Research Statement – Structure-aware Intelligence Learning

Hao FEI

https://haofei.vip

My current research direction lies in both Natural Language Processing (NLP) and Multimodal (MM) Learning (e.g., Language, Vision, Video, Speech). My interests broadly cover NLP and MM applications, such as Large Language Modeling (LLM), Information Extraction, Affective Computing, Syntax/Semantic Parsing, Text-to-Image/Video/Audio Generation, and Multimodal Reasoning. I am apt to construct learning models, with the fundamental goal of building systems capable of human-level understanding of the world. My ongoing research focuses on the particular angle of **Structure-aware Intelligence Learning** (**SAIL**), which aims at enhancing the semantics understanding of varied modalities with the intrinsic data structure modeling. The SAIL idea works effectively for the deep learning-based AI, and also holds for the current LLMs, which will ultimately help achieve AGI of universal modalities (world modeling). Also, I believe so much that the key to realizing human-level AGI lies in two fundamental aspects simultaneously, **A)** human-level complex reasoning ability, and **B)** mastering of the world knowledge, with one not doing without the other.

My research papers have been published in top-tier ML/NLP/DM venues, including, ICML, NeurIPS, ACL, CVPR, AAAI, WWW, SIGIR, IJCAI, EMNLP, etc. Over the past 5 years, my published papers have garnered over 3,000 citations, with an h-index of 33, according to my Google Scholar profile. I won the award of *World AI Conference* (WAIC'23) Rising Star in 2023. My research received the *Paper Award Nomination* at ACL 2023. My NeurIPS 2022 paper was elected as a spotlight paper. I won more than ten honors and awards for excellent research performance in my PhD stage. My PhD thesis was awarded the *Excellent Doctoral Thesis of Chinese Information Processing Society* in 2022 (5 persons per year&country).

Following, I will elaborate in detail on my research status.

Table of Content

1 Research Envision: Structure-aware Semantics Understanding

The semantics of the world can be essentially organized in structured formats, and different data of modalities comes with structural representations. For example, the understanding of almost all the NLP applications can be seen as a hierarchy with different levels. For the understanding of other modal information (e.g., visions), the key also lies in the comprehension of semantic structure, such as with the scene graph representation. Figure 1 exemplifies the linguistic syntax structures in NLP of dependency tree & constituency grammar, and also the visual scene graph structure in CV. Besides, the world knowledge has been represented in structured formats, i.e., knowledge graph. Also human-level reasoning largely follows structural manner.

(a) Linguistic syntactic structures

(b) Visual scene graph structure

Figure 1: Examples of linguistic structures in NLP and scene graph structures in CV.

With the idea of SAIL, I generally divide my research into three key branches: **structure-aware NLP**, **structure-aware MM** and **structure-aware LLM**. Starting with deep learning-based semantics understanding in NLP area, I engaged in the exploration of structure-aware NLP. During my postdoctoral research stage, I have extended the SAIL idea to structure-aware MM. The recent rise and great triumph of LLM have revealed the great potential of leading AGI via this path. Correspondingly, latest I proactively integrate the idea of structural awareness into the LLM for semantics understanding, i.e., structure-aware LLM. And the ultimate goal is thus to realize human-level AGI for universal modalities by modeling the semantic structures of the world. To achieve the AGI goal via SAIL that aligns the most with human society, these targets also should and will be achieved, including efficacy, interpretability, robustness (generalizability), efficiency (scalability) and trustworthiness. In Figure 2, I summarize and illustrate the big picture of my research goal. In what follows I will expand in detail each of the three research branches with the explorations in my existing work, one by one.

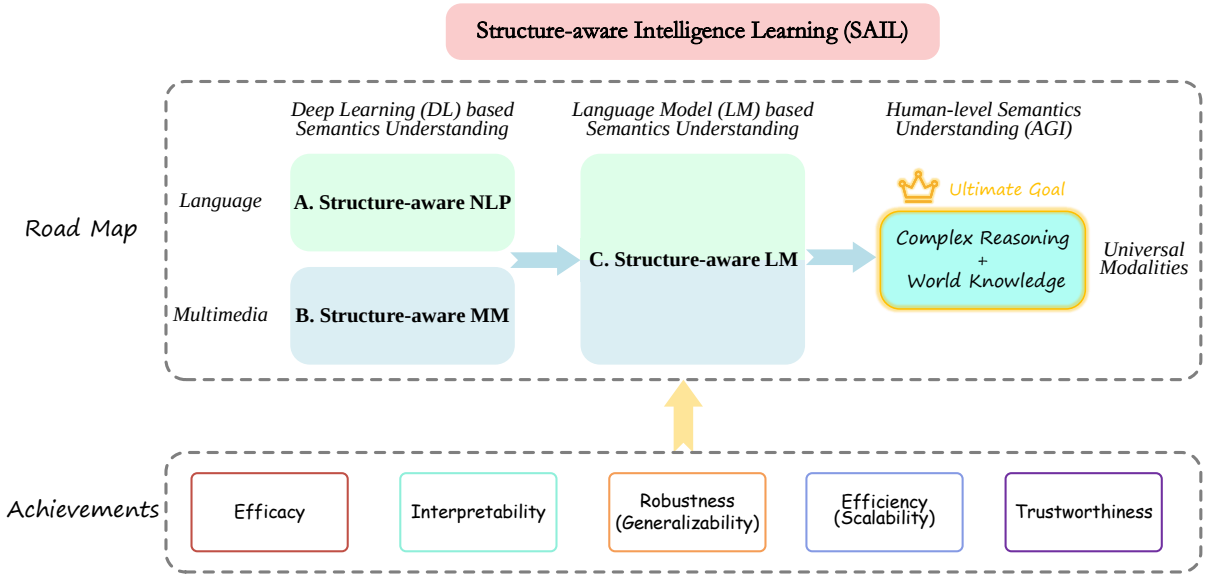


Figure 2: The big picture of my research to achieve the ultimate AGI goal.

2 Research Branch-A: Structure-aware NLP

Empowering machines with the ability of understanding the deep semantics of the natural language is the ultimate goal of NLP, which is also the key to the realization of AGI. In essence, most NLP tasks can be conceptualized as a hierarchical structure consisting of different levels, which, from bottom to top, encompass *lexical structures*, *syntactic structures* and *semantic (or discourse) structures*. For humans it would be effortless to understand the language and various NLP tasks, while for machines it is non-trivial to understand the structural representations of syntax, semantics and discourse. In my research, I have dedicated to developing effective models and tools to parse languages into these structures. More specifically, according to the genre and scenario of languages, I split the structural modeling into **1) sentence level**, **2) dialogue level**, and **3) document level**.

2.1 Sentence-level Structural Modeling

Most of the natural languages and texts come in the sentence piece, which makes sentence-level structural modeling the majority of the analysis.

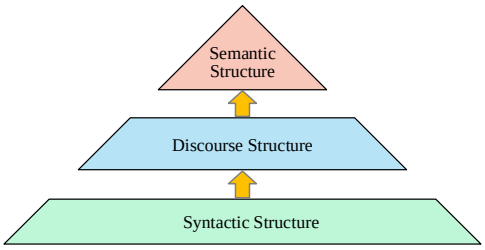


Figure 3: Hierarchy of language modeling.

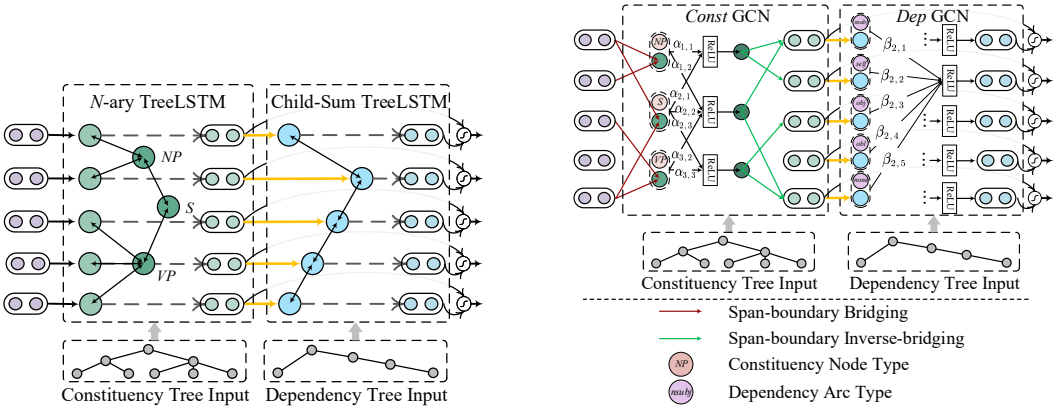
• **Syntax Parsing and Grammar Induction** Constituency and dependency are the two most representative syntactic structures but with distinct formalisms [1, 2], which are created to depict the grammatical structure of a sentence. Constituency trees, also known as phrase-structure trees, arrange words and phrases into hierarchically nested constituents. In contrast, dependency trees establish a direct connection between words, linking them with directed dependencies. Due to the foundational position of syntactic structure in the hierarchy of language understanding, it provides direct support to higher-level semantic comprehension. Consequently, much of my prior research has utilized these two fundamental language structures.

• **Structured Information Extraction** Information Extraction involves automatically extracting structured information from unstructured or semi-structured text data, which serves as a very crucial and initial step in building downstream applications, e.g., the knowledge graph. IE consists of multiple tasks, representatively including Named Entity Recognition (NER), Relation Extraction (RE) and Event Extraction (EE). In [3], we present an innovative model for discontinuous NER based on pointer networks. In [4], we present a novel alternative by modeling the unified NER as word-word relation classification, namely W^{2}NER. In [5], we investigate the integration of rich knowledge to prompt tuning for stronger few-shot NER. In [6, 7], we propose an end-to-end neural model for overlapping RE by treating the task as a quintuple prediction problem. In [8] we design a simple yet effective tagging scheme and model to formulate EE as word-word relation recognition. In [9], we propose a tree-based model to automatically learn the features from the syntactic dependency tree for trigger detection. In [10], we devise a novel syntax-based dynamic latent graph model for event temporal relation extraction and subevent relation extraction tasks. Also in [11], we investigate an unbiased universal dependency (UD) based cross-lingual RE transfer by constructing a type of code-mixed UD forest.

• **Structured Sentiment Analysis** Sentiment analysis and opinion mining is one of the most important research topics of NLP, which aims to infer the sentiments and attitudes towards goods and services behind social media texts, thus having fundamental impacts on real-world society. My prior research in this track mostly focuses on structured sentiment analysis, including Aspect-based Sentiment Analysis (ABSA) and Syntax-aided Sentiment Analysis. In [12], we propose a token graph model with a novel labeling strategy, consisting of the whole and essential label sets, to extract sentiment tuples for structured sentiment analysis. In [13], we consider modeling all the ABSA subtasks as a hierarchical dependency, multiplexing the information of low-level tasks to the higher-level tasks one by one. In [14, 15] we build syntax-enriched frameworks for encoding syntactic features, modeling both the dependency edges and labels, as well as the POS tags unifiedly, and a local-attention module encoding POS tags for better sentiment term extraction.

• **Semantic Parsing** Semantic parsing is a type of fundamental task in NLP that involves mapping natural language expressions to their corresponding formal representations of underlying semantics, typically in the form of logical forms. Semantic Role Labeling (SRL) is a key specific subtask of semantic parsing for offering deeper insights into the meaning of the sentence. SRL focuses on converting natural language sentences into structured *predicate–argument* representations, such as 'who did what to whom, when and where'. In [16], we present the first work of transition-based neural models for end-to-end SRL. In [17], we propose a novel label-aware graph convolutional network to encode both the syntactic dependent arcs and labels.. In [18], we further explore the

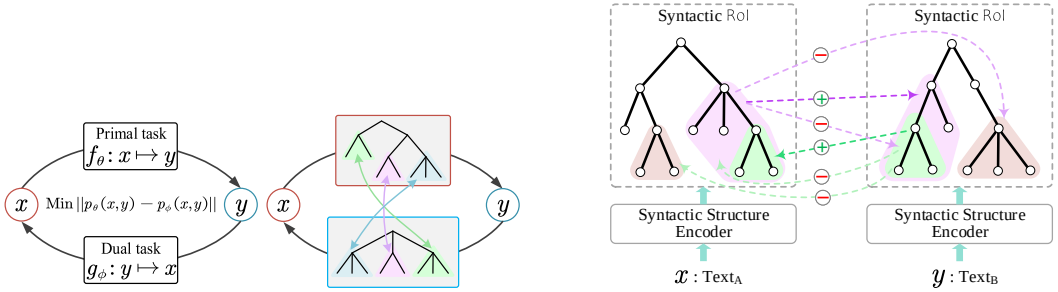
integration of heterogeneous syntactic representations for SRL (HeSyFu), as shown in Figure 4. We verify that SRL and both kinds of syntactic structures have strong associations and should be exploited for mutual benefits.



