THE 2022 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING

EMNLP 2022

Robustness of Fusion-based Multimodal Classifiers to Cross-Modal Content Dilutions

Gaurav Verma, Vishwa Vinay, Ryan A. Rossi, and Srijan Kumar

Georgia Insitute of Technology, Adobe Research




Georgia Institute of Technology

Adobe

Fusion-based Multimodal Classifiers

Safety-critical Applications

Individual

Family mourns 11 dead after church falls at baptism during Mexico earthquake.

Rescue

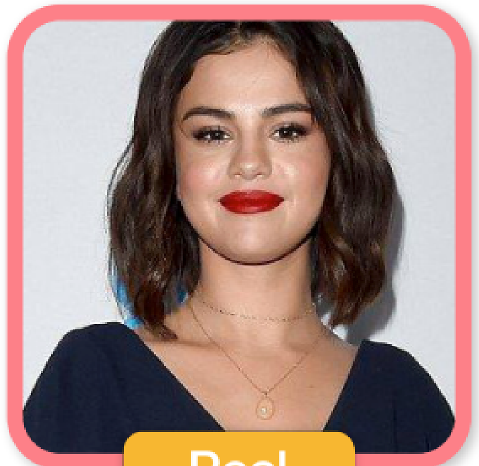
Blood donation lines in Tehran to help earthquake survivors in west of Iran.

Humanitarian information in crises

Alam et al., 2018
Ofli et al., 2020

Fake

Prince William may not attend wedding leaving Harry without a best man.



Real

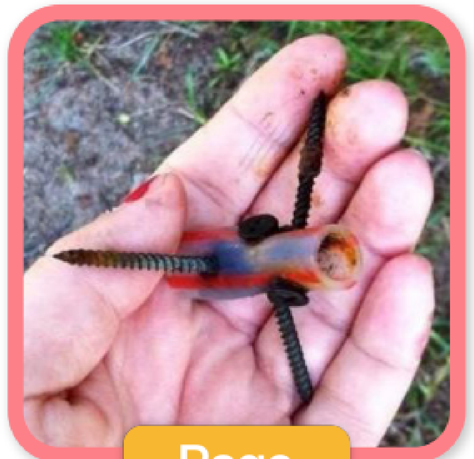
Selena Gomez says she'll protect her children like no one's business.

Fake news and hate speech detection

Shu et al., 2018
Kiela et al., 2020

Creepy

If you believe in life after death trespass here...



Rage

Someone have been throwing these into water near my home.

Emotional indicators for mental health

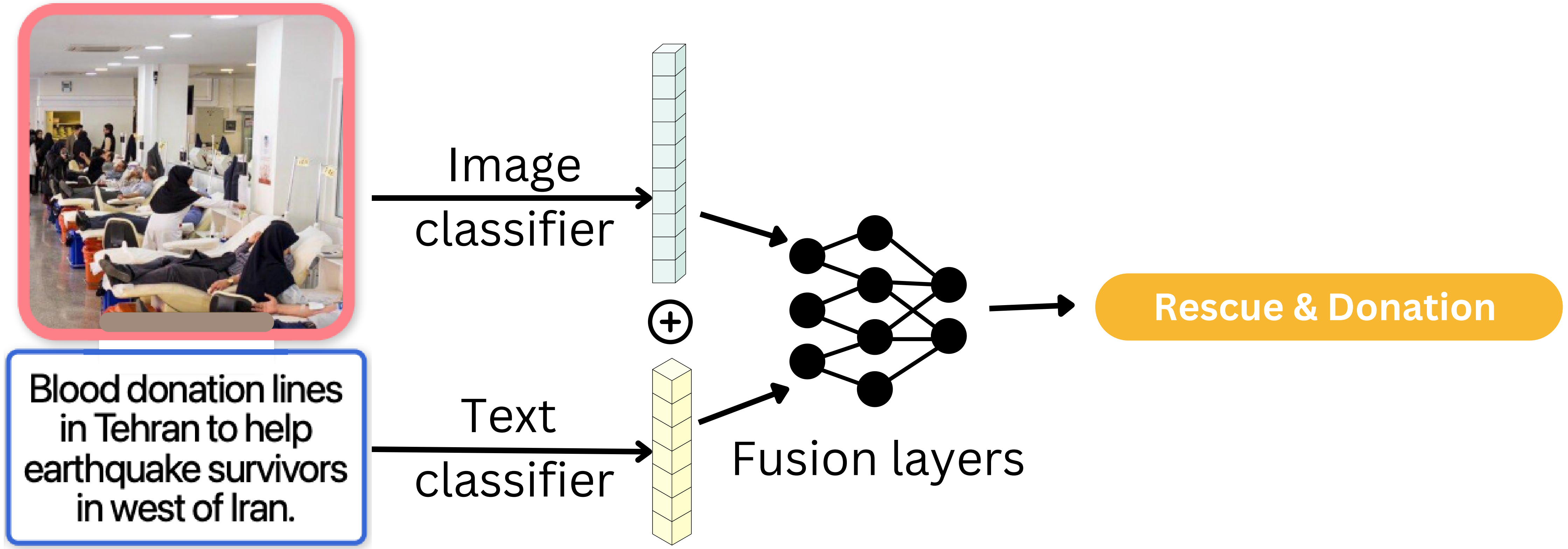
Duong et al., 2018
Xu et al., 2020

Fusion-based Multimodal Classifiers

Blood donation lines in Tehran to help earthquake survivors in west of Iran.

Humanitarian information in crises

Fusion-based Multimodal Classifiers



Humanitarian information in crises

Are multimodal classifiers robust to plausible variations?

@user1

Blood donation lines in Tehran to help earthquake survivors in west of Iran.

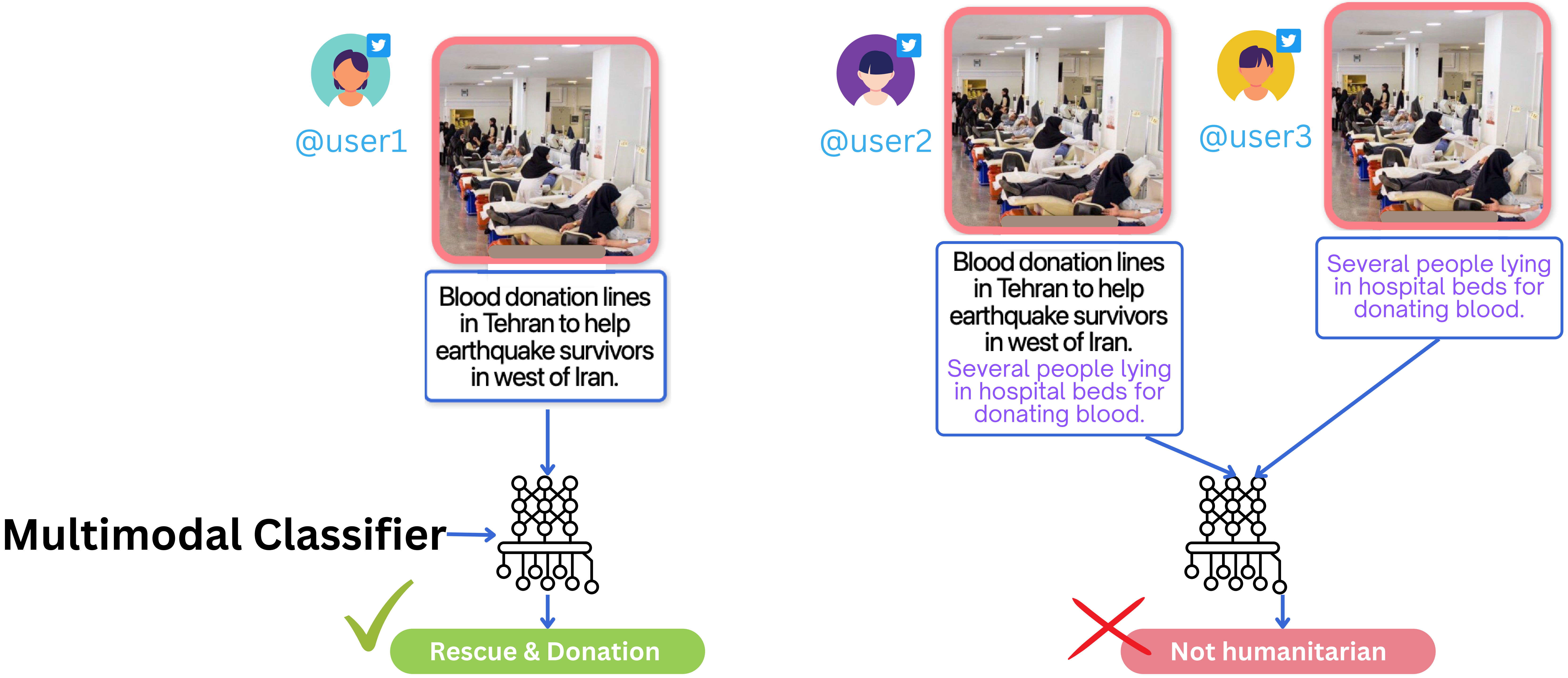
@user2

Blood donation lines in Tehran to help earthquake survivors in west of Iran.
Several people lying in hospital beds for donating blood.

