



東北大學
Northeastern University



TIGER: A Unified Generative Model Framework for Multimodal Dialogue Response Generation

——LREC-COLING 2024

Fanheng Kong, Peidong Wang, Shi Feng[†], Daling Wang, Yifei Zhang
School of Computer Science and Engineering, Northeastern University, China



東北大學
Northeastern University

PATR 01

Motivation & Contribution

Open-domain conversational agents have shown great performance on text-only dialogue generation.
(e.g., DialoGPT^[1], Menna^[2], BlenderBot^[3]...)

Relying on text-only modality falls short of the rich visual perception of the real world.



Multimodal dialogue → interaction and expressiveness.

Input: Dialogue Context
Output: Multimodal Response (Text+Image)

[1] Yizhe Zhang, et al. 2020. DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278.

[2] Daniel Adiwardana, et al. 2020. Towards a Human-like Open-Domain Chatbot. *arXiv preprint arXiv:2001.09977*.

[3] Stephen Roller, et al. 2021. Recipes for Building an Open-Domain Chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325.

Multimodal dialogue

Input: Dialogue Context

Output: Multimodal Response (Text+Image)

A: I am doing great. Do you have anything exciting planned for the weekend?

B: no, what about you?

A: I am helping my uncle Enzo. He is an attorney. He has a court date for next week.

B: oh! interesting

A: He is a defense attorney. Do you know anyone who works in the judicial system?

B: doesn't your uncle a lawyer? No



He is the blurred man in the front. That is my uncle, Enzo.

Existing research on multimodal dialogues primarily focuses on:

- (1) textual response generation that ground the conversation on a given image; (Output: Text-only)
 - (2) visual response selection based on the dialogue context. (Retrieval Methods)
-

Rare multimodal dialogue models with the ability to generate multimodal content:

- (1) which modal content should be generated as the response?
- (2) inability to guarantee consistently high-quality and contextually appropriate images generated from the dialogue context

- We propose **mulTI**modal **GE**nerator for dialogue **R**esponse (**TIGER**), a unified generative framework designed for multimodal dialogue response generation. Notably, this framework is capable of handling conversations involving any combination of modalities.
- We implement a system for multimodal dialogue response generation, incorporating both text and images, based on TIGER.
- Extensive experiments show that TIGER achieves new state-of-the-art results on both automatic and human evaluations, which validate the effectiveness of our system in providing a superior multimodal conversational experience.



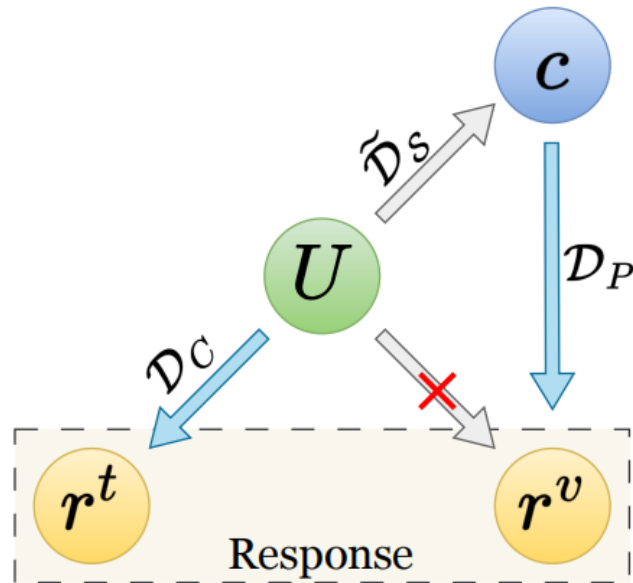
東北大學
Northeastern University

PATR 02

Approach

2.1 Low-resource Setting

the limitation of less available multimodal dialogue instances \longrightarrow low-resource setting



Input: a dialogue context U

Output: a textual response r_t or a visual response r_v

$U \rightarrow r_t$

$U \rightarrow r_v$ (less effective)

A feasible approach:

