

# 공황장애 상담을 위한 멀티모달 과제 지향 대화 태스크 구축

김수빈<sup>1</sup>, 이근배<sup>1,2</sup>  
포항공과대학교 인공지능대학원<sup>1</sup>, 컴퓨터공학과<sup>2</sup>  
{kimsubin, gblee}@postech.ac.kr

## Building a multimodal task-oriented dialogue task for panic disorder counseling

Subin Kim<sup>1</sup>, Gary Geunbae Lee<sup>1,2</sup>  
Pohang University of Science and Technology,  
Graduated School of Artificial Intelligence<sup>1</sup>, Computer Science and Engineering<sup>2</sup>

### 요약

과제 지향 대화 시스템은 발화 의도 및 요구사항을 파악하여 사용자가 원하는 과제를 달성한다는 점에서 유용하다. 대화 상태 추적은 과제 지향 대화 시스템의 핵심 모듈이며, 최근에는 텍스트뿐만 아니라 시각 정보까지 활용하여 대화 상태를 추적하는 멀티모달 대화 상태 추적 연구가 활발히 진행되는 중이다. 본 논문에서는 멀티모달 공황장애 상담 대화 속 내담자의 상태를 추적하는 과제를 제안하였다. ChatGPT를 통한 멀티모달 공황장애 상담 과제 지향 대화 데이터셋 구축 프레임워크와, 구축한 데이터셋의 품질을 증명하기 위한 분석도 함께 제시하였다. 사전학습 언어 모델인 GPT-2를 벤치마크 데이터셋에 대해 학습한 성능을 측정함으로써 향후 멀티모달 대화 추적 성능이 능가해야 할 베이스라인 성능을 제시하였다.

**주제어:** 멀티모달, 멀티모달 대화 상태 추적, 멀티모달 과제 지향 대화 시스템

### 1. 서론

과제 지향 대화(Task-oriented dialogue, TOD) 시스템은 사용자와의 대화를 통해서 특정 목적의 과제를 달성하는 것을 목표로 한다. 그 예시로 사용자와의 대화를 통해 호텔 및 항공편 예약, 일정 관리 등을 수행할 수 있으며, 이는 실제 비즈니스에 많이 활용된다. 대화를 통해 사용자가 원하는 과제를 수행하기 위해서는 매 턴마다 사용자의 발화 의도 및 요구사항을 추적하며 업데이트해야 하는데, 이는 대화 상태 추적(Dialogue state tracking, DST) 모듈을 통해 수행된다. 대화 상태 추적은 과제 지향 대화 시스템의 필수 구성 요소이다. 이는 이전 대화 기록과 현재 사용자의 발화가 주어졌을 때, 대화 목적에 걸맞은 슬롯(slot)과 그에 상응하는 벨류(value)를 추출한다.

최근에는 이미지나 비디오를 보고 사용자와의 대화를 수행하는 멀티모달 과제 지향 대화(Multi-modal task-oriented dialogue, MM-TOD)에 관한 연구가 활발히 진행 중이다 [1, 2, 3]. [1]의 연구에서는 가상현실 속 쇼핑 장면을 기반으로 하는 멀티모달 과제 지향 대화 데이터셋 SIMMC2.0을 제안하였으며, Dialogue State Tracking Challenge 10 (DSTC10) 멀티모달 대화 상태 추적에 트랙에서는 SIMMC2.0 데이터를 통해 진행되었다. [2]의 연구에서는 여러가지 형태를 띤 사물이 등장하는 비디오 입력에 대하여 과제 지향 대화를 수행하는 데이터셋을 도입하였으며, 이에 대한 베이스라인 연구도 제시하였다.

2021년 보건복지부 정신질환실태 역학조사에 따르면 정신장애 평생유병률이 27.8%에 달하며, 일반 인구 4명 중 1명은 살면서 한 번 이상 정신질환을 경험한 것으로 조사되었다 [4].

정신질환은 흔한 질환이지만, 사회적 편견으로 인하여 정신과 진료를 기피하는 현상이 존재한다. 또한 상담이 필요한 인원에 비하여 전문성을 갖춘 상담자가 부족하다는 문제도 존재한다 [5]. 이러한 니즈를 충족시키기 위해, 정신건강의학과 분야에서는 인공지능 챗봇을 이용한 정신질환 상담 연구를 수행하였다. 그 중에서도 사용자 감정을 토대로 상담을 수행한 연구가 제안되기도 하였다 [6, 7].

