MV-CLAM: Multi-View Molecular Interpretation with Cross-Modal Projection via Language Model

Anonymous ACL submission

Abstract

Human expertise in chemistry and biomedicine relies on contextual molecular understanding, a capability that large language models (LLMs) can extend through fine-grained alignment between molecular structures and text. Recent multimodal learning advances focus on cross-modal alignment, but existing molecule-text models ignore complementary information in different molecular views and rely on single-view representations, limiting molecular understanding. Moreover, naïve multi-view alignment strategies face two challenges: (1) *separate aligned spaces* with inconsistent mappings between molecule and text embeddings, and that (2) existing loss objectives *fail to preserve* complementary information for fine-grained alignment. This can limit the LLM's ability to fully understand the molecular properties. To address these issues, we propose MV-CLAM, a novel framework that aligns multi-view molecular representations into a unified textual space using a multi-query transformer (MQ-Former). Our approach ensures cross-view consistency while a token-level contrastive loss preserves diverse molecular features across textual queries. MV-CLAM enhances molecular reasoning, improving retrieval and captioning accuracy. The source code of MV-CLAM is available in https://anonymous.4open.science/r/mv-clam-4827.

1 Introduction

A profound contextual understanding of both molecular structures and biomedical text is crucial in chemistry and biomedicine. For large language models to capture these relationships, fine-grained alignment between textual and molecular representations is required to harness their high-context reasoning ability. In vision-language models, researchers have moved beyond coarse image-text matching toward precise region-word alignment, ensuring detailed semantic correspondence between textual descriptions and visual features (Li et al., 2022; Lavoie et al., 2024). Recent studies have leveraged large language models (LLMs) for molecular understanding by integrating sequential representations (1D SMILES strings) and structural features (2D molecular graphs and 3D conformers) (Edwards et al., 2022; Liu et al., 2023a). This approach mitigates the inherent limitations of LLMs which are primarily trained on textual data, that lacks native reasoning over molecular structures. To enable LLMs to further understand molecule information, Q-former based models (Liu et al., 2023b; Li et al., 2024) align molecular structures into text space (Figure 2B).

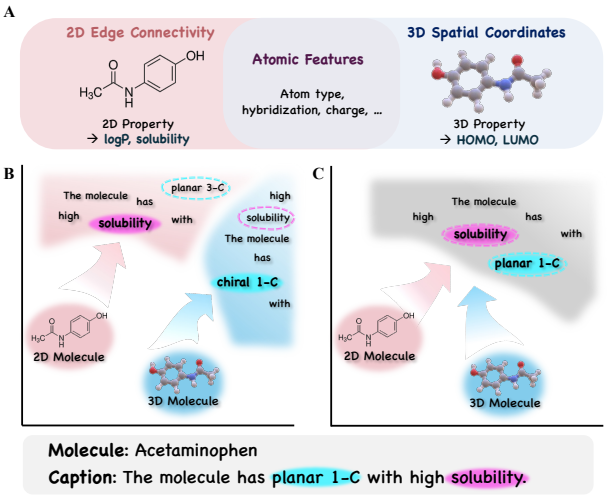


Figure 1: Motivations of MV-CLAM. (A) Complementary molecular information captured by 2D and 3D representations, where 2D graph encodes edge connectivity, and 3D conformers captures spatial coordinate structures. (B) Inconsistent mappings between molecule (2D and 3D) and property tokens (e.g., 2D property token like *solubility* and 3D structural information like *chiral 3-C*) in distinct text spaces. (C) A unified alignment with a Multi-Querying Transformer (MQ-Former) allows all text tokens share a single text space.

Combining multi-view molecular features simultaneously is essential, as their complementary nature provides a more complete understanding of

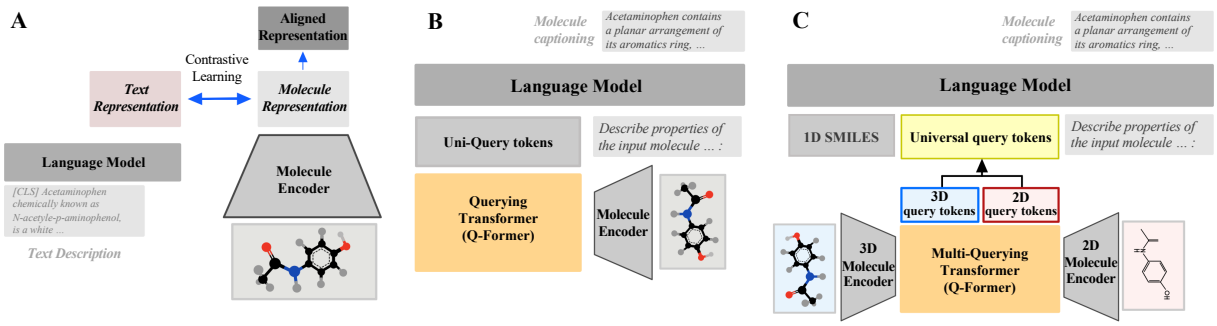


Figure 2: Methods for molecular language modeling. (A) Contrastive learning aligns two modalities via a contrastive objective, excelling in retrieval but lacking generative capabilities. (B) The Q-Former framework uses learnable query tokens for caption generation but is limited to a single molecular representation. (C) MV-CLAM extends this by integrating multiple representations with modality-specific queries, enabling fine-grained knowledge integration.

molecular characteristics. For example, as shown in Figure 1A, 2D molecular graphs primarily capture atomic bonding patterns, absent in 3D point clouds. Hence, 2D graphs focus on properties highly affected by atomic bond patterns (eg.,log P, solubility) (Guo et al., 2022) while 3D molecular conformations encode spatial atomic coordinates that influence molecular interactions and quantum properties such as HOMO and LUMO (Kim et al., 2024; Zhou et al., 2023; Du et al., 2023). In the context of molecule understanding, aligning both molecular views into the unified text space of LLMs enables the model to capture all relevant molecular details effectively.

