A Survey on Multi-modal Intent Recognition: *Recent Advances and New Frontiers*

Zhihong Zhu[1] **Fan Zhang**[2] **Yunyan Zhang**[1] **Jinghan Sun**[1]
Zhiqi Huang[3] **Qingqing Long**[3] **Bowen Xing**[4] **Xian Wu**[1,*]
[1]Tencent Jarvis Lab [2]The Chinese University of Hong Kong
[3]Peking University [4]University of Science and Technology Beijing
{profzhu, kevinxwu}@tencent.com

Abstract

Multi-modal intent recognition (MIR) requires integrating non-verbal cues from real-world contexts to enhance human intention understanding, which has attracted substantial research attention in recent years. Despite promising advancements, a comprehensive survey summarizing recent advances and new frontiers remains absent. To this end, we present a thorough and unified review of MIR, covering different aspects including (1) *Extensive survey*: we take the first step to present a thorough survey of this research field covering textual, visual (image/video), and acoustic signals. (2) *Unified taxonomy*: we provide a unified framework including evaluation protocol and advanced methods to summarize the current progress in MIR. (3) *Emerging frontiers*: We discuss some future directions such as multi-task, multi-domain, and multi-lingual MIR, and give our thoughts respectively. (4) *Abundant resources*: we collect abundant open-source resources, including relevant papers, data corpora, and leaderboards. We hope this survey can shed light on future research in MIR.

1 Introduction

Intent recognition (IR)[1] has achieved remarkable success in unimodal settings, particularly in textual (Chong et al., 2023; Zou et al., 2022) and visual domains (Jia et al., 2021; Ye et al., 2023). However, traditional unimodal approaches are inherently limited in capturing the complexity of real-world communication, where intent is often conveyed through a combination of verbal and non-verbal signals. This limitation arises because human communication is inherently multi-modal, relying not only on explicit textual content but also on prosodic variations, facial expressions, and body gestures.

Figure 1: An example of multi-modal intent recognition (MIR), where intent cannot be easily inferred from text alone. By combining a man's smirking expression with an exaggerated tone, it can be classified as 'Joke'.

To this end, multi-modal intent recognition (MIR) has emerged as a key research direction for enhancing intent understanding by systematically integrating diverse modalities, as illustrated in Figure 1. Leveraging textual semantics, acoustic features (*e.g.*, tone and prosody), and visual cues (*e.g.*, gestures and facial expressions), MIR facilitates a more holistic interpretation of human intent (Zhou et al., 2024; Zhu et al., 2024b; Zhang et al., 2024b), with broad implications for applications such as human-computer interaction (Zhang et al., 2024b).

Despite rapid advancements, there is still a lack of a comprehensive survey that summarizes recent advances and new frontiers. To bridge this gap, we present the first survey on MIR, reviewing over 60 cutting-edge studies published between 2019 and 2024. In a nutshell, our contributions can be summarized as follows: ❶ *Extensive survey*: we categorize existing studies based on their modality combinations, encompassing Textual-Visual, Textual-Acoustic, and Textual-Visual-Acoustic intent recognition. ❷ *Unified taxonomy*: we provide a systematic review of existing progress from evaluation protocol and advanced methods perspectives, establishing three leaderboards under unified metrics. ❸ *Emerging frontiers*: we highlight key challenges

*Corresponding author.

[1]It is also referred to as intent detection or classification; in this paper, we consistently use the term intent recognition.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 15223–15236
November 4-9, 2025

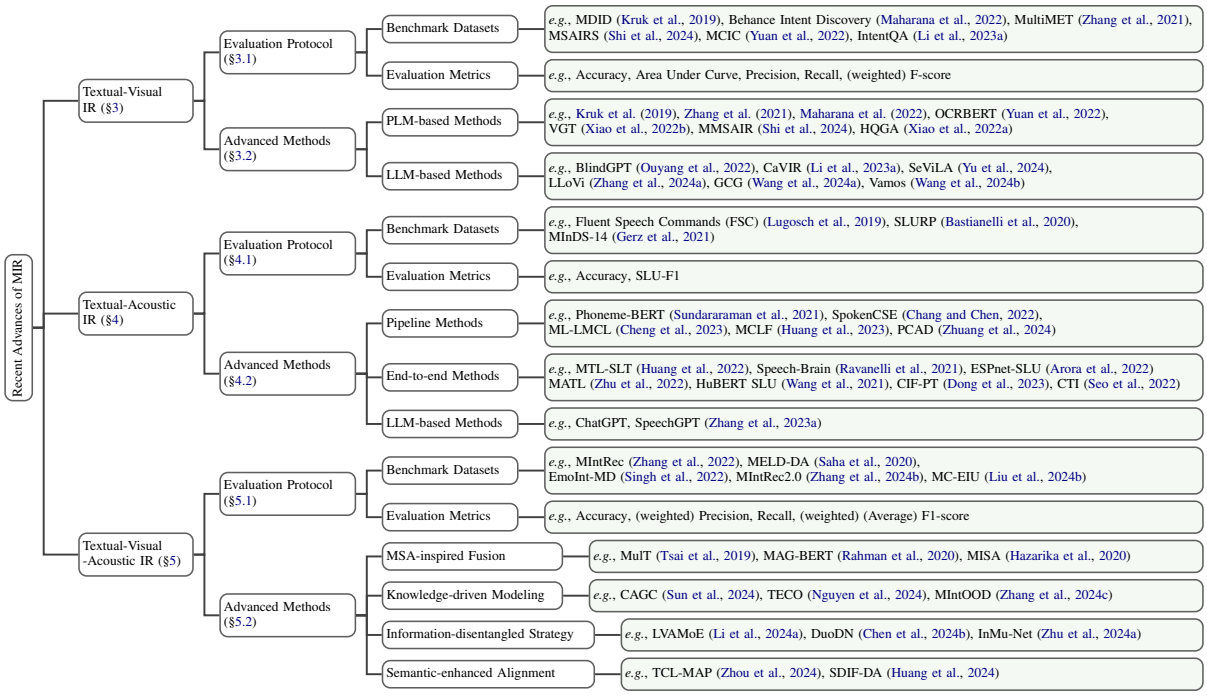

Figure 2: Recent advances of Multi-modal Intent Recognition (MIR). Extended discussions on emerging modality combinations (*e.g.*, Visual-Acoustic IR) are systematically analyzed in Appendix A.

in MIR, including multi-task, multi-domain, and multi-lingual MIR, outlining potential research directions to advance this research field. ❹ *Abundant resources*[2]: we attempt to organize open-source resources, including open-source software, diverse corpora, and a curated list of relevant publications.

2 Background and Preliminary

This section first outlines the background of MIR, and then provides an overview of MIR.

Background. MIR enhances intent understanding by integrating multiple modalities, distinguishing it from unimodal IR. However, research on MIR remains in its early stages due to:
❶ *Dataset*: While numerous multi-modal language datasets have been introduced for tasks such as sentiment analysis and emotion recognition, high-quality datasets specifically annotated for MIR remain scarce (Zhang et al., 2024b).
❷ *Methodology*: IR is inherently more abstract than tasks involving explicit emotional expression (Zhu et al., 2024a), making effective multi-modal fusion considerably more complex.