사용자의 감정 상태는 언어적 뿐만 아니라 비언어적 표현까지 고려해야 파악할 수 있으며, 공황장애 상담의 경우에는 내담자가 발작으로 인해 응답을 제대로 수행하지 못할 수 있다. [7]의 연구에서는 사용자 얼굴 이미지로부터 감정 레이블을 추출하여, 대화 상태 추적에 활용하는 시스템을 제안하였다.

정신질환 상담 목적의 대화 시스템은 언어적 표현과 더불어 표정 및 제스처와 같은 비언어적 표현을 모두 파악해야 하며, 해당 도메인의 과제 지향 대화 시스템은 사용자의 언어적 및 비언어적 표현을 고려할 수 있는 능력을 벤치마크 데이터셋을 통해 평가해야 한다. 그러나 기존에는 정신질환 상담 도메인의 멀티모달 과제 지향 대화 벤치마크는 존재하지 않았다.

본 논문에서는 ChatGPT [8]를 이용하여 1만 4천여개의 멀티모달 공황장애 상담 과제 지향 대화 벤치마크 (Multimodal task-oriented dialogue in psychiatric counseling, MMTOD-PC) 데이터셋을 구축하였다. 사용자 얼굴 이미지로부터 생성된 이미지 캡션을 시각 맥락(visual context)으로 활용하여 학습한 GPT-2 [9]의 성능을 베이스라인 성능으로 제시하였다. 본 연구의 기여는 아래와 같다.

- ChatGPT를 이용한 공황장애 상담 도메인의 멀티모달 과

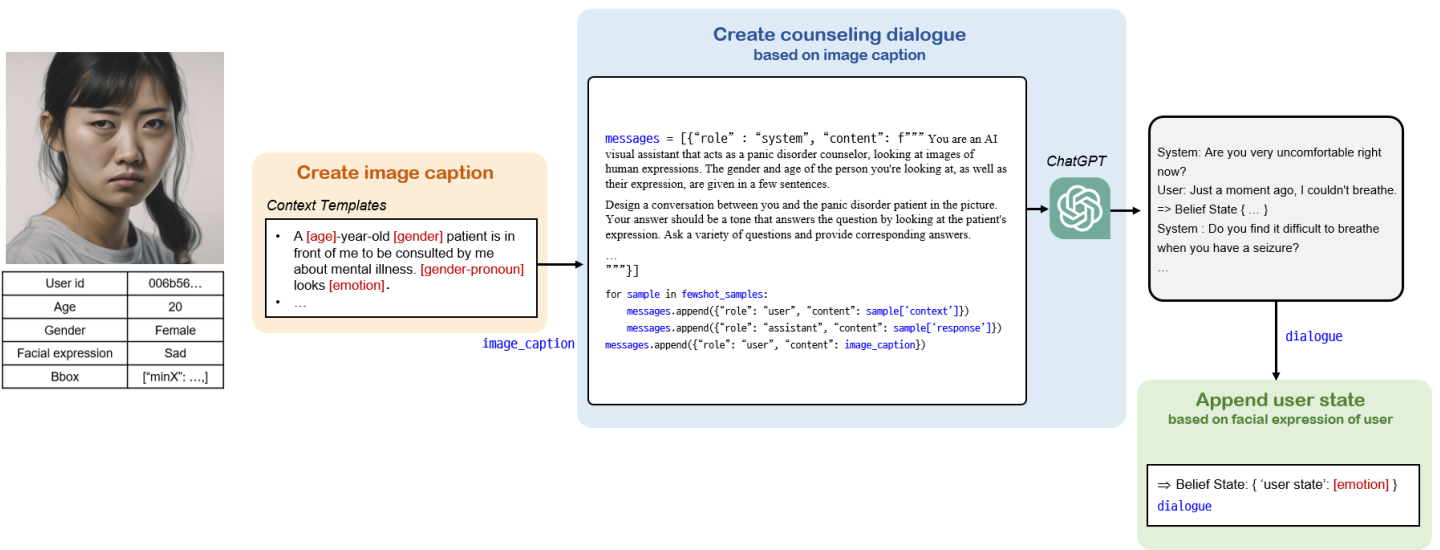


그림 1. 공황장애 상담을 위한 멀티모달 TOD 데이터셋 구축 프레임워크

제 지향 대화 벤치마크 MMTOD-PC 데이터셋의 구축 프레임워크를 제안하였다.