However, existing molecule-text modeling focuses on the alignment of a single molecular view as shown in Figure 2A and 2B (Cao et al., 2023; Li et al., 2024; Liu et al., 2023b,a). Naïve approaches to multi-view alignment might be to independently map each molecular view to text using separate alignment modules. However, this leads to several issues. (1) *Separated aligned spaces*. Aligning 2D and 3D molecular representations separately to text results in distinct aligned spaces for the same molecule. As shown in Figure 1B, *"solubility"* and *"chiral 3-C"* correspond to 2D and 3D molecular properties, but each has redundant embeddings in its own space. This inconsistency can prevent the LLM from fully understanding molecular properties, as it lacks a unified representation of 2D and 3D structures. (2) *Insufficient fine-grained molecule-text alignment*. Existing Q-Former-based approaches (Li et al., 2024; Liu et al., 2023b) for aligning molecule queries into a unified text space select the most similar query-to-single token pairs for contrastive learning (Figure 4). This coarse alignment overlooks structural diversities across molecular views (Appendix Figure 6B), failing to preserve complementary information necessary for fine-grained alignment and limiting the LLM's ability to fully understand molecular properties.

To address this, we propose MV-CLAM, a novel framework that aligns multi-view molecule features using a multi-query transformer, MQ-Former (Figure 2C). Specifically, our approach jointly integrates multi-view molecular representations into a unified textual space, where *"solubility"* and *"chiral 3-C"* have unique unified embedding. Such helps generate universal query tokens with more semantic information. Additionally, we propose a multi-token contrastive loss to refine alignment by considering all text tokens within the description, rather than a single CLS token. Such multi-token contrasting ensures that molecular structures are contextualized with finer, token-level associations, capturing both atomic and functional relevance. MV-CLAM enhances molecular reasoning in LLMs, improving both retrieval and captioning accuracy.

Our main contributions are as follows:

- We propose a novel framework, MV-CLAM, that simultaneously aligns multiple molecular views (1D smiles, 2D graphs, and 3D conformers) to a unified textual space to enhance LLM-based molecular reasoning.
- We present a novel contrastive learning loss in molecule-language modeling for fine-grained alignment, considering all text tokens with enriched molecular query tokens.
- We achieve state-of-the-art performance in molecule-text retrieval and molecule captioning tasks while improving the interpretability of molecular representations.

2 MV-CLAM

MV-CLAM provides molecule captions given multi-view structural information. 2D and 3D molecular structural information is extracted from specialized encoders and processed through MQ-Former's cross-attention layers to update learnable query tokens for each dimension. The shared self-attention layer enables information sharing across all modalities. 2D and 3D queries are combined to create a universal query, which is trained with our modified multi-objective loss for fine-grained alignment with textual descriptions. The learned universal query is then passed with the prompt and SMILES strings to the language model for caption generation. The overall framework of MV-CLAM shown in Figure 2C is comprised of three main components: (1) Molecule structural graph encoders for 2D and 3D molecular structures, (2) MQ-Former as a cross-modal projector, and (3) LLaMA2 as the language model.

2.1 Molecular Graph Encoder

To capture structural information from multiple views, we used molecular embeddings from both 3D and 2D structural encoders. For the 3D encoder f_{3d}, we deployed **Uni-Mol** (Zhou et al., 2023), a SE(3)-transformer based model pretrained on 209 million 3D molecular conformations using two tasks: 3D position recovery and masked atom prediction. Input 3D molecule for Uni-Mol is denoted as $m_{3d} = (\mathcal{V}, \mathbf{f}, \mathbf{P})$, where $\mathcal{V}$ and $\mathbf{f}$ each represents atomic nodes and their features, and $\mathbf{P} \in \mathbb{R}^{|\mathcal{V}|\times 3}$ represents 3D coordinates of atoms. Pair representations are initialized by invariant spatial positional encoding from atom coordinates and interact with atom representations. The output atomic representation $H_{3d} \in \mathbb{R}^{|\mathcal{V}|\times d_{3d}}$, where h_i corresponds to the i-th atom and d_{3d} denotes hidden dimension size of H_{3d}, updates learnable 3D query tokens through the cross-attention layers in MQ-Former's 3D molecular transformer block.

$$H_{3d} = [h_1, h_2, ..., h_{|\mathcal{V}|}] = f_{3d}(m_{3d}) \quad (1)$$

For the 2D molecular encoder f_{2d}, we adopted **Molecule Attention Transformer (MAT)** (Maziarka et al., 2020), pretrained on two million molecule samples from ZINC15 dataset (Irwin et al., 2012). Given 2D molecule $m_{2d} = (\mathcal{V}, \mathbf{f}, \mathbf{A})$ where A represents edges within the molecule as adjacency matrix, MAT generates atomic representations $H_{2d} \in \mathbb{R}^{|\mathcal{V}|\times d_{2d}}$ using a specialized molecule-specific attention mechanism that considers edges, atomic distances and atomic features. The atomic representations interact with the learnable 2D query tokens via cross-attention layers in 2D molecular transformer block.

$$H_{2d} = [h_1, h_2, ..., h_{|\mathcal{V}|}] = f_{2d}(m_{2d}) \quad (2)$$

2.2 MQ-Former: Multi-Querying Transformer

Previous studies applying Q-Former to the molecular domain projects single-dimensional structural embeddings into the textual space (Li et al., 2024; Zhang et al., 2024). These models consist of a single molecule transformer and a text transformer. However, this approach is inherently limited in preserving molecular information when aligning with text embeddings for two main reasons: (1) separate aligned spaces with inconsistent mappings between molecule and text embeddings, and (2) information loss caused by single-token contrastive learning. MQ-Former addresses this limitation by introducing a novel architecture capable of aligning multiple modalities to a unified aligned space using a refined multi-objective loss for better information preservation (Figure 3).

