MOLVISION: MOLECULAR PROPERTY PREDICTION WITH VISION LANGUAGE MODELS

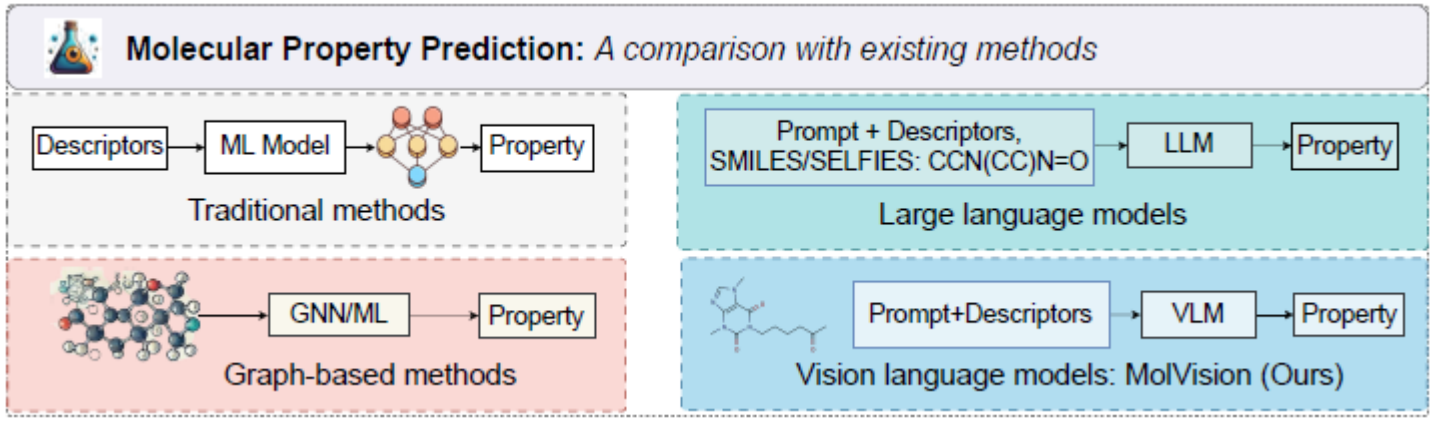
Deepan Adak[1], Yogesh Singh Rawat[2], Shruti Vyas[2]

NIT Kurukshetra[1], University of Central Florida[2]

MOTIVATION

WHY VISION FOR MOLECULES?

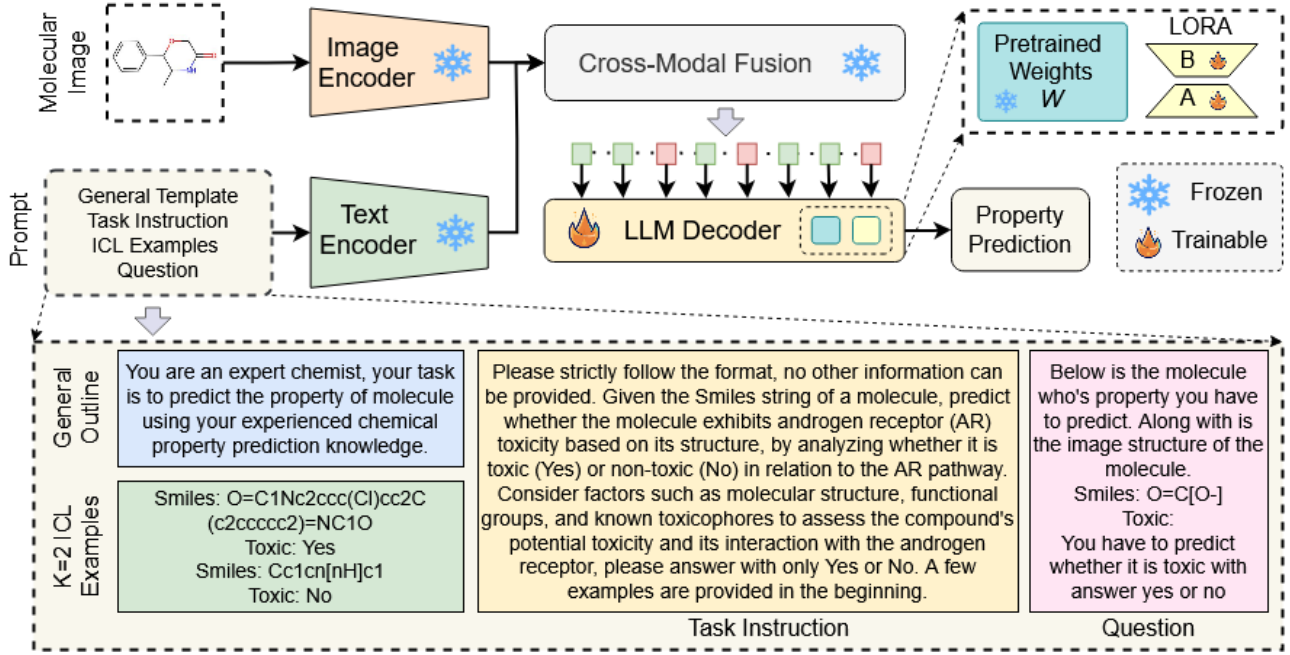
- **Problem:** Current LLM approaches rely only on SMILES/SELFIES text Ambiguous: Same molecule → different SMILES strings
  - Loses structural information
- Key Insight: Chemists analyze molecules visually
  - Stereochemistry, geometry, electron delocalization
- Our Approach: Leverage Vision-Language Models (VLMs)