@user3

Several people lying in hospital beds for donating blood.

Are multimodal classifiers robust to plausible variations?



What do we know?

About NLP and Multimodal robustness

NLP Robustness

CHECKLIST (Ribeiro et al., 2020)
Rule-based dilutions/distractions

- Random URLs
- phrases to fool the model

(Naik et al., 2018;
Ribeiro et al., 2020)

What do we know?

About NLP and Multimodal robustness

NLP Robustness

CHECKLIST (Ribeiro et al., 2020)

Rule-based dilutions/distractions

- Random URLs
- phrases to fool the model

(Naik et al., 2018; Ribeiro et al., 2020)

Multimodal Robustness

Imperceptible unimodal changes

(Li et al., 2020; Chen et al., 2020)

Adversarial examples for VQA

(Sheng et al., 2021; Li et al., 2021)

Imperceptibility doesn't constrain the plausible action space in human-facing applications.

(Gilmer et al., 2018)

What do we know?

About NLP and Multimodal robustness

NLP Robustness

CHECKLIST (Ribeiro et al., 2020)
Rule-based dilutions/distractions

- Random URLs
- phrases to fool the model

(Naik et al., 2018; Ribeiro et al., 2020)

Multimodal Robustness

Imperceptible unimodal changes
(Li et al., 2020; Chen et al., 2020)

Adversarial examples for VQA
(Sheng et al., 2021; Li et al., 2021)

Imperceptibility doesn't constrain the plausible action space in human-facing applications.
(Gilmer et al., 2018)

What are plausible variations in user-generated multimodal data?

Goal

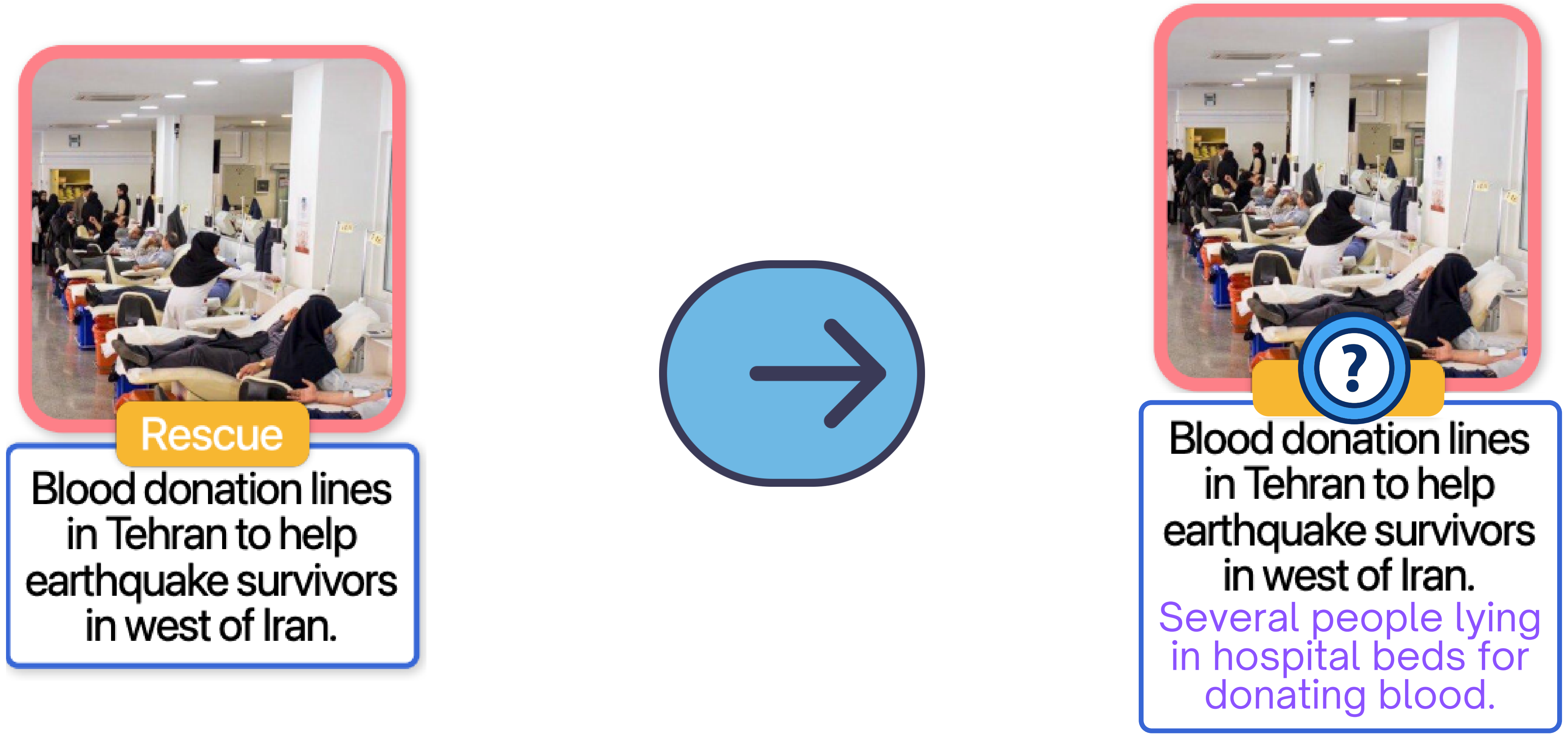
Are multimodal classifiers robust to user-generated plausible variations?

Research Question

Our work: We introduce and study robustness of multimodal classifiers to cross-modal dilutions!

Cross-Modal Dilutions

Information from one modality is added to the other corresponding modality (image → text), leading to dilution.



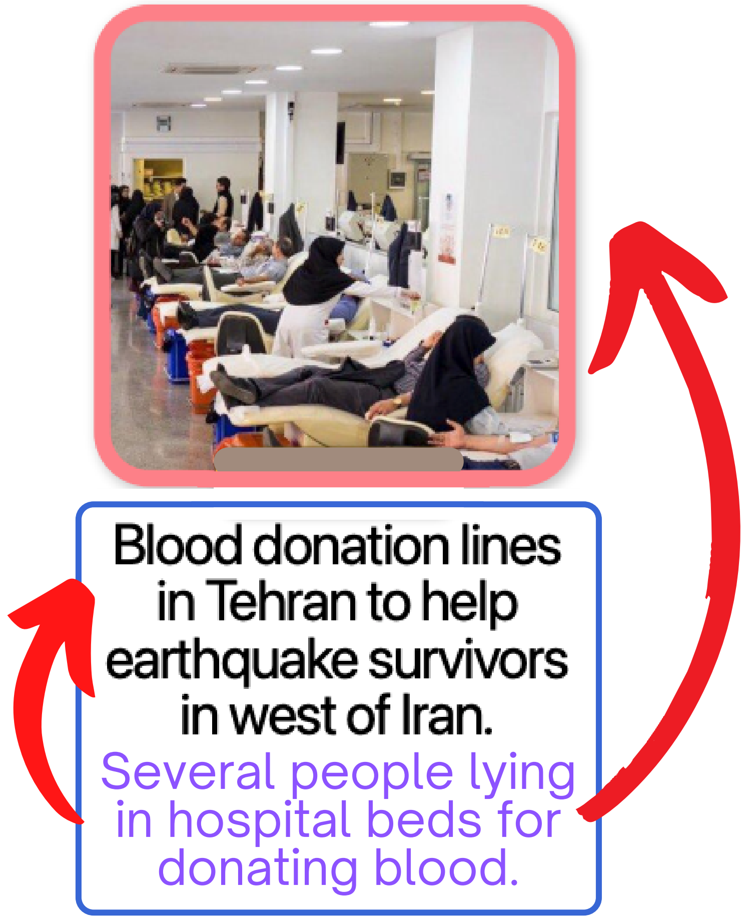
Desirable Properties of Cross-Modal Dilutions

