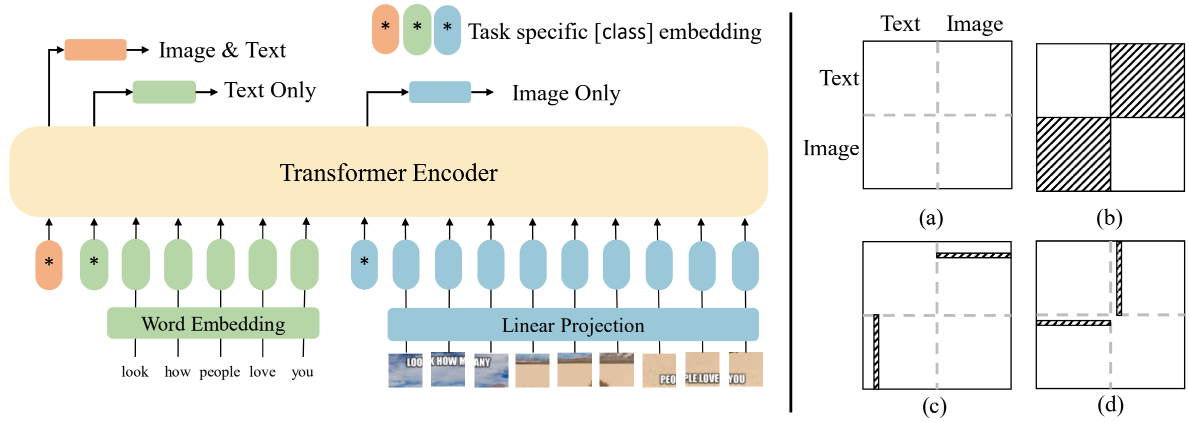

Figure 1. *Left*: Overview of our model. *Right*: Attention masks for different tasks: (a) Original attention without masking; (b) Mask-out cross-modal attention; (c) Mask-out image attention for text only [class] token; (d) Mask-out text attention for image only [class] token.

sampled from the policy parameters.

4.1. Improve Robustness via Multi-task Learning

On a bimodal dataset, *e.g.,* image and text, multi-task learning can have up to three distinct tasks: full-modal (image + text) task, image-only task, and text-only task. Let f_θ denote a Transformer parameterized by $\boldsymbol{\theta}$. The total loss function is defined as follows:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{img}(\boldsymbol{x}^1; \boldsymbol{\theta}) + \lambda_2 \mathcal{L}_{txt}(\boldsymbol{x}^2; \boldsymbol{\theta}) + \lambda_3 \mathcal{L}_{it}(\boldsymbol{x}^1, \boldsymbol{x}^2; \boldsymbol{\theta}), \quad (1)$$

where $\mathcal{L}_{img}$ is the loss for image only task; $\mathcal{L}_{txt}$ is the loss for text only task; $\mathcal{L}_{it}$ is the loss for image + text task; λ_1, λ_2, and λ_2 are hyperparameters to balance each loss.

Transformer models leverage classification tokens [18, 26, 40] to generate embeddings for classification. For the three tasks, we add three classification tokens to the Transformer model. Each classification token will output task-specific embedding for the target task. The model overview is shown in Figure 1 Left. For multi-task learning, each task is expected to use only corresponding modalities for classification, *e.g.*, text modality for the text-only task. Therefore we apply masks on the attention matrix, ensuring that the output embedding of each classification token contains only information from corresponding modalities. For instance, in the text-only task, we mask out all the self-attention to the image and the cross-attention between image and text. The attention masks are shown in Figure 1 Right.

4.2. Search for the Optimal Fusion Strategy

We first introduce the formulation of the search problem. Let $\boldsymbol{\alpha} = \{\alpha_m\}_{m=1}^{M}$ denote the policy parameters, where M is the total number of layers. To learn the optimal policy parameters, we formulate the parameter learning into a bi-level optimization problem. The object of optimization is to minimize the loss on a validation set $\mathcal{L}^{val}(\boldsymbol{\alpha}, \boldsymbol{\theta}^*)$, where

Algorithm 1: Search for Optimal Fusion Policy.