Preliminary. Given a multi-modal input that may comprise any combination of textual (t), visual (v), and acoustic (a) modalities, MIR aims to determine the most appropriate intent label(s) of the input as:

$$y = f(\{X_m\}_{m \in M}), \tag{1}$$

where $f(\cdot)$ denotes the MIR model; $M \subseteq \{\mathtt{t}, \mathtt{v}, \mathtt{a}\}$ represents the available set of modalities; X_m corresponds to the input features from modality m; and $y \in \mathcal{Y} = \{y_1, y_2, \ldots, y_K\}$ is the predicted intent label(s) among K predefined classes.

We next summarize recent advances in MIR across three major modality combinations as shown in Figure 2. We also discuss emerging combinations in Appendix A. Each combination poses distinct data characteristics and modeling challenges, reflecting varying levels of methodological maturity. Accordingly, we adopt tailored categorization strategies for each, inspired by similar practices in sentiment analysis (Das and Singh, 2023).

3 Textual-Visual Intent Recognition

Text on social media or e-commerce platforms is often accompanied by visual signals (*i.e.*, images or videos), which are common ways for users to express their intentions. In the following, we detail the collected Textual-Visual IR benchmarks and their corresponding metrics (§3.1), as shown in Table 1. Additionally, we summarize some advanced methods tailored for Textual-Visual IR (§3.2).

[2] https://github.com/Zhihong-Zhu/MIR-Survey

Dataset Name	Source	#Intent	Modality			Evaluation Metric	Additional Remarks
			v	t	a		
MDID (Kruk et al., 2019) EMNLP	Instagram	7	✓	✓	✗	ACC, AUC	Annotated manually via consensus
MultiMET (Zhang et al., 2021) ACL	Twitter, Facebook and (Ye et al., 2019)	4	✓	✓	✗	ACC	Annotates metaphor authorial intent
Behance Intent Discovery (Maharana et al., 2022) NAACL	Behance Livestreams	2	✓	✓	✗	P, R, F	Manually annotated via crowdsourcing; each sample contains a transcribed phrase
MCIC (Yuan et al., 2022) NLPCC	JD.com	212	✓	✓	✗	ACC	30,716 multi-modal dialogues with images and OCR texts (85% images contain text)
MSAIRS (Shi et al., 2024) arXiv	WeChat, TikTok and QQ	20	✓	✓	✗	ACC, wF1	Human annotation combined with GPT-4V review
IntentQA (Li et al., 2023a) ICCV	NExT-QA (Xiao et al., 2021)	-	✓	✓	✗	ACC	Annotated via Amazon Mechanical Turk (AMT) with contrastive samples (same action, different intents)
SLURP (Bastianelli et al., 2020) EMNLP	Home Assistant	18×46	✗	✓	✓	ACC, SLU-F1	Contains 72k audio recordings (58 hours); supports both pipeline (ASR+NLU) and end-to-end SLU approaches
Fluent Speech Commands (FSC) (Lugosch et al., 2019) INTERSPEECH	Crowdsourcing	31	✗	✓	✓	ACC	Contains 30,043 audio utterances (19 hours); designed for end-to-end SLU
MInDS-14 (Gerz et al., 2021) EMNLP	Crowdsourcing	14	✗	✓	✓	ACC	For the e-banking domain across 14 languages; includes spoken data and ASR transcriptions
MIntRec (Zhang et al., 2022) MM	TV series Superstore	20	✓	✓	✓	ACC, P, R, F1	First tri-modal intent dataset; includes automatic speaker annotation
EMOTyDA (Saha et al., 2020) ACL	MELD (Poria et al., 2019) IEMOCAP (Busso et al., 2008)	11	✓	✓	✓	ACC, P, R, F1	emotion-aware multi-modal dialogue act (DA) classification dataset; joint learning of DAs and emotions
EmoInt-MD (Singh et al., 2022) TASLP	Movies (drama, action, fantasy, etc.)	15	✓	✓	✓	ACC, F1	32k dialogues annotated with 15 empathetic intents
MIntRec2.0 (Zhang et al., 2024b) ICLR	TV series Superstore, The Big Bang Theory, and Friends	30	✓	✓	✓	ACC, P, R F1, wP, wF1	tri-modal dataset with 15,040 samples (9,304 in-scope, 5,736 out-of-scope); supports multi-turn, multi-party conversations
MC-EIU (Liu et al., 2024b) arXiv	TV series	9	✓	✓	✓	Weighted Average F	Emotion and intent joint understanding dataset; covers two languages (English and Mandarin)

Table 1: Major datasets for multi-modal intent recognition (MIR) over the past six years (2019 - 2024), covering visual (v), textual (t), and acoustic (a) modalities. '-' denotes information not reported in the original publication.

3.1 Evaluation Protocol

Benchmark Datasets. Recent textual–visual IR datasets are largely derived from social media, highlighting the importance of modeling non-literal cross-modal complementarity. For instance, MDID (Kruk et al., 2019) compiles 1,299 Instagram posts with annotations spanning three taxonomies: authorial intent, contextual relations, and semiotic relations. MultiMET (Zhang et al., 2021) further explores metaphor understanding with 10,437 text–image pairs, introducing intent labels such as descriptive, persuasive, and expressive.

Beyond social media, domain-specific applications have expanded dataset design. MCIC (Yuan et al., 2022) provides a large-scale Chinese e-commerce corpus of 30,000+ multi-modal dialogues, where 80% of images contain OCR-recognizable text. Similarly, the Behance Intent Discovery dataset (Maharana et al., 2022) focuses on instructional videos, offering 20,011 annotated clips for procedural intent identification.

More recently, novel modalities and interaction paradigms have been introduced. MSAIRS (Shi et al., 2024) investigates sticker-centric retrieval; IntentQA (Li et al., 2023a) extends intent reasoning to video narratives, comprising 16,297 QA pairs across 4,303 videos and requiring fine-grained temporal alignment between actions and goals.

Evaluation Metrics. In Textual-Visual IR, accuracy (ACC) emerges as the predominant evaluation metric, adopted by five of the six collected benchmarks. Besides, the MDID dataset further introduces macro-averaged AUC as a complementary metric to address potential class skew.

Method	Dataset	ACC	P	R	F
PLM-based Methods					
Kruk et al. (2019) EMNLP	MDID	56.7	-	-	-
Zhang et al. (2021) ACL	MultiMET	72.45	-	-	-
Maharana et al. (2022) NAACL	BID	-	62/30	61/31	62/30
OCRBERT (Yuan et al., 2022) NLPCC	MCIC	87.41	-	-	-
MMSAIR (Shi et al., 2024) arXiv	MSAIRS	69.82	-	-	69.82
HQGA (Xiao et al., 2022a) AAAI	IntentQA	47.7	-	-	-
VGT (Xiao et al., 2022b) ECCV	IntentQA	51.3	-	-	-
LLM-based Methods					
BlindGPT (Ouyang et al., 2022) NeurIPS	IntentQA	51.6	-	-	-
CaVIR (Li et al., 2023a) ICCV	IntentQA	57.6	-	-	-
SeViLA (Yu et al., 2024) NeurIPS	IntentQA	60.9	-	-	-
LLoVi (Zhang et al., 2024a) EMNLP	IntentQA	67.1	-	-	-
Vamos (Wang et al., 2024b) ECCV	IntentQA	71.7	-	-	-
GCG (Wang et al., 2024a) MM	IntentQA	73.1	-	-	-

Table 2: Leaderboard in Textual-Visual IR. Note that Behance Intent Discovery (BID) reports the results based on the defined two intents.