Relevance with text

Relevance with image

Fluent

Effective



Different Approaches for Dilutions

1

Simple Dilutions

- Random URL
- Keywords from
  - Image
  - Text
  - both Image and Text

Different Approaches for Dilutions

(Cross-modal dilutions & text-only)

1

Simple Dilutions

- Random URL
- Keywords from
  - Image
  - Text
  - both Image and Text

Different Approaches for Dilutions

(Cross-modal dilutions & text-only)

1

Simple Dilutions

- Random URL
- Keywords from
    - Image
    - Text
    - both Image and Text

seen from above. entire california communities reduced to ash.

+ https:t/co/gXvDrs
+ earth rock
+ earth rock california communities reduced

Different Approaches for Dilutions

(Cross-modal dilutions & text-only)

1 Simple Dilutions

- Random URL
- Keywords from
  - Image
  - Text
  - both Image and Text

seen from above. entire california communities reduced to ash.

+ https:t/co/gXvDrs

+ earth rock

+ earth rock california communities reduced

2 Off-the-shelf Generation

- GPT-2's generation
- Fine-tuned GPT-2's gen.
- Image captioning models

Different Approaches for Dilutions

(Cross-modal dilutions & text-only)

1
Simple Dilutions

- Random URL
- Keywords from
  - Image
  - Text
  - both Image and Text

seen from above. entire california communities reduced to ash.

⊕ https:t/co/gXvDrs

⊕ earth rock

⊕ earth rock california communities reduced

2
Off-the-shelf Generation

- GPT-2's generation
- Fine-tuned GPT-2's gen.
- Image captioning models

seen from above. entire california communities reduced to ash.

⊕ 2 days of rain in one day.

⊕ broken rocks lying on the ground.

Different Approaches for Dilutions

(Cross-modal dilutions & text-only)

1 Simple Dilutions

- Random URL
- Keywords from
  - Image
  - Text
  - both Image and Text

seen from above. entire california communities reduced to ash.

+ https:t/co/gXvDrs

+ earth rock

+ earth rock california communities reduced

2 Off-the-shelf Generation

- GPT-2's generation
- Fine-tuned GPT-2's gen.
- Image captioning models

seen from above. entire california communities reduced to ash.

+ 2 days of rain in one day.

+ broken rocks lying on the ground.

3 Cross-Modal Dilution Generator (XMD; Ours)

Train a language model to perform constrained generation using image and text keywords **and** encourage misclassification

seen from above. entire california communities reduced to ash.
the devastation in California: why have entire communities either been destroyed or reduced to a few bare earth bare rock formation?

Desirable Properties

Relevance with text | Relevance with image | Fluent | Effective

1

Simple Dilutions

- Random URL
- Keywords from
  - Image
  - Text
  - both Image and Text

2

Off-the-shelf Generation

- GPT-2's generation
- Fine-tuned GPT-2's gen.
- Image captioning models

3

Cross-Modal Dilution Generator (XMD; Ours)

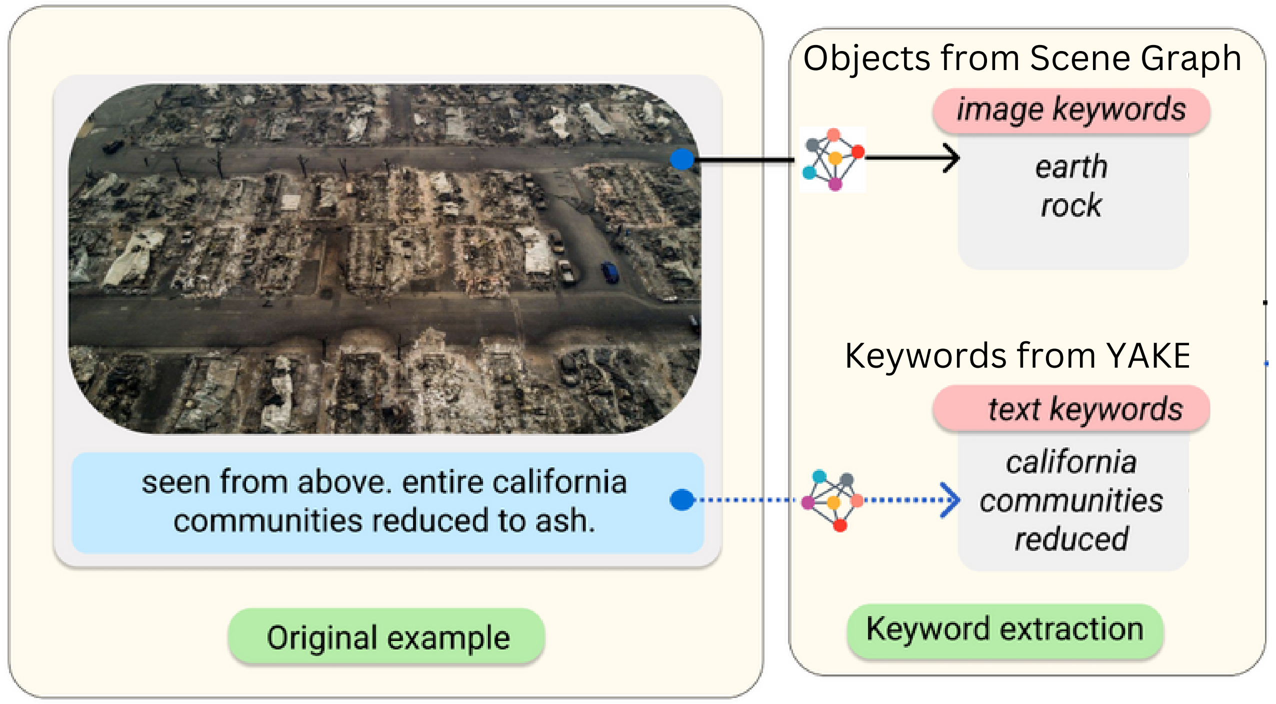
Desirable Properties

	Relevance with text	Relevance with image	Fluent	Effective
1 Simple Dilutions				
• Random URL				
• Keywords from				
○ Image		✓		
○ Text	✓			
○ both Image and Text	✓	✓		
2 Off-the-shelf Generation				
• GPT-2's generation	✓		✓	
• Fine-tuned GPT-2's gen.	✓		✓	
• Image captioning models		✓	✓	
3 Cross-Modal Dilution Generator (XMD; Ours)	✓	✓	✓	✓

Cross-Modal Dilutions Generator (XMD)

Two-stage, multi-task adversarial fine-tuning

Proposed model:

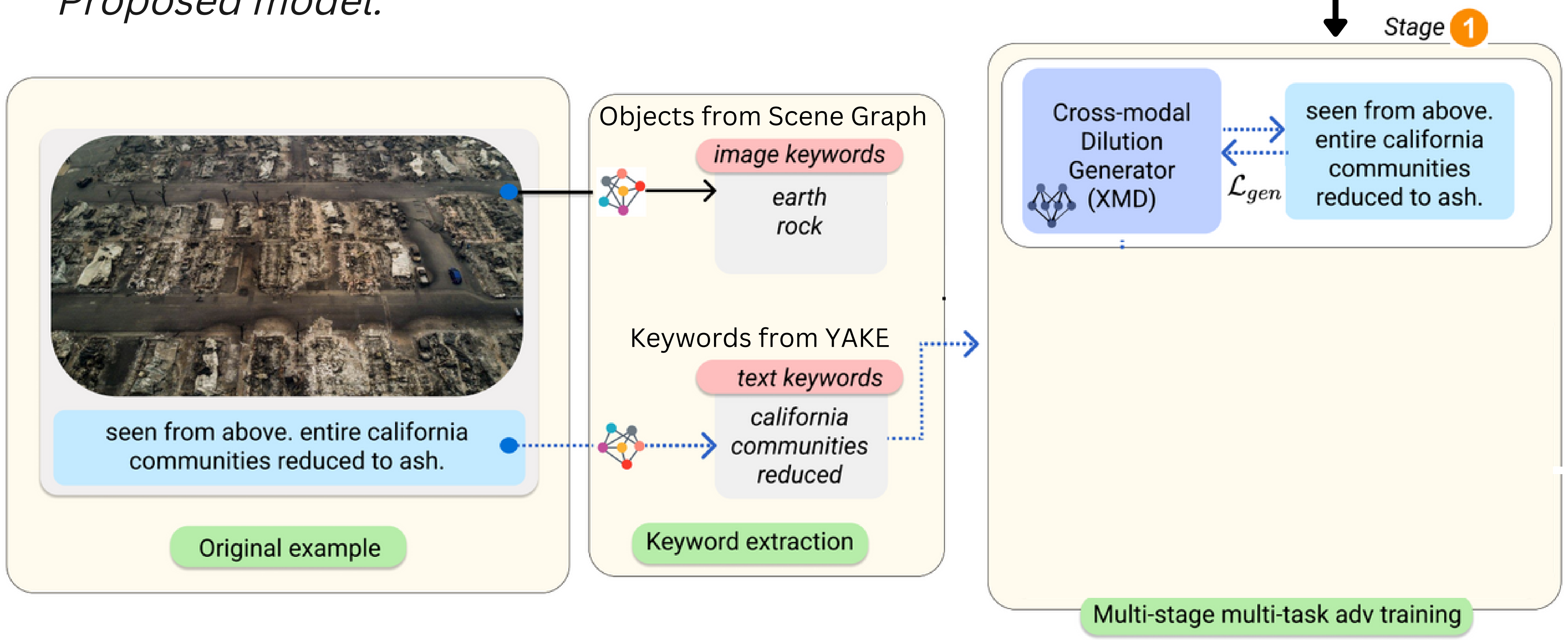


Cross-Modal Dilutions Generator (XMD)

Two-stage, multi-task adversarial fine-tuning

Base generation model: hard-constrained generation model trained on Wikipedia (Zhang et al., 2020)

Proposed model:



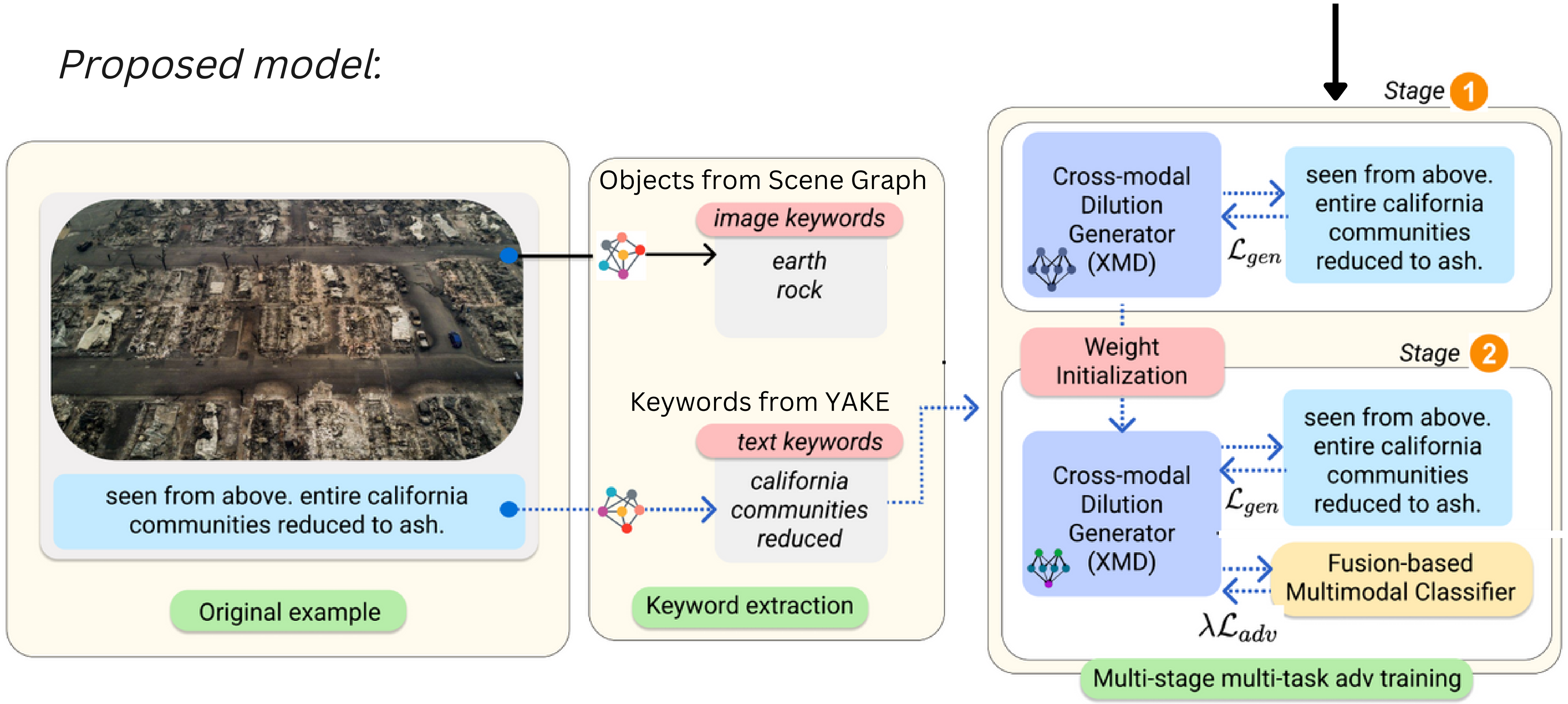
Domain adaptation
model learns to generate domain-specific text from keywords

Cross-Modal Dilutions Generator (XMD)

Two-stage, multi-task adversarial fine-tuning

Base generation model: hard-constrained generation model trained on Wikipedia (Zhang et al., 2020)

Proposed model:



Domain adaptation
model learns to generate domain-specific text from keywords

$$\mathcal{L}_{combined} = \mathcal{L}_{gen} + \lambda\mathcal{L}_{adv}$$

$$\mathcal{L}_{adv} = -(y\log(\hat{y}) + (1-y)\log(1-\hat{y}))$$

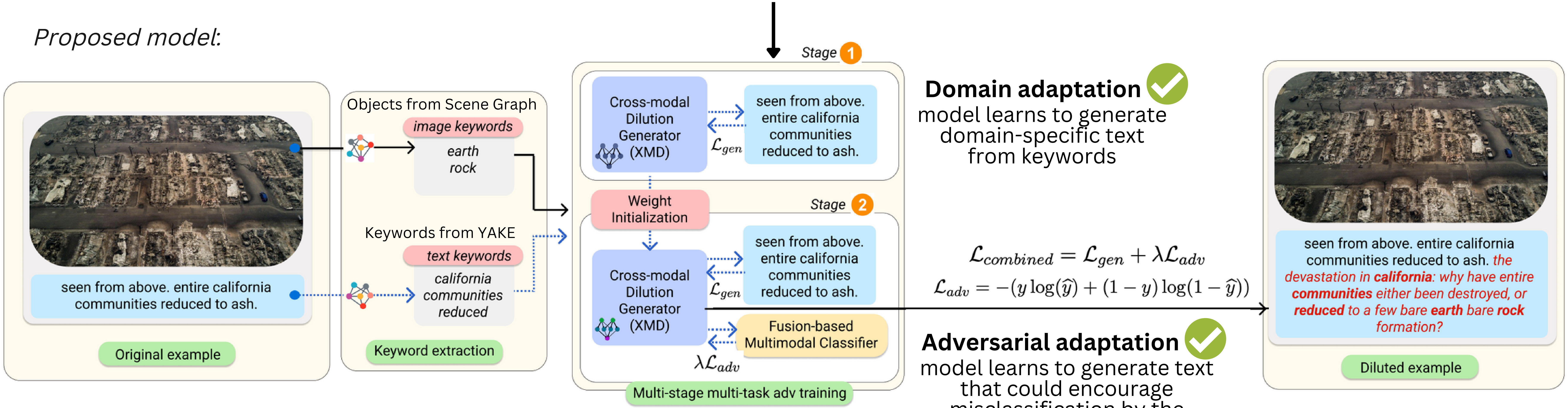
Adversarial adaptation
model learns to generate text that could encourage misclassification by the multimodal classifier

Cross-Modal Dilutions Generator (XMD)

Two-stage, multi-task adversarial fine-tuning

Base generation model: hard-constrained generation model trained on Wikipedia (Zhang et al., 2020)

Proposed model:



Domain adaptation
model learns to generate domain-specific text from keywords

Adversarial adaptation
model learns to generate text that could encourage misclassification by the multimodal classifier

How robust are multimodal classifiers?