Input: Multimodal dataset D^{tr}, D^{val}; inner-level learning rate γ; outer-level learning rate β; initialized policy parameter $\boldsymbol{\alpha}$; number of iterations K.

while *not converged* **do**
- $\{\mathbf{x}_i^1, \mathbf{x}_i^2, y_i\} \sim D^{tr}$; $\{\mathbf{x}_j^1, \mathbf{x}_j^2, y_j\} \sim D^{val}$
- $\boldsymbol{\theta}_0 \leftarrow \boldsymbol{\theta}$
- **Lower-Level**:
- **for** $k = 0$ to $K-1$ **do**
  - Sample policy $\boldsymbol{s}$ with $\boldsymbol{\alpha}$ using Eqn. 3
  - $\boldsymbol{\theta}_{k+1} \leftarrow \boldsymbol{\theta}_k - \gamma \nabla_{\boldsymbol{\theta}_k} \mathcal{L}^{tr}(\mathbf{x}_i^1, \mathbf{x}_i^2, \boldsymbol{s}; \boldsymbol{\theta})$
- **end**
- $\boldsymbol{\theta}^* \leftarrow \boldsymbol{\theta}_K$
- **Upper-Level:**
- Sample policy $\boldsymbol{s}$ with $\boldsymbol{\alpha}$ using Eqn. 3
- $\boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha} - \beta \nabla_{\boldsymbol{\alpha}} \mathcal{L}^{val}(\mathbf{x}_j^1, \mathbf{x}_j^2, \boldsymbol{s}; \boldsymbol{\theta}^*)$

end

the optimal weights $\boldsymbol{\theta}^*$ are obtained by minimizing the the training loss $\mathcal{L}^{tr}(\boldsymbol{\theta}, \boldsymbol{\alpha}^*)$. The optimization problem is formulated as follows:

$$\begin{aligned} &\min_{\boldsymbol{\alpha}} \ \mathcal{L}^{val}(\boldsymbol{\theta}^*, \boldsymbol{\alpha}), \\ &s.t. \ \ \boldsymbol{\theta}^* = \operatorname*{argmin}_{\boldsymbol{\theta}} \mathcal{L}^{tr}(\boldsymbol{\theta}, \boldsymbol{\alpha}^*). \end{aligned} \quad (2)$$

Next, we describe how to generate the fusion strategy using policy parameters. Existing work on policy learning generally assumes α_m to be bivariate [36], resulting in a search space with 2^M possible policies[2]. However, the search space can be significantly reduced. In multimodal fusion, one usually conducts fusion starting from a certain

[2] Two actions, fusion or not fusion, for each layer and M layers in total.

layer until the last layer. Following this convention, we set α_m to be univariate, leading to a search space with M policies. Let $\boldsymbol{Q}$, a lower triangular matrix with all non-zero elements equal to one and with the size of $M \times M$, denote all the policies. Let $\boldsymbol{s}$ denote the final policy. To obtain a policy, we first softmax the policy parameter to get a soft policy: $\boldsymbol{s}_s = \text{softmax}(\boldsymbol{\alpha})$. Then, we convert $\boldsymbol{s}_s$ into a hard policy using one-hot encoding with the differential trick[3]: $\boldsymbol{s}_h = \text{onehot}(\boldsymbol{s}_s)$. The final policy can be obtained by sampling from $\boldsymbol{Q}$ using the hard policy $\boldsymbol{s}_h$.

$$\boldsymbol{s} = \langle \boldsymbol{Q}, \boldsymbol{s}_h \rangle . \tag{3}$$

Our method significantly reduces the search space, resulting in an differentiable and easy-to-train policy learning process. The overall method is shown in Algorithm 1. Once the optimal policy is learned, we fixed the policy to retrain the model $\boldsymbol{\theta}$ using the whole training set.