THE MOLVISION FRAMEWORK

MULTIMODAL MOLECULAR PROPERTY PREDICTION

- **Key Components:**
  - Visual: Skeletal molecular structures
  - Textual: SMILES + task descriptions
  - Efficient adaptation via LoRA



MOLVISION BENCHMARK

COMPREHENSIVE EVALUATION ACROSS 10 DATASETS

Task Type	Datasets	Properties
Classification	BACE-V, BBBP-V, HIV-V, ClinTox-V, Tox21-V	Bioactivity, Toxicity, BBB Penetration
Regression	ESOL-V, LD50-V, QM9-V, PCQM4Mv2-V	Solubility, Toxicity, Quantum Properties
Molecular Description	ChEBI-V	Molecular Descriptions

EXPERIMENTAL SETUP

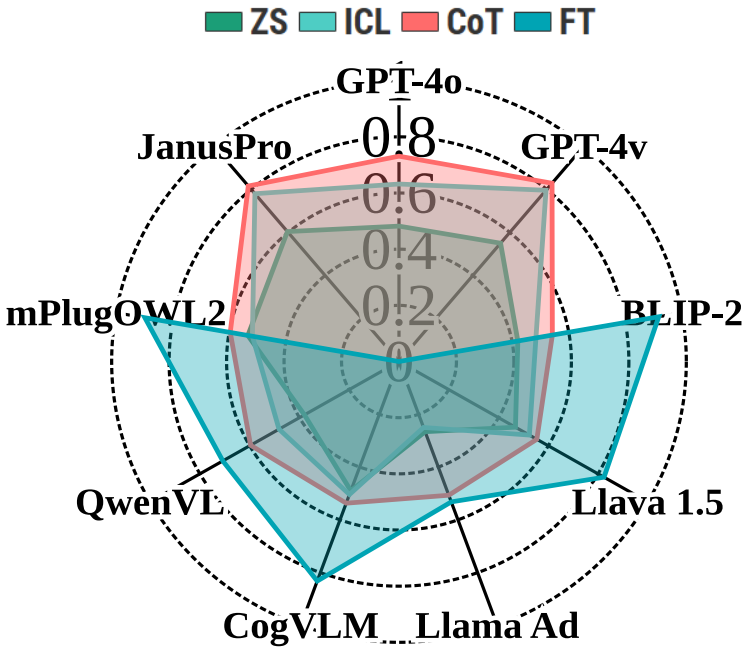
- Evaluation Setting
  - Zero-Shot
  - Few-Shot
  - Chain of Thought
  - Fine Tuning
- Models Evaluated
  - Open-Sourced - Janus-Pro, BLIP-2, LLaVA, CogVLM, QwenVL, mPlugOWL2, Llama Adapter v2
  - Closed-Sourced – GPT-4o, GPT-4v