And how well can we generate cross-modal dilutions?

Datasets

Dilution Baselines

Metrics

How robust are multimodal classifiers?

And how well can we generate cross-modal dilutions?

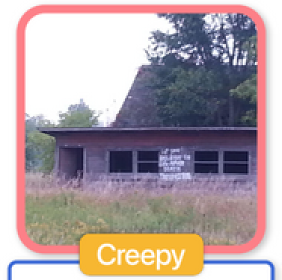
Datasets

Crisis Humanitarianism Dataset
(7,216 examples, 5 classes)
(Alam et al., 2018; Ofli et al., 2020)

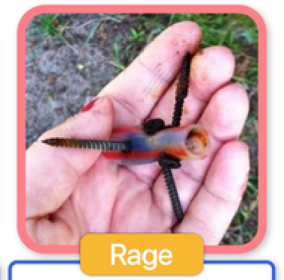
Family mourns 11 dead after church falls at baptism during Mexico earthquake.

Blood donation lines in Tehran to help earthquake survivors in west of Iran.

Emotion Detection Dataset
(3,207 examples; 4 classes)
(Duong et al., 2018)



If you believe in life after death trespass here...



Someone have been throwing these into water near my home.

Dilution Baselines

Metrics

How robust are multimodal classifiers?

And how well can we generate cross-modal dilutions?

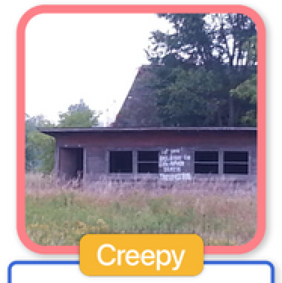
Datasets

Crisis Humanitarianism Dataset
(7,216 examples, 5 classes)
(Alam et al., 2018; Ofli et al., 2020)

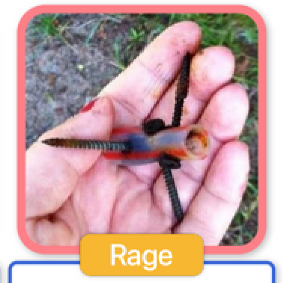
Family mourns 11 dead after church falls at baptism during Mexico earthquake.

Blood donation lines in Tehran to help earthquake survivors in west of Iran.

Emotion Detection Dataset
(3,207 examples; 4 classes)
(Duong et al., 2018)



If you believe in life after death trespass here...



Someone have been throwing these into water near my home.

Dilution Baselines

Rule-based
- Random URL
- Image KW
- Text KW
- Text + Image KW

Model-based
- GPT-2
- GPT-2 Fine-tuned
- SCST Captions
- XLAN Captions

XMD (Ours)

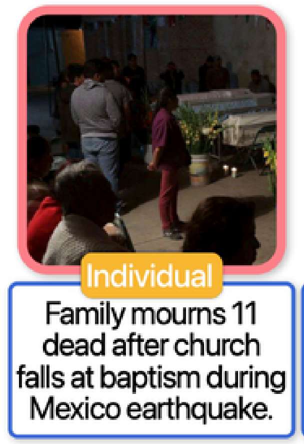
Metrics

How robust are multimodal classifiers?

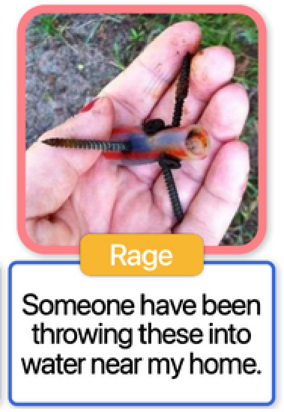
And how well can we generate cross-modal dilutions?

Datasets

Crisis Humanitarianism Dataset
(7,216 examples, 5 classes)
(Alam et al., 2018; Ofli et al., 2020)



Emotion Detection Dataset
(3,207 examples; 4 classes)
(Duong et al., 2018)



Dilution Baselines

Rule-based
- Random URL
- Image KW
- Text KW
- Text + Image KW

Model-based
- GPT-2
- GPT-2 Fine-tuned
- SCST Captions
- XLAN Captions

XMD (Ours)

Metrics

How do dilutions impact classification performance?
- F1 score,
- Precision,
- Recall, and
- Accuracy

How relevant are dilutions to...
- original text (BERT similarity)
- image (CLIP similarity)

Are generated dilutions topically coherent?
- KL Divergence

Are generated dilutions realistic?
- Human evaluation

How robust are multimodal classifiers?

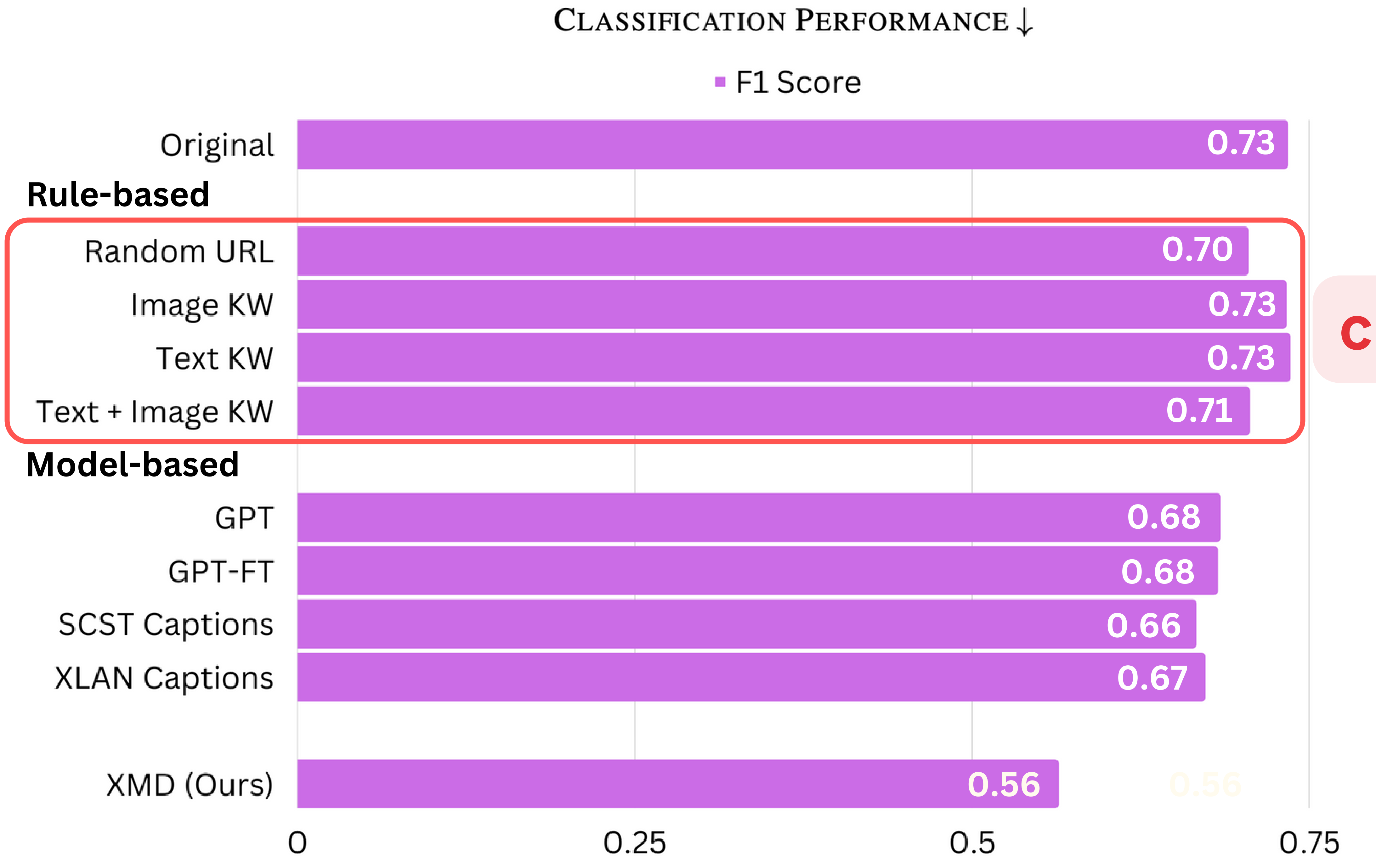
And how well can we generate cross-modal dilutions?