Our approach combines structural representations of two dimensions, but the architecture can be extended using multiple molecule transformers and a single text transformer. Each molecule transformer, based on the BERT architecture with additional cross-attention layer, processes K learnable query tokens specific to their respective views. Following previous studies (Li et al., 2024; Liu et al., 2023b), we adopt the SciBERT (Beltagy et al., 2019) architecture for the text transformer and initialize all blocks with SciBERT's pretrained weights. Hence, textual descriptions S of length L are tokenized with SciBERT's tokenizer f_{sci} to $X_{\text{text}} = \{x_1, x_2, ..., x_T\}$ (T: number of tokens in text) before being processed through MQ-Former's text transformer. The cross-attention mechanism extracts relevant information from embeddings into the query tokens, and shared self-attention layers enable information exchange across all embeddings, over-passing the limitation of separated aligned spaces.

Figure 3 illustrates MQ-Former generating a universal query tokens for a molecule given two different views. Two molecule transformer modules each updates distinct K query tokens $Q_{2d} \in \mathbb{R}^{K\times 768}$ and $Q_{3d} \in \mathbb{R}^{K\times 768}$, which are randomly initialized. The learned query tokens, $\hat{Q}_{2d}$ and $\hat{Q}_{3d}$ of

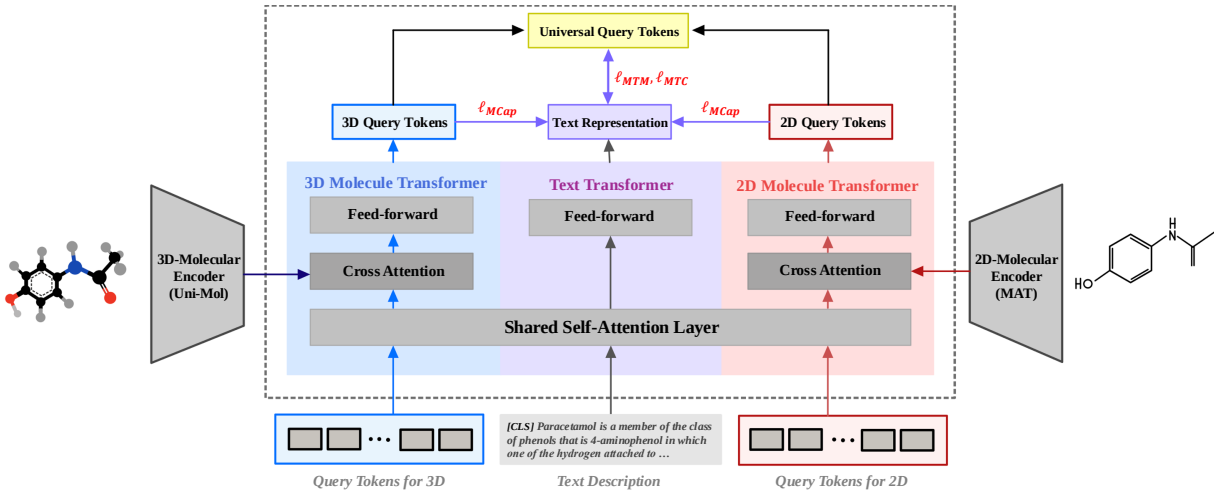


Figure 3: Training scheme of MQ-Former. The proposed MQ-Former enhances molecular language modeling by incorporating multi-token contrasting and amplified molecule captioning losses to the prior multi-objective loss (Li et al., 2023, 2024; Liu et al., 2023b). (1) The novel multi-token contrasting loss ℓ_{MTC} replaces conventional molecule-text contrastive learning, encouraging diverse query-token alignment. (2) The molecule captioning loss ℓ_{MCap} is amplified to improve text generation quality. The molecule-text matching loss ℓ_{MTM} remains unchanged.

same size, are updated representations of these initial tokens, refined through the alignment of multiple molecule views and textual descriptions $X_{\text{text}} \in \mathbb{R}^{L\times 768}$. Updated query tokens are concatenated to create a single universal query $\hat{Q} \in \mathbb{R}^{2K\times 768}$, containing complementary structural information aligned to textual space. The resulting universal query tokens are then used as inputs for the language model, along with 1D SMILES string and task prompt as depicted in Figure 2C.

$$\begin{aligned}\hat{Q} &= f_{\text{concat}}(\hat{Q}_{2d}, \hat{Q}_{3d}) \\ &= f_{\text{MQformer}}(H_{2d}, H_{3d}, X_{\text{text}}, Q_{2d}, Q_{3d})\end{aligned} \tag{3}$$

2.3 LLaMA2 & LoRA

The pretraining corpus of LLaMA2 (Touvron et al., 2023) includes a vast amount of biomedical literature and thereby exerts powerful text generation capability with internal chemistry knowledge. This allows LLaMA2 to effectively interpret 1D molecular sequences and address tasks related to molecular comprehension. Despite its inherent capabilities, the language model necessitates fine-tuning to effectively address the universal queries posed by MQ-Former, particularly due to the modifications in the tokenizer resulting from changes in module processing of textual descriptions. To facilitate efficient fine-tuning, we implemented low-rank adaptation (LoRA, (Hu et al., 2021)).

3 Training MV-CLAM

The training of MV-CLAM consists of two stages. (1) Guiding MQ-Former to align both multi-view molecular representations to a consistent textual space, and (2) Refining query tokens to be effectively soft-prompted by LLaMA2. Molecular encoders are frozen during the entire pipeline.

3.1 Stage 1: Training MQ-Former

Two sets of K learnable query tokens are updated by each molecule transformer block in Stage 1. Molecule transformer blocks hold self-attention, cross-attention and feed-forward layers. Specifically, the self attention layers in all blocks of MQ-Former are shared to exchange information between modalities and view. The 2D and 3D query tokens $Q_{2d}(i)$, $Q_{3d}(i)$ for i-th molecule are processed through their respective molecule transformers. Our $2K$ universal query token $\hat{Q}(i)$ is formed by concatenating the learned query sets. The objective is to train MQ-Former to learn a unified latent space for all molecular embeddings and obtain highly informed molecular soft-prompt $\hat{Q}(i)$ without any inconsistencies.