A distinct evaluation paradigm is introduced in the Behance Intent Discovery dataset, which employs a 75% partial match-based F-score metric for span prediction tasks, alleviating ASR transcription errors and imperfect modality alignment. Meanwhile, MSAIRS incorporates weighted F1 scores alongside accuracy, potentially addressing multi-class imbalance through class-aware weighting.

3.2 Advanced Methods

With the evolution of benchmarks, Textual-Visual IR has also witnessed the emergence of methods, which can generally be classified into: ❶ pre-trained language model (PLM)-based and ❷ large language model (LLM)-based methods.

❶ ***PLM-based Methods.*** Early methods such as Kruk et al. (2019) and Zhang et al. (2021) used modality-specific encoders (e.g., ResNet for images, BERT for text) with handcrafted fusion strategies. While effective as a starting point, these methods were limited in handling complex intent scenarios that require fine-grained cross-modal alignment. Later work incorporated auxiliary signals; for example, Yuan et al. (2022) used OCR-extracted text to resolve ambiguities in user utterances.

Video data further raises modeling challenges. Xiao et al. (2022a,b) introduced graph-based hierarchies and dynamic spatio-temporal graphs to align objects and actions with textual queries.

This line of work reflects a shift toward contextualized intent modeling, emphasizing hierarchical structure and intra-modal relations.

❷ ***LLM-based Methods.*** LLMs have recently been adapted for MIR. Yu et al. (2024) and Zhang et al. (2024a) illustrate this trend, employing BLIP-2 (Li et al., 2023b) and GPT variants (Ouyang et al., 2022) for self-chained localization–answering and long-range video reasoning. A common approach is to decompose tasks into localized captioning followed by LLM-based aggregation, reducing reliance on costly temporal annotations and enabling weakly supervised training (Wang et al., 2024a,b).

Leaderboard. We summarize the collected Textual-Visual IR methods in Table 2. Note that the first five methods are not directly comparable, as their benchmark datasets are inconsistent.

Highlight. Current visual–textual IR methods have shifted from end-to-end fusion toward semantic distillation to support LLM-based reasoning. This direction remains constrained by two issues: ❶ dependence on weak supervision, which may amplify errors, and ❷ limited interpretability stemming from LLM black-box characteristics.

4 Textual-Acoustic Intent Recognition

Acoustic signals in voice-based platforms (*e.g.*, voice assistants or spoken dialogue systems) often serve as the primary modality for intent expression, with textual content derived through automatic speech recognition (ASR) to complement paralinguistic information. Below, we introduce the curated Textual-Acoustic IR benchmark datasets and their associated evaluation metrics (§4.1), as summarized in Table 1. We also discuss state-of-the-art methods specifically designed to address the unique challenges of Textual-Acoustic IR (§4.2).

4.1 Evaluation Protocol

Benchmark Datasets. Unlike Textual-Visual IR, textual-acoustic IR has to deal directly with noisy signals and speaker variation, which makes robustness a central concern. Early benchmarks such as ATIS (Hemphill et al., 1990) and SNIPS (Coucke et al., 2018) provided useful testbeds, but they were restricted to narrow domains: airline travel (21 intents, 5k utterances) and virtual assistants (7 intents, 14k utterances). Their distributions are also highly skewed; over 70% of ATIS queries are about flights, which limits transferability to other domains.

More recent datasets began to push toward realistic usage scenarios. SLURP (Bastianelli et al., 2020) contains around 72k utterances across 18

domains and 46 action types, collected in everyday home and office settings with natural acoustic variability such as background noise, speaker movement, and mismatched microphones. Fluent Speech Commands (FSC) (Lugosch et al., 2019), though smaller (30k commands, about 19 hours), targets smart-home interactions with similar emphasis on recording diversity. In parallel, multilingual datasets have been introduced to broaden linguistic coverage. MInDS-14 (Gerz et al., 2021) spans 14 banking intents across 14 languages (roughly 50 examples per intent), capturing both dialectal differences (e.g., British vs. Australian English) and typologically distant languages (e.g., Slavic vs. Asian). By contrast, classic resources like TREC (Li and Roth, 2002) remain confined to coarse-grained English-only classification.

Evaluation Metrics. Like Textual-Visual IR, Textual-Acoustic IR also predominantly adopts accuracy as its primary evaluation metric, where semantic correctness depends on exact matches between predicted and gold-standard intents.

4.2 Advanced Methods

Based on different model architectures, existing Textual-Acoustic IR methods can be categorized into three main types as follows:

❶ ***Pipeline Methods.*** These methods aim to reduce cascading errors from ASR transcripts by improving representation learning. Phoneme-BERT (Sundararaman et al., 2021) jointly modeled phoneme sequences and transcripts with BERT-style pre-training. SpokenCSE (Chang and Chen, 2022) applied contrastive pre-training to improve robustness to ASR noise. ML-LMCL (Cheng et al., 2023) used mutual learning between clean and noisy transcripts to reduce intra-class variation. PCAD (Zhuang et al., 2024) introduced prototype-calibrated decoupling, which uses label priors to separate error-prone semantics. MCLF (Huang et al., 2023) advanced multi-grained contrastive learning with localized error-aware augmentation, aligning features from phoneme to utterance level.

Overall, pipeline methods focus on disentangling ASR-induced noise from semantic content through contrastive and representation-based strategies.

❷ ***End-to-End Methods.*** These methods focus on direct speech-to-intent mapping by jointly modeling textual and acoustic signals. MTL-SLT (Huang et al., 2022) integrates pre-trained ASR and language models under a multi-task learning framework to support cross-task knowledge transfer. CTI (Seo et al., 2022) connects ASR and NLU networks with vocabulary-aligned representations and trains them jointly for noise-robust intent recognition. With the availability of large speech models, HuBERT SLU (Wang et al., 2021) explored partial fine-tuning of transformer layers for intent decoding, while CIF-PT (Dong et al., 2023) introduced a continuous integrate-and-fire mechanism to achieve frame-to-token alignment during pre-training. MATL (Zhu et al., 2022) extended this line by applying token-frame cross-attention and sentence-level contrastive regularization for multi-grained alignment. In addition, toolkits like SpeechBrain (Ravanelli et al., 2021) and ESPnet-SLU (Arora et al., 2022) provide modular implementations that support rapid development.

Overall, end-to-end methods advance Textual-Acoustic IR by improving cross-modal pre-training, refining temporal alignment, and simplifying model design through unified architectures.