5. Experiments

In this section, we analyze the performance of our approach on three multimodal datasets and aim to answer the following questions: (1) Does the Transformer model perform well with modal-complete data? (Sec. 5.3) (2) Does the proposed method improve the robustness of backbone against missing-modal data? (Sec. 5.4) (3) Why different datasets prefer different layers for multimodal fusion? (Sec. 5.5) (4) What factors affect the effectiveness of our method? (Sec. 5.6)

5.1. Datasets and Metrics

Datasets. *MM-IMDb* [2] has two modalities: image and text. The target task is to predict the genres of a movie using image, text, or both. This task is multi-label classification, as each movie might have multiple genres. This dataset contains 25,956 image-text pair and 23 classes.

UPMC Food-101 [43] is a classification dataset composed of text and images. The UPMC Food-101 categories are identical to one of the largest publicly available food image datasets: the ETHZ Food-101 [4]. In the UPMC Food-101, image and text pairs are noisy since all the images are obtained in an uncontrolled environment. This dataset contains 90,704 image-text pairs and 101 classes.

Hateful Memes [17] is another challenging multimodal dataset that focuses on identifying hate speech in memes. It is constructed to fail models that rely on single modality and multimodal models are likely to perform well: challenging samples ("benign confounders") are added to the dataset to make relying on unimodal signals more difficult. Hateful Memes contain exactly 10k memes.

Metrics. For *MM-IMDb* dataset, following previous works [15, 27, 41], we use F1 Micro, F1 Macro, F1 Samples, and F1 Weighted to evaluate multi-label classification. For *UPMC Food-101*, similar to previous works [9, 43], we compute the classification accuracy. For *Hateful Memes*, following [17], we use Area Under the Receiver Operating Characteristic Curve (AUROC) to evaluate model performance.

[3]Differentiable trick: $\boldsymbol{s}_h = \text{onehot}(\boldsymbol{s}_s) - \boldsymbol{s}_s.\text{detach}() + \boldsymbol{s}_s$.

5.2. Implementation Details

Multimodal backbone. We use ViLT as the backbone since it represents the common design of multimodal transformer. ViLT [18] is a pure Transformer-based model that does not rely on modality-specific sub-models to extract features, and multiple objectives are used to pre-train the model, *e.g.,* Image Text Matching (ITM) and Masked Language Modeling (MLM).

Inputs. For image modality, we resize the input image into 384×384. Following [7], we extract 32×32 patches from the input images, yielding a total of $12 \times 12 = 144$ patches per image. For text modality, we adopt *bert-base-uncased* tokenizer to tokenize text inputs. The maximum length of the text sequences is various across different datasets: 1024 (MM-IMDb), 512 (Food-101), and 128 (Hateful Memes).

Network training. We use Adam optimizer [19] in all experiments with different learning rates for network training and policy learning. For network training, the base learning rate is 3×10^{-5} and weight decay is 2×10^{-2}. For policy learning, the base learning rate is 3×10^{-3} and weight decay is 3×10^{-5}. Model parameters are initialized using the pre-trained weights provided by ViLT [18].

5.3. Performance on the Full Test Set

We compare our model with other baseline models under the "full" evaluation setting, where all modalities are observed.

In Tables 3, 4, and 5, we report results on MM-IMDb, UPMC Food-101, and Hateful Memes, respectively. Our results are either state-of-the-art or on par with other methods: on MM-IMDb, compared to CentralNet, we achieve an F1 Weighted of 64.4 in comparison to 63.1; on UPMC Food-101, compared to MMBT, the most similar model to ours, we achieve an accuracy of 92.0 in comparison to 92.1; on Hateful Memes, we achieve comparable performance to MMBT (70.2 vs. 72.2). The results demonstrate the superiority of our model.