For training, we introduce the following key modifications to the multi-objective loss in previous works inspired by the BLIP-2 framework (Li et al., 2023, 2024), designed to maximize the diversity of queries. In order to preserve complementary chemical aspects embedded in each dimension, we introduce the following key modifications: (1) a

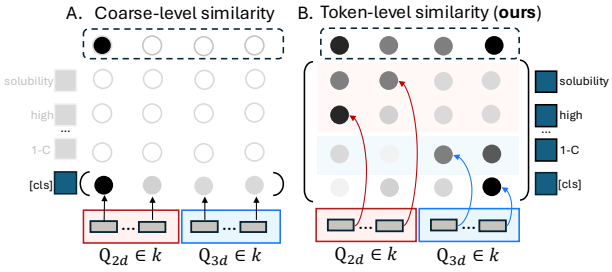


Figure 4: Molecule-text similarity for query-token contrasting. (A) Previous approach compute coarse-level similarity between molecule queries and CLS text token. (B) We propose a new approach to compute token-level similarity between molecule queries and all text tokens, which preserves molecule query diverse information.

novel multi-token contrasting loss ℓ_{MTC} in replacement to single-token (molecule-text) contrasting, and (2) amplification of the molecule captioning loss ℓ_{MCap}. Molecule-text matching is used without further modifications ℓ_{MTM}. This allows our model to capture and preserve both fine-grained atomic interactions and high-level chemical semantics, enhancing interpretability and expressiveness in molecular language modeling. Overall, the total loss for training MQ-Former ℓ_{MQ} in Stage 1 is as follows:

$$\ell_{MQ} = \ell_{MTC} + \ell_{MTM} + \alpha * \ell_{MCap} \quad (4)$$

Multi-Token Contrasting. Unlike the previous approach that retrieved only the maximum similarity between a query token and CLS text token(Figure 4A), we introduce a refined similarity computation where each molecule token is matched against all text tokens, retrieving the maximum similarity for each token against all T text tokens (Figure 4B). The average loss over all k tokens represents a fine-grained similarity calculation between molecule-text pairs, preventing query collapse, where a single query token with high similarity dominates the training process by aligning only with easily capturable text concepts. By distributing alignment across multiple queries and text tokens, we achieve richer molecule-text representations, improving cross-modal association.

ℓ_{MTC} is measured as the batch mean of the sum of molecule-to-text loss ℓ_{g2t} and text-to-molecule loss ℓ_{t2g}. For each query in the universal query token, we calculate the maximum cosine similarity it has against all text tokens $x(i) \in X_{\text{text}}(i)$ with temperature scaling for precision. The average of the calculated similarity for $2K$ queries represents pairwise similarity in a more precise manner. Similarly, ℓ_{t2g} aligns the text representation with its matching molecular query while contrasting it against all other queries within the batch. The similarity calculation can be formulated as the following:

$$\begin{aligned} S(i,j) &= \frac{1}{2K}\sum_{2K}\max_t cos(\hat{Q}_k(i), x_t(j)) \\ S'(i,j) &= \frac{1}{T}\sum_{T}\max_k cos(x_t(i), \hat{Q}_k(j)) \end{aligned} \quad (5)$$

Together ℓ_{MTC} form a bidirectional alignment between molecular features and textual descriptions in a detailed token-wise manner. ℓ_{g2t} and ℓ_{t2g} is as written below, where M is the size of the batch and τ is the temperature parameter.

$$\begin{aligned} \ell_{g2t} &= -\sum_{i=1}^{M}\log\frac{\exp(S(i,i)/\tau)}{\sum_{j=1}^{M}\exp(S(i,j)/\tau)} \\ \ell_{t2g} &= -\sum_{i=1}^{M}\log\frac{\exp(S'(i,i)/\tau)}{\sum_{j=1}^{M}\exp(S'(i,j)/\tau)} \end{aligned} \quad (6)$$

Molecule-text Matching. ℓ_{MTM} is for a binary classification task to predict matching molecule-text pairs. Universal query tokens are obtained then processed through a linear classifier after mean pooling. Let $\rho(\hat{Q}(i), X_{\text{text}}(i))$ denote the predicted probability that universal query $\hat{Q}(i)$ matches its corresponding text description $X_{\text{text}}(i)$. ℓ_{MTM} is calculated as follows:

$$\begin{aligned} \ell_{\text{MTM}} = \frac{1}{M}\sum_{i=1}^{M}\Big(&-\log\rho(\hat{Q}(i), X_{\text{text}}(i)) \\ &+\log\rho(\hat{Q}(i), X_{\text{text}}(j)) + \log\rho(\hat{Q}(r), X_{\text{text}}(i))\Big) \end{aligned} \quad (7)$$

where $X_{\text{text}}(j)$, $\hat{Q}(r)$ are randomly selected negative samples from the batch. Overall, ℓ_{MTM} aids MQ-Former to maximize the likelihood of matched pairs and minimize mismatches, enhancing its ability to differentiate between true and false pairs.

Molecule Captioning. ℓ_{MCap} is designed to generate accurate text descriptions based on multi-view query tokens. Text is generated autoregressively, where each token is predicted sequentially based on the corresponding molecular queries. Instead of harnessing universal queries, ℓ_{MCap} sums up separate losses for 2D and 3D query tokens, ensuring that each query token retains its unique dimensional information for high captioning ability. The ℓ_{MCap} is defined as follows:

$$\begin{aligned} \ell_{MCap} = &-\frac{1}{M}\sum_{i=1}^{M}\log p(X_{\text{text}}(i)|\hat{Q}_{2d}(i)) \\ &-\frac{1}{M}\sum_{i=1}^{M}\log p(X_{\text{text}}(i)|\hat{Q}_{3d}(i)) \end{aligned} \quad (8)$$

where $p(X_{\text{text}}|\hat{Q}_{2d})$ and $p(X_{\text{text}}|\hat{Q}_{3d})$ represents the probability of generating the text description based independently on 2D or 3D molecular queries, respectively. While the other two losses focus on aligning or matching molecule-text pairs, the ℓ_{MCap} directly impacts the ability to generate new text based on molecular representations, encouraging further diverse feature learning in correspondence to our modified multi-token contrasting loss. Given its critical role, we assigned a greater weight α, guiding MQ-Former to generate quality tokens for text-generation tasks.