Ranking Table

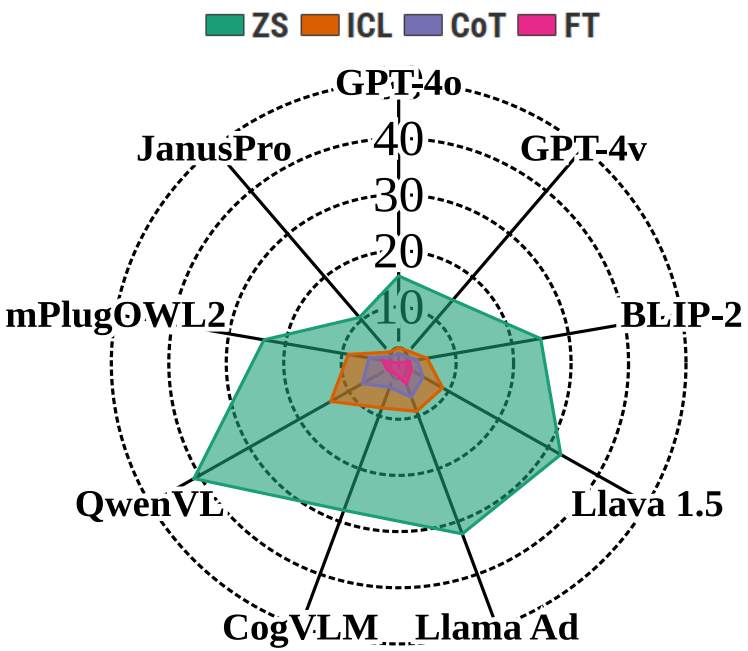
Models	Zero	ICL	CoT	LoRA
GPT-4v	2/1/2	1/3/1	1/1/2	-
GPT-4o	5/3/1	3/2/2	3/2/1	-
JanusPro 7B	1/2/3	2/1/3	2/3/3	-
BLIP-2	7/5/8	7/4/6	7/3/7	1/1/3
Llava1.5 13B	4/8/4	4/6/7	6/6/4	4/4/2
Llama v2 7B	9/7/9	9/8/9	9/8/9	6/6/6
CogVLM	4/6/5	6/5/4	8/5/5	3/2/1
Qwen VL	8/9/7	8/9/5	5/9/6	5/3/4
mPlugOWL2	3/4/6	5/7/8	4/7/8	2/5/5