- 구축한 MMTOD-PC 데이터셋에 대한 분석을 통해 품질을 증명하였다.

## 2. 멀티모달 공황장애 상담 과제 지향 대화 데이터셋

### 2.1 데이터셋 구축 프레임워크

본 연구에서는 공황장애 상담 목적의 멀티모달 과제 지향 대화 데이터셋인 MMTOD-PC를 새롭게 구축하였다. 멀티모달 공황장애 상담 과제 지향 대화 데이터셋 구축 프레임워크는 그림 1와 같다. 기존에 ChatGPT를 활용하여 이미지로부터 멀티모달 대화 데이터셋을 생성하는 선행 연구들이 존재하였으며 [10, 11], 이 선행 연구들을 따라 프레임워크를 설계하였다.

ChatGPT는 텍스트 모달리티(modality)의 입력만 다루므로, 사용자 얼굴 이미지를 캡션으로 변환하여 활용하였다. 사용자 얼굴 이미지는 AI-Hub에 공개된 '한국인 감정인식을 위한 복합영상' 데이터 [12]를 활용하였다. 해당 데이터셋에는 얼굴 이미지와 그의 성별, 나이, 감정이 레이블링 되어 있다. 사전에 정의된 템플릿을 통해 얼굴 이미지로부터 사용자의 정보 및 현재 상태를 나타내는 캡션을 생성한 후, 이를 ChatGPT의 프롬프트(prompt)로 사용하였다.

선행 연구와 달리, 본 연구에서는 공황장애 상담 목적의 슬롯과 벨류가 레이블링 된 상담 대화를 생성하였다. 이를 위해 ChatGPT가 공황장애 상담 목적의 과제 지향 대화를 적절히 생성할 수 있도록, 레이블링 해야 할 슬롯에 관한 정보와 공황장애에 대한 설명을 포함한 시스템 메시지(system message)를 설계하였다. 또한 12가지 예제를 포함한 퓨샷 학습(few-shot

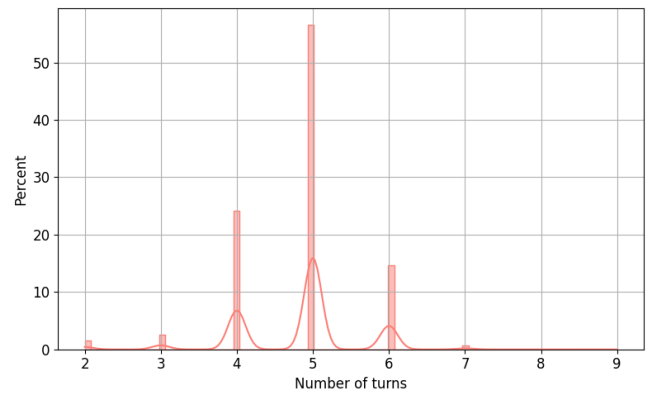


그림 2. MMTOD-PC 데이터셋의 대화 당 턴 분포

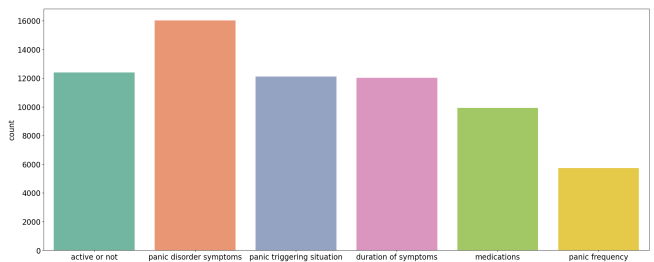


그림 3. MMTOD-PC 데이터셋의 각 슬롯 별 레이블링 된 횟수

learning)을 수행하였다. 인지행동치료에 필요한 '사용자 감정 상태', '발작 여부', '공황장애 증상', '공황 촉발 상황', '발작 주기', '증상 지속 시간', '복용 약물'의 총 7가지 대화 상태 (belief state)를 정의하였으며, ChatGPT에게는 '사용자 감정 상태'를 제외한 6가지 슬롯을 레이블링하도록 지시하였다. ChatGPT를 통해 대화를 생성한 후에 메타 데이터에 따라 '사용자 감정 상태' 슬롯을 추가하였다. '사용자 감정 상태' 슬롯은 대화 상태