3.2 Stage 2: Specializing LLaMA2 for Molecule Captioning

In Stage 2, MQ-Former is further trained alongside LLaMA2 to generate molecular descriptions. The goal is to enhance MQ-Former's ability to produce universal queries that are not only aligned with the textual space but better interpretable by LLaMA2. In this stage, textual descriptions are tokenized and decoded using LLaMA tokenizer. Universal query tokens, 1D SMILES are given as input with prompt. Autoregressive generation loss of LLaMA2 is used for training the framework with LoRA (Hu et al., 2021). Detailed LoRA setting are in Appendix A3.

4 Experiments

4.1 Datasets

PubChem324K. For molecule-text alignment and molecule captioning, we collected 324k molecular SMILES-text pairs from PubChem (Kim et al., 2021). 2D graph features were constructed using (Maziarka et al., 2020), and 3D conformers were generated with ETKDG and optimized using the MMFF algorithm in RDKit (Landrum et al., 2013). We follow dataset construction as provided in 3D-MoLM (Li et al., 2024) which also requires 3D molecular conformations. High-quality subset of 15k pairs with text longer than 19 words are sampled for train, valid, test datasets. Shorter pairs are used for pretraining. The statistics for the final PubChem324k dataset used in this study are presented in Appendix Table 4.

4.2 Benchmark models

Baseline models include (1) pretrained language models for science: Sci-BERT (Beltagy et al., 2019), (2) models with molecule-text contrastive learning: KV-PLM (Zeng et al., 2022), MoMu (Su et al., 2022), MoleculeSTM (Liu et al., 2023a) and (3) models with Q-Former modules: MolCA (Liu et al., 2023b), 3D-MoLM (Li et al., 2024), Uni-MoT (Zhang et al., 2024). For molecule captioning, we also benchmark Llama2-7B and 2D-MoLM, each as a variant of 3D-MoLM using 1D and 2D information along with MolT5 (Edwards et al., 2022) and InstructMol (Cao et al., 2023).

5 Results

5.1 Molecule-Text Retrieval

We evaluate MV-CLAM for molecule-text retrieval on the PubChem324k dataset. We perform two rounds of evaluation on molecule-to-text and text-to-molecule retrieval tasks, using Accuracy and Recall@20 metrics: within batch size of 64 and is across the entire test set. We report baseline performances as written in literature (Li et al., 2024; Zhang et al., 2024).

As shown in Table 1, MV-CLAM outperforms baseline approaches that represent molecules as 1D SMILES strings, 2D graphs, or 3D conformers. We attribute our superior performance to (1) our use of a universal query that aligns both 2D and 3D molecular representations to a consistent text, and (2) a modified multi-objective loss, designed to maximize query diversity and prevent over-reliance on dominant alignment patterns.

5.2 Molecule Captioning

Following previous studies(Li et al., 2024), we use BLEU, ROUGE, METEOR metrics to evaluate molecule captioning on the PubChem324k dataset. Table 2 shows MV-CLAM consistently outperforms all baselines with notable performance gain from our modified multi-objective loss. PubChem324k dataset includes molecular nomenclature, which our model accurately generates in addition to information on clinical usage and chemical properties. Appendix Table 5 highlights the model's ability to correctly identify International Union of Pure and Applied Chemistry (IUPAC) nomenclature and generic drug names that differ significantly in language model processing. IUPAC names follow systematic chemical rules, making them complex and highly structured, while generic drug names are more standardized and commonly used in clinical contexts. Despite these differences, MV-CLAM successfully identifies both types of names, showcasing its ability to handle a range of linguistic and chemical complexities. Moreover, MV-CLAM demonstrates its capacity to generate

Table 1: Molecule-Text retrieval performance in batch and test set for different models. The highest value in each category is indicated in bold, and the second highest value is underlined. For MoleculeSTM* and MolCA*, we report results from UniMoT (Zhang et al., 2024).

Model	Retrieval in batch				Retrieval in test set			
	M2T		T2M		M2T		T2M	
	ACC	R@20	ACC	R@20	ACC	R@20	ACC	R@20
1D SMILES								
Sci-BERT(Beltagy et al., 2019)	85.32	98.74	84.20	98.43	41.67	87.31	40.18	86.77
KV-PLM(Zeng et al., 2022)	86.05	98.63	85.21	98.47	42.80	88.46	41.67	87.80
2D Graph								
MoMu-S(Su et al., 2022)	87.58	99.24	86.44	99.38	47.29	90.77	48.13	89.92
MoMu-K(Su et al., 2022)	88.23	99.41	87.29	99.42	48.47	91.64	49.46	90.73
MoleculeSTM* (Liu et al., 2023a)	90.50	99.60	88.60	99.50	52.70	92.90	53.20	92.50
MolCA* (Liu et al., 2023b)	92.60	99.80	91.30	99.50	67.90	94.40	68.60	93.30
2D Graph + Tokenizer								
UniMoT(Zhang et al., 2024)	93.60	**100.0**	92.70	99.40	69.50	96.30	69.80	94.40
3D Conformer								
3D-MoLM(Li et al., 2024)	93.50	**100.0**	92.89	99.59	69.05	95.91	70.13	94.88
2D Graph + 3D Conformer								
MV-CLAM w/ SINGLE-TOKEN CONTRASTING	96.57	99.95	97.03	**99.95**	76.32	96.57	77.03	96.42
MV-CLAM w/ MULTI-TOKEN CONTRASTING	**97.34**	99.95	**97.19**	99.90	**78.67**	**96.98**	**79.34**	**96.93**

literature-matching captions absent in ground truth, as seen in the case of *Rifapentine* (Appendix Table 5), highlighting the ability to produce highly informed outputs.

