
Universal Visuo-Tactile Video Understanding for Embodied Interaction

Yifan Xie¹, Mingyang Li¹, Shoujie Li¹, Xingting Li¹,
Guangyu Chen², Fei Ma³, Fei Richard Yu³, Wenbo Ding¹

¹ Tsinghua University

² Sun Yat-sen University

³ Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ)

Abstract

Tactile perception is essential for embodied agents to understand physical attributes of objects that cannot be determined through visual inspection alone. While existing approaches have made progress in visual and language modalities for physical understanding, they fail to effectively incorporate tactile information that provides crucial haptic feedback for real-world interaction. In this paper, we present VTV-LLM, the first multi-modal large language model for universal Visuo-Tactile Video (VTV) understanding that bridges the gap between tactile perception and natural language. To address the challenges of cross-sensor and cross-modal integration, we contribute VTV150K, a comprehensive dataset comprising 150,000 video frames from 100 diverse objects captured across three different tactile sensors (GelSight Mini, DIGIT, and Tac3D), annotated with four fundamental tactile attributes (hardness, protrusion, elasticity, and friction). We develop a novel three-stage training paradigm that includes VTV enhancement for robust visuo-tactile representation, VTV-text alignment for cross-modal correspondence, and text prompt finetuning for natural language generation. Our framework enables sophisticated tactile reasoning capabilities including feature assessment, comparative analysis, scenario-based decision making and so on. Experimental evaluations demonstrate that VTV-LLM achieves superior performance in tactile video understanding tasks, establishing a foundation for more intuitive human-machine interaction in tactile domains.

1 Introduction

Touch is a fundamental sensory modality that provides humans with physical information unattainable through vision alone, such as material attributes, surface texture, and compliance. This tactile feedback enables sophisticated physical reasoning and interaction in our environment [1, 2, 3]. While recent advances in vision-language models [4, 5, 6, 7, 8] have demonstrated impressive capabilities in visual reasoning, these models remain fundamentally limited by their inability to perceive tactile attributes, restricting their effectiveness in scenarios requiring physical interaction and reasoning about material characteristics that cannot be reliably inferred from visual cues alone.

Visuo-tactile sensors [9], like GelSight [10], DIGIT [11], and Tac3D [12], have emerged as promising technologies for capturing tactile information, generating image-like representations that encode physical properties such as pressure distribution, surface geometry, and friction characteristics. However, there remains a significant challenge in bridging the domain gap between these tactile

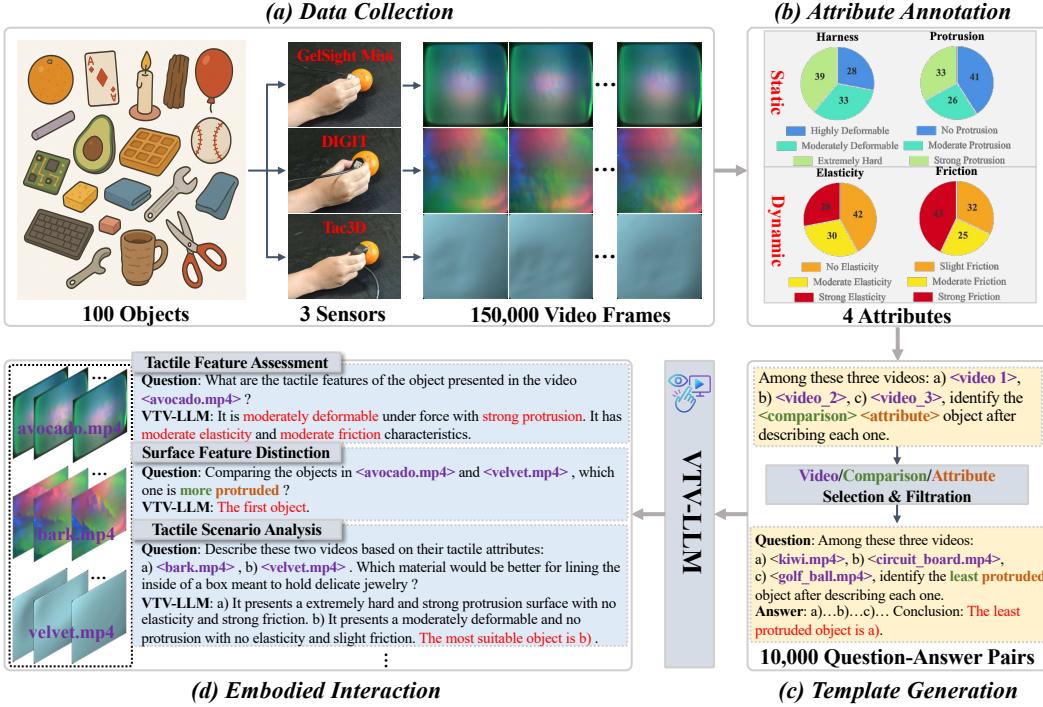


Figure 1: The workflow consists of four key components: (a) Data Collection, which includes 100 diverse objects recorded by 3 different tactile sensors, resulting in 150,000 video frames; (b) Attribute Annotation, where objects are systematically categorized across 4 static and dynamic tactile attributes: hardness, protrusion, elasticity, and friction; (c) Template Generation, which generates 10,000 question-answer pairs using structured templates for various comparative analyses; and (d) Embodied Interaction, demonstrating VTV-LLM’s capability to perform tactile feature assessment, surface feature distinction, tactile scenario analysis and so on. Through this integrated approach, VTV-LLM enables multi-modal reasoning about physical attributes that cannot be determined through visual inspection alone, creating a foundation for more sophisticated human-machine interaction in tactile understanding domains.

representations and natural language understanding. The inherent differences between tactile data captured across various sensor types further complicates this integration, as each sensor produces distinct data formats with varying resolutions and physical property encodings.