- introduce text-to-image into text-only dialogue
- large-scale text dialogue data \mathcal{D}_C and image-text pairs \mathcal{D}_P assist \implies small-scale multimodal dialogue data $\tilde{\mathcal{D}}_s$

\mathcal{D}_C : large-scale text dialogue data (e.g., Reddit comments)

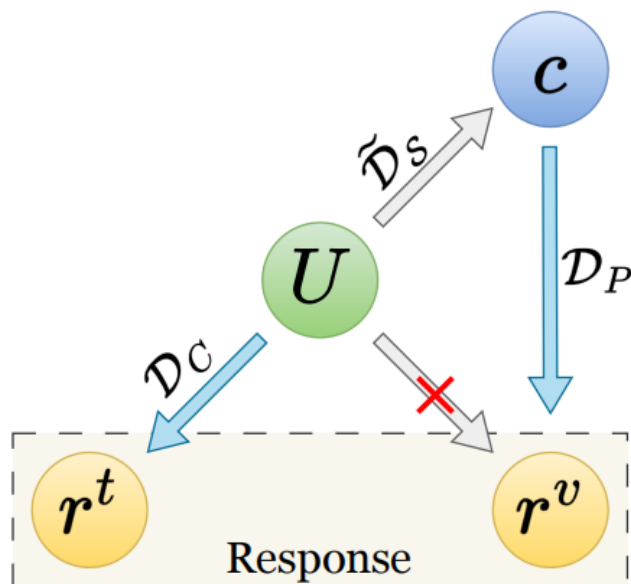
\mathcal{D}_P : large-scale image-text pair data (e.g., LAION-5B)

$\tilde{\mathcal{D}}_s$: small-scale multimodal dialogue data

2.1 Low-resource Setting

Target: learn a **generative** multimodal dialogue model $P(R | U; \theta)$

with $D = \{D_C, D_P, \tilde{D}_S\}$

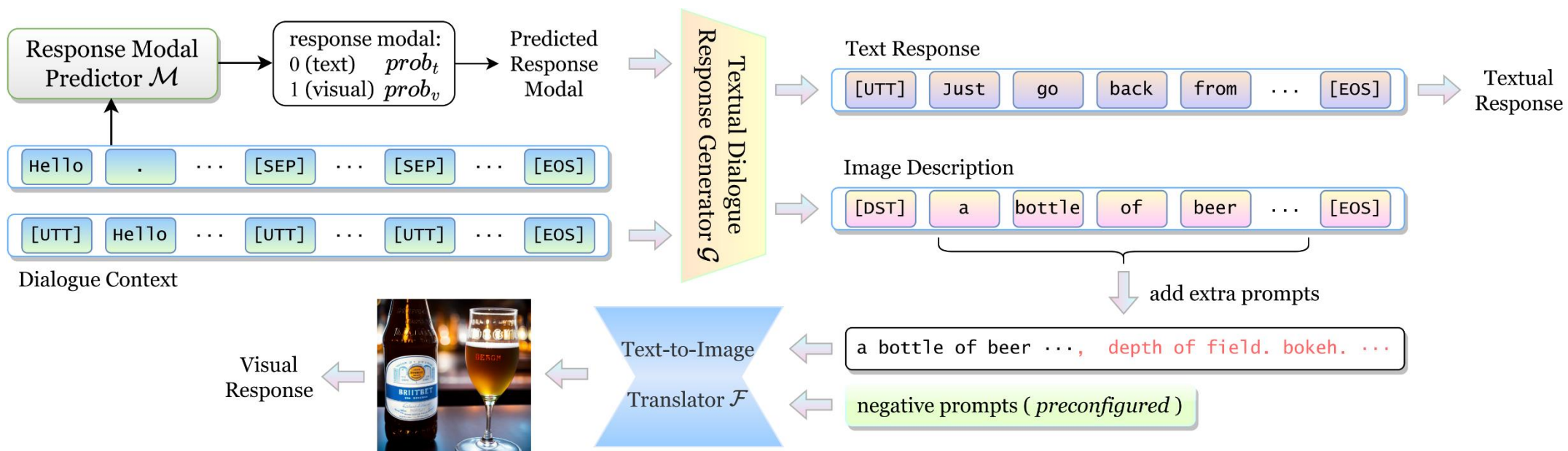


D_C : large-scale text dialogue data (e.g., Reddit comments)