❸ ***LLM-based Methods.*** LLMs extend beyond architectural integration by leveraging large-scale pre-trained knowledge for zero-shot generalization, especially when combined with cross-modal instruction tuning. SpeechGPT (Zhang et al., 2023a) illustrates this direction with a three-phase pipeline: modality adaptation aligns speech tokens with textual semantics through continuation tasks, instruction tuning introduces multi-modal task awareness using synthesized speech-text command data, and

Method	SLURP	ATIS	TREC
Pipeline Methods			
Phoneme-BERT (Sundararaman et al., 2021) arXiv	83.78	94.83	85.96
SpokenCSE (Chang and Chen, 2022) INTERSPEECH	85.26	95.10	86.36
ML-LMCL (Cheng et al., 2023) ACL	88.52	96.52	89.24
MCLF (Huang et al., 2023) EMNLP	85.39	95.22	87.00
PCAD (Zhuang et al., 2024) ACL	90.58	97.64	91.25
End-to-End Methods			
MATL (Zhu et al., 2022) INTERSPEECH	78.72	-	-
MTL-SLT (Huang et al., 2022) ACL	83.10	97.13	-
Speech-Brain (Ravanelli et al., 2021) arXiv	85.34	-	-
ESPnet-SLU (Arora et al., 2022)	86.3	-	-
CTI (Seo et al., 2022) ICASSP	86.92	-	-
HuBERT SLU (Wang et al., 2021) arXiv	89.38	-	-
CIF-PT (Dong et al., 2023) ACL	91.32	-	-
LLM-based Methods			
ChatGPT (gpt-3.5-turbo-0125)	73.96	84.13	73.68
SpeechGPT (Zhang et al., 2023a) EMNLP	72.84	83.21	71.34

Table 3: Leaderboard for SLURP, ATIS and TREC datasets in Textual-Acoustic IR. Results are reported in terms of accuracy.

parameter-efficient methods such as LoRA (Hu et al., 2022) enhance cross-modal reasoning. These developments reflect a broader trend of positioning LLMs as universal semantic interfaces.

Leaderboard. As shown in Table 3, we report the performance of advanced methods on three popular datasets (SLURP, ATIS and TREC).

Highlight. Overall, *Pipeline methods* prioritize hierarchical noise disentanglement through contrastive learning and error-aware augmentation, yet face scalability bottlenecks; whereas, *end-to-end methods* streamline cross-modal integration via structural synergy but remain data-hungry. The rise of *LLM-based methods* shifts focus toward semantic distillation for zero-shot generalization, although performance gaps still exist.

5 Textual-Visual-Acoustic Intent Recognition

5.1 Evaluation Protocol

Benchmark Datasets. Benchmarks across Textual-Visual-Acoustic modalities are bringing IR closer to real-world scenarios. MIntRec (Zhang et al., 2022) introduced a tri-modal dataset with 2,224 text-video-audio samples annotated across 20 fine-grained intents. MIntRec 2.0 (Zhang et al., 2024b) expands to 15,040 samples (9,304 in-scope and 5,736 out-of-scope) covering 30 intents. EMOTyDA (Saha et al., 2020) provides the first multi-modal dialogue act dataset, repurposing 13,000 utterances from Friends episodes with dialogue act labels, which can be treated as coarse-grained intents (Firdaus et al., 2021).

More recent datasets incorporate affective dimensions alongside intent. MC-EIU (Liu et al., 2024b) combines 9 intent classes with 7 emotion categories across 45,009 English and 11,003 Mandarin utterances, offering bilingual coverage and affective diversity. EmoInt-MD (Singh et al., 2022) links 15 intents with 32 emotions over 32,000 dialogues from movies. Despite these advances, multilingual support remains limited, with only MC-EIU and EmoInt-MD extending beyond English.

Evaluation Metrics. Accuracy (ACC) and macro-averaged F1 are widely adopted, addressing class imbalance in multi-class settings. MIntRec 2.0 (Zhang et al., 2024b) added weighted metrics such as wF1 and wP, and MC-EIU (Liu et al., 2024b) employed a Weighted Average F-score to better reflect skewed distributions.

Textual-Visual IR has applied span prediction metrics, while Textual-Acoustic IR often reports SLU-F1 to account for ASR errors. Textual-Visual-Acoustic IR, however, remains centered on utterance-level metrics. The adoption of weighted variants across datasets highlights cross-domain recognition of class imbalance as a persistent issue.

5.2 Advanced Methods

Based on the different objectives pursued by the model design, we categorize existing methods in Textual-Visual-Acoustic IR into four types:

❶ ***MSA-inspired Fusion.*** Given the recent emergence of Textual-Visual-Acoustic IR, it draws inspiration from advanced cross-modal interaction mechanisms in multi-modal sentiment analysis (MSA) as competitive baselines. For example, MulT (Tsai et al., 2019) introduces six bidirectional cross-modal Transformers to explicitly model pairwise interactions between modalities. Building upon Transformer architectures, MAG-BERT (Rahman et al., 2020) addresses the integration challenge in PLMs through its multi-modal adaptation gate, which dynamically adjusts textual representations through weighted displacements derived from acoustic and visual features. To address the tension between cross-modal alignment and modality fidelity, MISA (Hazarika et al., 2020) advances modality representation learning by explicitly separating shared and unique characteristics,

❷ ***Knowledge-driven Modeling.*** Rather than relying solely on isolated data or intrinsic model features, recent Textual-Visual-Acoustic IR approaches address intent ambiguity by integrating external or contextual knowledge. For example, CAGC (Sun et al., 2024) shifts from isolated video modeling to cross-video contextual reasoning through intra- and cross-video contrastive learning.

TECO (Nguyen et al., 2024) tackles semantic sparsity by infusing commonsense knowledge through a hybrid retrieval-generation mechanism. By extracting relational features from external knowledge and fusing them with multi-modal inputs via dual-perspective learning, TECO bridges the gap between implicit multi-modal cues and explicit world knowledge. MIntOOD (Zhang et al., 2024c) synthesizes pseudo-OOD data through convex combinations of ID samples, enabling joint optimization of coarse-grained OOD detection and fine-grained ID classification.