(a) TreeLSTM-based HeSyFu system. (b) GCN-based HeSyFu system.

Figure 4: Two explorations of the heterogeneous syntactic representation integrations for SRL.

• **Structure-guided Text Generation** Text generation is another one of the key venues of NLP, which covers many specific applications, e.g., text summarization, machine translation, headline generation, and text paraphrase. Many natural language generation (NLG) tasks appear in dual forms, which can be solved by the dual learning technique [19] to model the dualities between the two coupled tasks. In [20], we propose to enhance dual learning with *structure matching*. As shown in Figure 5(a), based on the vanilla dual learning framework of text↔text generation, we perform structural alignment unsupervisedly between the primal and dual tasks, bridging them with explicitly structure connections. We perform dually-syntactic structure co-echoing of the region of interest (RoI) between the task pair, together with a syntax cross-reconstruction at the decoding side, cf. Figure 5(b).



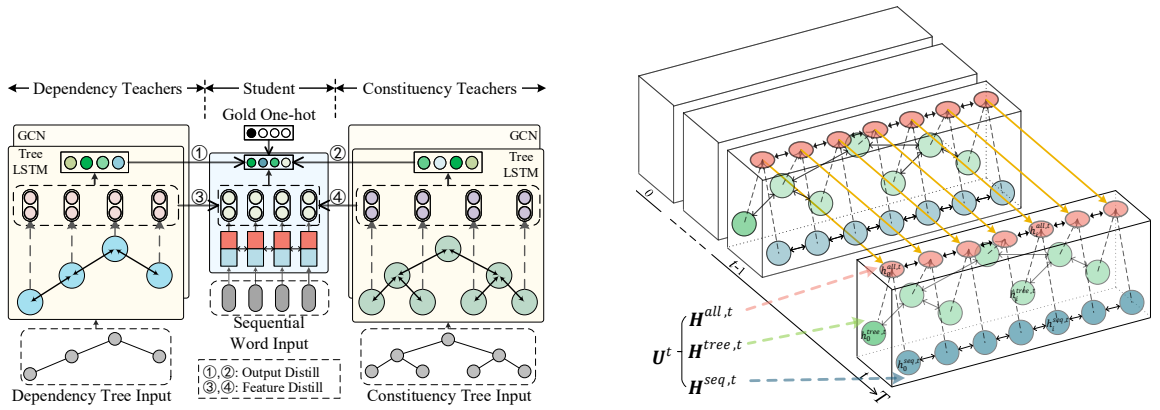
(a) **Left**: dual learning framework. **Right**: dual learning with alignment of structural supervision.

(b) Dually-syntactic RoI alignment.

Figure 5: Natural language generation with dual learning.

• **Coreference Chain Resolution** Coreference chain resolution, often referred to as coreference resolution, aims at identifying expressions in a text that refer to the same entity or concept, forming what is known as a coreference chain. The resolution of the chain is essentially the structure modeling. In [21], we introduce a new benchmark of nominal compound chain, which is characterized by the semantic-enriched lexicons within the chain, compared with the standard lexical chain. In [22] we study the coreference resolution for the legal texts, where different from the texts in the general domain, we study how to encode legal texts and incorporate reference relations between entities.

• **Syntax-aided Semantics Modeling** Syntax has been shown useful for various NLP tasks. Besides the above work on one specific NLP task, we also explore the general-purpose syntax-aided semantics modeling for broad-covered NLP tasks. In [23], we investigate a simple and effective method, Knowledge Distillation, to integrate heterogeneous structure knowledge into a unified sequential LSTM encoder, cf. Figure 6(a). Experiments on four typical syntax-dependent tasks show the efficacy of integrating rich heterogeneous structure syntax with syntax structure distillation. While prior syntax-based NLP research employs shallow integration of syntax and semantics, in [24] we propose a deep neural communication system between syntax and semantics to improve the performance of text understanding. As shown in Figure 6(b), global communication is performed to ensure comprehensive information propagation, and local communication between syntactic tree encoder and sequential semantic encoder enhances the mutual learning of information exchange.



(a) Distilling heterogeneous syntax.

(b) Deep syntax-semantics communication system.

Figure 6: Two explorations on general-purpose syntax-aided text semantics modeling.

• **Universal Structured NLP** Many NLP tasks can be essentially understood as the comprehension of the underlying semantic or syntactic structure of texts, e.g., syntactic parsing, information extraction, coreference resolution, and sentiment&opinion extraction, all of which can be summarized as *structured NLP* (XNLP). Essentially, as the key common characteristics, all the XNLP tasks have revolved around predicting two key elements from input: **1) textual spans** and **2) relations between spans**, as depicted in the Figure 7 (lower part). Thus, by unifying various XNLP tasks, we can take advantage of the shared characteristics among tasks, leading to better model generalization and improved performance in realistic scenarios of product deployment. Despite certain recent efforts to explore universal solutions for specific categories of XNLP tasks, a comprehensive and effective approach for unifying all XNLP tasks long remains underdeveloped. In the meanwhile, while XNLP demonstration systems are vital for researchers exploring various XNLP tasks, existing platforms can be limited to, e.g., supporting few XNLP tasks, lacking interactivity and universalness. To this end, in [25] we propose an advanced XNLP demonstration platform, where we propose leveraging LLM to achieve universal XNLP, *with one model for all* with high generalizability. We deploy the demo system online at https://xnlp.haofei.vip/. Please visit the webpage to gain a direct use experience.

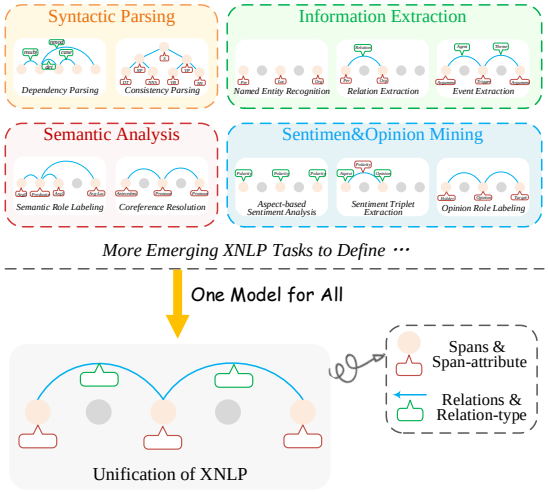
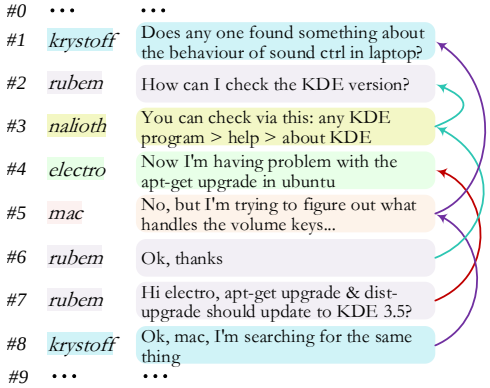


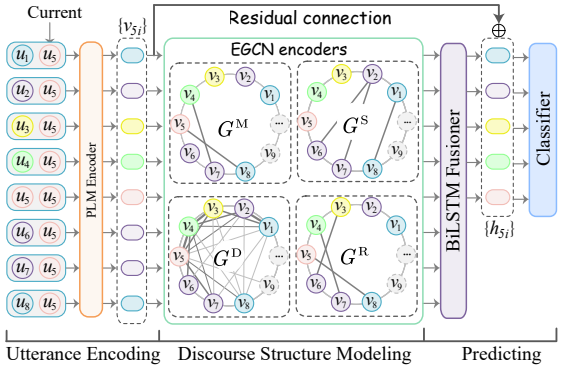
Figure 7: Universal Structured NLP (XNLP).

2.2 Dialogue-level Structural Modeling

Dialog can also be a ubiquitous scenario of NLP, where texts are organized into utterance pieces raised by different speakers and parties. As dialogs come with conversational discourses, modeling the dialogue-level structure becomes one of the essential topics of conversational NLP.



(a) A conversation snippet with the replying structure.



(b) Modeling conversation discourse structures for dialogue disentanglement.

Figure 8: Modeling dialogue-level structures.

• **Conversation Discourse Structure Parsing** Dialogue disentanglement, aka., conversation discourse structure parsing, aims to detach the chronologically ordered utterances into several independent sessions and replying relations, cf. Figure 8(a). Conversation utterances are essentially organized and described by the underlying discourse, and thus dialogue disentanglement requires the full understanding and harnessing of the intrinsic discourse attribute. In [26], we thus propose enhancing dialogue disentanglement by taking full advantage of the dialogue discourse characteristics (cf. Figure 8(b)).

• **Conversational Information Extraction** There is also an emergency of information extraction under the conversational context, especially the dialogue-level relation extraction, i.e., DiaRE. Piror DiaRE methods either simply concatenate the utterances in dialogue into a long piece of text, while the structural characteristics in dialogues have not been fully utilized. In [27], we investigate a novel dialogue-level mixed dependency graph (D^2G, cf. Figure 9) and an argument reasoning graph (ARG) for DiaRE with a global relation reasoning mechanism.

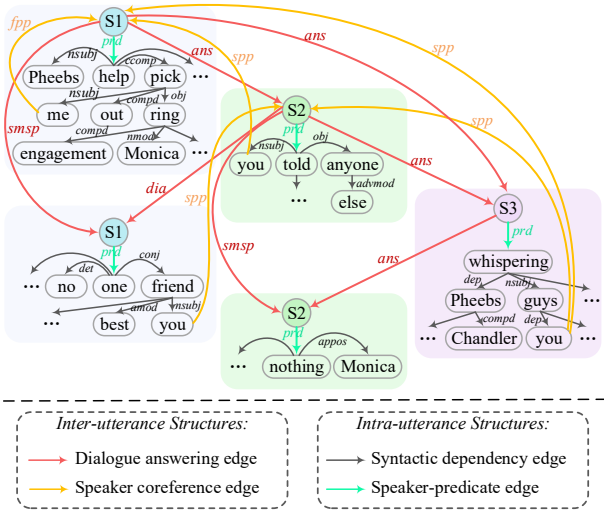


Figure 9: Dialogue-level mixed dependency graph (D^2G) for DiaRE.

• **Conversational Semantic Role Labeling** Conversational SRL, aka., ConvSRL, extends the regular SRL into multi-turn dialogue scenario. While few attempts have been made for ConvSRL, unfortunately, several key ConvSRL characteristics remain unexplored, such as the dialogue discourse structural information integration. In [28], we investigate the integration of a latent graph for ConvSRL. We propose to automatically induce a predicate-oriented latent graph with a predicate-centered Gaussian mechanism, by which the nearer and more informative words to the predicate will be allocated with more attention. Our system outperforms best-performing baselines on three benchmark datasets with big margins.

• **Conversation Sentiment Analysis** The sentiment analysis in the conversation context has also gained increasing attention. The sentiment classification (SC) often intertwines with the dialogue act recognition (DAR), which necessitates the joint DAR and DCN analysis. In [29], we improve the task by fully modeling the local contexts of utterances. First, we employ the dynamic convolution network as the utterance encoder to capture the dialogue contexts. While the current fine-grained sentiment analysis research is mostly limited to the scenario of a single text piece, the study under the conversation should also have critical values in real-world society. Thus, we explore this topic in two works. First, in [30] we introduce a novel task of conversational aspect-based sentiment quadruple analysis, namely DiaASQ, aiming to detect the quadruple of *target-aspect-opinion-sentiment* in dialogues. Second, in [31] we extend the emotion-cause pair extraction task with a broader definition and scenario, presenting a new task, Emotion-Cause Quadruple Extraction in Dialogs.

2.3 Document-level Structural Modeling

Texts also sometimes occur in the form of long documents, and these documents are composed of sentences that are structured in a coherent discourse manner, creating a flow of information. Given the natural organization of documents, understanding the structure at the document level is a fundamental aspect of NLP.