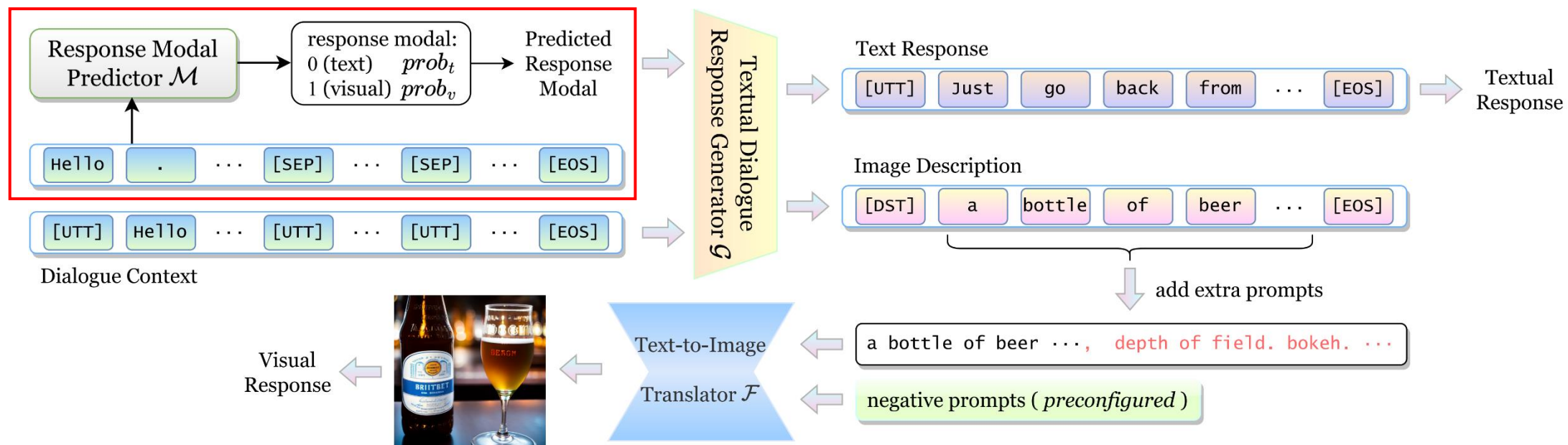
D_P : large-scale image-text pair data (e.g., LAION-5B)