Existing research on tactile learning has made progress in representation learning [13, 14, 15, 16, 17, 18], but these approaches often focus either exclusively on static attributes or fail to develop comprehensive frameworks that integrate both tactile perception and language understanding. Most critically, they lack the ability to ground tactile perceptions in natural language descriptions and reasoning, which is essential for human-machine communication about physical properties and interactions [19, 20]. Additionally, the temporal dimension of tactile interactions, which captures how surfaces respond to pressing, sliding, and rotational movements, remains underexplored in current approaches, despite containing crucial information about dynamic material attributes.

To address these challenges, we present VTV-LLM, the first multi-modal large language model for universal visuo-tactile video understanding. Our approach treats tactile perception as a cross-modal reasoning problem, where tactile videos are aligned with linguistic descriptions to enable sophisticated reasoning about physical attributes. As illustrated in Fig. 1(d), VTV-LLM supports a diverse range of embodied interaction capabilities, from basic tactile feature assessment to complex comparative analyses and scenario-based decision making. Additionally, we construct the VTV150K dataset, comprising 150,000 video frames collected from 100 common objects across three different tactile sensors. We systematically annotate these videos with four fundamental tactile attributes (hardness, protrusion, elasticity, and friction), creating a structured foundation for tactile reasoning. To bridge the substantial gap between tactile perception and language understanding, we develop a three-stage training paradigm: (1) VTV enhancement through optical flow-guided masking to learn

robust visuo-tactile representations, (2) VTV-text alignment to establish cross-modal correspondence, and (3) text prompt finetuning to optimize natural language generation about tactile attributes.

Our main contributions can be summarized as follows:

- We introduce VTV-LLM, the first multi-modal large language model capable of universal visuo-tactile video understanding, enabling sophisticated embodied reasoning through natural language interaction.
- We contribute VTV150K, a comprehensive dataset of 150,000 visuo-tactile video frames capturing 100 diverse objects across three tactile sensors, annotated with four fundamental tactile attributes.
- We develop a novel three-stage training paradigm that effectively bridges the domain gap between tactile perception and language understanding, providing a valuable reference for future cross-modal learning efforts.

2 Related Works

Tactile Perception Tactile perception has evolved significantly from early sensors measuring basic physical properties to sophisticated vision-based systems providing high-resolution contact information. Visuo-tactile sensors [9] such as GelSight [10], DIGIT [11], and Tac3D [12] have garnered widespread attention for their ability to capture detailed contact deformations through elastomeric gels and embedded cameras. These sensors have enabled numerous robotic applications including material classification [21], shape reconstruction [22, 23], and dexterous manipulation tasks [24, 14]. Recent research has focused on developing representation learning approaches for tactile data, progressing from task-specific models [25] to general-purpose representations using self-supervised techniques like contrastive multi-view coding [21] and masked autoencoders [26]. The integration of tactile sensing with other modalities has also emerged as a promising direction, with works like UniTouch [17] dynamically fusing tactile signals with visual and audio data to enhance cross-sensor knowledge transferability, Yu et al. [15] aligned tactile images with vision-language models for object property reasoning, and Fu et al. [16] used a touch-vision-language model for open-vocabulary classification. Unlike prior works, our method processes visuo-tactile video directly and focuses on sophisticated tactile reasoning.

Self-Supervised Video Representation Learning Self-supervised video representation learning has emerged as a critical area for developing robust visual features without manual annotations. VideoMAE [27] pioneered this approach by effectively adapting masked autoencoding strategies to the video domain, demonstrating significant performance improvements across various benchmark tasks. Subsequently, VideoMAEv2 [28] enhanced this framework through the introduction of dual masking mechanisms, which substantially improved computational efficiency while maintaining representational power. Recent advancements in this field have focused on sophisticated optimizations along both temporal and spatial dimensions [29, 30, 31, 32], addressing challenges unique to video understanding such as motion coherence and long-range dependencies. In the tactile domain, Sparsh [18] explored the ability of different existing self-supervised learning methods to characterize in tactile video. Feng et al. [13] utilized the tube masking strategy to process the tactile video. Our method builds upon these foundations by introducing optical flow-guided masking specifically designed for visuo-tactile videos, which addresses the unique challenges of capturing both spatial deformation and temporal dynamics in tactile interactions.

Multi-Modal Large Language Models Multimodal Large Language Models (MLLMs) have transformed AI research by enabling reasoning across textual and visual modalities. Early efforts integrated LLMs as agents for downstream tasks [33, 34, 35]. Later approaches focused on parameter-efficient tuning [36, 37] and instruction tuning [38, 39] to align visual semantics with language. Recent advances have incorporated video processing [40, 41] and diverse sensory inputs [42], enabling applications in robotics [43, 44]. In our work, we present the first visuo-tactile video large language model to bridge the gap between tactile perception and natural language.

3 Methods

In this section, we first introduce VTV150K, a large-scale dataset of video-question-answer pairs in Sec. 3.1. Subsequently, we present VTV-LLM, the first visuo-tactile video large language model designed for visuo-tactile video understanding and embodied interaction in Sec. 3.2.

3.1 VTV150K

Overview Visuo-tactile sensor technologies suffer from inadequate standardization and significant cross-sensor data discrepancies, which substantially impede the transferability of tactile representation models across different sensing platforms. Existing methods [14, 18, 45, 13] addressing these challenges exhibit notable limitations, as they either neglect the integration of both static and dynamic tactile attributes or fail to incorporate comprehensive visuo-tactile video understanding for embodied interaction.

In this work, we introduce VTV150K, a comprehensive large-scale dataset comprising video-question-answer pairs collected across three diverse visuo-tactile sensors, as illustrated in Fig. 1(a-c). The dataset construction methodology encompasses three sequential stages: data collection, attribute annotation, and template generation. We will delve into the specifics of these stages.