• **Document Discourse Structure Parsing** Rhetorical structure theory (RST) parsing is a typical task of analyzing the hierarchical and rhetorical structure of document texts, cf. Figure 10. RST involves identifying the relationships and dependencies between different textual elements, i.e., elementary discourse unit (EDU), to reveal how they contribute to the overall meaning and coherence of a document. RST discourse parsing methods span from transition-based models [32] to chart-based parsing [33, 34], most of which take a bottom-up manner. Further top-down RST parsing is

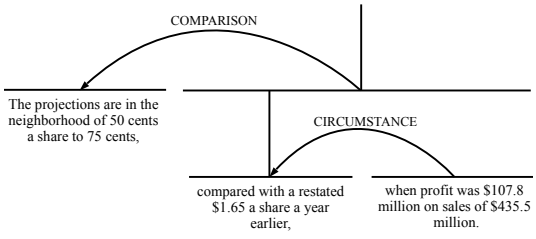


Figure 10: An example of RST discourse structure.

proposed to maintain an overall view of the input document. In [35], Stack-Pointer network is adopted for the parsing based on a seq2seq model.

• **Text Generation from Document** In [36], we explore the structural guidance for headline generation, a task to summarize a long document (such as news articles) with a short catchy title. By assembling the document-level rhetorical structure theory (RST) trees and the sentence-level abstract meaning representation (AMR) graphs, we construct the S^3 graphs (cf. Figure 11), where the hierarchical composition of the sentence, clause and word intrinsically characterizes the semantic meaning of the overall document.

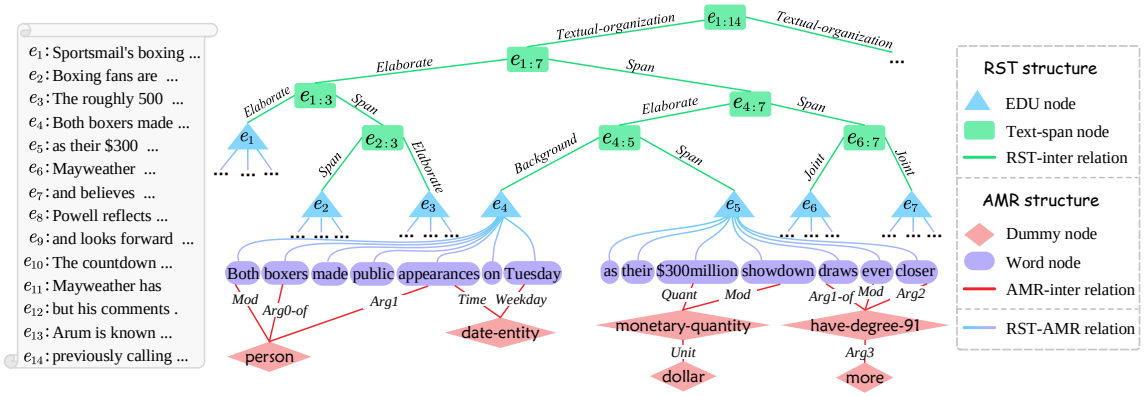


Figure 11: Illustration of our proposed unified semantic discourse structure (S^3) of the document.

• **Documental Information Extraction** Information extraction also has the scenario of documents. Compared with documented NER and EE, document-level RE has received the greatest research attention, since it requires complex reasoning using mentions, entities, local and global contexts. In [37], we propose a novel mention-based reasoning (MRN) module based on explicitly and collaboratively local and global reasoning. While most studies focus on extracting the relations within one single document, some researchers have begun to explore cross-document RE. In [38], we address the shortages of prior cross-document RE methods, e.g., noisy introduction and under-consideration of the connections between cross-document paths.

• **Documental Sentiment Analysis** Compared with sentence-level sentiment analysis, document-level sentiment classification deals with longer document texts (e.g., product reviews). Prior research simply treats the document as an overall text unit, performing feature extraction with various sophisticated model architectures. In [39], we draw inspiration from fine-grained sentiment analysis, proposing to first learn the latent target-opinion distribution behind the documents, and then leverage such fine-grained prior knowledge into the classification process.

3 Research Branch-B: Structure-aware MM

In a world filled with diverse data types, there is an urgent need to enable machines to understand and process information from multiple sources of sensory data, e.g., text, images, audio and video. This necessitates multimodal learning, an interdisciplinary field at the intersection of CV, NLP and machine learning. In this part, I will walk through my prior explorations on structure-aware MM. According to whether the multimodal structures are being parsed or being used, I divide the theme into 1) structural parsing and 2) structure-aware multimodal applications.

3.1 Structure Parsing

Representing the multimedia content and data into structured representations has been a fundamental task. In this topic, I mainly have done two types of explorations, as elaborated following.

• **Scene Graph Parsing** Scene Graph (SG) is a highly structured representation [40], which is defined as a grounded (textually, visually or in video) graph over the object instances within a specific scene at a semantic level. Generally, SG comprises three types of nodes, where the object nodes correspond to object bounding boxes with their category labels and attributes, and the edges represent pair-wise relationships between objects. As illustrated in Figure 12, SG representations can be used to represent multimodal data, e.g., text, image and video. Similar to language structures, SGs effectively depict the semantic relationships within various modalities, which can assist numerous downstream tasks by providing superior feature modeling. For this reason, my related research on multimodal learning heavily relies on modeling this multimodal structural information.

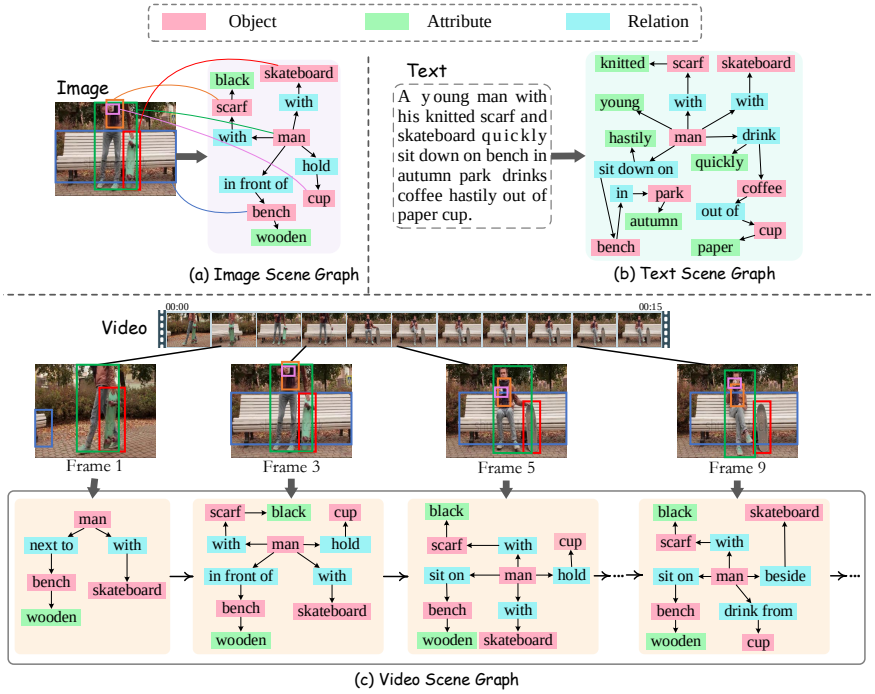


Figure 12: Illustration of (a) Image Scene Graph, (b) Text Scene Graph, and (c) Video Scene Graph.

• **Multimodal Grammar Induction** From the human cognition perspective, the language acquisition follows a phylogenetic manner, in which we perceive the world through rich heterogeneous signals, such as text, vision, and acoustics. In the process, features from distinct modalities essentially serve complementary roles to each other. With such intuition, in [41] we introduce a novel *unsupervised visual-audio-text grammar induction* task (namely **VAT-GI**), to induce the constituent grammar trees from parallel images, text, and speech inputs. Figure 13 illustrates the concept of VAT-GI. Inspired by the fact that toddlers first learn to communicate verbally and visually before acquiring textual reading skills, we further introduce a *textless* setting of VAT-GI, wherein the task solely relies on visual and auditory inputs. To approach the task, we propose a visual-audio-text inside-outside recursive autoencoder (VaTiora) framework, in which the rich modal-specific and complementary features are fully leveraged and integrated for effective grammar parsing.

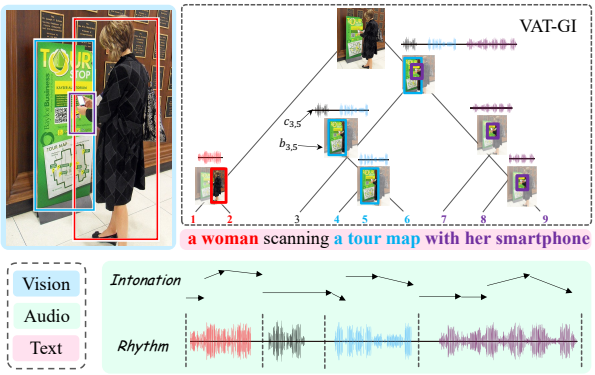


Figure 13: Unsupervised grammar induction with vision, audio and text modality sources, each of which contributes complementarily to the task.

3.2 Structure-aware Multimodal Applications

Essentially, SG representations can aid mulitmodal learning due to three key characteristics. 1) SG helps improve the cross-modal alignment for allowing more fine-grained modality-agnostic vision-text matching, compared with the existing mostly adopted instance-level multimodal alignment. 2) SG, representing the semantic relational scenes, helps enhance multimodal fusion by providing semantic-level features for cross-modal learning. 3) SG as the highly structured representation, also ensures more controllable end-task prediction when being integrated into multimodal applications. Following, I will elaborate on my explorations of the SG structure representation-enhanced multimodal applications.

• **Multimodal Sentiment Analysis** It has been a hot research topic to enable machines to understand human emotions in multimodal contexts. In [42] we study a more challenging scenario: multimodal emotion analysis in conversation (MM-ERC). After revisiting the characteristic of MM-ERC, we argue that both the *feature multimodality* and *conversational contextualization* should be properly modeled simultaneously. Thus, based on the contrastive learning technique, we devise a dual-level disentanglement mechanism to decouple the features into both the modality space and utterance space.

• **Multimodal Information Extraction** Information extraction under multimodal scenarios also has been an important topic. Among all multimodal IE tasks, multimodal relation extraction (MRE) has gained increasing attention. However, existing research on MRE faces with two co-existing challenges, *internal-information over-utilization* and *external-information under-exploitation*. To this end, in [43] we propose a novel framework that simultaneously implements the idea of *internal-information screening* and *external-information exploiting*. Also, the task of event extraction of vision or video, also having its name such as multimodal semantic role labeling, or vision/video situation recognition, has been a hot venue. In [44], we study the most challenging one, i.e., Video Semantic Role Labeling (VidSRL), detecting the salient events from given videos, by recognizing the predict-argument event structures and the interrelationships between events. This work explores a novel *holistic spatio-temporal scene graph* representation based on the existing dynamic scene graph structures, which well model both the fine-grained spatial semantics and temporal dynamics of videos for VidSRL.

• **Multimodal Machine Translation** Unsupervised multimodal machine translation (UMMT) is an important application, where the model is trained supervisedly in only the text-image pairs (<*text-img*>) without large volume of parallel sentences (<*src-tgt*>). In [45], we investigate a more realistic UMMT setup, *inference-time image-free* UMMT, where the model is trained with source-text image pairs, and tested with only source-text inputs, achieving the goal of not only avoiding *the parallel sentences during training*, but also *the text-image pairs during inference*. We propose a SG-pivoted UMMT system, cf. Figure 14, where we represent the input images and texts with the visual and language SGs, and such fine-grained vision-language features ensure a holistic understanding of the semantics.

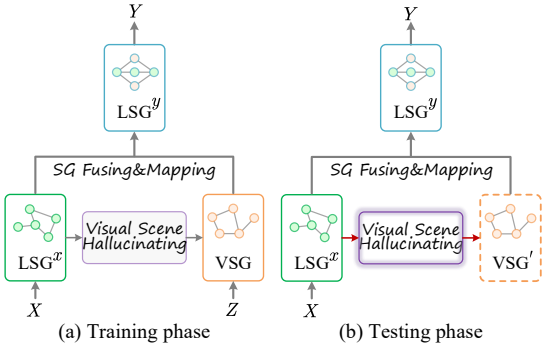


Figure 14: SG-pivoted UMMT system.