\tilde{D}_S : small-scale multimodal dialogue data

2.2 TIGER

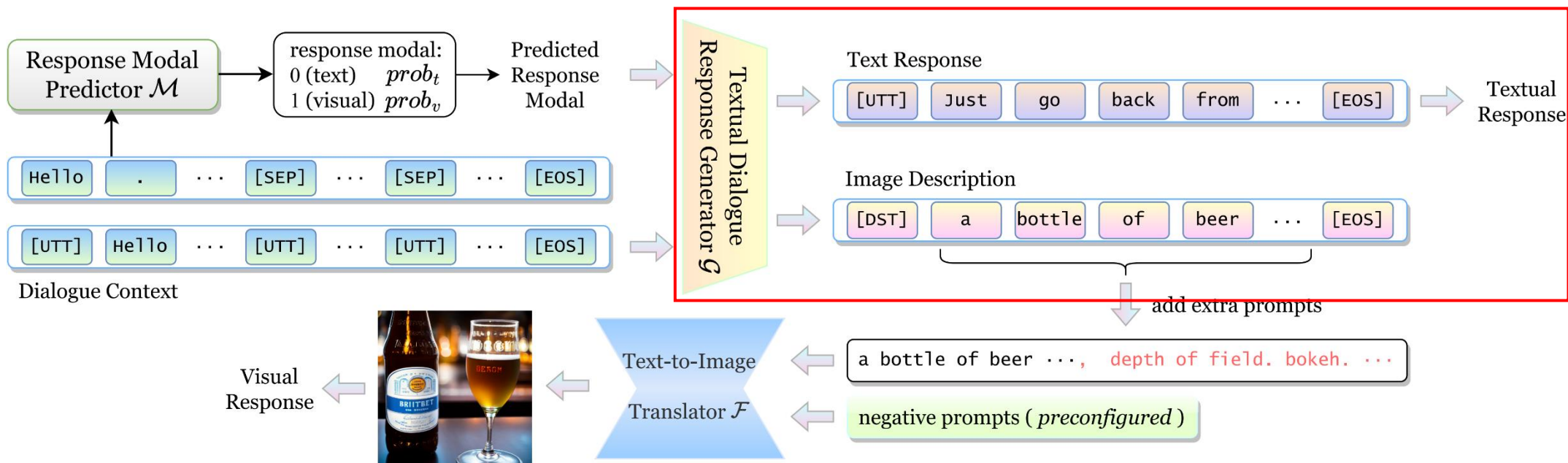


2.2 TIGER



TIGER framework:

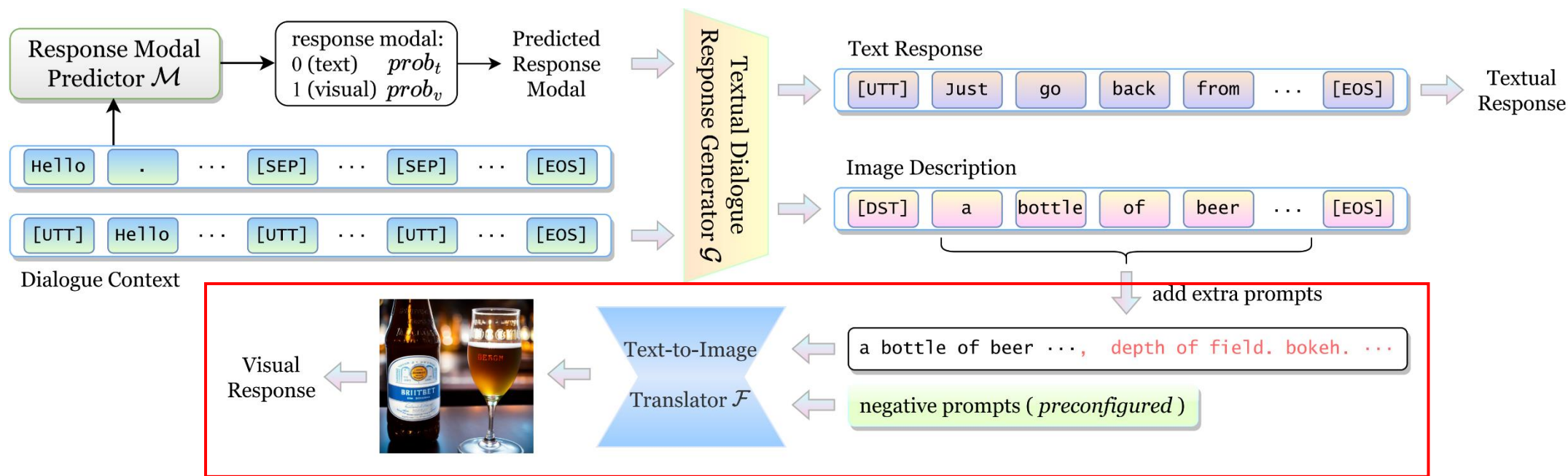
(1) Response Modal Predictor \mathcal{M}



TIGER framework:

(1) Response Modal Predictor \mathcal{M}

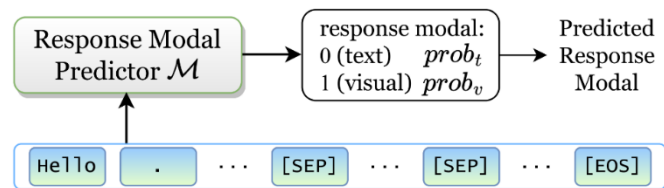
(2) Textual Dialogue Response Generator \mathcal{G}