Data Collection To facilitate the grounding of embodied interaction on tactile inputs, we collected a comprehensive dataset comprising 100 common objects, yielding a total of 150,000 visuo-tactile video frames. As illustrated in Fig. 1(a), we employed multiple visuo-tactile sensors to ensure style diversity: GelSight mini [10] and DIGIT [11] sensors for capturing high-resolution visuo-tactile information, and Tac3D [12] for measuring deformation force fields. Due to the relatively low resolution of Tac3D, we implemented the cubic spline interpolation algorithm [46] to reconstruct more detailed force field representations.

Data collection was performed manually to address the challenges associated with properly interacting with irregularly shaped objects. For each object, we systematically captured five visuo-tactile videos across different regions using various sensors. Our data collection process consisted of three sequential interactions: (1) normal pressing against the object surface to capture pressure distribution, (2) rotational movement to acquire shear information, and (3) sliding motion to obtain friction characteristics. This multi-interaction approach enables comprehensive tactile information extraction for embodied interaction.

Attribute Annotation To facilitate tactile reasoning, we annotated our dataset across four fundamental static and dynamic tactile attributes as shown in Fig. 1(b). Each attribute was categorized into three distinct levels, with harness classified as highly deformable (28%), moderately deformable (33%), and extremely hard (39%); protrusion categorized as absent (41%), moderate (26%), or strong (33%); elasticity measured as none (42%), moderate (30%), or strong (28%); and friction assessed as slight (32%), moderate (25%), or strong (43%). This structured annotation framework enables comprehensive tactile attribute analysis for downstream reasoning tasks.

Template Generation Template generation facilitates the creation of question-answer pairs for model training. We developed multiple problem templates encompassing various reasoning tasks: tactile feature assessment, surface feature distinction, texture optimal selection and so on. To instantiate these templates, we systematically integrated diverse visuo-tactile video combinations, comparison operators (*e.g.*, "more", "less", "most", "least"), and attribute selectors to generate a comprehensive dataset of 10,000 question-answer pairs. As illustrated in Fig. 1(c), our generation process follows a hierarchical framework: selection, filtration, and structured question formulation with corresponding ground-truth annotations. For more comprehensive details about attribute annotation and template generation, please refer to the Supplementary Material.

3.2 VTV-LLM

Overview VTV-LLM aims to serve as a multi-modal framework capable of integrating visual-tactile video data with large language models to facilitate tactile reasoning for embodied interaction. As illustrated in Fig. 2(a), VTV-LLM formulates tactile perception as a cross-modal approach to question answering and descriptive generation. By leveraging the rich sensory information inherent

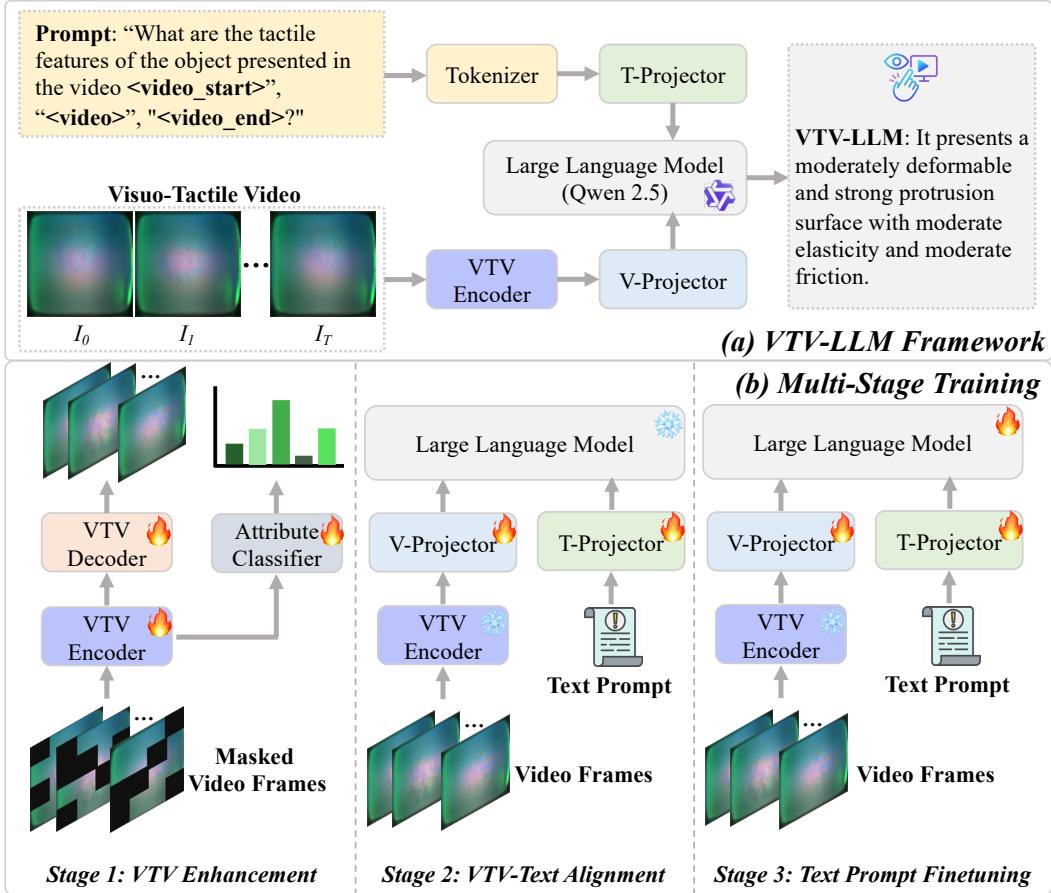


Figure 2: (a) VTV-LLM framework: A multi-modal system integrating visual-tactile video data with large language models to facilitate tactile reasoning for embodied interaction; (b) Multi-Stage Training: It consists of VTV enhancement, alignment between visuo-tactile video and text, and prompt-based finetuning to generate accurate tactile descriptions.

in visuo-tactile video data, VTV-LLM enhances understanding in scenarios traditionally challenging for standard vision-only models, particularly in applications requiring tactile attribute inference.