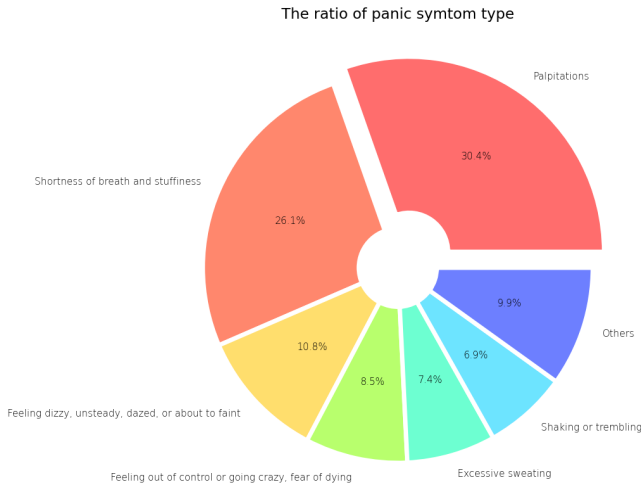


그림 4. 공황장애 증상 슬롯의 값 분포

추적 모델이 이미지 모달리티의 시각 맥락으로부터 추적해야 하기 때문에 단순히 텍스트 모달리티의 대화 맥락만으로 파악이 불가능한 슬롯이다. 따라서 본 연구에서 구축한 MMTOD-PC 데이터셋을 통해 대화 상태 추적 모델이 텍스트와 이미지 모달리티를 모두 고려할 수 있는지 파악할 수 있다.

## 2.2 데이터셋 분석

ChatGPT를 통해 총 1만 4천여개의 공황장애 상담을 위한 과제 지향 대화 데이터셋을 생성하였으며, 데이터셋의 통계 정보는 표 1과 같다. 구축한 데이터셋의 신뢰성을 증명하기 위해 대화의 턴 수, 각 슬롯의 레이블링 된 횟수 그리고 '공황장애 증상' 슬롯에 매핑된 여러 증상 분포를 분석하였다. 그림 2는 생성된 대화 당 턴 수 분포를 그래프로 나타낸 것이다. 본 연구에서 구축한 데이터셋은 대화 당 평균 4.82의 턴을 가지고 있으며, 이는 평균 5.26의 턴을 가지는 SIMMC2.0과 비교하였을 때, 기존 멀티모달 과제 지향 대화 벤치마크 데이터셋에 준하는 대화 길이를 가지는 것을 알 수 있다.

그림 3은 ChatGPT를 통해 레이블링하고자 했던 6가지 슬롯의 레이블링 된 횟수를 나타낸 것이다. 6가지 슬롯이 모두 레이블링 되었으며, 상담 시 필수적으로 수집해야 할 정보인 '발작 여부', '공황장애 증상', '공황 촉발 상황'이 가장 많이 레이블링 된 것을 확인할 수 있다. 그림 4는 '공황장애 증상' 슬롯에

표 1. 멀티모달 공황장애 TOD 데이터 셋의 통계정보

	Train	Validation	Test
대화 수	8,841	2,947	2,947
턴 수	42,633	14,104	14,168

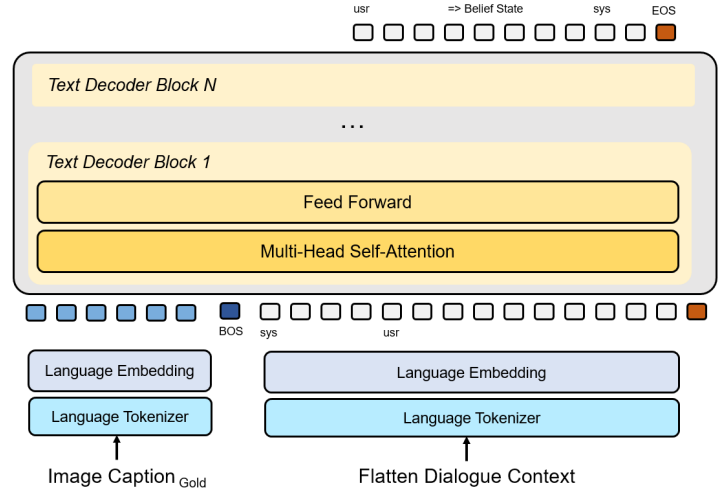


그림 5. PC TOD 베이스라인 구조

매핑된 증상 분포를 그래프로 나타낸 것이다. 심계항진과 호흡 곤란이 가장 많이 등장하였으며, 어지럽고 현기증이 날듯함이 그다음으로 많이 등장하였다. 이는 실제 공황장애 주요 증상과 상당 부분 일치함을 알 수 있다. ChatGPT를 통해 생성된 멀티모달 과제 지향 대화 예시는 그림 7와 같다.