Method	MIntRec						EMOTyDA					
	ACC	F1	wF1	P	wP	R	ACC	F1	wF1	P	wP	R
MSA-inspired Fusion												
MulT (Tsai et al., 2019) ACL	72.31	68.97	72.07	69.73	72.24	68.83	63.35	54.20	62.28	58.45	62.96	53.57
MAG-BERT (Rahman et al., 2020) ACL	72.00	68.36	71.78	69.01	72.45	68.92	64.50	54.30	63.16	58.81	63.14	53.51
MISA (Hazarika et al., 2020) MM	72.29	69.32	72.38	70.85	73.48	69.24	59.98	-	58.52	-	59.28	48.75
Knowledge-driven Modeling												
CAGC (Sun et al., 2024) CVPR	73.39	70.09	-	71.21	-	70.39	-	-	-	-	-	-
TECO (Nguyen et al., 2024) PACLIC	72.36	69.96	-	70.49	-	69.92	-	-	-	-	-	-
MIntOOD (Zhang et al., 2024c) arXiv	74.34	70.94	74.15	72.24	74.51	70.46	65.00	56.20	63.53	65.09	64.62	54.20
Information-disentangled Strategy												
LVAMoE (Li et al., 2024a) ICME	73.13	70.26	-	71.47	-	69.89	-	-	-	-	-	-
DuoDN (Chen et al., 2024b) EMNLP	75.28	-	75.09	-	75.80	71.77	62.86	-	60.90	-	62.13	51.63
InMu-Net (Zhu et al., 2024a) MM	76.05	-	75.96	-	76.18	73.93	63.78	-	61.64	-	63.40	52.31
Semantic-enhanced Alignment												
TCL-MAP (Zhou et al., 2024) AAAI	73.21	69.02	72.73	69.39	73.02	69.88	64.23	53.98	62.94	57.10	62.73	53.22
SDIF-DA (Huang et al., 2024) ICASSP	71.42	68.53	71.24	72.24	74.51	70.46	64.33	55.56	63.19	62.11	63.75	54.00

Table 4: Leaderboard for MIntRec and EMOTyDA datasets in Textual-Visual-Acoustic IR. Missing values indicate unreported or unreproducible metrics. For EMOTyDA, underlined results indicate evaluations conducted using different test splits.

❸ ***Information-disentangled Strategy.*** As multi-modal systems still grapple with entangled representations, disentanglement emerges as a critical strategy for balancing semantic coherence and modality fidelity. LVAMoE (Li et al., 2024a) adopts a dual-encoder architecture, decoupling modality-invariant and modality-specific features through dense-sparse encoding. DuoDN (Chen et al., 2024b) explicitly disentangles semantics-oriented and modality-oriented representations using counterfactual intervention. By introducing confounders to simulate causal effects, it isolates the impact of modality-specific noise on predictions. Besides, InMu-Net (Zhu et al., 2024a) adopts a similar fashion, which addresses redundancy and long-tailed distributions through an information bottleneck strategy, filtering out intent-irrelevant features via denoising modules while preserving saliency through kurtosis regularization.

❹ ***Semantic-enhanced Alignment.*** Aligning semantics among triple modalities remains pivotal yet challenging. As such, TCL-MAP (Zhou et al., 2024) establishes bidirectional modality-text synergy, whose modality-aware prompting generates context-rich textual embeddings, which then guide video/audio feature refinement through token-level contrastive learning. SDIF-DA (Huang et al., 2024) adopts a progressive alignment strategy, where shallow interactions initially harmonize low-level features before deep fusion captures higher-order correlations. Complemented by ChatGPT-generated synthetic data, it enhances model robustness against modality-specific perturbations.

Overall, both frameworks mitigate semantic asymmetry through adaptive interaction mechanisms. Concretely, TCL-MAP operates with token-level precision, whereas SDIF-DA hierarchically integrates cross-modal signals.

Leaderboard. To unify this tri-modal research direction, we also present a comprehensive leaderboard for two widely used MIR datasets (*i.e.*, MIntRec and EMOTyDA), as shown in Table 4.

Highlight. Textual-Visual-Acoustic IR methods emphasize interaction granularity, external knowledge grounding, representation purity, or alignment precision and have achieved promising results.

However, multi-modal large language models (MLLMs) remain in the early stages of development within Textual-Visual-Acoustic IR domain.

6 New Frontiers

§3, §4, and §5 introduced prominent achievements in intent recognition under different modality combinations. This section discusses some new frontiers of MIR below, aiming to inspire researchers and promote the advancement of this research field.

Multi-task MIR. A promising direction for MIR is integrating multi-task learning. Liu et al. (2024b) proposed emotion and intent joint understanding in multi-modal conversation. Zhang et al. (2023b) highlighted the close relationship between sarcasm, semantics, and emotion, constructing three tasks to

perform sarcasm detection, semantic classification, and emotion classification, respectively.

Future research could explore strategies such as adaptive task weighting (Chen et al., 2024a) and shared-private architectures (Wu et al., 2025) to enhance the effect of multitask learning for MIR.

Multi-lingual MIR. Although there is a significant amount of research on MIR, most of these models primarily support the English language, and there is limited research on multilingual MIR benchmarks (Gerz et al., 2021; Zhao et al., 2022; Liu et al., 2024b), which hinders their application in non-English-speaking countries and regions.

In natural language processing (NLP), multilingual research is relatively mature, and some works have demonstrated excellent performance on multilingual tasks (Qin et al., 2022; Mullick, 2023; Fan et al., 2021). Therefore, researchers can extend these approaches to MIR, which would help reduce the disparity between high-resource and low-resource languages, enabling the creation of more extensive MIR systems in the future.

Multi-domain MIR. Though existing MIR models have achieved strong results in single-domain settings, they remain heavily dependent on large amounts of annotated data, which limits their adaptability to new domains. In practice, collecting sufficiently rich labeled datasets for every domain is infeasible (Wu et al., 2024). Since out-of-scope utterances frequently arise, extending MIR to multi-domain scenarios is a promising direction and a key step toward improving model robustness.

In MIR, only MIntRec (Zhang et al., 2024b) and MIntOOD (Zhang et al., 2024c) have made progress toward this goal. It is non-trivial to directly extend previous IR methods to the multi-domain setting (Li et al., 2024b), as it requires effectively fusing and aligning heterogeneous multimodal data streams while preserving domain-relevant information. As such, multi-domain MIR is an area that warrants further exploration.

Multi-modal Large Language Models. Empowered by large language models (LLMs), the understanding and reasoning capabilities of multi-modal large language models (MLLMs) have reached unprecedented levels, demonstrating impressive capabilities in various tasks (Yin et al., 2023; Caffagni et al., 2024; Liang et al., 2024; Zhu et al., 2025).

However, MLLMs in MIR currently serve only as components for data augmentation or perform zero-shot generalization. Rather than relying on advanced encoders with extensive training (Liu et al., 2024a), a possible alternative is to leverage MLLMs in combination with emerging multimodal reasoning techniques such as Visual-CoT (Shao et al., 2024; Zhao et al., 2025) and Audio-CoT (Ma et al., 2025) to achieve accurate outputs.

Survey	Year	Discussion Modality		
		Visual	Textual	Acoustic
Brenes et al. (2009)	2009	✗	✓	✗
Kofler et al. (2016)	2016	✓	✗	✗
Hamroun and Gouider (2020)	2020	✓	✗	✗
Louvan and Magnini (2020)	2020	✗	✓	✗
Weld et al. (2022)	2022	✗	✓	✗
Qin et al. (2021)	2021	✗	✓	✗
Zailan et al. (2023)	2023	✗	✓	✗
Ours	2025	✓	✓	✓

Table 5: Comparison with existing intent related surveys including year and discussion modality.