5.3 Effectiveness of MQ-Former

In this section, we substantiate the effectiveness of incorporating multi-view chemical information within the MQ-Former architecture. We conduct both quantitative and qualitative analysis to compare our superiority to the prior single-view alignment using Q-Former. Molecular encoders are identically set for the ablation studies.

As a quantitative analysis, we compared our approach to prior works that independently align 2D embeddings or 3D embeddings with Q-Former. We also evaluated an alternative setup where multi-view molecular embeddings were pre-combined and aligned to text with Q-Former. We show that the combination of both modalities leads to a notable synergistic effect, improving the model's overall performance (Table 3). Coupled with our modified contrastive loss, the simultaneous alignment of both modalities using MQ-Former ensures that critical information is utilized, leading to more robust and detailed description predictions. Our framework outperforms the setting where multi-view embeddings are pre-combined and aligned to text using a single Q-Former module. Overall, the results supports the hypothesis that well-orchestrated multi-view fusion can surpass the limitations of single-view approaches to capture diverse complementary characteristics within molecules.

We exemplify two case studies to interpret how each transformer module and modality focus on distinct aspects of the molecule and its corresponding text. These qualitative studies provide insight into the alignment process by analyzing how different views contribute to the comprehensive understanding of molecular structures and their textual descriptions.

Case Study 1: Visualizing Attention Maps for 2D and 3D Query Tokens. Embedding grounded on different latent spaces and dimensions differently align molecular information to text. Visualization of the distinct alignment is performed by extracting and comparing the attention maps of the shared self-attention layers when processing 2D and 3D query tokens respectively with text tokens.

With multi-token contrasting loss, each query token attends distinctly to individual tokens in the captioning sentence, exhibiting diverse attention scores (Appendix Figure 6). While query maintaining diversity, 2D query tokens effectively capture 2D-related terms - such as *boiling point* - focusing on chemical and material properties that may be overlooked in 3D settings. Conversely, 3D query tokens capture 3D-specific structural information, such as *bis (2-dimethylamino)ethyl)*, informed by 3D spatial coordinates. In contrast, when MQ-Former is trained with the original contrastive loss, it not only lacks diversity among query tokens but also struggles to properly align with 2D- and 3D-related terms.

Case Study 2: Comparing molecule captions with 2D-Qformer and 3D-Qformer. We illustrates the difference in captioning results between the uni-modal Q-Former ablation models and ours

Table 2: Molecule captioning performance across models. The highest value in each category is bolded, and the second highest is underlined. Models marked with †were pretrained on larger datasets, as noted in their original papers. Results for InstructMol and MolCA are from UniMoT (Zhang et al., 2024), with MolCA evaluated in two variations using OPT-125M (small) and OPT-1.3B (large) as language models.

	BLEU-2	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
1D SMILES						
MolT5-Small(Edwards et al., 2022)	22.53	15.23	30.44	13.45	20.30	23.98
MolT5-Base(Edwards et al., 2022)	24.51	16.61	32.19	14.04	21.35	26.10
MolT5-Large(Edwards et al., 2022)	25.87	17.28	34.07	16.42	23.41	28.04
Llama2-7B†(Li et al., 2024)	27.01	20.94	35.76	20.68	28.88	32.11
2D Graph						
MoMu-Small(Su et al., 2022)	22.86	16.01	30.98	13.65	20.75	24.35
MoMu-Base(Su et al., 2022)	24.74	16.77	32.45	14.62	22.09	27.16
MoMu-Large(Su et al., 2022)	26.34	18.01	34.75	16.86	24.76	28.73
2D-MoLM†(Li et al., 2024)	27.15	21.19	36.02	20.76	29.12	32.28
InstructMol*(Cao et al., 2023)	18.90	11.70	27.30	11.80	17.80	21.30
MolCA-Small*(Liu et al., 2023b)	25.90	17.50	34.40	16.60	23.90	28.50
MolCA-Large*(Liu et al., 2023b)	28.60	21.30	36.20	21.40	29.70	32.60
2D Graph + Tokenizer						
UniMoT(Zhang et al., 2024)	31.30	23.80	37.50	23.70	33.60	34.80
3D Conformer						
3D-MoLM(Li et al., 2024)	30.32	22.52	36.84	22.32	31.23	33.06
2D Graph + 3D Conformer						
MV-CLAM W/ SINGLE-TOKEN CONTRASTING	31.75	24.48	40.43	25.72	33.79	36.54
MV-CLAM W/ MULTI-TOKEN CONTRASTING	**32.32**	**25.11**	**40.87**	**26.48**	**34.79**	**36.87**

Table 3: Captioning Performance Comparison. We compare the captioning performance using the original Q-Former module for each single-view and multi-view(pre-combined) molecular embeddings. MV-CLAM‡ denotes performance achieved using multi-token contrasting while the other, single-token contrasting.

Model	B-2	B-4	R-1	R-2	R-L	M
2D only	29.72	22.26	38.22	23.45	31.61	34.22
3D only	29.45	22.03	37.86	23.11	31.83	33.79
Multi-view	29.80	22.70	39.07	24.92	33.09	35.49
MV-CLAM	31.75	24.48	40.43	25.72	33.79	36.54
MV-CLAM‡	**32.32**	**25.11**	**40.87**	**26.48**	**34.79**	**36.87**

demonstrating the effects of utilizing multi-view molecular understanding in text generation (Appendix Figure 5). The 2D and 3D uni-modal ablations struggle to fully capture complex and large structures like '*(R)-3-hydroxytriacontanoyl-CoA*'. The ablation models fail to retain sufficient structural information required to differentiate long carbon chains with their functional groups. However, our model captures not only carboxylic acid but also phosphonate groups, which are often considered bioisosteric replacements for sulfonate acids in medicinal chemistry due to their structural similarity (Macchiarulo and Pellicciari, 2007). In comparison, the ablation models only managed to capture one of these groups, indicating that multi-view approach enables the generation of accurate nomenclature and richer descriptive information.