At the core of VTV-LLM lies a (Qwen 2.5 [4, 5]) that synthesizes complex multi-modal information from visuo-tactile videos, utilizing world knowledge to generate coherent, human-readable descriptions of tactile attributes. In general, a visuo-tactile video can be mathematically represented as a sequence of frames $\mathcal{V} = \{I_t\}_{t=0}^T$, where each frame I_t captures both visual and tactile information at timestamp t . Initially, high-dimensional features F_{VTV} are extracted from \mathcal{V} using a VTV encoder based on ViT-base architecture [47] adapted from VideoMAE [27, 28]:

$$F_{VTV} = f_{enc}(V) = \text{ViT} \left(\{\text{Patch}(I_t) + \text{TE}(t)\}_{t=0}^T \right), \quad (1)$$

where $\text{Patch}(\cdot)$ denotes the patch embedding operation and $\text{TE}(t)$ represents temporal embeddings. These features are then processed through a visual projector f_{V-proj} consisting of two linear layers with a GELU activation function [48] in between to produce the visual embedding E_V :

$$E_V = f_{V-proj}(F_{VTV}) = W_2 \cdot \text{GELU}(W_1 \cdot F_{VTV} + b_1) + b_2, \quad (2)$$

where W_1, W_2 are learnable weight matrices and b_1, b_2 are bias terms. Concurrently, the textual prompt is tokenized and processed through LLM's text projector to produce text embedding E_T . For effective multi-modal reasoning, we introduce special tokens <video_start>, <video>, and <video_end> to denote the beginning, content and end of the visuo-tactile video in the input sequence. These tokens serve as anchors for the model to properly align visual information with textual understanding during the inference process.

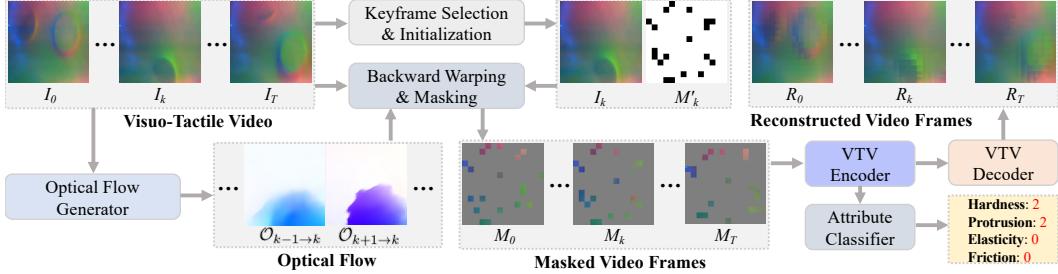


Figure 3: Training pipeline of VTV enhancement.

Given these aligned representations, the large language model f_{LLM} performs reasoning to generate a response A describing tactile attributes:

$$A = f_{LLM}(E_V, E_T) = \text{Qwen}(\text{Concat}([E_V; E_T])). \quad (3)$$

Given the complexity of integrating visuo-tactile information with language representations, we implement a staged training approach to develop our framework. As shown in Fig. 2(b), VTV-LLM adopts a three-stage training paradigm encompassing VTV enhancement, VTV-text alignment, and text prompt finetuning. This structured progression enables the model to first learn robust tactile-visual representations, then align them with textual descriptions, and finally optimize response generation, enhancing VTV-LLM’s capability for cross-modal understanding and tactile reasoning. In the following, we describe each of these stages in detail.

VTV Enhancement Existing multi-modal LLMs predominantly process natural images via unmodified Vision Transformer (ViT) encoders [47]. However, our research addresses visuo-tactile inputs, which exhibit fundamentally different characteristics from natural images, thus necessitating specialized fine-tuning to extract meaningful representations.

Furthermore, the temporal nature of our video data introduces challenges not present in static images. Unlike images, videos possess an inherent time dimension characterized by temporal redundancy and inter-frame correlations, requiring robust video representation methodologies. While VideoMAE [27, 28] offers a powerful masked video autoencoder with an asymmetric encoder-decoder architecture utilizing tube masking, this approach assumes minimal motion across large frame regions. This assumption proves problematic for visuo-tactile videos, which typically exhibit significant motion patterns. Direct application of tube masking to such inputs risks substantial information leakage, wherein the model can trivially reconstruct masked segments using visible tokens from temporally adjacent frames, which is a critical concern in masked video pre-training. To address these limitations, we propose a novel training pipeline specifically designed for visuo-tactile video representation, as illustrated in Fig. 3.

Given the visuo-tactile video sequence $\mathcal{V} = \{I_t \in \mathbb{R}^{H \times W \times C}\}_{t=0}^T$, where each frame I_t encodes both visual and tactile information at timestamp t with spatial dimensions $H \times W$ and C channels, we propose selecting the middle frame as the keyframe. This selection is motivated by empirical observations that the middle frame typically exhibits the maximum contact surface area, facilitating more robust optical flow warping in subsequent processing stages. For keyframe mask initialization, conventional binarization approaches [49] significantly degrade the spatial continuity of object surfaces, compromising the fidelity of the reconstructed tactile information. Therefore, we introduce a Gaussian mixture model [50] to obtain the keyframe mask. For the keyframe I_k , we formulate a probabilistic mask using localized Gaussian functions. We select a set of $N = \lceil \alpha \cdot HW / \beta^2 \rceil$ sampling points $\{p_i\}_{i=1}^N$ distributed across the frame, where $\alpha \in (0, 1)$ controls density and β is the sampling grid size. Each point p_i generates a Gaussian kernel $G_i(x, y) = \exp\left(-\frac{(x-p_{ix})^2 + (y-p_{iy})^2}{2\lambda^2}\right)$ with scale parameter λ . The final keyframe mask is defined as $M'_k = \min\left(1, \sum_{i=1}^N G_i\right)$, creating a continuous-valued mask that preserves spatial structure while enabling controlled sparsity for subsequent processing.