TIGER framework:

- (1) Response Modal Predictor \mathcal{M}
- (2) Textual Dialogue Response Generator \mathcal{G}
- (3) Text-to-Image Translator \mathcal{F}

2.2.1 Response Modal Prediction



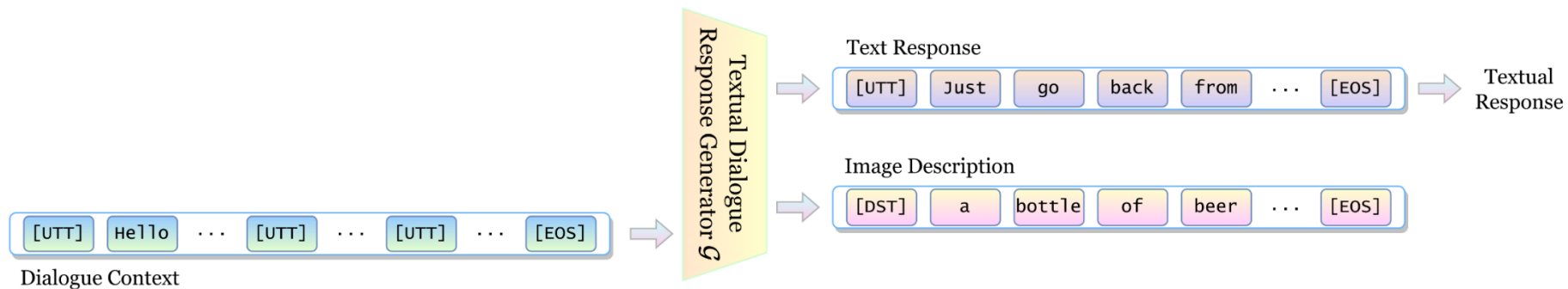
Given a dialogue context U , the target is to predict the modality $m \in \{t, v\}$ of the next response r_j .
(t : textual modality, v : visual modality)

Formulaically, response modal prediction is defined as a **binary classification** task:

$$\forall j \in [1, h], \mathcal{M}(U, R_{<j}) \in \{0, 1\}$$

predicted response modality $m = ?$ $\left\{ \begin{array}{l} m = t \rightarrow \text{generate textual response} \\ m = v \rightarrow \text{generate textual caption} \end{array} \right.$

2.2.2 Textual Dialogue Response Generation



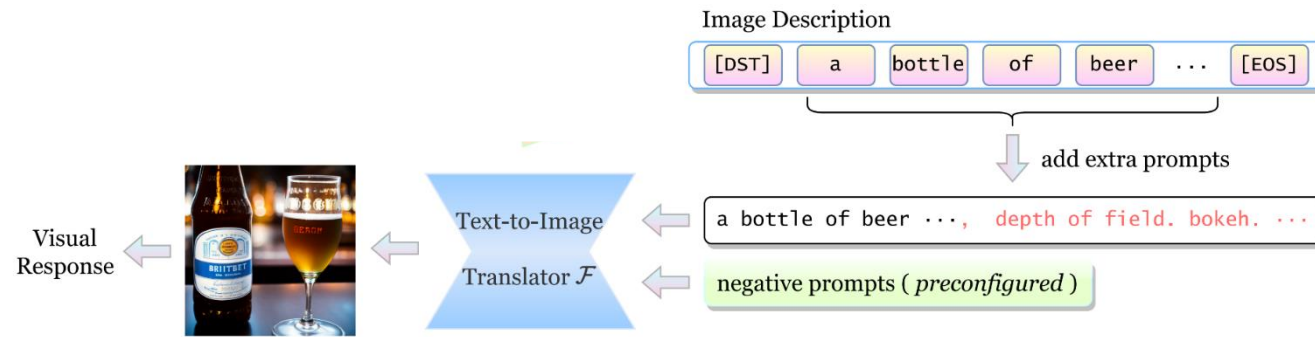
Textual Dialogue Response Generator \mathcal{G} : standard decoder-only causal transformer model $P(R_G|U, m; \theta_G)$