• **Vision Captioning** Generating captions for the given vision is one of the key directions in multimodal learning. Despite the impressive results achieved by deep learning in automatic image captioning, one performance bottleneck is the availability of large paired datasets because neural image captioning models are generally annotation-hungry. This motivates the task of Unpaired Cross-lingual Image Captioning, which however has long suffered from irrelevancy and disfluency issues, due to the inconsistencies of the semantic scene and syntax attributes during transfer. In [46], we propose to address the above problems by incorporating the SG structures and the syntactic constituency (SC) trees. Our captioner contains the *semantic structure-guided image-to-pivot captioning* and the *syntactic structure-guided pivot-to-target translation*, two of which are joined via pivot language (cf. Figure 15). We then take the SG and SC structures as pivoting, performing cross-modal semantic structure alignment and cross-lingual syntactic structure alignment learning.

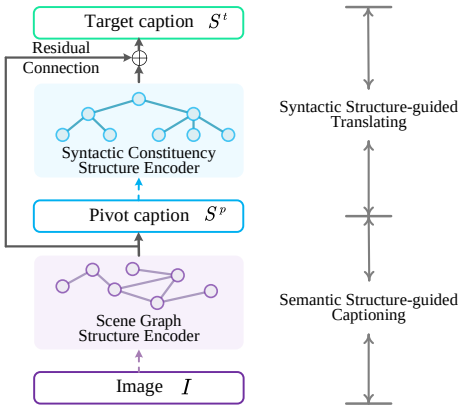
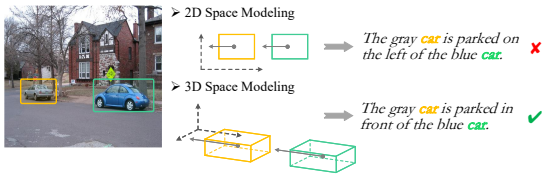
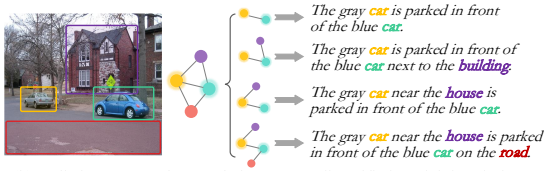


Figure 15: The cross-lingual captioner.

In addition to the captioning of visual semantics, enabling machines with the captioning for spatial understanding is also critical, i.e., the Visual spatial description (VSD) task. VSD aims to generate texts that describe the spatial relations of the given objects within images [47]. Existing VSD work merely models the 2D geometrical vision features, thus inevitably falling prey to the problem of skewed spatial understanding of target objects, cf. Figure 16. In [48], we investigate the incorporation of 3D scene features for VSD. With an external 3D scene extractor, we obtain the 3D objects and scene features for input images, based on which we construct a target object-centered 3D spatial scene graph (GO3D-S^{2}G), such that we model the spatial semantics of target objects within the holistic 3D scenes. Besides, we propose a scene subgraph selecting mechanism, sampling topologically diverse subgraphs from GO3D-S^{2}G, where the diverse local structure features are navigated to yield spatially diversified text generation.



(a) Modeling 3D scene features results in correct spatial understanding



(b) Holistic 3D scene features help generate diversified spatial descriptions

Figure 16: Modeling 3D scene graph for VSD.

• **Cross-modal Retrieval** Cross-modal retrieval is a multimodal task to retrieve the images/videos (resp. texts) that are relevant to a given textual (resp. image/video) query. The fundamental challenge in this task is to accurately learn the language-vision semantic grounding. While most of the text-image retrieval is limited to the instance (i.e., coarse-grained) level, in [49] we explore the fine-grained one, the Referring Image Segmentation (RIS) task. RIS aims to ground a given language query onto a specific region of an image, i.e., typically represented by a segmentation map. Building upon the existing RefCOCO and Visual Genome datasets, we propose a new RIS benchmark with complex queries, namely RIS-CQ, which challenges the existing RIS with enriched, specific and informative queries, and enables a more realistic scenario of RIS research. We also present a dual-modality graph alignment model (DUMOGA) for RIS-CQ, which helps outperform a series of RIS methods. Going beyond the image, we also study video-language retrieval. Given a descriptive language query, Video Moment Retrieval (VMR) aims to seek the corresponding semantic-consistent moment clip in the video. In [50] we design a new setting of VMR where users can easily point to small segments of non-controversy video moments and our proposed method can automatically fill in the remaining parts based on the video and query semantics. To support this, we propose a new framework named Video Moment Retrieval via Iterative Learning (VMRIL), treating the partial temporal region as the seed, then expanding the pseudo label by iterative training.

• **Audio/Speech Modeling** Audio or speech modeling is a subfield of multimodal learning, which generally encompasses a wide range of tasks related to processing and understanding audio data, such as Automatic Speech Recognition (ASR), Text-to-Speech Synthesis (TTS), Voice Activity Detection (VAD), etc. Among them, I mainly study the task of Voice Conversion (VC). VC aims to generate a new speech with the source content and a target voice style. In [51], we focus on one general setting, i.e., non-parallel many-to-many voice conversion, which is close to the real-world scenario. Inspired by the inherent structure of mel-spectrogram, we propose a new voice conversion framework, i.e., Subband-based Generative Adversarial Network for Voice Conversion (SGAN-VC), explicitly exploits the style spatial characteristics of different subbands to convert each subband content of source speech separately.

• **Text-to-Vision Generation** Text-conditioned visual generation is one of the currently hottest research topics of multimodal learning, where textual inputs are translated into corresponding image or video contents. One of the keys to successful text-to-vision generation lies in effectively bridging the gap between the language and vision modalities. Language has always been abstract and concise, compared with the vision (either image or video) which is specific and even redundant in content. Thus, it is critical to correctly understand the user input texts before generating high-quality vision. In this thread, my research methods revolve around integrating structural representations to bridge the gap between language and visions, such as SG representations of image and video, or layout information.

While existing Text-to-Image (T2I) synthesis methods achieve photorealistic image generation, still several misalignment issues hinder the high-faithfulness T2I performance, including problematic spatial relation understanding and numeration failure. In [52], we strive to synthesize high-fidelity images that are semantically aligned with a given textual prompt without any guidance. Toward this end, we propose a coarse-to-fine paradigm to

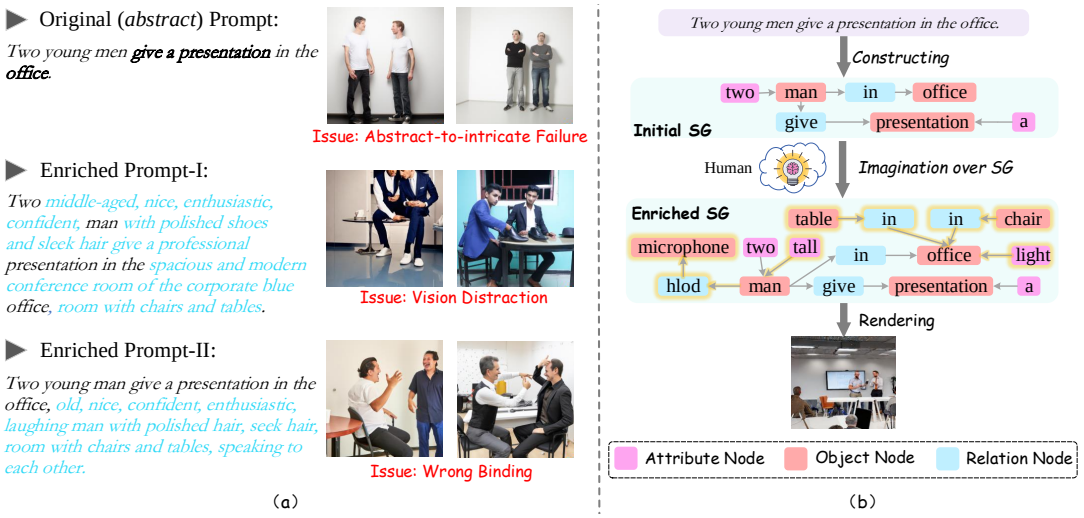


Figure 17: (a) Abstract-to-intricate T2I synthesis example. Enriched contexts are in *blue*. (b) Human intuition on the abstract-to-intricate T2I process: we always first grasp the semantic structure of the original prompt text, i.e., SG, and then carry out imagination with more complete scenes based on the SG.

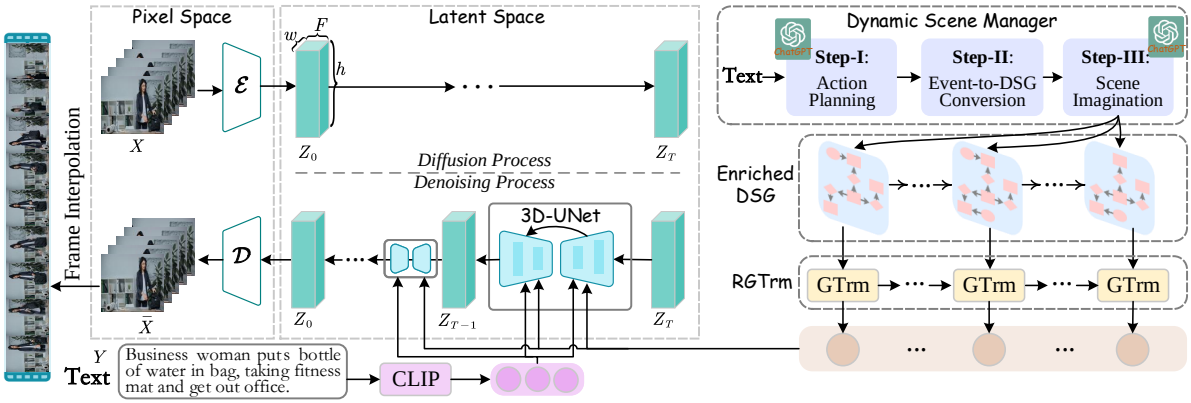


Figure 18: The architecture of the proposed dynamics-aware T2V diffusion framework (Dysen-VDM).

achieve image generation with **layout planning**. Concretely, we first generate the coarse-grained layout conditioned on a given textual prompt via in-context learning based on LLMs. Afterward, we propose a fine-grained object-interaction diffusion method to synthesize high-faithfulness images conditioned on the prompt and the automatically generated layout. In [53], we investigate the T2I synthesis under the abstract-to-intricate setting, i.e., *generating intricate visual content from simple abstract text prompts*, cf. Figure 17. Inspired by human imagination intuition, we propose a novel scene-graph hallucination (SGH) mechanism for effective abstract-to-intricate T2I synthesis. SGH carries out scene hallucination by expanding the initial SG of the input prompt with more feasible specific scene structures, in which the structured semantic representation of SG ensures high controllability of the intrinsic scene imagination.

Although achieving the current SoTA generative performance, diffusion-based Text-to-video (T2V) synthesis can still face several common yet non-negligible challenges, such as *unsmooth video transition*, *crude video motion* and *action occurrence disorder*, all of which are rooted in lacking the modeling of the *intricate video temporal dynamics*, the crux of video synthesis. In [54], we investigate strengthening the awareness of video dynamics for DMs, for high-quality T2V generation. Inspired by human intuition, we design an innovative dynamic scene manager (dubbed as **Dysen**) module. As shown in Figure 18, Dysen-VDM includes (**step-1**) extracting from input text the key actions with proper time-order arrangement, (**step-2**) transforming the action schedules into the dynamic scene graph (DSG) representations, and (**step-3**) enriching the scenes in the DSG with sufficient and reasonable details. See the project page here: `http://haofei.vip/Dysen-VDM/`.