5.4. Performance on the Missing Test Set

The "missing" test set. We follow a conventional setting [39] to evaluate the model robustness against missing-modal data, in which the training data are modal-complete, while the testing data are modal-incomplete. We denote the full-modal train/test set as 100% Image + 100% Text and the missing-modal test set as 100% Image + η% Text, where

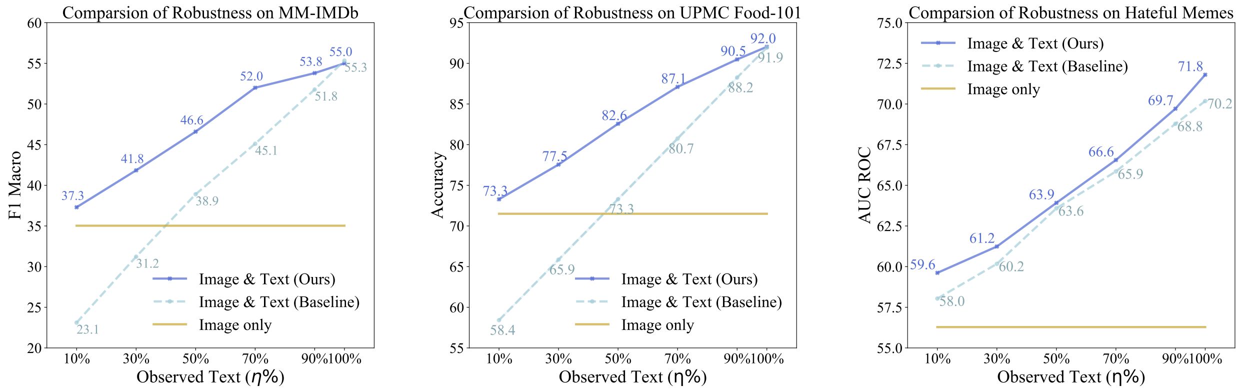

Figure 2. Comparison of Transformer robustness [18] on MM-IMDb [2] (*left*), UPMC Food-101 [43] (*middle*), and Hateful Memes dataset [17] (*right*). We adopt ViLT [18] as the backbone. Models are trained with 100% text + 100% image and tested with η% text + 100% image. "Image only" refers to the single modality setting – only image modality is used for training and testing. *Our method significantly improves model robustness, especially when the modality is severely missing.*

η% is percentage of observed modality and η indicates the severity of modality missing. The smaller the η, the severer the modality missing. When $\eta = 0$, we evaluate the model's unimodal performance.

Robustness to missing modality. We compare Transformer robustness on three multimodal datasets. Results are shown in Figure 2. As shown, our method improves Transformer robustness on datasets with dominant modalities. Specifically, on *MM-IMDb* and *UPMC Food-101*, the model performance degrades as η decreases. When text modality is severely missing, *i.e.,* only 10% are available, the multimodal performance is even worse than the unimodal one. For example, on MM-IMDb, the baseline model achieves an F1 Macro of 23.1, which is 34.0% lower than the unimodal one of 35.0; on UPMC Food-101, the baseline model achieves Accuracy of 58.4 which is 18.3% lower than the unimodal Accuracy of 73.3. However, in our method, the Transformer model still maintains good performance when modality is severely missing (only 10% text modality is observed), *i.e.,* on MM-IMDb, our model yields F1 macro of 37.3 which is 6.6% larger then unimodal F1 Macro of 35.0; on UPMC Food-101, our method achieves an accuracy of 73.3, which is 2.5% greater than the unimodal accuracy of 71.5.

Our method improves Transformer robustness on datasets with equally important modalities. Different from MM-IMDb and UPMC Food-101, *the Hateful Memes* dataset does not have dominant modalities. We observe that the multimodal performance is always superior to the unimodal one. In this dataset, our method outperforms the baseline model when tested with modal-complete and modal-incomplete data. As shown in Figure 2, when only 10% of text is observed, our model yields an AUROC of 59.6, which is 2.8% higher than the baseline (58.0); when full modalities are observed, our model outperforms the baseline by 2.3%.