Results on Crisis Humanitarianism Dataset

How robust are multimodal classifiers?

And how well can we generate cross-modal dilutions?

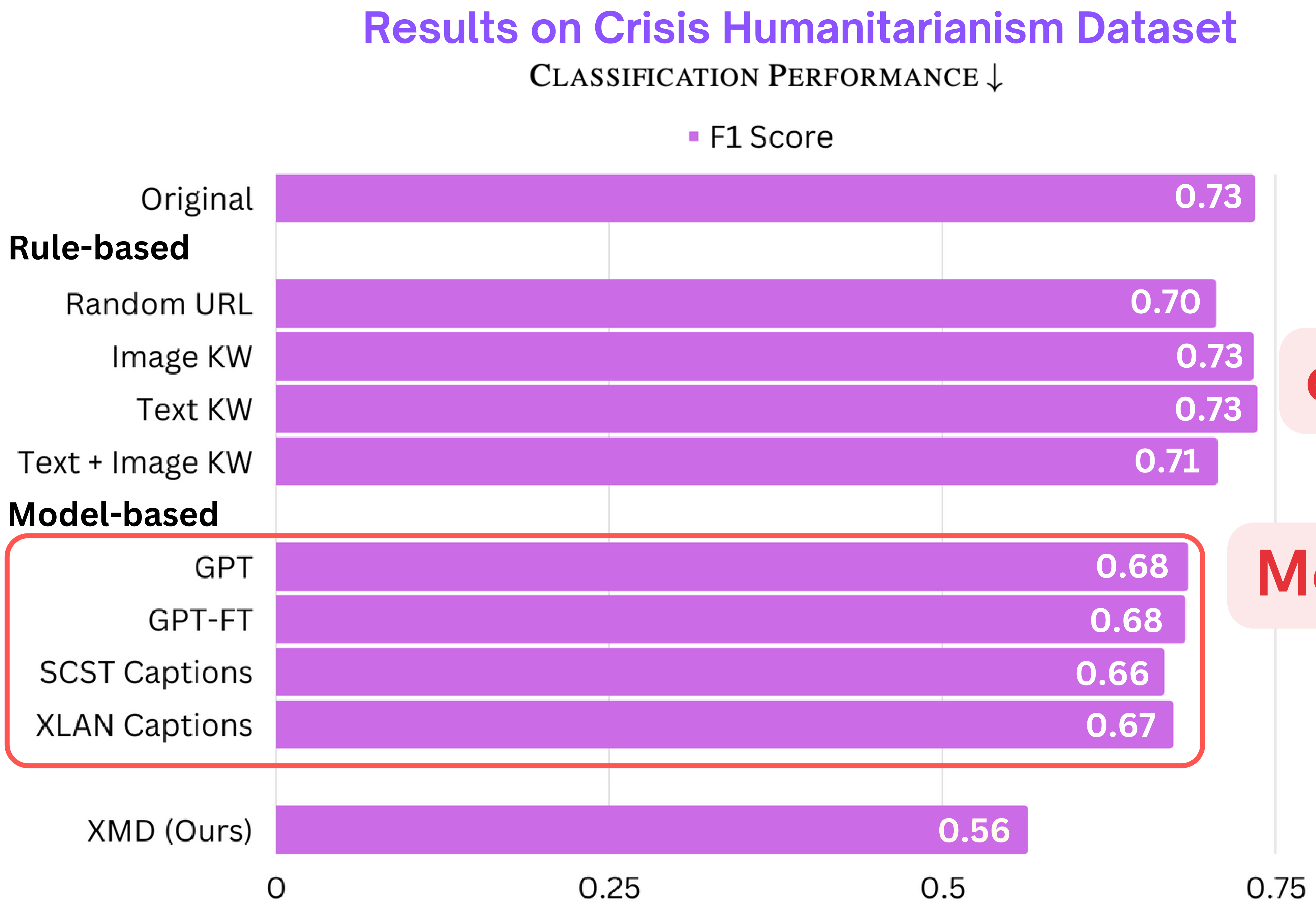
Results on Crisis Humanitarianism Dataset



No major effect on classification performance with rule-based

How robust are multimodal classifiers?

And how well can we generate cross-modal dilutions?



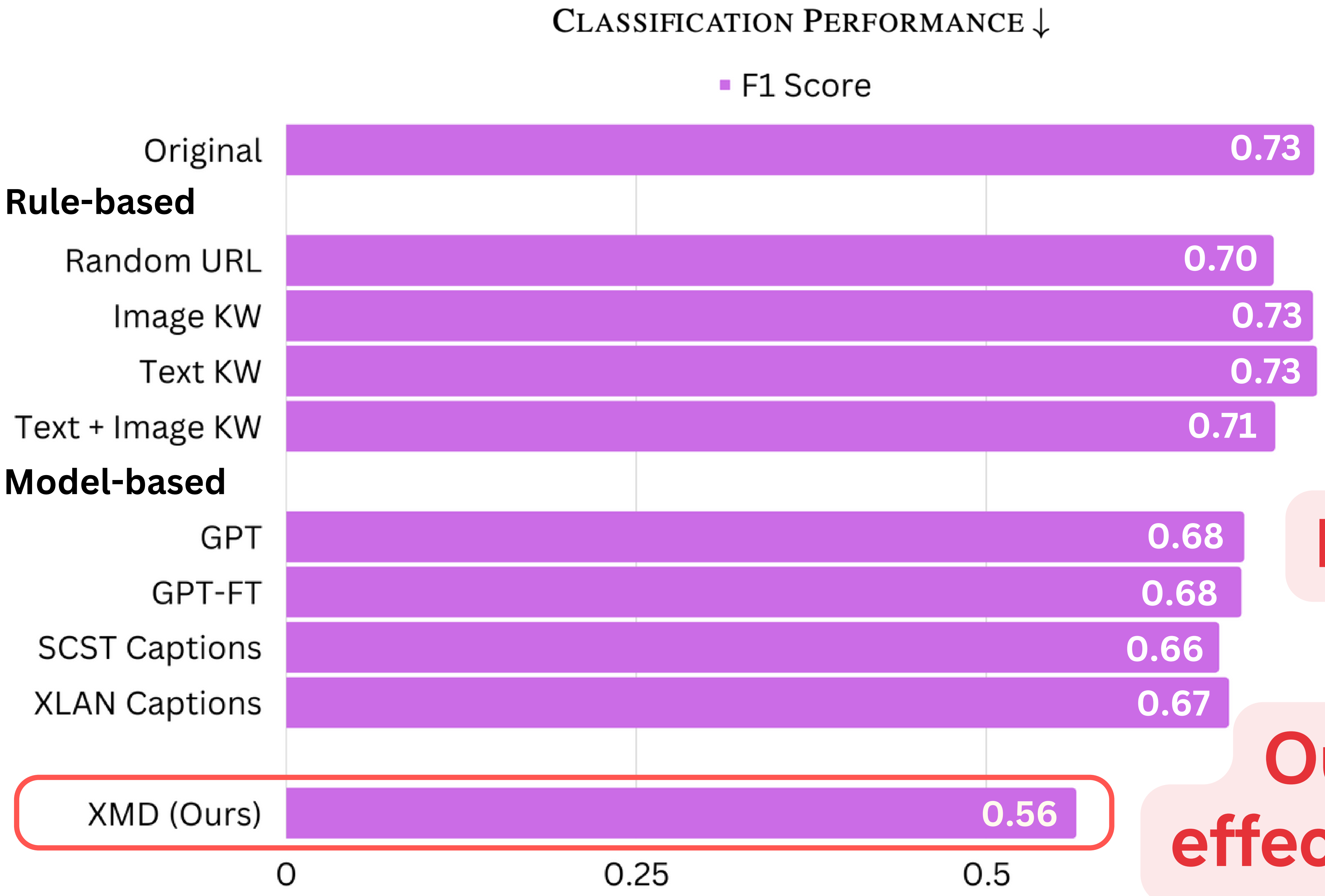
No major effect on classification performance with rule-based

Model-based baselines are effective!