Additionally, we employ dense motion estimation across the visuo-tactile video \mathcal{V} using the RAFT architecture [51]. We compute bidirectional optical flow fields between consecutive frames to capture

the continuous deformation patterns throughout the interaction process. For each adjacent frame pair, we define the forward flow field $\mathcal{O}_{t \rightarrow t+1} = \text{RAFT}(I_t, I_{t+1})$. Each flow field $\mathcal{O}_{t \rightarrow t+1} \in \mathbb{R}^{H \times W \times 2}$ encodes pixel-wise displacement vectors $(u_{x,y}, v_{x,y})$ for every spatial location (x, y) , mapping positions from frame I_t to their corresponding locations in frame I_{t+1} . The complete set of optical flows Φ for the sequence is formulated as:

$$\Phi = \bigcup_{t=0}^{k-1} \{\mathcal{O}_{t \rightarrow t+1}\} \cup \bigcup_{t=k+1}^T \{\mathcal{O}_{t \rightarrow t-1}\}. \quad (4)$$

This bidirectional flow representation tracks visuo-tactile features throughout the interaction, supporting warping operations and masked frame generation. We apply spatial normalization before flow computation to ensure scale invariance across different sequences.

After that, we utilize the backward warping [52, 53] to generate the temporal consistent masking map based on the keyframe and mask the corresponding video frames. The masked visuo-tactile frames $\mathcal{V}_m = \{M_t\}_{t=0}^T$ are fed into the VTV encoder-decoder architecture for reconstruction using the mean squared error loss [27, 28]. We also incorporate an attribute classifier to predict tactile attributes (hardness, protrusion, elasticity, and friction) using the cross-entropy loss [54]. Our total loss function combines both the reconstruction loss and the attribute classification loss, enabling simultaneous optimization of visuo-tactile reconstruction quality and tactile attribute classification accuracy.

VTV-Text Alignment In the VTV-text alignment stage, we focus on establishing cross-modal alignment between video and language representations. With the pretrained VTV Encoder from stage 1, we introduce both V-Projector and T-Projector modules while keeping the Large Language Model frozen. This stage leverages our initial constructed VTV150K dataset to bridge the representational gap between visual and textual modalities. The V-Projector maps video embeddings from the VTV Encoder into the language model’s embedding space, while the T-Projector processes corresponding text prompt representations. By training these projection modules exclusively while freezing other components, we establish foundational cross-modal understanding, enabling the model to associate visual content with appropriate textual descriptions. This alignment is critical for downstream video understanding and description tasks as it creates a shared semantic space between the video frames and natural language.

Text Prompt Finetuning In the text prompt finetuning stage, we enhance the model’s capacity to respond accurately to textual prompts about video content by implementing supervised fine-tuning across multiple components. The V-Projector, and T-Projector are jointly fine-tuned along with the LLM. Unlike previous stages where the LLM remained frozen, this stage employs parameter-efficient techniques [37, 36] to fine-tune the language model using 10,000 newly generated question-answer pairs. These pairs are created using the same template generation approach as our VTV150K dataset, featuring diverse video understanding tasks. By generating new data rather than reusing subsets, we significantly increase training diversity and model robustness. This end-to-end optimization enables the model to generate more coherent, accurate, and contextually relevant responses to text prompts about video content. The supervised nature of this phase significantly improves the model’s ability to comprehend complex video scenes and produce natural language descriptions that align with human expectations. This final stage integrates the previously aligned representations into a cohesive multi-modal understanding system, culminating in enhanced video-language capabilities.

4 Experiments

4.1 Setup

Our experiments utilize the proposed VTV150K dataset for both training and evaluation protocols. The training process follows our three-stage paradigm: Stage 1 employs multi-sensor visuo-tactile videos with their corresponding attribute annotations for representation learning. For Stage 2 and 3, we utilize two independently generated sets of 10,000 question-answer pairs to prevent data leakage between stages. To evaluate model performance, we create a separate test set comprising 600 question-answer pairs for novel objects not present in the training data, ensuring comprehensive coverage across various tactile reasoning tasks. Our LLM backbone is based on Qwen 2.5 [4, 5], experimenting with three model variants (3B, 7B, and 14B parameters). All experiments are conducted on 4 NVIDIA

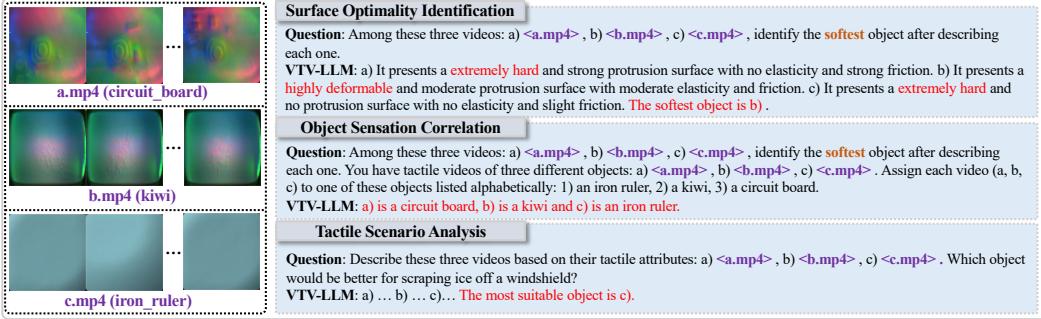


Figure 4: Several task examples from the proposed VTV150K along with predictions from VTV-LLM.

Table 1: Performance comparison of VTV-LLM-7B against seven state-of-the-art methods on the VTV150K dataset. The evaluation covers different tasks, with results reported in percentages (%) and the boldface indicates the best performance.