5.5. Analysis of Optimal Fusion Strategy

We visualize the optimal policy of three datasets in Figure 3. We observe that late fusion is preferred on MM-IMDb, while early fusion is preferred on Hateful Memes. The learned policy is consistent with the characteristics of each dataset. Recall that in Sec. 3.3, the depth of the fusion layer influences Transformer capacity to model cross-modality relations. The deeper the fusion layers, the lower the capacity. On MM-IMDb, the dominant modality text (plot descriptions) provides more details of the movie genres than the image modality (poster). It is reasonable that the model adopts a late fusion strategy since the prediction task can easily be addressed utilizing the dominant modality, and modeling cross-modal relations only brings marginal gains. In contrast, Hateful Memes dataset is constructed to fail models that rely on a single modality by adding challenging samples ("benign confounders") to the dataset. Therefore, to handle this dataset, the model should have enough capacity to model cross-modal relations. For a dataset that relies on both modalities to make an accurate prediction, it is reasonable for our method to learn an early fusion strategy.

5.6. Ablation Study

Comparison with a new baseline. Training with missing modalities is a simple way to improve model robustness. We implement this method as a new baseline. Results are shown in Table 7. Our experiments show that this simple method does not work. As shown, the performance of the new baseline is even worse than the unimodal baseline (Image only) on Food-101 and Hateful Memes.

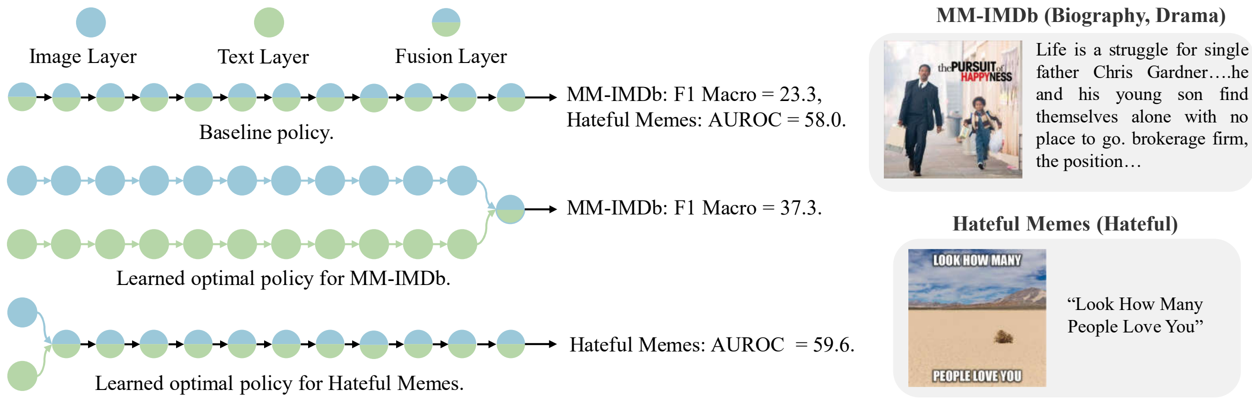

Figure 3. *Left*: Visualization of the learned policy. *Right*: Example sample from MM-IMDb and Hateful Memes. *Late fusion yields the best robustness on MM-IMDb, while early fusion leads to the most robust model on Hateful Memes.* The reported results were obtained using the following settings: training with 100% Image + 100% Text, testing with 100% Image + 10% Text.

Table 7. Results of the new baseline: training and testing with 100% image + 30% text.

Method	MM-IMDb	Food-101	Hateful Memes
Image only	31.2	65.9	60.2
New baseline	40.4	44.3	59.7
Ours	**46.6**	**77.5**	**61.2**

Table 8. Ablation study on multi-task learning and optimal fusion on MM-IMDb.