Given a dialogue context $U = \{u_1, u_2, \dots, u_K\}$, the target is to generate a textual output $R_G \in \{r^t, c\}$

Text response $r^t = \{[UTT], w_1, \dots, w_T\}$ and image description $c = \{[DST], w_1, \dots, w_L\}$

$$\mathcal{L}_{\text{text}} = - \sum_{i=1}^k \log p(w_i | U, w_1, \dots, w_{i-1}; \theta_G)$$

2.2.3 Text-to-Image Translation

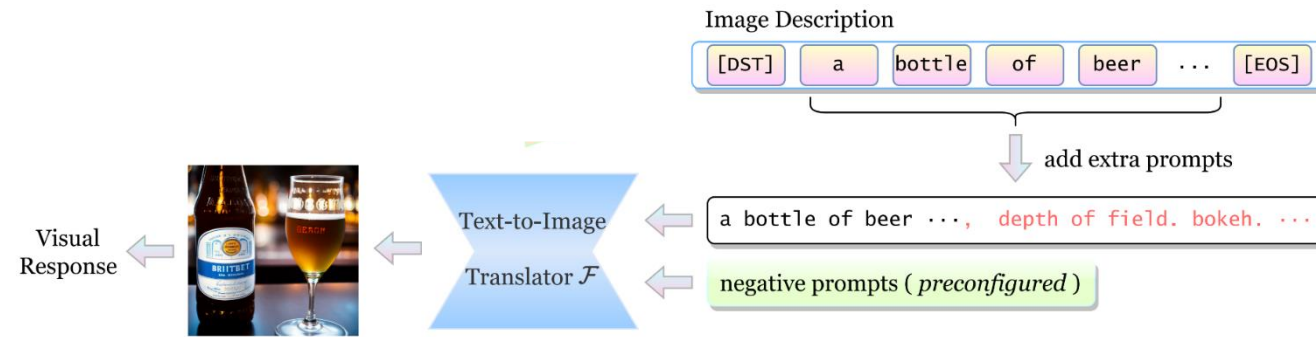


predicted response modality $m = ?$ $\left\{ \begin{array}{l} m = t \rightarrow \text{generate textual response} \\ m = v \rightarrow \text{generate textual caption} \rightarrow \text{generate visual response} \end{array} \right.$

text-to-image translator $P(r^v|c; \theta_{\mathcal{F}})$ generates a high-quality, high-resolution image r^v as a visual response, on conditional of the context-sensitive image description c .

$$\mathcal{L}_{t2i} = \mathbb{E}_{\epsilon(x), \epsilon \sim \mathcal{N}(0,1), t} [\|\epsilon - \epsilon_{\theta}(z_t, t)\|_2^2]$$

2.2.3 Text-to-Image Translation



Inference:

generated image descriptions c : brief texts (express the main image semantics)

richer prompts \rightarrow better visual responses

add extra prompts & adopt negative prompts

Low-resource Setting

- introduce text-to-image into text-only dialogue
- large-scale text dialogue data and image-text pairs $\xRightarrow{\text{assist}}$ small-scale multimodal dialogue data

Textual Dialogue Response Generation

1. Pre-training: large-scale text dialogue data \Rightarrow basic text dialogue generation capability.
2. Fine-tuning: **small-scale** multimodal dialogue data \Rightarrow textual response / image description.
(image \rightsquigarrow its textual description \Rightarrow ensure that the data used is pure text)

Text-to-Image Translation

1. Pre-training: large-scale image-text pair data \Rightarrow general image generation capability.
2. Fine-tuning: **very limited** image-text pairs from multimodal dialogue data \Rightarrow natural image.
(BLIP-2 refined caption)

Arbitrary Modal Compatibility:

TIGER framework is also suitable for multimodal dialogues that integrate arbitrary modalities (e.g., text and video, text and audio). To achieve this, one simply needs the **predicted response modal $m \in \{\text{text, target modal}\}$** of the response modal predictor and **replace the text-to-image translator with a Text-to-<Target Modal> translator**.