7 Related Work

Intent Recognition (IR) is one of the foundational tasks in natural language understanding (NLU), with early surveys dating back to Brenes et al. (2009), which reviewed automatic query intent detection. Recently, Kofler et al. (2016) focused on user intent in multimedia search, primarily involving visual intent in images and videos. On the other hand, Hamroun and Gouider (2020) summarized the methods and applications of textual intent detection. More recently, IR is typically surveyed in conjunction with slot filling (Louvan and Magnini, 2020; Weld et al., 2022; Zailan et al., 2023; Xia et al., 2025; Xing et al., 2025), as they are highly relevant in dialogue systems (Zhu et al., 2023).

However, there has yet to be a comprehensive survey on IR covering multiple modalities, which motivates this first work. As shown in Table 5, our survey covers three modalities: textual, visual (image/video), and acoustic, while ensuring the timeliness of the literature (from 2019 to 2024).

8 Conclusion

In this paper, we present the first comprehensive survey on the MIR task, which begins by systematically summarizing existing works that cover various modality combinations. Additionally, we compile and review currently available datasets and metrics while organizing three leaderboards to benchmark performance. Furthermore, we highlight emerging trends in this research field, provid-

ing insights into future directions. We hope this first survey with a website serves as a valuable resource to advance research in MIR.

Limitations

Although we strive to conduct a rigorous and comprehensive analysis of the existing literature on MIR, several limitations remain: (1) Some works may have been inadvertently omitted due to variations in search keywords. (2) Due to space constraints, our survey primarily focuses on the high-level aspects of the approaches, omitting fine-grained experimental comparisons. (3) Some representative MIR direction such as Textual-Visual IR are reported on distinct datasets (*e.g.*, IntentQA, MDID, MultiMET) using varying evaluation metrics. This fragmentation substantially hinders direct performance comparison across models. We note that this reflects the current landscape of the field rather than a design flaw of our survey. This situation underscores the urgent need for standardized benchmarks and unified evaluation protocols.

We will continuously track the latest MIR literature to promote the development of the field.

References

Siddhant Arora, Siddharth Dalmia, Pavel Denisov, Xuankai Chang, Yushi Ueda, Yifan Peng, Yuekai Zhang, Sujay Kumar, Karthik Ganesan, Brian Yan, et al. 2022. Espnet-slu: Advancing spoken language understanding through espnet. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7167–7171. IEEE.

Emanuele Bastianelli, Andrea Vanzo, Pawel Swietojanski, and Verena Rieser. 2020. Slurp: A spoken language understanding resource package. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7252–7262.

David J Brenes, Daniel Gayo-Avello, and Kilian Pérez-González. 2009. Survey and evaluation of query intent detection methods. In *Proceedings of the 2009 Workshop on Web Search Click Data*, pages 1–7.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. 2024. The revolution of multimodal large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13590–13618, Bangkok, Thailand. Association for Computational Linguistics.

Ya-Hsin Chang and Yun-Nung Chen. 2022. Contrastive learning for improving ASR robustness in spoken language understanding. In *Proc. of INTERSPEECH*.

Shijie Chen, Yu Zhang, and Qiang Yang. 2024a. Multitask learning in natural language processing: An overview. *ACM Computing Surveys*, 56(12):1–32.

Zhanpeng Chen, Zhihong Zhu, Xianwei Zhuang, Zhiqi Huang, and Yuexian Zou. 2024b. Dual-oriented disentangled network with counterfactual intervention for multimodal intent detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17554–17567, Miami, Florida, USA. Association for Computational Linguistics.

Xuxin Cheng, Bowen Cao, Qichen Ye, Zhihong Zhu, Hongxiang Li, and Yuexian Zou. 2023. ML-LMCL: Mutual learning and large-margin contrastive learning for improving ASR robustness in spoken language understanding. In *Proc. of ACL Findings*.

Ruining Chong, Cunliang Kong, Liu Wu, Zhenghao Liu, Ziye Jin, Liner Yang, Yange Fan, Hanghang Fan, and Erhong Yang. 2023. Leveraging prefix transfer for multi-intent text revision. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1219–1228.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*.

Ringki Das and Thoudam Doren Singh. 2023. Multimodal sentiment analysis: a survey of methods, trends, and challenges. *ACM Computing Surveys*, 55(13s):1–38.

Linhao Dong, Zhecheng An, Peihao Wu, Jun Zhang, Lu Lu, and Ma Zejun. 2023. CIF-PT: Bridging speech and text representations for spoken language understanding via continuous integrate-and-fire pre-training. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8894–8907, Toronto, Canada. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Mauajama Firdaus, Hitesh Golchha, Asif Ekbal, and Pushpak Bhattacharyya. 2021. A deep multi-task model for dialogue act classification, intent detection and slot filling. *Cognitive Computation*, 13:626–645.

Daniela Gerz, Pei-Hao Su, Razvan Kusztos, Avishek Mondal, Michał Lis, Eshan Singhal, Nikola Mrkšić, Tsung-Hsien Wen, and Ivan Vulić. 2021. Multilingual and cross-lingual intent detection from spoken data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7468–7475.

Mohamed Hamroun and Mohamed Salah Gouider. 2020. A survey on intention analysis: successful approaches and open challenges. *Journal of Intelligent Information Systems*, 55:423–443.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1122–1131.

Charles T Hemphill, John J Godfrey, and George R Doddington. 1990. The atis spoken language systems pilot corpus. In *Proceedings of the workshop on Speech and Natural Language*, pages 96–101.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Shijue Huang, Libo Qin, Bingbing Wang, Geng Tu, and Ruifeng Xu. 2024. Sdif-da: A shallow-to-deep interaction framework with data augmentation for multi-modal intent detection. *ICASSP*.

Zhiqi Huang, Dongsheng Chen, Zhihong Zhu, and Xuxin Cheng. 2023. MCLF: A multi-grained contrastive learning framework for ASR-robust spoken language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7936–7949, Singapore. Association for Computational Linguistics.

Zhiqi Huang, Milind Rao, Anirudh Raju, Zhe Zhang, Bach Bui, and Chul Lee. 2022. MTL-SLT: Multi-task learning for spoken language tasks. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 120–130, Dublin, Ireland. Association for Computational Linguistics.

Menglin Jia, Zuxuan Wu, Austin Reiter, Claire Cardie, Serge Belongie, and Ser-Nam Lim. 2021. Intentonomy: a dataset and study towards human intent understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12986–12996.

Christoph Kofler, Martha Larson, and Alan Hanjalic. 2016. User intent in multimedia search: a survey of the state of the art and future challenges. *ACM Computing Surveys (CSUR)*, 49(2):1–37.

Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. 2019. Integrating text and image: Determining multimodal document intent in instagram posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4622–4632.

Jiapeng Li, Ping Wei, Wenjuan Han, and Lifeng Fan. 2023a. Intentqa: Context-aware video intent reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11963–11974.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Tingyu Li, Junpeng Bao, Jiaqi Qin, Yuping Liang, Ruijiang Zhang, and Jason Wang. 2024a. Multi-modal intent detection with lvamoe: the language-visual-audio mixture of experts. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *Proc. of COLING*.

Yan Li, So-Eon Kim, Seong-Bae Park, and Soyeon Caren Han. 2024b. Midas: Multi-level intent, domain, and slot knowledge distillation for multi-turn nlu. *arXiv preprint arXiv:2408.08144*.