6 Conclusion

In this paper, we introduce MV-CLAM equipped with MQ-Former, a novel cross-modal projector. The essence of cross-modal projection lies in aligning the enriched molecular representation spaces with the text space of language models. Our architecture successfully retains complementary information from multiple dimension into a single universal token easily interpreted by large language models for molecule description tasks. Extensive experiments demonstrate that MV-CLAM has successfully fine-tunes large language models for molecule understanding, including molecule-text retrieval and molecule captioning tasks, with potential for broader applications.

7 Limitations

For future work, we aim to extend this framework to incorporate additional molecular representations, including other chemical structures, proteomics, and multiomics data. By aligning more views within MV-CLAM's architecture, we anticipate improved navigation of the drug space and a deeper understanding of molecular interactions across biological contexts. Additionally, curating larger molecule-text datasets is expected to enhance the model's performance and its ability to generalize to subtle molecular variations.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

He Cao, Zijing Liu, Xingyu Lu, Yuan Yao, and Yu Li. 2023. Instructmol: Multi-modal integration for building a versatile and reliable molecular assistant in drug discovery. *arXiv preprint arXiv:2311.16208*.

Kirill Degtyarenko, Paula De Matos, Marcus Ennis, Janna Hastings, Martin Zbinden, Alan McNaught, Rafael Alcántara, Michael Darsow, Mickaël Guedj, and Michael Ashburner. 2007. Chebi: a database and ontology for chemical entities of biological interest. *Nucleic acids research*, 36(suppl_1):D344–D350.

Wenjie Du, Xiaoting Yang, Di Wu, FenFen Ma, Baicheng Zhang, Chaochao Bao, Yaoyuan Huo, Jun Jiang, Xin Chen, and Yang Wang. 2023. Fusing 2d and 3d molecular graphs as unambiguous molecular descriptors for conformational and chiral stereoisomers. *Briefings in Bioinformatics*, 24(1):bbac560.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. 2022. Translation between molecules and natural language. *arXiv preprint arXiv:2204.11817*.

Xiaomin Fang, Lihang Liu, Jieqiong Lei, Donglong He, Shanzhuo Zhang, Jingbo Zhou, Fan Wang, Hua Wu, and Haifeng Wang. 2022. Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence*, 4(2):127–134.

Zhichun Guo, Kehan Guo, Bozhao Nan, Yijun Tian, Roshni G Iyer, Yihong Ma, Olaf Wiest, Xiangliang Zhang, Wei Wang, Chuxu Zhang, et al. 2022. Graph-based molecular representation learning. *arXiv preprint arXiv:2207.04869*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

John J Irwin, Teague Sterling, Michael M Mysinger, Erin S Bolstad, and Ryan G Coleman. 2012. Zinc: a free tool to discover chemistry for biology. *Journal of chemical information and modeling*, 52(7):1757–1768.

Ross Irwin, Spyridon Dimitriadis, Jiazhen He, and Esben Jannik Bjerrum. 2022. Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology*, 3(1):015022.

Seonghwan Kim, Jeheon Woo, and Woo Youn Kim. 2024. Diffusion-based generative ai for exploring transition states from 2d molecular graphs. *Nature Communications*, 15(1):341.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. 2021. Pubchem in 2021: new data content and improved web interfaces. *Nucleic acids research*, 49(D1):D1388–D1395.

Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. 2020. Self-referencing embedded strings (selfies): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1(4):045024.

Greg Landrum et al. 2013. Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum*, 8(31.10):5281.

Samuel Lavoie, Polina Kirichenko, Mark Ibrahim, Mahmoud Assran, Andrew Gordon Wilson, Aaron Courville, and Nicolas Ballas. 2024. Modeling caption diversity in contrastive vision-language pretraining. *arXiv preprint arXiv:2405.00740*.

Juncai Li and Xiaofei Jiang. 2021. Mol-bert: An effective molecular representation with bert for molecular property prediction. *Wireless Communications and Mobile Computing*, 2021(1):7181815.

Juncheng Li, Xin He, Longhui Wei, Long Qian, Linchao Zhu, Lingxi Xie, Yueting Zhuang, Qi Tian, and Siliang Tang. 2022. Fine-grained semantically aligned vision-language pre-training. *Advances in neural information processing systems*, 35:7290–7303.

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

Sihang Li, Zhiyuan Liu, Yanchen Luo, Xiang Wang, Xiangnan He, Kenji Kawaguchi, Tat-Seng Chua, and Qi Tian. 2024. Towards 3d molecule-text interpretation in language models. *arXiv preprint arXiv:2401.13923*.

Pengfei Liu, Yiming Ren, Jun Tao, and Zhixiang Ren. 2024. Git-mol: A multi-modal large language model for molecular science with graph, image, and text. *Computers in biology and medicine*, 171:108073.

Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Animashree Anandkumar. 2023a. Multi-modal molecule structure–text model for text-based retrieval and editing. *Nature Machine Intelligence*, 5(12):1447–1457.

Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. 2021. Pre-training molecular graph representation with 3d geometry. *arXiv preprint arXiv:2110.07728*.

Zhiyuan Liu, Sihang Li, Yanchen Luo, Hao Fei, Yixin Cao, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. 2023b. Molca: Molecular graph-language

modeling with cross-modal projector and uni-modal adapter. *arXiv preprint arXiv:2310.12798*.

Antonio Macchiarulo and Roberto Pellicciari. 2007. Exploring the other side of biologically relevant chemical space: insights into carboxylic, sulfonic and phosphonic acid bioisosteric relationships. *Journal of Molecular Graphics and Modelling*, 26(4):728–739.