## 3. 멀티모달 공황장애 상담 대화 상태 추적 시스템

본 연구에서는 멀티모달 공황장애 상담 대화 상태 추적 베이스라인 모델로 GPT-2 [9]를 사용하였다. GPT-2는 트랜스포머(transformer)의 디코더(decoder)로만 구성된 언어 모델로, 대량의 텍스트로 사전 학습되어 언어 모델링 및 생성 문제에서 우수한 성능을 보인다. 그림 5와 같이 시각 맥락과 대화 맥락(dialogue context)가 주어지면 모델은 사용자 발화에 대해 대화 상태를 추적하고, 시스템 응답을 생성한다.

GPT-2는 단일 모달리티를 다루기 때문에, 사용자 얼굴 이미지에 레이블링된 성별, 나이, 감정을 토대로 텍스트 모달리티의 이미지 캡션을 생성한 후 이를 시각 맥락으로 활용하였다. 그림 6은 이미지 캡션 생성에 활용한 템플릿이다.

A [age]-year-old [gender] patient is in front of me to be consulted by me about mental illness. [gender-pronoun] looks [emotion].

그림 6. 사용자 얼굴 이미지에 대한 캡션 생성을 위한 템플릿

사용자의  $t$  시점 발화를  $x_t$ , 시스템의  $t$  시점 발화를  $y_t$ 라 할 때,  $t$  시점의 대화 맥락  $C_t$ 는  $t-1$  시점까지의 이전 대화 기록과 현재  $t$  시점의 사용자 발화로 구성된다.

$$C_t = \{x_1, y_1, \dots, x_{t-1}, y_{t-1}, x_t\} \quad (1)$$



=> Belief State: {'user state': 'neutral'}

**System:** How are you feeling right now?

**User:** I'm feeling okay at the moment.

=> Belief State: {'active or not': False}

**System:** Have you been experiencing frequent panic attacks?

**User:** No, I haven't had a panic attack in a while. They occur only once in a blue moon.

=> Belief State: {'panic frequency': 'once in a while'}

**System:** Do you remember any specific triggers for your panic attacks?

**User:** It's hard to pinpoint a specific trigger, but it usually happens when I'm in a crowded place or when I feel overwhelmed.

=> Belief State: {'panic triggering situation': 'crowded place or feeling overwhelmed'}

**System:** How long do your panic attack symptoms typically last?

**User:** They usually last for about 15 to 20 minutes and then gradually subside.

=> Belief State: {'duration of symptoms': '15 to 20 minutes'}

**System:** Have you been prescribed any medications to help manage your panic attacks?

**User:** No, I haven't been taking any medications for my panic attacks. I try to manage them through relaxation techniques and self-care.

=> Belief State: {}

그림 7. 공황장애 상담을 위한 멀티모달 TOD 데이터셋 예시

#### 4. 실험 및 성능 평가

실험은 Ubuntu 18.04.5 LTS에 NVIDIA CUDA 12.1 환경에서 진행하였으며, 모델 학습은 Python 3.8, pytorch 2.0.1에서 진행하였다.

##### 4.1 평가 지표

대화 상태 추적 연구에서 활용되는 평가 지표인 Joint Goal Accuracy (JGA)와 Slot F1 점수를 이용하여 평가를 수행하였다.

JGA는 각 대화 턴마다 모델이 예측한 슬롯과 벨류 쌍이 정답 레이블과 전부 일치하는지 평가하는 혹독한 지표이다. 반면에 Slot F1 점수는 각 대화 턴에서 슬롯 별로 모델이 예측한 벨류와 일치하는지 평가하며, JGA보다는 유연한 기준으로 성능을 측정한다.