Method		Training		Testing		F1 Macro
Multi-task	Opt. Policy	Image	Text	Image	Text	
✓		100%	100%	100%	30%	31.2
	✓	100%	100%	100%	30%	28.6
✓	✓	100%	100%	100%	30%	**41.8**
✓		100%	100%	100%	10%	22.6
	✓	100%	100%	100%	10%	17.3
✓	✓	100%	100%	100%	10%	**37.3**

Analysis on the multi-task learning and optimal fusion layers. We conduct experiments to validate the effectiveness of each component under two different evaluation settings, *i.e.,* 30% or 10% text are available. Results are shown in Table 8. Both components improve Transformer robustness. Furthermore, we find that multi-task learning contributes more than the fusion strategy. In detail, when only 10% of text is available at test, multi-task learning outperforms the optimal fusion policy by 30%.

Analysis on the attention mask. In our method, we apply masks on the attention matrix to enforce the classification tokens to leverage information only from the corresponding modalities. We study the effect of attention masks. The results are shown in Table 9. We observe that it is important to make sure that each classification token is not peeking the information from other modalities.

Table 9. Ablation study on the effect of attention mask for multi-task learning on MM-IMDb.

Method	Training		Testing		F1 Macro
	Image	Text	Image	Text	
without masking	100%	100%	100%	10%	23.0
with masking	100%	100%	100%	10%	37.3

6. Conclusion

We empirically find that Transformer models are sensitive to missing-modal data. And surprisingly, the optimal fusion strategy is dataset-dependent; there does not exist a universal strategy that works in the presence of modal-incomplete data. Based on the findings, we build a robust Transformer via multi-task optimization. We develop an algorithm that automatically searches the optimal fusion strategies on different datasets. The searching for optimal fusion layers and network training are formulated into a bi-level optimization problem. Experiments across multiple benchmark datasets verify the superior robustness of our method. The limitation of our method is that multi-task learning can only ensure the multimodal performance is not worse than the unimodal one, which may not meet the requirements of safety-critical systems, such as autonomous driving. We plan to explore the effectiveness of generative-based methods, *e.g.*, reconstructing the missing tokens to improve Transformer robustness against missing modality.

Acknowledgements

This work is partially supported by NSF (CMMI-2039857) Research Grant and Snap Gift Research Grant.

References

[1] Hassan Akbari, Linagzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. VATT: Transformers for multimodal self-supervised learning from raw video, audio and text. In *NeurIPS*, 2021. 1

[2] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. In *ICLRW*, 2017. 1, 2, 3, 7

[3] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 1, 2, 4

[4] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *ECCV*, 2014. 6

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 4

[6] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, pages 104–120. Springer, 2020. 1

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1, 3, 4, 6

[8] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *ECCV*, pages 214–229. Springer, 2020. 1

[9] Ignazio Gallo, Gianmarco Ria, Nicola Landro, and Riccardo La Grassa. Image and text fusion for upmc food-101 using bert and cnns. In *IVCNZ*, pages 1–6, 2020. 3, 6

[10] Yunhui Guo, Honghui Shi, Abhishek Kumar, Kristen Grauman, Tajana Rosing, and Rogerio Feris. SpotTune: Transfer learning through adaptive fine-tuning. In *CVPR*, 2019. 3, 4

[11] Yizeng Han, Gao Huang, Shiji Song, Le Yang, Honghui Wang, and Yulin Wang. Dynamic neural networks: A survey. *arXiv preprint arXiv:2102.04906*, 2021. 3

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 2

[13] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 2

[14] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016. 3

[15] Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*, 2019. 3, 4, 6

[16] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Casey A. Fitzpatrick, Peter Bull, Greg Lipstein, Tony Nelli, Ron Zhu, Niklas Muennighoff, et al. The hateful memes challenge: Competition report. In *NeurIPS 2020 Competition and Demonstration Track*, pages 344–360, 2021. 4

[17] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. In *NeurIPS*, volume 33, 2020. 1, 2, 4, 6, 7