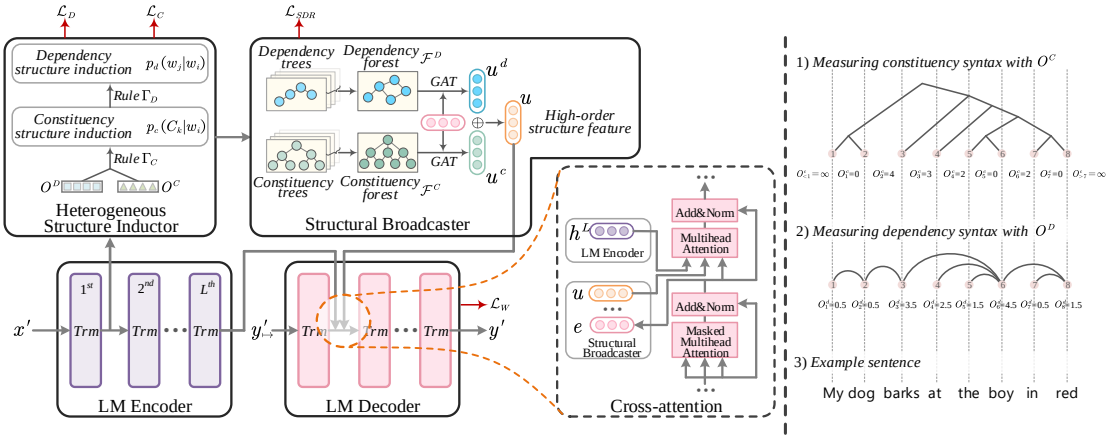
4 Research Branch-C: Structure-aware LLM

Recently, LLMs represent a groundbreaking area of AI research in ushering the human-like AI, exhibiting remarkable capabilities in understanding, generating, and processing human language at an unprecedented scale. At the core of this research is the development of the large-scale pre-training and various fine-tuning techniques. LLMs have revolutionized various learning tasks across multiple areas, e.g., NLP, CV and MM. In this theme, I extend the SAIL idea to the LLM for enhanced semantics understanding.

4.1 Language Modeling

I have explored language modeling with various external information integration and structural modeling.

• **Structure-aided Language Modeling** It is an established finding that the syntactic structure signals help improve many end tasks that rely on syntax features. It has also been shown that the language models (LMs), after being pre-trained, capture certain implicit language structure knowledge within it. However, such structure features learned via the vanilla Transformer LM are insufficient for task improvement. In [55], we retrofit the structure-aware Transformer language model for facilitating end tasks by proposing to exploit syntactic distance to encode both the phrasal constituency and dependency connection into the language model. Further in [56], we devise a novel structure-aware generative language model (GLM), fully unleashing the power of syntactic knowledge for universal information extraction (UIE). The framework is named LasUIE system, cf. Figure 19. A heterogeneous structure inductor is explored to unsupervisedly induce rich heterogeneous structural representations by post-training an existing GLM. In particular, a structural broadcaster is devised to compact various latent trees into explicit high-order forests, helping to guide a better generation during decoding. We finally introduce a task-oriented structure fine-tuning mechanism, further adjusting the learned structures to most coincide with the end-task's need. Over 12 IE benchmarks across 7 tasks our system shows significant improvements over the baseline UIE system.

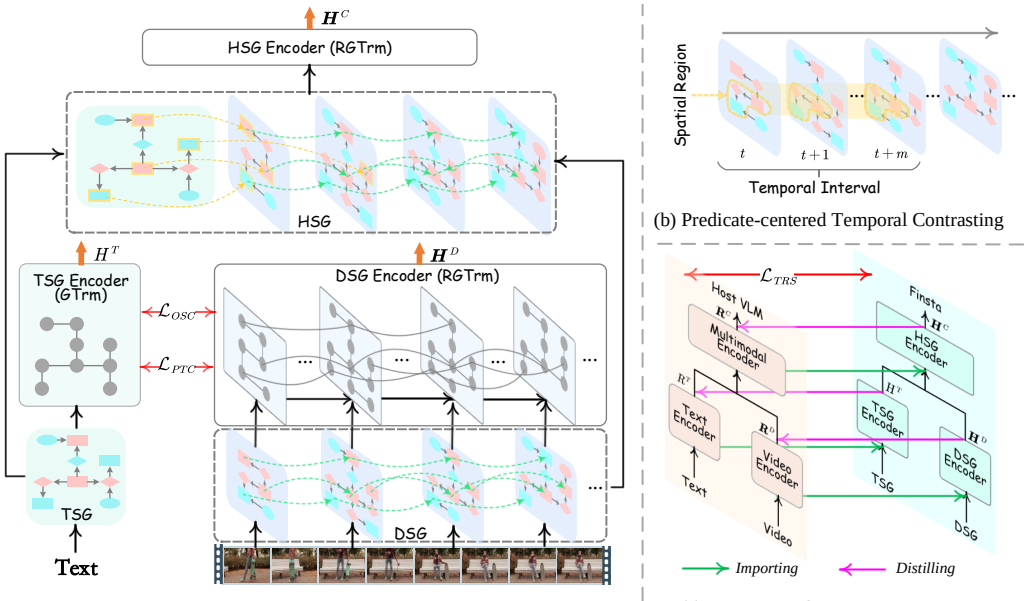


(*a*) Unsupervised structure-aware post-training (*b*) Heterogeneous syntax measurements

Figure 19: Our LasUIE framework under (*a*) unsupervised structure-aware post-training. Heterogeneous structure inductor module generates both constituency and dependency structures via (*b*) two heterogeneous syntax measurements.

• **KG-enriched Language Modeling** Injecting the knowledge from the knowledge graph (KG) into LM would intuitively enhance the capability of LM of a specific domain. This is a direct manner to compensate for the lack of domain knowledge of the general-purpose LM. In [57], we explore the integration of external biomedical knowledge graph into an LM, so as to enhance the performance of an array of biomedical information extraction (BioIE) tasks. We enrich a contextual LM by integrating a set of large-scale biomedical knowledge graphs.

• **Vision Multimodal Language Modeling** To date, the LLMs show amazing capability in language understanding. Yet, our world is inherently multimodal, and humans perceive the world with different sensory organs for varied modal information, such as language, images, videos, and sounds. With such intuition, the purely text-based LLMs have further been endowed with understanding and perception capabilities of other modalities, such as visual, video, audio, etc. By associating the text-based LLMs with other modalities, researchers extend the LLMs into multimodal ones, i.e., MLLMs. Among various MLLMs, the image-based vision-language LLMs (VL-LLMs) have gained the most attention. While training a VL-LLM from scratch is prohibitively expensive, in [58] we investigate the visual prompt generator (VPG) transferability across LLMs. We for the first time show that effective VPG transfer across LLMs can be achieved, suggesting that it is possible to build a high-performance VL-LLM with considerably lower cost. We design a two-stage transfer framework named VPGTrans, which is simple yet highly effective. With our VPGTrans, we customize two novel VL-LLMs, VL-LLaMA and VL-Vicuna, at a significantly lower cost. See the project page here: https://vpgtrans.github.io/.



(a) The High-level View of Our Framework (b) Predicate-centered Temporal Contrasting (c) Registering Our Finsta into a VLM

Figure 20: (a) Fine-grained structural spatio-temporal alignment learning (Finsta) based on the dual-stream framework with three SG encoders. (b) Extracting the spatial region and temporal interval for the predicate-centered temporal alignment. (c) Injecting our Finsta representations into a host LVM.

Pre-training video-language model (VLM) has also received increasing research attention. Compared with image-language modeling which focuses mainly on individual image semantic understanding, video understanding goes beyond static images, requiring both the comprehension of spatial semantics and temporal dynamics, due to the nature of a sequence of frames over time. Nevertheless, existing VLMs can suffer from certain common limitations, e.g., coarse-grained cross-model aligning, under-modeling of temporal dynamics, detached video-language view. In [59], we target enhancing VLMs with a *fine-grained structural spatio-temporal alignment learning* method (namely Finsta, cf. Figure 20). First, we represent the input texts and videos with fine-grained SG structures, both of which are further unified into a holistic SG (HSG) for bridging two modalities. Then, an SG-based framework is built, where the textual SG (TSG) is encoded with a graph transformer, while the video dynamic SG (DSG) and the HSG are modeled with a novel recurrent graph transformer for spatial and temporal feature propagation. Next, based on the fine-grained structural features of TSG and DSG, we perform object-centered spatial alignment and predicate-centered temporal alignment respectively, enhancing the VL correspondence in both the spatiality and temporality. We design our method as a plug&play module, which can be integrated into existing well-trained VLMs for further representation augmentation without training from scratch, or relying on SG annotations in downstream applications.

Developing LMMs that can comprehend, reason, and plan in complex and diverse 3D environments remains a challenging topic, especially considering the demand for understanding permutation-invariant point cloud 3D representations of the 3D scene. Existing works seek help from multi-view images, and project 2D features to 3D space as 3D scene representations, while leads to huge computational overhead and performance degradation. In [60] we present **LL3DA** (cf. Figure 21), a 3D MLLM that takes point cloud as direct input and responds to both textual instructions and visual prompts. This help LMMs better comprehend human interactions and further help to remove the ambiguities in cluttered 3D scenes. Experiments show that LL3DA achieves remarkable results, and surpasses various 3D vision-language models on both 3D Dense Captioning and 3D Question Answering. See the project page here: https://ll3da.github.io/.

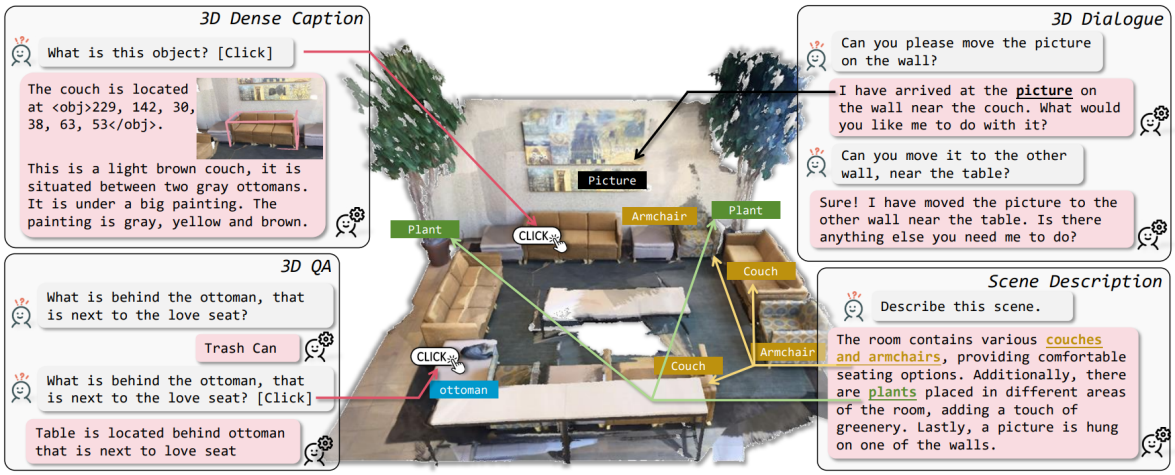


Figure 21: LL3DA is a 3D MLLM that demonstrates mighty instruction-following capacities of understanding, reasoning, and planning in complex 3D environments.