KEY RESULTS

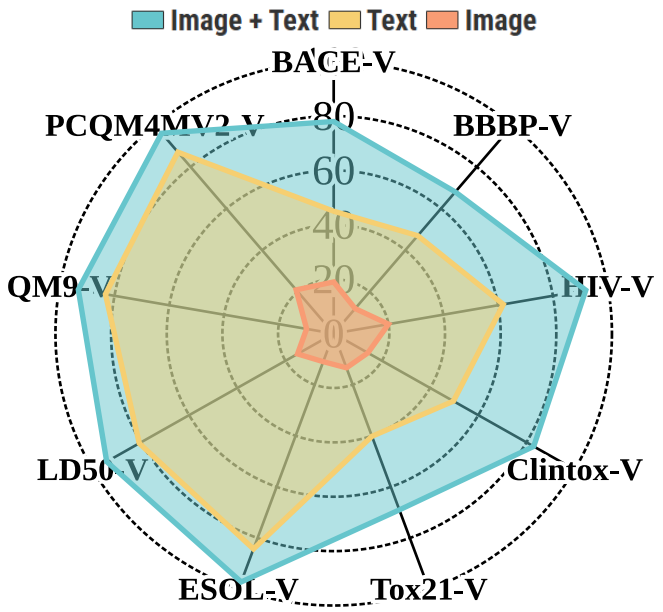
Model Performance Across Settings



Classification



Regression

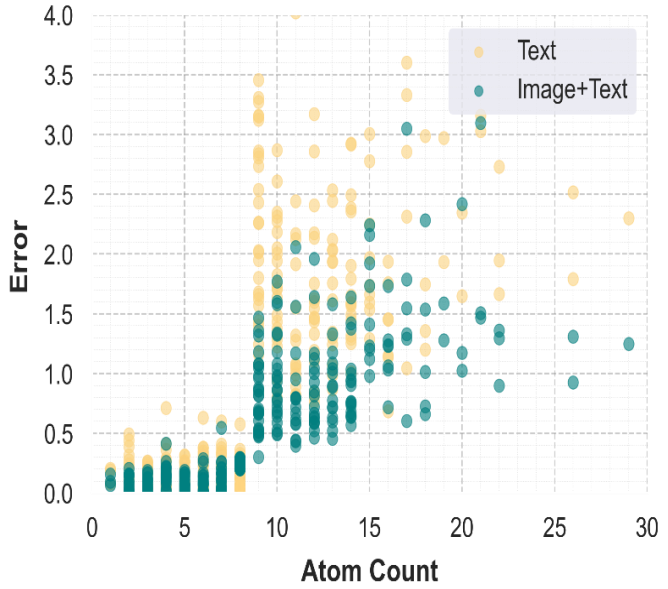


Impact of Visual information using JanusPro

IMPACT OF VISUAL INFORMATION

DOES VISION REALLY HELP?

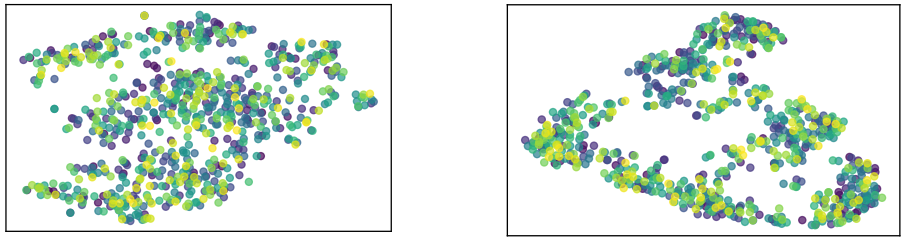
Input Mode	BACE-V Acc	HIV-V Acc	LD50-V MAE
Text Only	0.71	0.69	7.80
Image Only	0.15	0.10	31.23
Image + Text	**0.86**	**0.92**	**0.49**



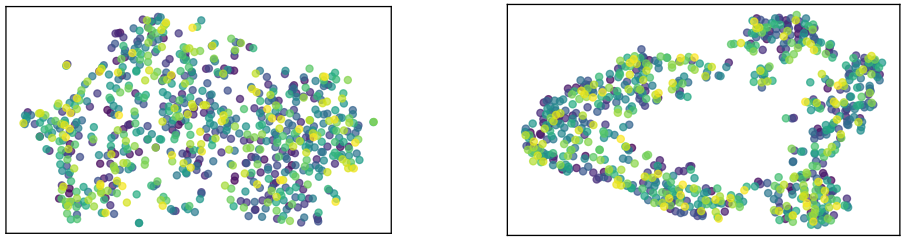
VISION ENCODER ADAPTATION

DOMAIN-AWARE VISION ADAPTATION

- Two t-SNE Visualizations Side-by-Side:
- **Before:** Poorly clustered, non-discriminative



- **After (T-Aug):** Clear clusters, pharmacophoric patterns
  - **Method:** Contrastive learning with Tanimoto similarity (>0.85)



- **Performance Gains:**
- ESOL RMSE: ↓35%
- LD50 MAE: ↓51%
- Classification: +2-4% accuracy

THANK YOU

EXHIBIT HALL C,D,E POSTER SESSION
5TH DECEMBER, FRIDAY
4:30PM – 7:30PM

Project Page

https://molvision.github.io/MolVision/