Zijing Liang, Yanjie Xu, Yifan Hong, Penghui Shang, Qi Wang, Qiang Fu, and Ke Liu. 2024. A survey of multimodel large language models. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*, pages 405–409.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024a. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*.

Rui Liu, Haolin Zuo, Zheng Lian, Xiaofen Xing, Björn W Schuller, and Haizhou Li. 2024b. Emotion and intent joint understanding in multimodal conversation: A benchmarking dataset. *arXiv preprint arXiv:2407.02751*.

Samuel Louvan and Bernardo Magnini. 2020. Recent neural methods on slot filling and intent classification for task-oriented dialogue systems: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 480–496.

Loren Lugosch, Mirco Ravanelli, Patrick Ignoto, Vikrant Singh Tomar, and Yoshua Bengio. 2019. Speech model pre-training for end-to-end spoken language understanding. *Interspeech*.

Ziyang Ma, Zhuo Chen, Yuping Wang, Eng Siong Chng, and Xie Chen. 2025. Audio-cot: Exploring chain-of-thought reasoning in large audio language model. *arXiv preprint arXiv:2501.07246*.

Adyasha Maharana, Quan Hung Tran, Franck Dernoncourt, Seunghyun Yoon, Trung Bui, Walter Chang, and Mohit Bansal. 2022. Multimodal intent discovery from livestream videos. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 476–489.

Ankan Mullick. 2023. Exploring multilingual intent dynamics and applications. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 7087–7088. International Joint Conferences on Artificial Intelligence Organization. Doctoral Consortium.

Quynh-Mai Thi Nguyen, Lan-Nhi Thi Nguyen, and Cam-Van Thi Nguyen. 2024. Teco: Improving multimodal intent recognition with text enhancement through commonsense knowledge extraction. *PACLIC*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Libo Qin, Qiguang Chen, Tianbao Xie, Qixin Li, Jian-Guang Lou, Wanxiang Che, and Min-Yen Kan. 2022. Gl-clef: A global–local contrastive learning framework for cross-lingual spoken language understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2677–2686.

Libo Qin, Tianbao Xie, Wanxiang Che, and Ting Liu. 2021. A survey on spoken language understanding: Recent advances and new frontiers. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4577–4584. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2020, page 2359. NIH Public Access.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, et al. 2021. Speechbrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624*.

Tulika Saha, Aditya Patra, Sriparna Saha, and Pushpak Bhattacharyya. 2020. Towards emotion-aided multimodal dialogue act classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4361–4372.

Seunghyun Seo, Donghyun Kwak, and Bowon Lee. 2022. Integration of pre-trained networks with continuous token interface for end-to-end spoken language understanding. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7152–7156. IEEE.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Yuanchen Shi, Biao Ma, and Fang Kong. 2024. Impact of stickers on multimodal chat sentiment analysis and intent recognition: A new task, dataset and baseline. *arXiv preprint arXiv:2405.08427*.

Gopendra Vikram Singh, Mauajama Firdaus, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Emoint-trans: A multimodal transformer for identifying emotions and intents in social conversations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:290–300.

Kaili Sun, Zhiwen Xie, Mang Ye, and Huyin Zhang. 2024. Contextual augmented global contrast for multimodal intent recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26963–26973.

Mukuntha Narayanan Sundararaman, Ayush Kumar, and Jithendra Vepa. 2021. Phoneme-BERT: Joint language modelling of phoneme sequence and ASR transcript. *CoRR*.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569.

Haibo Wang, Chenghang Lai, Yixuan Sun, and Weifeng Ge. 2024a. Weakly supervised gaussian contrastive grounding with large multimodal models for video question answering. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5289–5298.

Shijie Wang, Qi Zhao, Minh Quan Do, Nakul Agarwal, Kwonjoon Lee, and Chen Sun. 2024b. Vamos: Versatile action models for video understanding. In *European Conference on Computer Vision*, pages 142–160. Springer.

Yingzhi Wang, Abdelmoumene Boumadane, and Abdelwahab Heba. 2021. A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding. *arXiv preprint arXiv:2111.02735*.

Henry Weld, Xiaoqi Huang, Siqu Long, Josiah Poon, and Soyeon Caren Han. 2022. A survey of joint intent detection and slot filling models in natural language understanding. *ACM Computing Surveys*, 55(8):1–38.

Wenteng Wu, Wen Peng, JinYun Liu, XuDong Li, Dianhua Zhang, and Jie Sun. 2025. An attention-based weight adaptive multi-task learning framework for slab head shape prediction and optimization during the rough rolling process. *Journal of Manufacturing Processes*, 133:408–429.

Xian Wu, Yutian Zhao, Yunyan Zhang, Jiageng Wu, Zhihong Zhu, Yingying Zhang, Yi Ouyang, Ziheng Zhang, Huimin Wang, Jie Yang, et al. 2024. Medjourney: Benchmark and evaluation of large language models over patient clinical journey. *Advances in Neural Information Processing Systems*, 37:87621–87646.

Ying Xia, Zhen Xiong, Kefan Shen, Zhihong Zhu, Shaorong Xie, and Wei Liu. 2025. Rethinking decoding in multi-intent spoken language understanding. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786.

Junbin Xiao, Angela Yao, Zhiyuan Liu, Yicong Li, Wei Ji, and Tat-Seng Chua. 2022a. Video as conditional graph hierarchy for multi-granular question answering. In *AAAI*.

Junbin Xiao, Pan Zhou, Tat-Seng Chua, and Shuicheng Yan. 2022b. Video graph transformer for video question answering. In *European Conference on Computer Vision*, pages 39–58. Springer.

Bowen Xing, Libo Qin, Zhihong Zhu, Zhou Yu, and Ivor W Tsang. 2025. Dxa-net: Dual-task cross-lingual alignment network for zero-shot cross-lingual spoken language understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Keren Ye, Narges Honarvar Nazari, James Hahn, Zaeem Hussain, Mingda Zhang, and Adriana Kovashka. 2019. Interpreting the rhetoric of visual advertisements. *IEEE transactions on pattern analysis and machine intelligence*, 43(4):1308–1323.

Mang Ye, Qinghongya Shi, Kehua Su, and Bo Du. 2023. Cross-modality pyramid alignment for visual intention understanding. *IEEE Transactions on Image Processing*, 32:2190–2201.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.

Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. 2024. Self-chained image-language model for video localization and question answering. *Advances in Neural Information Processing Systems*, 36.

Shaozu Yuan, Xin Shen, Yuming Zhao, Hang Liu, Zhiling Yan, Ruixue Liu, and Meng Chen. 2022. Mcic: multimodal conversational intent classification for e-commerce customer service. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 749–761. Springer.

Anis Syafiqah Mat Zailan, Noor Hasimah Ibrahim Teo, Nur Atiqah Sia Abdullah, and Mike Joy. 2023. State of the art in intent detection and slot filling for question answering system: A systematic literature review. *International Journal of Advanced Computer Science & Applications*, 14(11).

Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. 2024a. A simple LLM framework for long-range video question-answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21715–21737, Miami, Florida, USA. Association for Computational Linguistics.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023a. SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15757–15773, Singapore. Association for Computational Linguistics.

Dongyu Zhang, Minghao Zhang, Heting Zhang, Liang Yang, and Hongfei Lin. 2021. Multimet: A multimodal dataset for metaphor understanding. 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 3214–3225.

Hanlei Zhang, Xin Wang, Hua Xu, Qianrui Zhou, Kai Gao, Jianhua Su, jinyue Zhao, Wenrui Li, and Yanting Chen. 2024b. MIntrec2.0: A large-scale benchmark dataset for multimodal intent recognition and out-of-scope detection in conversations. In *The Twelfth International Conference on Learning Representations*.

Hanlei Zhang, Hua Xu, Xin Wang, Qianrui Zhou, Shaojie Zhao, and Jiayan Teng. 2022. Mintrec: A new

dataset for multimodal intent recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1688–1697.

Hanlei Zhang, Qianrui Zhou, Hua Xu, Jianhua Su, Roberto Evans, and Kai Gao. 2024c. Multimodal classification and out-of-distribution detection for multimodal intent understanding. *arXiv preprint arXiv:2412.12453*.

Yazhou Zhang, Jinglin Wang, Yaochen Liu, Lu Rong, Qian Zheng, Dawei Song, Prayag Tiwari, and Jing Qin. 2023b. A multitask learning model for multimodal sarcasm, sentiment and emotion recognition in conversations. *Information Fusion*, 93:282–301.

Jinming Zhao, Tenggan Zhang, Jingwen Hu, Yuchen Liu, Qin Jin, Xinchao Wang, and Haizhou Li. 2022. M3ed: Multi-modal multi-scene multi-label emotional dialogue database. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5699–5710.

Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, et al. 2025. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1702–1713.

Qianrui Zhou, Hua Xu, Hao Li, Hanlei Zhang, Xiaohan Zhang, Yifan Wang, and Kai Gao. 2024. Token-level contrastive learning with modality-aware prompting for multimodal intent recognition. *AAAI*.

Yi Zhu, Zexun Wang, Hang Liu, Peiying Wang, Mingchao Feng, Meng Chen, and Xiaodong He. 2022. Cross-modal transfer learning via multi-grained alignment for end-to-end spoken language understanding. In *Interspeech 2022*, pages 1131–1135.

Zhihong Zhu, Xuxin Cheng, Zhaorun Chen, Yuyan Chen, Yunyan Zhang, Xian Wu, Yefeng Zheng, and Bowen Xing. 2024a. Inmu-net: advancing multi-modal intent detection via information bottleneck and multi-sensory processing. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 515–524.

Zhihong Zhu, Weiyuan Xu, Xuxin Cheng, Tengtao Song, and Yuexian Zou. 2023. A dynamic graph interactive framework with label-semantic injection for spoken language understanding. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Zhihong Zhu, Yunyan Zhang, Xianwei Zhuang, Fan Zhang, Zhongwei Wan, Yuyan Chen, QingqingLong QingqingLong, Yefeng Zheng, and Xian Wu. 2025. Can we trust AI doctors? a survey of medical hallucination in large language and large vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6748–6769, Vienna, Austria. Association for Computational Linguistics.

Zhihong Zhu, Xianwei Zhuang, Yunyan Zhang, Derong Xu, Guimin Hu, Xian Wu, and Yefeng Zheng. 2024b. Tfcd: Towards multi-modal sarcasm detection via training-free counterfactual debiasing. In *Proc. of IJCAI*.

Xianwei Zhuang, Xuxin Cheng, Liming Liang, Yuxin Xie, Zhichang Wang, Zhiqi Huang, and Yuexian Zou. 2024. PCAD: Towards ASR-robust spoken language understanding via prototype calibration and asymmetric decoupling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5235–5246, Bangkok, Thailand. Association for Computational Linguistics.

Yicheng Zou, Hongwei Liu, Tao Gui, Junzhe Wang, Qi Zhang, Meng Tang, Haixiang Li, and Daniell Wang. 2022. Divide and conquer: Text semantic matching with disentangled keywords and intents. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3622–3632.

A Discussion of IR with Other Modality Combinations

While this work comprehensively reviews intent recognition (IR) systems involving Text-Visual, Text-Acoustic, and Text-Visual-Acoustic modalities, the omission of Visual-Acoustic modality combinations warrants discussion. This exclusion stems from the following two factors:

(1) Data Scarcity. Visual-Acoustic IR lacks established benchmarks due to the absence of large-scale, intent-annotated datasets that exclude textual signals. (2) Utility Gaps. The practical relevance of Visual-Acoustic IR remains niche compared to text-inclusive multi-modal systems.

Overall, Visual-Acoustic IR presents an untapped potential for scenarios where textual signals are absent or unreliable. Addressing above issues could establish Visual-Acoustic IR as a viable subfield, complementing text-centric multi-modal IR.

B Application and Availability

The applications of MIR range from individual use to organizations. As the majority of these applications are similar to those in IR, this is not the focus of our survey. Nevertheless, we summarize the applications and data availability in Table 6.

Dataset Name	Potential Application or Task Setting	Data Link
MDID (Kruk et al., 2019) EMNLP	Social media event detection and user engagement prediction	https://ksikka.com/document_intent.html
MultiMET (Zhang et al., 2021) ACL	Multi-modal metaphors understanding in communicative environments	-
Behance Intent Discovery (Maharana et al., 2022) NAACL	Instructional video understanding	https://github.com/adymaharana/VideoIntentDiscovery
MCIC (Yuan et al., 2022) NLPCC	E-commerce customer service	-
MSAIRS (Shi et al., 2024) arXiv	Chatting applications, social platforms, and media comment sections	-
IntentQA (Li et al., 2023a) ICCV	Inference video question answering	https://github.com/JoseponLee/IntentQA
SLURP (Bastianelli et al., 2020) EMNLP	Spoken language understanding, task-oriented dialogue systems	https://github.com/pswietojanski/slurp
Fluent Speech Commands (FSC) (Lugosch et al., 2019) INTERSPEECH	Spoken language understanding, task-oriented dialogue systems	fluent.ai/research/fluent-speech-commands/
MInDS-14 (Gerz et al., 2021) EMNLP	Multilingual task-oriented dialogue systems	https://huggingface.co/datasets/PolyAI/minds14
MIntRec (Zhang et al., 2022) MM	Conversational interactions	https://github.com/thuiar/MIntRec
EMOTyDA (Saha et al., 2020) ACL	Intelligent dialogue systems, conversational speech transcription	https://github.com/thuiar/MIntRec
EmoInt-MD (Singh et al., 2022) TASLP	Social Conversations	-
MIntRec2.0 (Zhang et al., 2024b) ICLR	Human-computer interaction	https://github.com/thuiar/MIntRec2.0
MC-EIU (Liu et al., 2024b) arXiv	Multi-modal conversation	https://github.com/MC-EIU/MC-EIU

Table 6: Existing MIR benchmarks in terms of applications and availability. '-' denotes not released.