• **Universal Vision Language Modeling** Existing vision LLMs might still encounter challenges such as *superficial instance-level understanding*, *lack of unified support for both images and videos*, and *insufficient coverage across various vision tasks*. To fill the gaps, in [61] we present **VITRON** (cf. Figure 22), a universal pixel-level vision LLM, designed for comprehensive understanding (perceiving and reasoning), generating, segmenting (grounding and tracking), editing (inpainting) of both static image and dynamic video content. Utilizing an LLM backbone, VITRON incorporates specialized encoders for images, videos, and pixel-level regional visuals within its frontend architecture, while as its backend, employing a text-centric invocation strategy for integrating diverse state-of-the-art off-the-shelf modules tailored for an array of vision-related end tasks. Via this, VITRON supports a spectrum of vision end tasks, spanning visual understanding to visual generation, from low level to high level. Through joint vision-language alignment and fine-grained region-aware instruction tuning, VITRON achieves precise pixel-level perception. We further enhance its capabilities with invocation-oriented instruction tuning, allowing for flexible and precise module invocation for downstream vision tasks. Demonstrated over 12 visual tasks and evaluated across 22 datasets, VITRON showcases its extensive capabilities in the four main vision task clusters, e.g., segmentation, understanding, content generation, and editing. Various demonstrations also illustrate VITRON's fortes in visual manipulation and user interactivity. With **VITRON**, we aspire to create

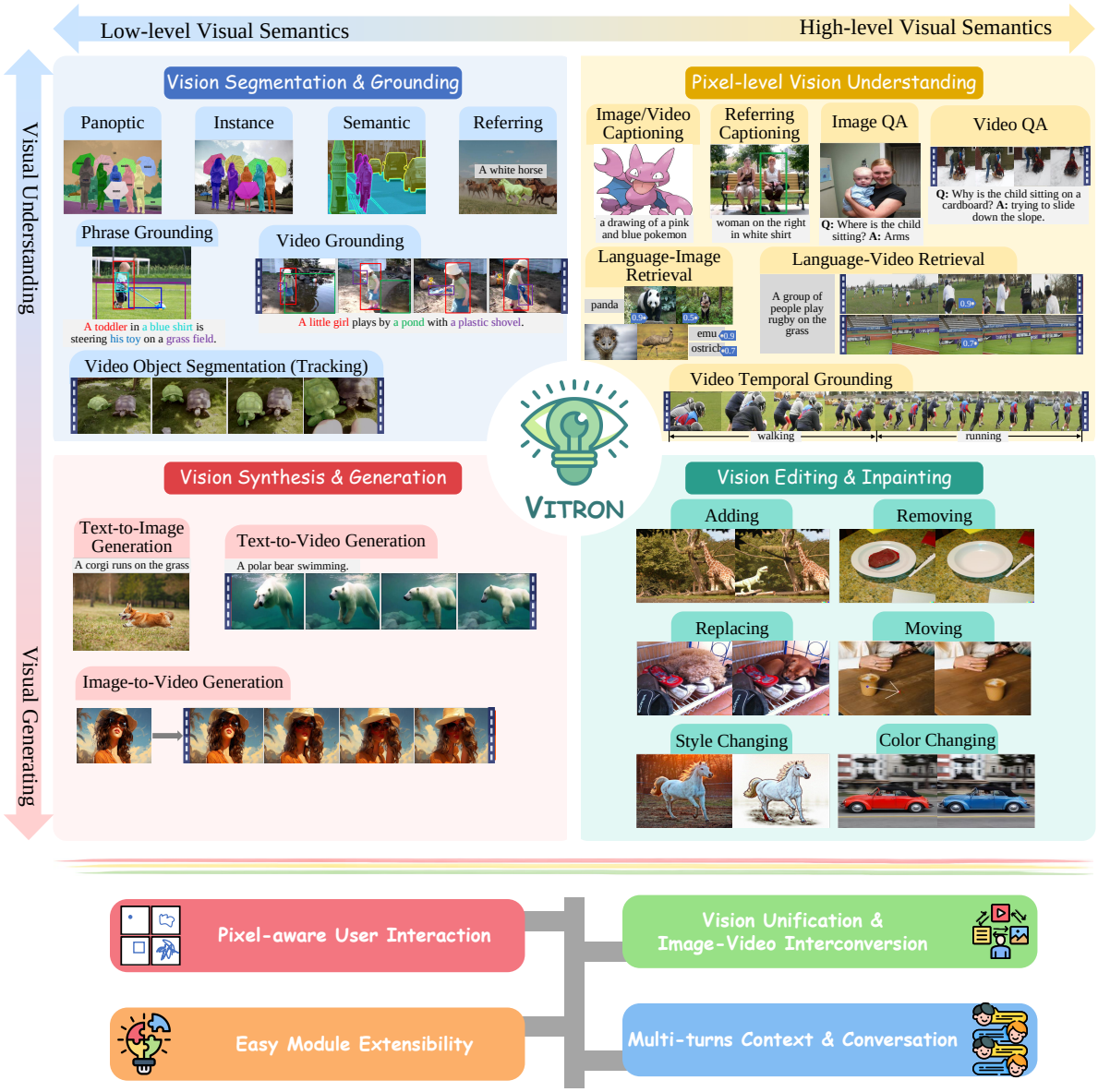


Figure 22: **VITRON**, a universal pixel-level vision LLM that supports four main task clusters of visions, and also advances in four key features.

a powerful open-sourced, interactive vision system that can compete with industry-level vision-language systems like OpenAI's DALL-E series and the Midjourney, thereby aiding the advancement of academic research. See the project page here: https://vitron-llm.github.io/.

• **Universal Multimodal Language Modeling** While existing multimodal LLMs show promising capability in perceiving various information more than texts, unfortunately, most of these efforts pay attention to the multimodal content understanding at the input side, lacking the ability to output content in multiple modalities more than texts. We emphasize that real human cognition and communication indispensably require seamless transitions between any modalities of information. This makes the exploration of any-to-any MLLMs critical to achieving real human-level AI, i.e., accepting inputs in any modality and delivering responses in the appropriate form of any modality. In [62] we pioneer this research gap by presenting an end-to-end general-purpose any-to-any MLLM system, **NExT-GPT**. We connect an LLM with multimodal adaptors and different diffusion decoders, enabling NExT-GPT to perceive inputs and generate outputs in arbitrary combinations of text, images, videos, and audio. By leveraging the existing well-trained highly-performing encoders and decoders, NExT-GPT is tuned with only a small amount of parameter (1%) of certain projection layers, which not only benefits low-cost training but also facilitates convenient expansion to more potential modalities. This research showcases the promising possibility of building a unified AI generalist capable of modeling universal modalities, paving the way for more human-like AI research in the community. Thus, NExT-GPT has been gaining huge attention both from academia and industry

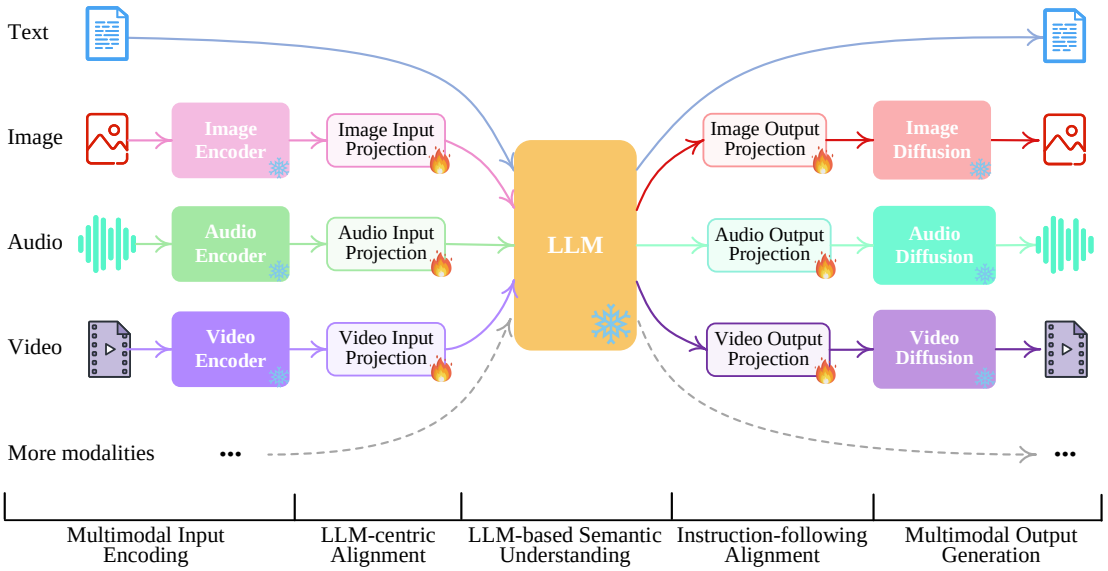


Figure 23: By connecting LLM with multimodal adaptors and diffusion decoders, **NExT-GPT** achieves universal multimodal understanding and any-to-any modality input and output.

upon its publication, i.e., receiving over 2.8k GitHub stars for the released code and model. Please visit the project homepage (https://next-gpt.github.io/) to access the online demo.

4.2 LLM-empowered Machine Learning

LLMs have sparked several concurrent techniques, such as prompt/instruction tuning, in-context learning, etc. On the other hand, as LLMs have exhibited the human-level semantics understanding ability to certain extent, utilizing LLMs to empower a wide range of downstream learning applications will be a promising approach, e.g., via LLM-empowered reasoning.

• **Prompt Tuning** Prompt tuning [63] is a simple yet effective mechanism for learning "soft prompts" to condition frozen LMs to perform specific downstream tasks. Unlike discrete text prompts, soft prompts are learned through back-propagation and can be tuned to incorporate signals from any number of labeled examples. While the vanilla prompt tuning can be limited to the drawback of knowledge-scarce characteristics, which may fall short of certain downstream tasks. In [5], we propose incorporating the deep prompt tuning framework with threefold knowledge, including the internal 1) *context knowledge* and the external 2) *label knowledge* & 3) *sememe knowledge*. TKDP encodes the three feature sources and incorporates them into the soft prompt embeddings, which are further injected into existing pre-trained LMs to facilitate predictions.

Figure 24: Illustration of modality-switching instruction tuning.

• **Instruction Tuning** As the manner of interactions with LLMs has been shifted into a more human-centric 'query-answer' style, the learning of LLMs has also been changed, where instruction tuning (IT) has been introduced as a major approach for LLMs fine-tuning. With IT, LLMs are trained to faithfully follow explicit, human-provided instructions to generate responses or complete tasks accurately. IT involves additional training of overall MM-LLMs using '*(INPUT, OUTPUT)*' pairs, where '*INPUT*' represents the user's instruction, and '*OUTPUT*' is the desired output that conforms to the given instruction. In [64] we introduce a generic and efficient approach with a parameter-isolated architecture, ControlRetriever, which makes use of instructions that explicitly describe retrieval intents in natural language, for controlling dense retrieval models to directly perform varied retrieval tasks.

For instruction tuning the multimodal LLMs, multimodal IT is then introduced, where the input instructions contain information in multiple modalities, including text, images, audio, etc. However, for the broader scenario of any-to-any multimodal LLMs, where users and LLMs involve diverse and dynamically changing modalities in

their inputs and outputs, the existing multimodal IT datasets fail to provide multimodal responses (i.e., output). To fill the gap, we in [62] propose a novel Modality-switching Instruction Tuning (`MosIT`, cf. Figure 24). `MosIT` not only supports complex cross-modal understanding and reasoning but also enables sophisticated multimodal content generation.

• **In-context Learning** With the increasing ability of LLMs, in-context learning (ICL) has become a new paradigm for NLP and related fields, where LLMs make predictions only based on contexts augmented with a few examples. It has been a new trend to explore ICL to evaluate and extrapolate the ability of LLMs. With ICL, we reach the goal of better eliciting knowledge and insights from LLMs for downstream tasks. In [52] we resort to ICL to activate LLMs for layout generation, as the preliminary step to the image generation. ICL employs a natural language prompt that includes a task description (`Instruction`), a few examples (`In-context examples`) selected from the training dataset as demonstrations, and a test instance (`Test instance`), as depicted in Figure 25. Also, previous studies have shown that the effectiveness of ICL is highly influenced by the selection of demonstrations. Thus, we devise an adaptive sampler based on the task feedback to select examples in a reinforcement learning framework. We also adopt the similar ICL strategy for the video synthesis for modeling the video dynamics [54], where we elicit the spatio-temporal insights from LLMs. We design three ICL examples for action planning, step-wise scene imagination and global scene polishment, respectively.

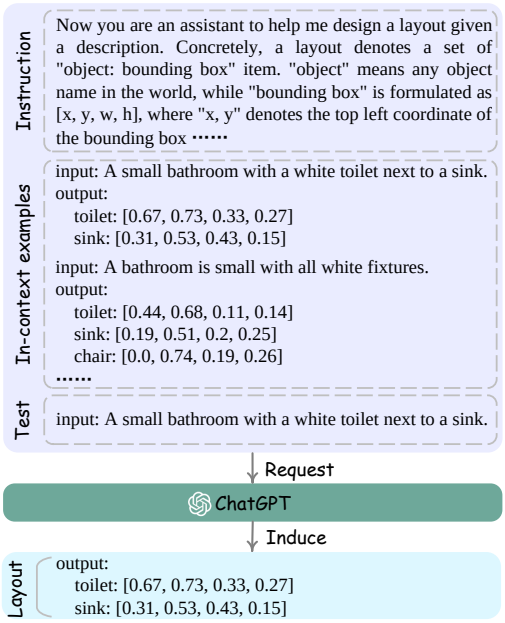


Figure 25: Example ICL for layout generation.