How robust are multimodal classifiers?

And how well can we generate cross-modal dilutions?

Results on Crisis Humanitarianism Dataset



No major effect on classification performance with rule-based

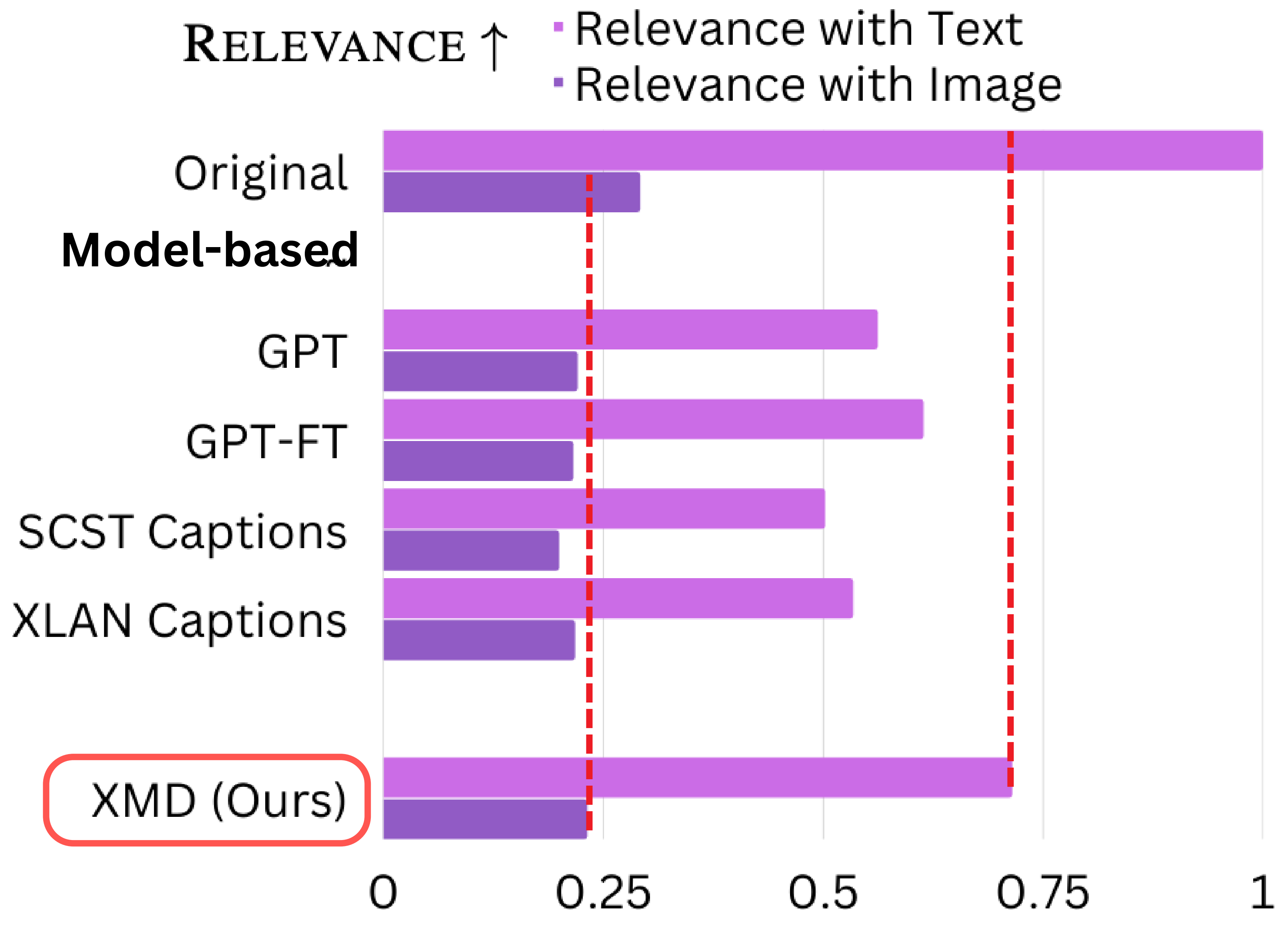
Model-based baselines are effective!

Our approach is most effective; ~23% drop in F1!

How robust are multimodal classifiers?

And how well can we generate cross-modal dilutions?

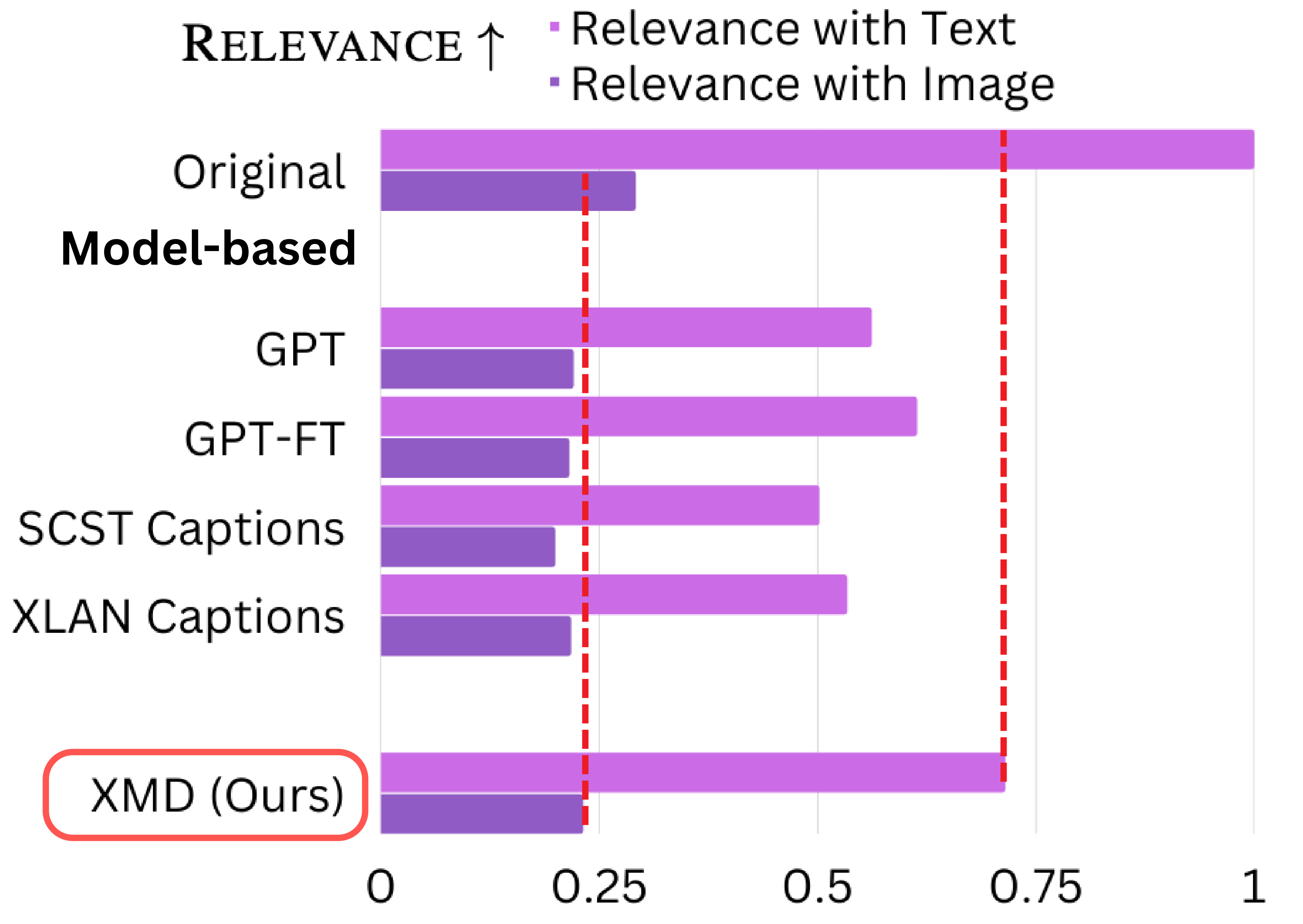
Results on Crisis Humanitarianism Dataset