##### 4.2 결과 분석

실험 결과, 베이스라인 성능이 JGA 74.188%, Slot F1 점수 75.444%의 준수한 성능을 달성하였다. MMTOD-PC 데이터셋은 멀티모달 모델의 공황장애 상담 대화 상태 추적 성능을 파악하는 것을 목표로 한다. 베이스라인 모델은 사용자 메타 데이터로부터 생성한 정답 이미지 캡션을 시각 맥락으로 활용하였으며, 멀티모달 모델이 이미지 모달리티로부터 올바른 사용자 성별, 나이, 감정 이상의 피처를 추출한 시각 맥락을 활용한다면 베이스라인 성능을 능가할 수 있을 것으로 기대된다.

표 2. 공황장애 상담 대화 상태 추적 성능

Method	JGA (%)	Slot F1 (%)	Slot Stderr
PC - TOD <sub>Capgold</sub>	74.188	75.444	0.004

## 5. 결론

본 논문에서는 공황장애 상담 목적의 멀티모달 과제 지향 대화 데이터셋 구축 프레임워크를 제시하였으며, 해당 데이터셋의 품질을 증명하기 위한 분석을 제공하였다. 대화 맥락과 얼굴 이미지로부터 생성된 이미지 캡션인 시각 맥락을 통해 대화 상태 추적을 수행하는 GPT-2 기반 베이스라인 성능을 제시하였다. 하지만 사용자 이미지로부터 얻을 수 있는 정보는 성별, 나이, 감정 뿐만이 아니기 때문에 이러한 점에서 베이스라인 연구는 시각 정보 손실을 피할 수 없다는 한계점을 지니고 있다. 향후에는 이미지 또는 비디오와 텍스트 모달리티를 함께 처리하는 멀티모달 모델을 활용하여 베이스라인 성능 개선 연구를 심층적으로 수행할 계획이다.

## 감사의 글

이 논문은 경찰청이 지원한 ‘스마트 건강관리 (www.kipot.or.kr)’의 지원을 받아 수행된 연구결과입니다. [과제명: 경찰관 맞춤형 건강관리 서비스를 위한 지능형 빅데이터 통합플랫폼 개발 / 과제번호: 220222M01]

또한 이 연구는 2021년도 산업통상자원부 및 산업기술평가관리원(KEIT) 연구비 지원에 의한 연구임(20015007)

## 참고문헌

- [1] S. Kottur, S. Moon, A. Geramifard, and B. Damavandi, “SIMMC 2.0: A task-oriented dialog dataset for immersive multimodal conversations,” *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4903–4912, Nov. 2021. [Online]. Available: <https://aclanthology.org/2021.emnlp-main.401>
- [2] H. Le, N. F. Chen, and S. C. Hoi, “Multimodal dialogue state tracking,” *arXiv preprint arXiv:2206.07898*, 2022.
- [3] T.-L. Wu, S. Kottur, A. Madotto, M. Azab, P. Rodriguez, B. Damavandi, N. Peng, and S. Moon, “SIMMC-VR: A task-oriented multimodal dialog dataset with situated and immersive VR streams,” *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6273–6291, Jul. 2023. [Online]. Available: <https://aclanthology.org/2023.acl-long.345>
- [4] 보건복지부, “국가 정신건강현황 보고서 2021,” 2021.
- [5] 이상훈, “정신건강인력의 현황과 과제,” *보건복지포럼*, Vol. 2020, No. 4, pp. 43–54, 2020.
- [6] K.-J. Oh, D. Lee, B. Ko, and H.-J. Choi, “A chatbot for psychiatric counseling in mental healthcare service based on emotional dialogue analysis and sentence generation,” *2017 18th IEEE international conference on mobile data management (MDM)*, pp. 371–375, 2017.
- [7] 옥태경 and 이근배, “정신질환 상담을 위한 멀티모달 과제 지향 대화 시스템,” *한국정보과학회 학술발표논문집*, pp. 271–273, 2023.
- [8] OpenAI, “Chatgpt,” <https://openai.com/blog/chatgpt>, 2022 (accessed September 7, 2023).
- [9] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, “Language models are unsupervised multitask learners,” *OpenAI blog*, Vol. 1, No. 8, p. 9, 2019.
- [10] H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual instruction tuning,” *arXiv preprint arXiv:2304.08485*, 2023.
- [11] K. Li, Y. He, Y. Wang, Y. Li, W. Wang, P. Luo, Y. Wang, L. Wang, and Y. Qiao, “Videochat: Chat-centric video understanding,” *arXiv preprint arXiv:2305.06355*, 2023.
- [12] AI-Hub, “한국인 감정인식을 위한 복합 영상,” <https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=82>, 2022.