[18] Wonjae Kim, Bokyung Son, and Ildoo Kim. ViLT: Vision-and-language transformer without convolution or region supervision. In *ICML*, pages 5583–5594, 2021. 1, 2, 3, 4, 5, 6, 7

[19] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[20] Gueorgi Kossinets. Effects of missing data in social networks. *Social networks*, 28(3):247–268, 2006. 1

[21] Sangho Lee, Youngjae Yu, Gunhee Kim, Thomas Breuel, Jan Kautz, and Yale Song. Parameter efficient multimodal transformers for video representation learning. In *ICLR*, 2021. 1, 2, 4

[22] Junnan Li, Ramprasaath R Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *arXiv preprint arXiv:2107.07651*, 2021. 2

[23] Linjie Li et al. A closer look at the robustness of vision-and-language pre-trained models. *arXiv*, 2020. 1

[24] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 3, 4

[25] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064*, 2018. 2

[26] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *NeurIPS*, volume 32, 2019. 4, 5

[27] Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. SMIL: Multimodal learning with severely missing modality. In *AAAI*, 2021. 2, 3, 6

[28] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016. 3

[29] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. In *NeurIPS*, 2021. 1, 2

[30] Juan-Manuel Perez-Rua, Valentin Vielzeuf, Stephane Pateux, Moez Baccouche, and Frederic Jurie. MFAS: Multimodal fusion architecture search. In *CVPR*, 2019. 3

[31] Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. Found in translation: Learn-

ing robust joint representations by cyclic translations between modalities. In *AAAI*, pages 6892–6899, 2019. 2

[32] Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain. Convolutional mkl based multimodal emotion recognition and sentiment analysis. In *ICDM*, pages 439–448. IEEE, 2016. 2

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 4

[34] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *ICCV*, 2021. 1, 2

[35] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *ICCV*, pages 7464–7473, 2019. 1

[36] Ximeng Sun, Rameswar Panda, Rogerio Feris, and Kate Saenko. Adashare: Learning what to share for efficient deep multi-task learning. In *NeurIPS*, 2020. 3, 4, 5

[37] Qiuling Suo, Weida Zhong, Fenglong Ma, Ye Yuan, Jing Gao, and Aidong Zhang. Metric learning on healthcare data with incomplete modalities. In *IJCAI*, pages 3534–3540, 2019. 1

[38] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *ACL*, volume 2019, page 6558. NIH Public Access, 2019. 1

[39] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning factorized multimodal representations. In *ICLR*, 2019. 2, 6

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017. 2, 3, 5

[41] Vielzeuf et al. Centralnet: A multilayer approach for multimodal fusion. In *ECCVW*, 2018. 3, 6

[42] Haohan Wang, Aaksha Meghawat, Louis-Philippe Morency, and Eric P Xing. Select-additive learning: Improving generalization in multimodal sentiment analysis. In *ICME*, pages 949–954. IEEE, 2017. 2

[43] Xin Wang, Devinder Kumar, Nicolas Thome, Matthieu Cord, and Frédéric Precioso. Recipe recognition with large multimodal food dataset. In *ICMEW*, 2015. 1, 2, 3, 6, 7

[44] Zuxuan Wu, Tushar Nagarajan, Abhishek Kumar, Steven Rennie, Larry S Davis, Kristen Grauman, and Rogerio Feris. Blockdrop: Dynamic inference paths in residual networks. In *CVPR*, pages 8817–8826, 2018. 3, 4

[45] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *CVPR*, pages 10502–10511, 2019. 2

[46] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *EMNLP*, pages 1103–1114, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. 2

[47] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. MERLOT: Multimodal neural script knowledge models. *arXiv preprint arXiv:2106.02636*, 2021. 1, 4

[48] Zhu Zhang, Jianxin Ma, Chang Zhou, Rui Men, Zhikang Li, Ming Ding, Jie Tang, Jingren Zhou, and Hongxia Yang. UFC-BERT: Unifying multi-modal controls for conditional image synthesis. In *NeurIPS*, 2021. 1