東北大學
Northeastern University

PATR 03

Experiment

| Models | Modal | Text Response Generation | | | Image Description Generation | | | Image Generation | |
|---------------------|-------------|--------------------------|-------------|-------------|------------------------------|--------------|--------------|--------------------------------|------------------|
| | F1 | BLEU-1 | BLEU-2 | ROUGE-L | BLEU-1 | BLEU-2 | ROUGE-L | IS \uparrow | FID \downarrow |
| Divter [†] | 56.2 | 6.52 | 1.66 | 5.69 | 15.08 | 11.42 | 15.81 | 15.8 \pm 0.6 | 29.16 |
| TIGER | 61.9 | 6.02 | 1.72 | 8.42 | 40.95 | 25.64 | 37.15 | 22.3\pm0.9 | 42.30 |

Automatic evaluation results of TIGER and baseline Divter^[1] on PhotoChat^[2] test set.

Evaluation focus on four aspects:

- (1) Response Modal Prediction;
- (2) Text Response Generation;
- (3) Image Description Generation;
- (4) Image Generation.

[1] Qingfeng Sun, et al. 2022. Multimodal Dialogue Response Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2854–2866

[2] Xiaoxue Zang, et al. 2021. PhotoChat: A Human-Human Dialogue Dataset With Photo Sharing Behavior For Joint Image-Text Modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6142–6152,

| Models | Modal | Text Response Generation | | | Image Description Generation | | | Image Generation | |
|---------------------|-------------|--------------------------|-------------|-------------|------------------------------|--------------|--------------|--------------------------------|------------------|
| | F1 | BLEU-1 | BLEU-2 | ROUGE-L | BLEU-1 | BLEU-2 | ROUGE-L | IS \uparrow | FID \downarrow |
| Divter [†] | 56.2 | 6.52 | 1.66 | 5.69 | 15.08 | 11.42 | 15.81 | 15.8 \pm 0.6 | 29.16 |
| TIGER | 61.9 | 6.02 | 1.72 | 8.42 | 40.95 | 25.64 | 37.15 | 22.3\pm0.9 | 42.30 |

Automatic evaluation results of TIGER and baseline Divter^[1] on PhotoChat^[2] test set.

Evaluation focus on four aspects:

- (1) Response Modal Prediction
- (2) Text Response Generation
- (3) Image Description Generation
- (4) Image Generation

(i) TIGER can accurately judge the timing of response with images.

| Models | F1 | Precision | Recall |
|---------------------|-------------|-------------|-------------|
| ALBERT-base* | 52.2 | 44.8 | 62.7 |
| BERT-base* | 53.2 | 56.1 | 50.6 |
| T5-base* | 58.1 | 58.2 | 57.9 |
| T5-3b* | 58.9 | 54.1 | 64.6 |
| Divter [†] | 56.2 | - | - |
| T5-base Encoder | 61.9 | 57.8 | 66.6 |
| T5-large Encoder | 60.0 | 61.5 | 58.5 |

Performance of response modal prediction on PhotoChat test set.

[1] Qingfeng Sun, et al. 2022. Multimodal Dialogue Response Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2854–2866

[2] Xiaoxue Zang, et al. 2021. PhotoChat: A Human-Human Dialogue Dataset With Photo Sharing Behavior For Joint Image-Text Modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6142–6152,

| Models | Modal | Text Response Generation | | | Image Description Generation | | | Image Generation | |
|---------------------|-------------|--------------------------|-------------|-------------|------------------------------|--------------|--------------|--------------------------------|------------------|
| | F1 | BLEU-1 | BLEU-2 | ROUGE-L | BLEU-1 | BLEU-2 | ROUGE-L | IS \uparrow | FID \downarrow |
| Divter [†] | 56.2 | 6.52 | 1.66 | 5.69 | 15.08 | 11.42 | 15.81 | 15.8 \pm 0.6 | 29.16 |
| TIGER | 61.9 | 6.02 | 1.72 | 8.42 | 40.95 | 25.64 | 37.15 | 22.3\pm0.9 | 42.30 |

Automatic evaluation results of TIGER and baseline Divter^[1] on PhotoChat^[2] test set.