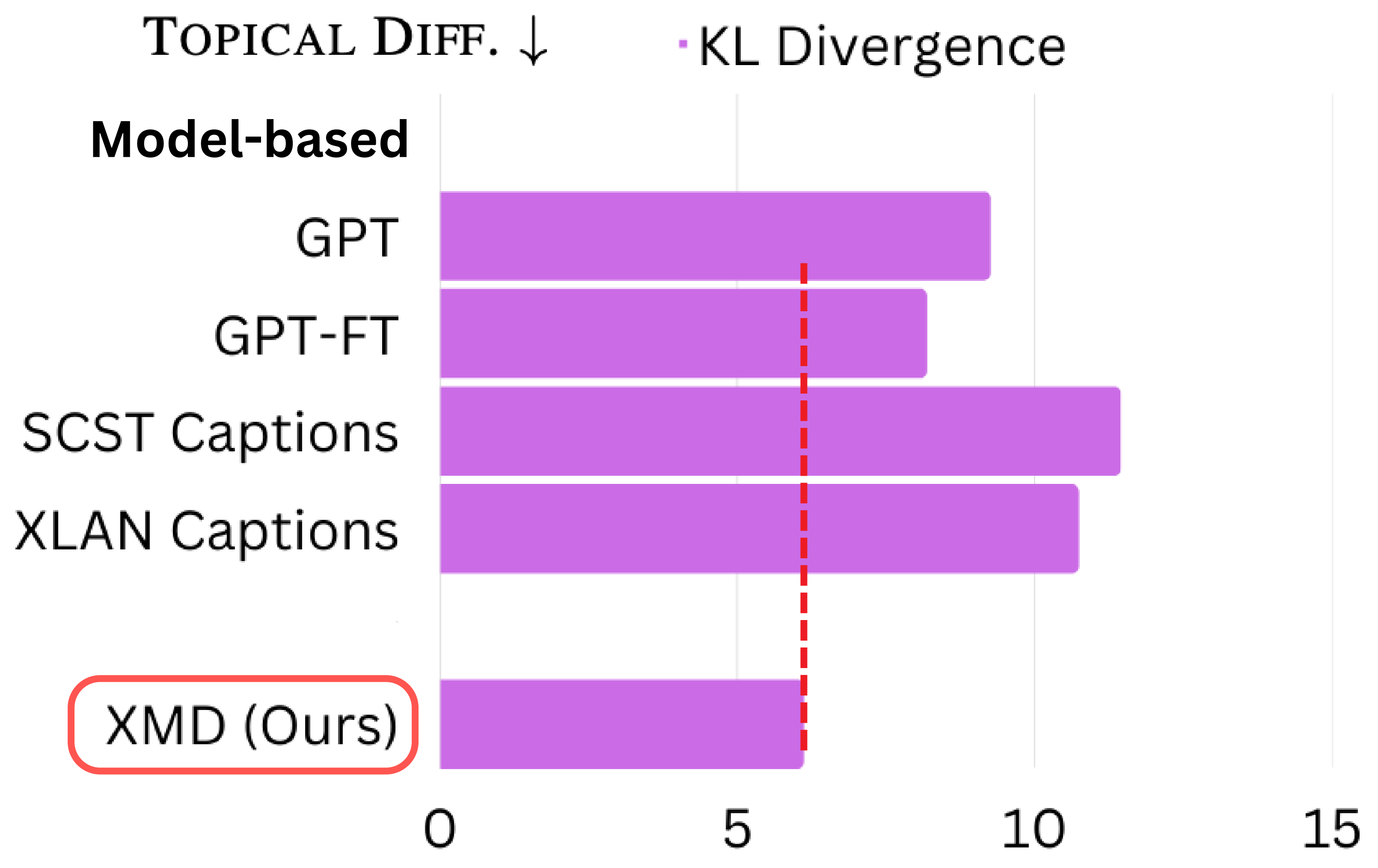


How robust are multimodal classifiers?

And how well can we generate cross-modal dilutions?

Results on Crisis Humanitarianism Dataset





Most relevant and topically coherent dilutions!

How robust are multimodal classifiers?

And how well can we generate cross-modal dilutions?

Results on Crisis Humanitarianism Dataset

seen from above. entire california communities reduced to ash.
it has caused communities to distribute food due to the heavy rains and fly out to neighboring counties with their children.

Most competitive baseline (GPT-FT)

VS

seen from above. entire california communities reduced to ash.
the devastation in California: why have entire communities either been destroyed or reduced to a few bare earth bare rock formation?

XMD (Ours)

Human evaluation

For 78.5% of examples, the majority of annotators consider our dilutions to be better!

How robust are multimodal classifiers?

And how well can we generate cross-modal dilutions?

Results on Crisis Humanitarianism Dataset

seen from above. entire california communities reduced to ash.

vs

seen from above. entire california communities reduced to ash. the devastation in California: why have entire communities either been destroyed or reduced to a few bare earth bare rock formation?

Unmodified multimodal post

Diluted using XMD (Ours)

Human evaluation

Annotators fail to distinguish diluted examples from unmodified examples!

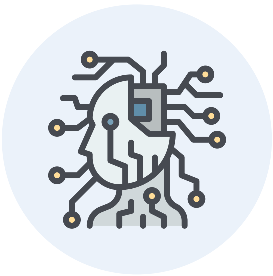
Robustness of Fusion-based Multimodal Classifiers to Cross-Modal Content Dilutions

EMNLP 2022

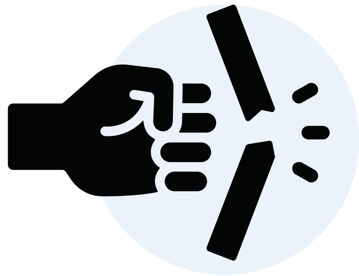
Gaurav Verma, Vishwa Vinay, Ryan A. Rossi, and Srijan Kumar

As multimodal learning is used for AI for **Social Good** applications, we must think about its **robustness**.

Consider not just **imperceptible** but also **plausible variations** in user-generated data!



XMD: Method to generate relevant and **realistic dilutions** that **effectively** highlight vulnerabilities



Fusion-based multimodal classifiers are **not** robust to realistic cross-modal content dilutions

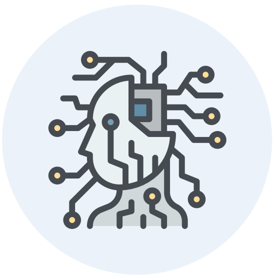
Robustness of Fusion-based Multimodal Classifiers to Cross-Modal Content Dilutions

EMNLP 2022

Gaurav Verma, Vishwa Vinay, Ryan A. Rossi, and Srijan Kumar

As multimodal learning is used for AI for **Social Good** applications, we must think about its **robustness**.

Consider not just **imperceptible** but also **plausible variations** in user-generated data!



XMD: Method to generate relevant and **realistic dilutions** that **effectively** highlight vulnerabilities

Fusion-based multimodal classifiers are **not** robust to realistic cross-modal content dilutions

gverma@gatech.edu

Project Webpage: *claws-lab.github.io/multimodal-robustness/* with **Code** and **Colab** notebook