Models	Hardness	Protrusion	Elasticity	Friction	Combined	SFD	SOI	OSC	TSA	Average
GPT-4o [55]	34.7	32.6	32.6	18.7	2.1	40.9	38.4	16.6	36.0	28.0
Gemini-2.5-Pro-Exp [56]	36.2	34.7	39.1	21.0	4.3	42.6	29.4	18.5	40.0	29.5
LLaVA-OneVision-7B [57]	27.5	32.6	26.0	20.2	0.7	40.9	28.2	11.7	30.0	24.2
LLaVA-Video-Qwen2-7B [58]	30.4	29.7	28.9	18.1	2.1	33.6	29.4	17.2	36.0	25.0
InternVL2.5-VL-8B [59]	18.1	23.9	21.0	13.7	0.0	24.5	17.9	11.1	24.0	17.1
VideoLLaMA3-7B [41]	15.2	21.7	14.4	10.8	0.0	11.4	12.8	7.4	20.0	12.6
Qwen2.5-VL-7B [60]	25.3	28.9	17.3	15.9	1.4	22.9	28.2	16.0	30.0	20.6
VTV-LLM-7B (Ours)	73.9	75.0	67.3	56.5	35.6	71.3	57.6	43.2	64.0	60.4

RTX 6000 Ada GPUs. Additional implementation details and hyperparameter configurations are provided in the Supplementary Material.

4.2 Results

To verify the effectiveness of our VTV-LLM, we compare it against two strong proprietary models, such as GPT-4o [55] and Gemini-2.5-Pro-Exp [56], as well as five open-source video-based VLMs, including LLaVA-OneVision-7B [57], LLaVA-Video-Qwen2-7B [58], InternVL2.5-VL-8B [59], VideoLLaMA3-7B [41] and Qwen2.5-VL-7B [60]. Since most of the video-based VLM models have parameters around 7B, we only use the VTV-LLM-7B model for fair comparison. To guarantee the robustness of the experimental results, we report the average results of the triplicate test with random seeds.

Our first experiment focuses on tactile feature assessment, which evaluates the model’s ability to perceive and describe physical sensory attributes of objects in visuo-tactile videos. As illustrated in Fig. 1(d), when presented with a visuo-tactile video and a question prompt, VTV-LLM generates descriptions of the four key tactile attributes. The results presented in Tab. 1 demonstrate that our method consistently outperforms all baseline models across both individual attribute and combined attribute settings. The performance gap is particularly notable in the combined attribute setting, which we attribute to our three-stage training paradigm that effectively bridges the domain gap between tactile perception and natural language understanding.

In addition, we conduct high-level tactile reasoning experiments, including surface feature distinction (SFD), surface optimality identification (SOI), object sensation correlation (OSC), and tactile scenario analysis (TSA). SFD involves comparing tactile qualities between objects to determine relative differences, SOI entails analyzing multiple surfaces to determine which exhibits the highest degree of a particular quality, OSC aims at relating tactile perceptual information to the identity of a particular real-world object, and TSA addresses applying haptic knowledge to real-world situations that require physical reasoning. It is worth noting that the TSA task is not included in the training set. The qualitative results presented in Fig. 1(d) and Fig. 4 demonstrate that VTV-LLM can generate reasonable outputs. The quantitative experimental results in Tab. 1 further confirm that VTV-LLM

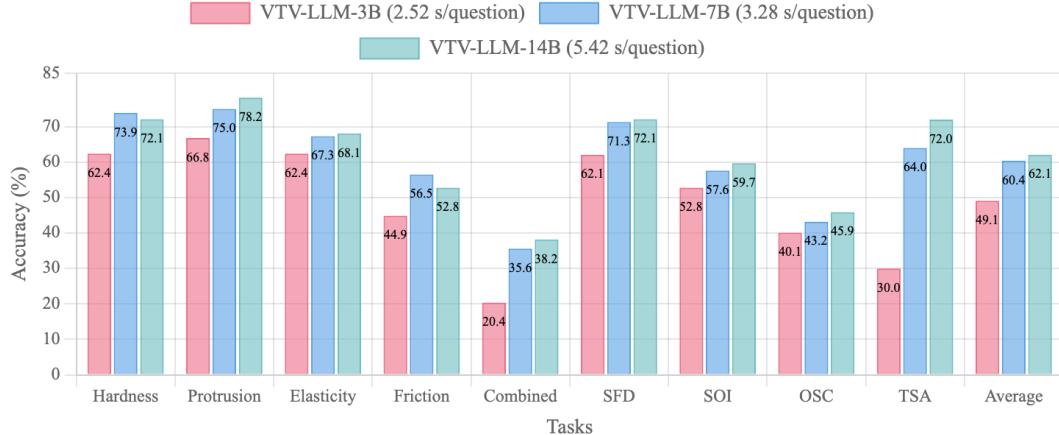


Figure 5: Performance comparison of VTV-LLM on the different parameters.

Table 2: Ablation study on VTV encoder settings using the VTV-LLM-7B model.

Settings	SFD	SOI	OSC	TSA	Average
VideoMAE (w/o train)	37.5	29.7	8.5	16.0	22.9
VideoMAE (w/ train)	52.4	46.1	28.3	38.0	41.2
Ours (w/o cls)	62.2	48.7	40.1	55.0	51.5
Ours	71.3	57.6	43.2	64.0	59.0

Table 3: Ablation study on three-stage training paradigm settings using the VTV-LLM-7B model.

Settings	SFD	SOI	OSC	TSA	Average
w/o stage 2	58.1	50.0	35.2	60.0	50.8
w/o stage 3	50.8	42.3	29.0	52.0	43.5
Same dataset	61.4	53.8	33.9	58.0	51.7
Ours	71.3	57.6	43.2	64.0	59.0

achieves superior performance across these complex reasoning tasks, highlighting its potential for embodied interaction.