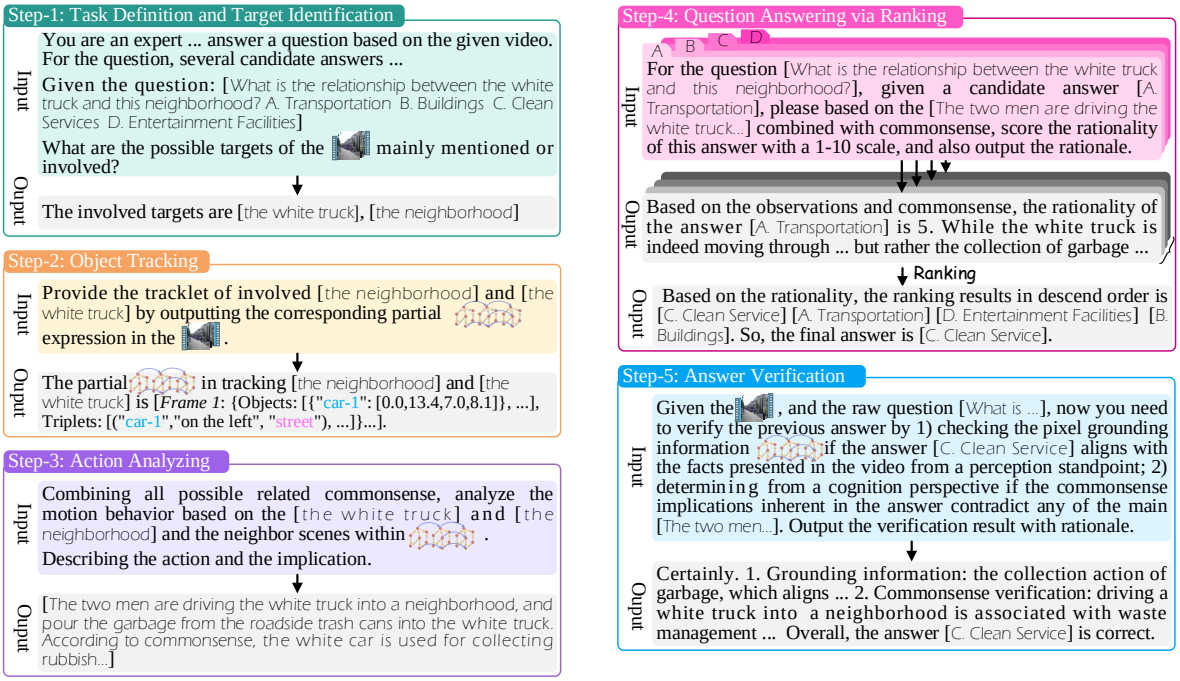


Figure 26: Overview of Video-of-Thought (VoT) reasoning framework.

• **Reasoning** Achieving human-like logical reasoning capabilities is crucial for realizing AGI, which plays a pivotal role in enabling intelligent systems to engage in problem-solving, decision-making, and critical thinking. In both the NLP and CV communities, previously simple perception and recognition problems have been perfectly addressed by powerful LLMs. However, for more complex tasks, especially those requiring in-depth reasoning, equipping models with cognitive-oriented reasoning capabilities remains a gap for further research. Fortunately, researchers have designed the Chain-of-Thought (CoT) concept [65], which offers a solution enabling LLMs with complex problem-solving abilities for various tasks. In [66] we investigate the CoT method for implicit

sentiment analysis (ISA), wherein the opinion cues come in an implicit and obscure manner, and thus require complex reasoning ability to infer the latent intent of opinion. We introduce a *Three-hop Reasoning* (THOR) CoT framework to mimic the human-like reasoning process for ISA. While the vanilla CoT might still struggle in handling logical reasoning that relies much on symbolic expressions and rigid deducing rules, in [67], we propose a novel Symbolic Chain-of-Thought (**SymbCoT**), a fully LLM-based framework that integrates symbolic expressions and logic rules with CoT prompting to strengthen the logical reasoning capability of LLMs.

Videos naturally entail the temporal consecution over frames, where the video dynamic logic requires reasoning. Existing research of video understanding still struggles to achieve in-depth comprehension and reasoning in complex videos, primarily due to the under-exploration of two key bottlenecks: *fine-grained spatial-temporal perceptive understanding* and *cognitive-level video scene comprehension*. In [68] we develop a Video-of-Thought (**VoT**) reasoning framework, as shown in Figure 26. VoT inherits the CoT core, breaking down a complex task into simpler and manageable sub-problems, and addressing them step-by-step from a low-level pixel perception to high-level cognitive interpretation. To support the fine-grained pixel-level spatial-temporal video grounding, we also propose a novel video MLLM, MotionEpic, which integrates video spatial-temporal scene graph (STSG) representation.

5 Ongoing and Future Research Plan

The idea of SAIL has great potential and is still worth further exploration. Also, the community has fully entered the era of LLMs, and we all have witnessed the immense potential of LLMs leading towards AGI. The following aspects and topics w.r.t. SAIL and LLMs will be considered in my near future research.

5.1 MLLM Unification with Universal Tokenization

I have investigated two universal LLMs, Vitron [61] and NExT-GPT [62], where Vitron is used to unify fine-grained image and video vision modeling, and NExT-GPT aims to unify all modalities. While achieving good unification, there is still a vast array of aspects to explore. All existing MLLMs are fundamentally language-based, i.e., pre-trained based solely on a pure language corpus. It is thus easy to question that such purely language-based pre-training could lead to perception bias, as humans intuitively understand and interact with the world by incorporating all different modalities without bias. For example, infants can learn about entities and concerns through vision before they learn language. Additionally, current MLLMs achieve multimodal perception and output by using external visual encoders (such as CLIP, Imagebind, Q-former) and visual decoders (such as Diffusion, VQ-VAE). However, different modalities, such as vision and language, have different implications when encoded in existing MLLMs. For instance, a token in text can directly represent an entity or concept; in contrast, a visual token inputted into an LLM is just a patch simply divided from an image, with each token actually not semantically encapsulated, failing to fully represent a visual entity or concept. This asymmetry in tokenization leads to an unequal status between language and other modal signals (such as vision) in LLMs, increasing the difficulty of cross-modal understanding and reducing the precision of LLMs in cross-modal applications. As a next step, I plan to explore more effective cross-modal tokenization techniques to achieve a more powerful MLLM unification.

5.2 Cognition-oriented Cross-Modal Logical Reasoning

So far, based on the CoT technology, I have explored reasoning techniques for text [66], symbolic reasoning [67], and video reasoning [68]. It is worth considering that in the real world, humans are actually capable of seamlessly reasoning across multimodal data. I plan to investigate a more powerful system for arbitrary cross-modal reasoning next. There are two main aspects to consider as solutions. First, how to integrate discrete and logical symbolic reasoning methods into general scene reasoning, achieving a human-like, multi-hop, unrestricted reasoning process. Second, similar to the VoT, we might consider simulating the human process of intuition to gradually progress from perception to cognition, thereby achieving cross-modal reasoning.

5.3 4D Vision Generation

Previously, I've explored image and video generation and 3D MLLM construction. I plan to take this topic a step further by researching the synthesis of 4D video. 3D is a very important topic in the field of vision, and most current research focuses on static 3D generation or editing. However, in reality, we experience the world as a dynamic interaction of 3D spatial and temporal changes. My next step is to investigate the generation of 4D content, which involves adding a temporal dimension to the existing 3D models. Considering the promising potential shown by current 3D generation methods based on Gaussian Splashing, I plan to use this approach for 4D generation. Most importantly, I plan to follow the workflow of Dysen-VDM [54] and draw inspiration from

human cognition by simulating human 4D cognitive patterns. This approach aims to integrate the dynamic and complex nature of human perception and understanding into the synthesis of 4D video, offering a more nuanced and realistic representation of time and space in digital media.

5.4 World Knowledge Modeling with Universal Scene Graph Representations

Currently the SG representations, e.g., images SG, text SGs, and video SGs, have their own separate definitions and benchmarks. But in fact, all of the information in different modalities of this world can be quite shared, and mutually beneficial and complementary. Thus the separation of SGs might be an obstruction for a unified view of world knowledge modeling. This makes the research of Universal SG (USG) indispensable. The USG can be defined as a topological structure of a scene description from text, image, video, or any combination of modalities, cf. Figure 27. Based on the motivation for constructing USG, here are some constraints for USG:

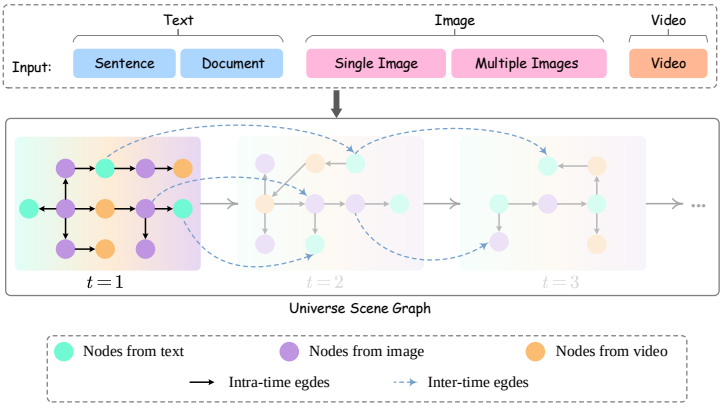


Figure 27: Illustration of a universe scene graph.

- In instances where a scene description comprises solely a single modality, the USG simplifies into an unimodal SG.
- The nodes in USG can be visual or textual objects.
- Within USG, redundancy of nodes is strictly avoided. For example, it is common in MSG to establish links between a textual entity, 'man', and a visual object 'man' are connected. However, in USG, when these two references pertain to the same objects, the textual object serves as the category label for the visual object, rather than existing as an independent object.
- The edges in USG include two types: *intra-time edges*, signifying concurrent relationships, and *inter-time edges*, representing coreference relationships between objects across different time points. In situations where relationships remain static over time, only intra-time edges are present. For example, only intra-time edges are applicable in a static image.

With USG representation, one potential and promising move is to further build more universal LLM or generalist over the USG. Thus, I plan to enhance the universal LLMs by (large-scale) pre-training them over such Universal SG representation of multimodal data.

References

[1] Chris Dyer, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah A Smith. Transition-based dependency parsing with stack long short-term memory. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, ACL*, pages 334–343, 2015.

[2] Timothy Dozat and Christopher D Manning. Deep biaffine attention for neural dependency parsing. In *Proceedings of International Conference on Learning Representations, ICLR*, 2016.

[3] Hao Fei, Donghong Ji, Bobo Li, Yijiang Liu, Yafeng Ren, and Fei Li. Rethinking boundaries: End-to-end recognition of discontinuous mentions with pointer networks. In *Proceedings of the AAAI conference on artificial intelligence, AAAI*, pages 12785–12793, 2021.

[4] Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. Unified named entity recognition as word-word relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence, AAAI*, pages 10965–10973, 2022.

[5] Jiang Liu, Hao Fei, Fei Li, Jingye Li, Bobo Li, Liang Zhao, Chong Teng, and Donghong Ji. Tkdp: Threefold knowledge-enriched deep prompt tuning for few-shot named entity recognition. *arXiv preprint arXiv:2306.03974*, 2023.

[6] Hao Fei, Yafeng Ren, and Donghong Ji. Boundaries and edges rethinking: An end-to-end neural model for overlapping entity relation extraction. *Information Processing & Management*, 57(6):102311, 2020.

[7] Hao Fei, Yue Zhang, Yafeng Ren, and Donghong Ji. A span-graph neural model for overlapping entity relation extraction in biomedical texts. *Bioinformatics*, 37(11):1581–1589, 2021.

[8] Hu Cao, Jingye Li, Fangfang Su, Fei Li, Hao Fei, Shengqiong Wu, Bobo Li, Liang Zhao, and Donghong Ji. Oneee: A one-stage framework for fast overlapping and nested event extraction. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING*, pages 1953–1964, 2022.

[9] Hao Fei, Yafeng Ren, and Donghong Ji. A tree-based neural network model for biomedical event trigger detection. *Information Sciences*, 512:175–185, 2020.

[10] Ling Zhuang, Hao Fei, and Po Hu. Syntax-based dynamic latent graph for event relation extraction. *Information Processing & Management*, 60(5):103469, 2023.

[11] Hao Fei, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Constructing code-mixed Universal Dependency forest for unbiased cross-lingual relation extraction. In *Findings of the Association for Computational Linguistics: ACL*, pages 9395–9408, 2023.