Evaluation focus on four aspects:

- (1) Response Modal Prediction
- (2) **Text Response Generation**
- (3) Image Description Generation
- (4) Image Generation

- (i) TIGER can accurately judge the timing of response with images.
- (ii) **TIGER achieves comparable performance on text response generation with Divter.**

[1] Qingfeng Sun, et al. 2022. Multimodal Dialogue Response Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2854–2866

[2] Xiaoxue Zang, et al. 2021. PhotoChat: A Human-Human Dialogue Dataset With Photo Sharing Behavior For Joint Image-Text Modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6142–6152,

| Models | Modal | Text Response Generation | | | Image Description Generation | | | Image Generation | |
|---------------------|-------------|--------------------------|-------------|-------------|------------------------------|--------------|--------------|--------------------------------|------------------|
| | F1 | BLEU-1 | BLEU-2 | ROUGE-L | BLEU-1 | BLEU-2 | ROUGE-L | IS \uparrow | FID \downarrow |
| Divter [†] | 56.2 | 6.52 | 1.66 | 5.69 | 15.08 | 11.42 | 15.81 | 15.8 \pm 0.6 | 29.16 |
| TIGER | 61.9 | 6.02 | 1.72 | 8.42 | 40.95 | 25.64 | 37.15 | 22.3\pm0.9 | 42.30 |

Automatic evaluation results of TIGER and baseline Divter^[1] on PhotoChat^[2] test set.

Evaluation focus on four aspects:

- (1) Response Modal Prediction
- (2) Text Response Generation
- (3) **Image Description Generation**
- (4) **Image Generation**

- (i) TIGER can accurately judge the timing of response with images.
- (ii) TIGER achieves comparable performance on text response generation with Divter.
- (iii) the generated image descriptions are more detailed and contextualized, and the generated images have better clarity and diversity.

| | Win (%) | Tie (%) | Lose (%) |
|-------------|---------|---------|----------|
| Fidelity | 71.5 | 24.5 | 4.0 |
| Clarity | 98.5 | 1.5 | 0.0 |
| Consistency | 33.0 | 46.5 | 20.5 |

Human evaluation results.

[1] Qingfeng Sun, et al. 2022. Multimodal Dialogue Response Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2854–2866

[2] Xiaoxue Zang, et al. 2021. PhotoChat: A Human-Human Dialogue Dataset With Photo Sharing Behavior For Joint Image-Text Modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6142–6152,

3.2 Demonstration

TIGER

Hi friend!

How are you doing?

Nothing much. Just go back from a museum.

That sounds fun. What museum was it?

A botanical museum.


Cool. What kind of botanical plant did you see?

I looked at a lot of flowers, including peonies, lilies and carnations.

That sounds pretty neat. I love peonies. I bet they are beautiful!

Of course, they are beautiful. Unfortunately, I didn't take any photos.

Oh well! That would be a shame. I bet you'd like it though.



TIGER

My daughter had her birthday today.


Happy Birthday to her!

Thank you. Really wish you could be with us to celebrate my daughter's birthday party.

I know. I really wish I could be with you.

I bought a birthday cake for her. She was very happy.

That sounds awesome.



Of course. We had a great time.

She looks so happy.



東北大學
Northeastern University

PATR 04

Conclusion

- We introduce TIGER, a unified generative model framework for multimodal dialogue response generation.
- We incorporate text-to-image into text-only dialogues, enabling the original dialogue model to generate multimodal responses without the need for extensive multimodal dialogue data.
- Through sufficient experiments, we demonstrate that TIGER outperforms other models on various automatic evaluation metrics and is also preferred by humans, offering a more satisfying conversational experience.



東北大學
Northeastern University

Thanks

TIGER: A Unified Generative Model Framework for Multimodal Dialogue Response Generation

Fanheng Kong, Peidong Wang, Shi Feng[†], Daling Wang, Yifei Zhang