4.3 Ablation Studies

LLM Backbone To examine the effect of model scale on visuo-tactile understanding, we compare different parameter sizes of our LLM backbone. Fig. 5 shows performance results for VTV-LLM using three Qwen 2.5 variants (3B, 7B, and 14B parameters). We observe consistent performance improvements with increasing model size. This improvement is most significant for complex reasoning tasks like TSA, indicating larger models better integrate cross-modal information. However, larger models also require substantially more computation time during inference.

VTV Encoder We conduct an ablation study on our VTV encoder design as shown in Tab. 2. Baseline VideoMAE [27, 28] without training achieves only 22.9% average performance, while training with our VTV150K dataset improves it to 41.2%. Our method without the attribute classifier reaches 51.5%, showing the effectiveness of our optical flow-guided masking strategy. The full method with the attribute classifier further improves to 59.0%, confirming that joint reconstruction and attribute classification significantly enhances tactile understanding.

Three-Stage Training Paradigm Tab. 3 validates our three-stage training paradigm through ablation studies. Removing stage 2 (VTV-text alignment) drops average performance to 50.8%, while omitting stage 3 (text prompt finetuning) causes a steeper decline to 43.5%. Using identical datasets across stages also underperforms at 51.7%, confirming that independent datasets for each stage significantly improve model robustness.

5 Conclusion

In this work, we presented VTV-LLM, the first multi-modal large language model for universal visuo-tactile video understanding. We contributed VTV150K, a comprehensive dataset of visuo-tactile videos across multiple sensors, and developed a novel three-stage training paradigm that effectively bridges the gap between tactile perception and natural language. Experimental results demonstrate

that VTV-LLM consistently outperforms state-of-the-art methods across various tactile reasoning tasks, establishing a foundation for more intuitive human-machine interaction in embodied domains.

References

- [1] Qiang Li, Oliver Kroemer, Zhe Su, Filipe Fernandes Veiga, Mohsen Kaboli, and Helge Joachim Ritter. A review of tactile information: Perception and action through touch. *IEEE Transactions on Robotics*, 36(6):1619–1634, 2020.
- [2] Müge Cavidan, Katja Doerschner, and Knut Drewing. Task and material properties interactively affect softness explorations along different dimensions. *IEEE Transactions on Haptics*, 14(3):603–614, 2021.
- [3] Mudassir Ibrahim Awan, Waseem Hassan, and Seokhee Jeon. Predicting perceptual haptic attributes of textured surface from tactile data based on deep cnn-lstm network. In *Proceedings of the 29th ACM Symposium on Virtual Reality Software and Technology*, pages 1–9, 2023.
- [4] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.
- [5] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.
- [6] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- [7] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465, 2024.
- [8] Yifan Xie, Jingge Wang, Tao Feng, Fei Ma, and Yang Li. Ccis-diff: A generative model with stable diffusion prior for controlled colonoscopy image synthesis. *arXiv preprint arXiv:2411.12198*, 2024.
- [9] Shoujie Li, Zihan Wang, Changsheng Wu, Xiang Li, Shan Luo, Bin Fang, Fuchun Sun, Xiao-Ping Zhang, and Wenbo Ding. When vision meets touch: A contemporary review for visuotactile sensors from the signal processing perspective. *IEEE Journal of Selected Topics in Signal Processing*, 2024.
- [10] Wenzhen Yuan, Siyuan Dong, and Edward H Adelson. Gelsight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors*, 17(12):2762, 2017.
- [11] Mike Lambeta, Po-Wei Chou, Stephen Tian, Brian Yang, Benjamin Maloon, Victoria Rose Most, Dave Stroud, Raymond Santos, Ahmad Byagowi, Gregg Kammerer, et al. Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. *IEEE Robotics and Automation Letters*, 5(3):3838–3845, 2020.
- [12] Lunwei Zhang, Yue Wang, and Yao Jiang. Tac3d: A novel vision-based tactile sensor for measuring forces distribution and estimating friction coefficient distribution. *arXiv preprint arXiv:2202.06211*, 2022.
- [13] Ruoxuan Feng, Jiangyu Hu, Wenke Xia, Ao Shen, Yuhao Sun, Bin Fang, Di Hu, et al. Anytouch: Learning unified static-dynamic representation across multiple visuo-tactile sensors. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [14] Jialiang Zhao, Yuxiang Ma, Lirui Wang, and Edward Adelson. Transferable tactile transformers for representation learning across diverse sensors and tasks. In *8th Annual Conference on Robot Learning*, 2024.
- [15] Samson Yu, Kelvin Lin, Anxing Xiao, Jiafei Duan, and Harold Soh. Octopi: Object property reasoning with large tactile-language models. *arXiv preprint arXiv:2405.02794*, 2024.