[12] Wenxuan Shi, Fei Li, Jingye Li, Hao Fei, and Donghong Ji. Effective token graph modeling using a novel labeling strategy for structured sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, ACL*, pages 4232–4241, 2022.

[13] Hao Fei, Fei Li, Chenliang Li, Shengqiong Wu, Jingye Li, and Donghong Ji. Inheriting the wisdom of predecessors: A multiplex cascade framework for unified aspect-based sentiment analysis. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, pages 4096–4103, 2022.

[14] Shengqiong Wu, Hao Fei, Yafeng Ren, Donghong Ji, and Jingye Li. Learn from syntax: Improving pair-wise aspect and opinion terms extraction with rich syntactic knowledge. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI*, pages 3957–3963, 2021.

[15] Shengqiong Wu, Hao Fei, Yafeng Ren, Bobo Li, Fei Li, and Donghong Ji. High-order pair-wise aspect and opinion terms extraction with edge-enhanced syntactic graph convolution. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2396–2406, 2021.

[16] Hao Fei, Meishan Zhang, Bobo Li, and Donghong Ji. End-to-end semantic role labeling with neural transition-based model. In *Proceedings of the AAAI conference on artificial intelligence, AAAI*, pages 12803–12811, 2021.

[17] Hao Fei, Fei Li, Bobo Li, and Donghong Ji. Encoder-decoder based unified semantic role labeling with label-aware syntax. In *Proceedings of the AAAI conference on artificial intelligence, AAAI*, volume 35, pages 12794–12802, 2021.

[18] Hao Fei, Shengqiong Wu, Yafeng Ren, Fei Li, and Donghong Ji. Better combine them together! integrating syntactic constituency and dependency representations for semantic role labeling. In *Findings of the Association for Computational Linguistics: ACL*, pages 549–559, 2021.

[19] Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. Dual learning for machine translation. In *Proceedings of Annual Conference on Neural Information Processing Systems, NeurIPS*, pages 820–828, 2016.

[20] Hao Fei, Shengqiong Wu, Yafeng Ren, and Meishan Zhang. Matching structure for dual learning. In *Proceedings of the International Conference on Machine Learning, ICML*, pages 6373–6391, 2022.

[21] Bobo Li, Hao Fei, Yafeng Ren, and Donghong Ji. Nominal compound chain extraction: A new task for semantic-enriched lexical chain. In *Proceedings of the Natural Language Processing and Chinese Computing - 9th CCF International Conference, NLPCC*, pages 119–131, 2020.

[22] Donghong Ji, Jun Gao, Hao Fei, Chong Teng, and Yafeng Ren. A deep neural network model for speakers coreference resolution in legal texts. *Inf. Process. Manag.*, 57(6):102365, 2020.

[23] Hao Fei, Yafeng Ren, and Donghong Ji. Mimic and conquer: Heterogeneous tree structure distillation for syntactic NLP. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 183–193, 2020.

[24] Hao Fei, Yafeng Ren, and Donghong Ji. Improving text understanding via deep syntax-semantics communication. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 84–93, 2020.

[25] Hao Fei, Meishan Zhang, Min Zhang, and Tat-Seng Chua. XNLP: an interactive demonstration system for universal structured NLP. *CoRR*, abs/2308.01846.

[26] Bobo Li, Hao Fei, Fei Li, Shengqiong Wu, Lizi Liao, Yinwei Wei, Tat-Seng Chua, and Donghong Ji. Revisiting conversation discourse for dialogue disentanglement. *arXiv preprint arXiv:2306.03975*, 2023.

[27] Hao Fei, Jingye Li, Shengqiong Wu, Chenliang Li, Donghong Ji, and Fei Li. Global inference with explicit syntactic and discourse structures for dialogue-level relation extraction. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, pages 4082–4088, 2022.

[28] Hao Fei, Shengqiong Wu, Meishan Zhang, Yafeng Ren, and Donghong Ji. Conversational semantic role labeling with predicate-oriented latent graph. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, pages 4114–4120, 2022.

[29] Jingye Li, Hao Fei, and Donghong Ji. Modeling local contexts for joint dialogue act recognition and sentiment classification with bi-channel dynamic convolutions. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING*, pages 616–626, 2020.

[30] Bobo Li, Hao Fei, Fei Li, Yuhan Wu, Jinsong Zhang, Shengqiong Wu, Jingye Li, Yijiang Liu, Lizi Liao, Tat-Seng Chua, and Donghong Ji. DiaASQ: A benchmark of conversational aspect-based sentiment quadruple analysis. In *Findings of the Association for Computational Linguistics: ACL*, pages 13449–13467, 2023.

[31] Li Zheng, Donghong Ji, Fei Li, Hao Fei, Shengqiong Wu, Jingye Li, Bobo Li, and Chong Teng. Ecqed: Emotion-cause quadruple extraction in dialogs. *arXiv preprint arXiv:2306.03969*, 2023.

[32] Nan Yu, Meishan Zhang, and Guohong Fu. Transition-based neural RST parsing with implicit syntax features. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING*, pages 559–570, 2018.

[33] Yangfeng Ji and Jacob Eisenstein. Representation learning for text-level discourse parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL*, pages 13–24, 2014.

[34] Yizhong Wang, Sujian Li, and Houfeng Wang. A two-stage parsing method for text-level discourse analysis. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL*, pages 184–188, 2017.

[35] Thanh-Tung Nguyen, Xuan-Phi Nguyen, Shafiq Joty, and Xiaoli Li. RST parsing from scratch. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL*, pages 1613–1625, 2021.

[36] Minghui Xu, Hao Fei, Fei Li, Shengqiong Wu, Rui Sun, Chong Teng, and Donghong Ji. Modeling unified semantic discourse structure for high-quality headline generation. *arXiv preprint arXiv:2309.04534*, 2023.

[37] Jingye Li, Kang Xu, Fei Li, Hao Fei, Yafeng Ren, and Donghong Ji. Mrn: A locally and globally mention-based reasoning network for document-level relation extraction. In *Findings of the Association for Computational Linguistics: ACL*, pages 1359–1370, 2021.

[38] Fengqi Wang, Fei Li, Hao Fei, Jingye Li, Shengqiong Wu, Fangfang Su, Wenxuan Shi, Donghong Ji, and Bo Cai. Entity-centered cross-document relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 9871–9881, 2022.

[39] Hao Fei, Yafeng Ren, Shengqiong Wu, Bobo Li, and Donghong Ji. Latent target-opinion as prior for document-level sentiment classification: A variational approach from fine-grained perspective. In *Proceedings of the web conference, WWW*, pages 553–564, 2021.

[40] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition, CVPR*, pages 3668–3678, 2015.

[41] Yu Zhao, Hao Fei, Bobo Li, Meishan Zhang, Jianguo Wei, Min Zhang, and Tat-Seng Chua. Phylogenetic language acquiring: Grammar induction from visual, speech and text. *arXiv preprint arXiv:2311.05385*, 2023.

[42] Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Chong Teng, Tat-Seng Chua, Donghong Ji, and Fei Li. Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition. In *Proceedings of the 31th ACM International Conference on Multimedia, MM*, 2023.

[43] Shengqiong Wu, Hao Fei, Yixin Cao, Lidong Bing, and Tat-Seng Chua. Information screening whilst exploiting! multimodal relation extraction with feature denoising and multimodal topic modeling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, ACL*, pages 14734–14751, 2023.

[44] Yu Zhao, Hao Fei, Yixin Cao, Bobo Li, Meishan Zhang, Jianguo Wei, Min Zhang, and Tat-Seng Chua. Constructing holistic spatio-temporal scene graph for video semantic role labeling. In *Proceedings of the 31th ACM International Conference on Multimedia, MM*, 2023.

[45] Hao Fei, Qian Liu, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Scene graph as pivoting: Inference-time image-free unsupervised multimodal machine translation with visual scene hallucination. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, ACL*, pages 5980–5994, 2023.

[46] Shengqiong Wu, Hao Fei, Wei Ji, and Tat-Seng Chua. Cross2StrA: Unpaired cross-lingual image captioning with cross-lingual cross-modal structure-pivoted alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, ACL*, pages 2593–2608, 2023.

[47] Yu Zhao, Jianguo Wei, Zhichao Lin, Yueheng Sun, Meishan Zhang, and Min Zhang. Visual spatial description: Controlled spatial-oriented image-to-text generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 1437–1449, 2022.

[48] Yu Zhao, Hao Fei, Wei Ji, Jianguo Wei, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Generating visual spatial description via holistic 3D scene understanding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, ACL*, pages 7960–7977, 2023.

[49] Wei Ji, Li Li, Hao Fei, Xiangyan Liu, Xun Yang, Juncheng Li, and Roger Zimmermann. Towards complex-query referring image segmentation: A novel benchmark. *arXiv preprint arXiv:2309.17205*, 2023.

[50] Wei Ji, Renjie Liang, Lizi Liao, Hao Fei, and Fuli Feng. Partial annotationbased video moment retrieval via iterative learning. In *Proceedings of the 31th ACM international conference on Multimedia, MM*, 2023.

[51] Jian Ma, Zhedong Zheng, Hao Fei, Feng Zheng, Tat-seng Chua, and Yi Yang. Subband-based generative adversarial network for non-parallel many-to-many voice conversion. *arXiv preprint arXiv:2207.06057*, 2022.

[52] Leigang Qu, Shengqiong Wu, Hao Fei, Liqiang Nie, and Tat-Seng Chua. Layoutllm-t2i: Eliciting layout guidance from llm for text-to-image generation. In *Proceedings of the 31th ACM international conference on Multimedia, MM*, 2023.

[53] Shengqiong Wu, Hao Fei, Hanwang Zhang, and Tat-Seng Chua. Imagine that! abstract-to-intricate text-to-image synthesis with scene graph hallucination diffusion. In *Proceedings of Annual Conference on Neural Information Processing Systems, NeurIPS*, 2023.

[54] Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, and Tat-Seng Chua. Empowering dynamics-aware text-to-video diffusion with large language models. *arXiv preprint arXiv:2308.13812*, 2023.

[55] Hao Fei, Yafeng Ren, and Donghong Ji. Retrofitting structure-aware transformer language model for end tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 2151–2161, 2020.

[56] Hao Fei, Shengqiong Wu, Jingye Li, Bobo Li, Fei Li, Libo Qin, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Lasuie: Unifying information extraction with latent adaptive structure-aware generative language model. In *Proceedings of Annual Conference on Neural Information Processing Systems, NeurIPS*, pages 15460–15475, 2022.

[57] Hao Fei, Yafeng Ren, Yue Zhang, Donghong Ji, and Xiaohui Liang. Enriching contextualized language model from knowledge graph for biomedical information extraction. *Briefings in bioinformatics*, 22(3): bbaa110, 2021.

[58] Ao Zhang, Hao Fei, Yuan Yao, Wei Ji, Li Li, Zhiyuan Liu, and Tat-Seng Chua. Transfer visual prompt generator across llms. In *Proceedings of Annual Conference on Neural Information Processing Systems, NeurIPS*, 2023.

[59] Hao Fei, Shengqiong Wu, Meishan Zhang, Shuicheng Yan, Min Zhang, and Tat-Seng Chua. Enhancing video-language representations with structural spatio-temporal alignment. *arXiv preprint arXiv:2310.18212*, 2023.

[60] Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. Ll3da: Visual interactive instruction tuning for omni-3d understanding, reasoning, and planning. *arXiv preprint arXiv:2311.18651*, 2023.

[61] Hao Fei, Shengqiong Wu, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Vitron: A unified pixel-level vision llm for understanding, generating, segmenting, editing, 2024.

[62] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023.

[63] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 3045–3059, 2021.

[64] Kaihang Pan, Juncheng Li, Hongye Song, Hao Fei, Wei Ji, Shuo Zhang, Jun Lin, Xiaozhong Liu, and Siliang Tang. Controlretriever: Harnessing the power of instructions for controllable retrieval. *arXiv preprint arXiv:2308.10025*, 2023.

[65] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. pages 24824–24837, 2022.

[66] Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. Reasoning implicit sentiment with chain-of-thought prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, ACL*, pages 1171–1182, 2023.

[67] Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. Faithful logical reasoning via symbolic chain-of-thought, 2024.

[68] Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Mong-Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition, 2024.