- [16] Letian Fu, Gaurav Datta, Huang Huang, William Chung-Ho Panitch, Jaimyn Drake, Joseph Ortiz, Mustafa Mukadam, Mike Lambeta, Roberto Calandra, and Ken Goldberg. A touch, vision, and language dataset for multimodal alignment. In *International Conference on Machine Learning*, pages 14080–14101. PMLR, 2024.
- [17] Fengyu Yang, Chao Feng, Ziyang Chen, Hyoungseob Park, Daniel Wang, Yiming Dou, Ziyao Zeng, Xien Chen, Rit Gangopadhyay, Andrew Owens, et al. Binding touch to everything: Learning unified multimodal tactile representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26340–26353, 2024.
- [18] Carolina Higuera, Akash Sharma, Chaithanya Krishna Bodduluri, Taosha Fan, Patrick Lancaster, Mrinal Kalakrishnan, Michael Kaess, Byron Boots, Mike Lambeta, Tingfan Wu, et al. Sparsh: Self-supervised touch representations for vision-based tactile sensing. In *8th Annual Conference on Robot Learning*, 2024.
- [19] Neha Sunil, Shaoxiong Wang, Yu She, Edward Adelson, and Alberto Rodriguez Garcia. Visuo-tactile affordances for cloth manipulation with local control. In *Conference on Robot Learning*, pages 1596–1606. PMLR, 2023.
- [20] Ying Zheng, Lei Yao, Yuejiao Su, Yi Zhang, Yi Wang, Sicheng Zhao, Yiyi Zhang, and Lap-Pui Chau. A survey of embodied learning for object-centric robotic manipulation. *arXiv preprint arXiv:2408.11537*, 2024.
- [21] Fengyu Yang, Chenyang Ma, Jiacheng Zhang, Jing Zhu, Wenzhen Yuan, and Andrew Owens. Touch and go: learning from human-collected vision and touch. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 8081–8103, 2022.
- [22] Ruohan Gao, Yen-Yu Chang, Shivani Mall, Li Fei-Fei, and Jiajun Wu. Objectfolder: A dataset of objects with implicit visual, auditory, and tactile representations. In *Conference on Robot Learning*, 2021.
- [23] Ruohan Gao, Zilin Si, Yen-Yu Chang, Samuel Clarke, Jeannette Bohg, Li Fei-Fei, Wenzhen Yuan, and Jiajun Wu. Objectfolder 2.0: A multisensory object dataset for sim2real transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10598–10608, 2022.
- [24] Hao Li, Yizhi Zhang, Junzhe Zhu, Shaoxiong Wang, Michelle A Lee, Huazhe Xu, Edward Adelson, Li Fei-Fei, Ruohan Gao, and Jiajun Wu. See, hear, and feel: Smart sensory fusion for robotic manipulation. In *Conference on Robot Learning*, pages 1368–1378. PMLR, 2023.
- [25] Wenzhen Yuan, Yuchen Mo, Shaoxiong Wang, and Edward H Adelson. Active clothing material perception using tactile sensing and deep learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4842–4849. IEEE, 2018.
- [26] Justin Kerr, Huang Huang, Albert Wilcox, Ryan Hoque, Jeffrey Ichnowski, Roberto Calandra, and Ken Goldberg. Self-supervised visuo-tactile pretraining to locate and follow garment features. *arXiv preprint arXiv:2209.13042*, 2022.
- [27] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.
- [28] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14549–14560, 2023.
- [29] Mohammadreza Salehi, Michael Dorkenwald, Fida Mohammad Thoker, Efstratios Gavves, Cees GM Snoek, and Yuki M Asano. Sigma: Sinkhorn-guided masked video modeling. In *European Conference on Computer Vision*, pages 293–312. Springer, 2024.
- [30] Bingkun Huang, Zhiyu Zhao, Guozhen Zhang, Yu Qiao, and Limin Wang. Mgmae: Motion guided masking for video masked autoencoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13493–13504, 2023.

- [31] Yunze Liu, Peiran Wu, Cheng Liang, Junxiao Shen, Limin Wang, and Li Yi. Videomap: Toward scalable mamba-based video autoregressive pretraining. *arXiv preprint arXiv:2503.12332*, 2025.
- [32] Gensheng Pei, Tao Chen, Xiruo Jiang, Huafeng Liu, Zeren Sun, and Yazhou Yao. Videomac: Video masked autoencoders meet convnets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22733–22743, 2024.
- [33] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueling Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36:38154–38180, 2023.
- [34] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11888–11898, 2023.
- [35] Rui Yang, Lin Song, Yanwei Li, Sijie Zhao, Yixiao Ge, Xiu Li, and Ying Shan. Gpt4tools: Teaching large language model to use tools via self-instruction. *Advances in Neural Information Processing Systems*, 36:71995–72007, 2023.
- [36] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.
- [37] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *ICLR*, 2024.
- [38] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.
- [39] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.
- [40] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12585–12602, 2024.
- [41] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.
- [42] Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. Llama-omni: Seamless speech interaction with large language models. *arXiv preprint arXiv:2409.06666*, 2024.
- [43] Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. *arXiv preprint arXiv:2410.07864*, 2024.
- [44] Joshua Jones, Oier Mees, Carmelo Sferrazza, Kyle Stachowicz, Pieter Abbeel, and Sergey Levine. Beyond sight: Finetuning generalist robot policies with heterogeneous sensors via language grounding. *arXiv preprint arXiv:2501.04693*, 2025.
- [45] Samanta Rodriguez, Yiming Dou, Miquel Oller, Andrew Owens, and Nima Fazeli. Touch2touch: Cross-modal tactile generation for object manipulation. *arXiv preprint arXiv:2409.08269*, 2024.
- [46] Sky McKinley and Megan Levine. Cubic spline interpolation. *College of the Redwoods*, 45(1):1049–1060, 1998.
- [47] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, G Heigold, S Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

- [48] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.
- [49] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.
- [50] Douglas A Reynolds et al. Gaussian mixture models. *Encyclopedia of biometrics*, 741(659-663):3, 2009.
- [51] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020.
- [52] Joi Shimizu, Heming Sun, and Jiro Katto. Forward and backward warping for optical flow-based frame interpolation. In *2022 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pages 082–086. IEEE, 2022.
- [53] Sung In Cho and Suk-Ju Kang. Extrapolation-based video retargeting with backward warping using an image-to-warping vector generation network. *IEEE Signal Processing Letters*, 27:446–450, 2020.
- [54] Anqi Mao, Mehryar Mohri, and Yutao Zhong. Cross-entropy loss functions: Theoretical analysis and applications. In *International conference on Machine learning*, pages 23803–23828. PMLR, 2023.
- [55] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- [56] Google DeepMind. gemini-2.5-pro-preview-05-06. <https://ai.google.dev/gemini-api/docs/models#gemini-2.5-pro-preview-05-06>, 2025.
- [57] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.
- [58] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.
- [59] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, ZhaoYang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.
- [60] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.