Łukasz Maziarka, Tomasz Danel, Sławomir Mucha, Krzysztof Rataj, Jacek Tabor, and Stanisław Jastrzębski. 2020. Molecule attention transformer. *arXiv preprint arXiv:2002.08264*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Bing Su, Dazhao Du, Zhao Yang, Yujie Zhou, Jiangmeng Li, Anyi Rao, Hao Sun, Zhiwu Lu, and Ji-Rong Wen. 2022. A molecular multimodal foundation model associating molecule graphs with natural language. *arXiv preprint arXiv:2209.05481*.

Jiabin Tang, Yuhao Yang, Wei Wei, Lei Shi, Lixin Su, Suqi Cheng, Dawei Yin, and Chao Huang. 2024a. Graphgpt: Graph instruction tuning for large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 491–500.

Xiangru Tang, Andrew Tran, Jeffrey Tan, and Mark B Gerstein. 2024b. Mollm: a unified language model for integrating biomedical text with 2d and 3d molecular representations. *Bioinformatics*, 40(Supplement_1):i357–i368.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Sheng Wang, Yuzhi Guo, Yuhong Wang, Hongmao Sun, and Junzhou Huang. 2019. Smiles-bert: large scale unsupervised pre-training for molecular property prediction. In *Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics*, pages 429–436.

Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. 2022. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3):279–287.

Fang Wu, Dragomir Radev, and Stan Z Li. 2023. Molformer: Motif-based transformer on 3d heterogeneous molecular graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5312–5320.

Zheni Zeng, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2022. A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals. *Nature communications*, 13(1):862.

Juzheng Zhang, Yatao Bian, Yongqiang Chen, and Quanming Yao. 2024. Unimot: Unified molecule-text language model with discrete token representation. *arXiv preprint arXiv:2408.00863*.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. 2023. Uni-mol: A universal 3d molecular representation learning framework. In *The Eleventh International Conference on Learning Representations*.

A Appendix

A.1 Related Works

Molecule-Text Modeling. Early approaches utilize 1D SMILES molecular sequences to treat molecules as text sequences by adapting Transformer models (Vaswani, 2017) designed for natural language processing (Irwin et al., 2022; Wang et al., 2019). KV-PLM (Zeng et al., 2022) specifically employs a masked language modeling loss to pretrain on biomedical texts with 1D SMILES representation. MolT5 (Edwards et al., 2022) specializes T5 model (Raffel et al., 2020) and tokenizer for SMILES-to-text and text-to-SMILES translations. Further enhancements represent molecules as 2D graphs. In particular, MoMu (Su et al., 2022) and MoleculeSTM (Liu et al., 2023a) leverage cross-modal contrastive learning to align the molecule graph representation to text. Current approaches to use multi-view representations of molecules primarily rely on contrastive learning, as demonstrated in models like GIT-Mol (Liu et al., 2024) and MolLM (Tang et al., 2024b). Additionally, aided with the development of vision large language models (VLLMs), molecular large language models with multi-modal learning architectures have been developed. Simple projection layers were used in prior works, InstructMol (Cao et al., 2023) and GraphGPT (Tang et al., 2024a), to project molecular graph representations to LLM's input text token space. Recent works have been concentrated on utilizing Q-Former (Li et al., 2023) suggested in vision domain to bridge the gap between molecule and text modality. MolCA (Liu et al., 2023b) and 3D-MoLM (Li et al., 2024) aligns 2D graph and 3D conformer molecular representations to text in purpose to generate effective soft-prompts for large language models. UniMoT (Zhang et al., 2024) employs a vector quantization-driven tokenizer with a Q-Former. Current methods for utilizing multi-view representations of molecules are limited to contrastive learning or usage of specialized tokenizers, failing to achieve simultaneous alignment across all views and text, thereby neglecting the core principle of cross-modal alignment.

Molecular representation learning. Recent research in representation learning for molecules has seen significant advancements, particularly in leveraging large-scale unlabeled molecular data. SMILES-BERT (Wang et al., 2019), MolBERT (Li and Jiang, 2021) adapts the BERT architecture on SMILES string for molecular property prediction tasks. To better focus on structural information of molecules, various graph-based representation learning models were presented. MolCLR (Wang et al., 2022) specifically tailored contrastive learning for molecular graphs using data augmentation while MAT (Maziarka et al., 2020) reinterpreted the attention mechanism of transformers to consider distance and edges. More recent works concentrate on employing 3D geometry, mostly to exploit 3D spatial coordinates. GraphMVP (Liu et al., 2021) proposed a contrastive learning framework that bridges 2D topological and 3D geometric views of molecules. GEM (Fang et al., 2022) incorporated 3D geometric information by using bond angles and lengths as additional edge attributes in molecular graphs. Uni-Mol is a SE(3)-transformer based model pretrained via 3D position recovery and masked atom prediction. Additionally, MolFormer (Wu et al., 2023) integrates SMILES, graph, and 3D conformer information in a unified transformer architecture for molecular property prediction. These recent advancements demonstrate a trend towards incorporating more diverse and rich molecular information to improve the quality and applicability of learned representations, validating the approach of our research.

A.2 Datasets Statistics

PubChem. We gathered 324k SMILES-text pairs from PubChem, generating 2D graphs and 3D conformations using existing methods (Maziarka et al., 2020; Landrum et al., 2013). Molecules with valid structures were used, with 15k longer-text pairs for training, and shorter ones for pretraining.

Table 4: PubChem324k dataset statistics

Subset	#Molecule-Text Pairs	#Min Words	#Avg Words
Pretrain	290,507	1	17.84
Train	11,753	20	57.24
Valid	977	20	58.31
Test	1,955	20	55.21

For the molecule captioning task, we chose not to use ChEBI-20 dataset (Degtyarenko et al., 2007) due to two main considerations (Li et al., 2024). First, ChEBI-20 is a curated subset of PubChem, which introduces potential issues of data redundancy and leakage given the overlap between the two datasets. Second, ChEBI-20 replaces molecular names with generic terms like 'the molecule', limiting the evaluation of the model's ability to

associate structural features with accurate molecular names. Therefore, we utilized the PubChem dataset, which retains molecular names and offers a broader variety of structures, ensuring a more comprehensive evaluation of our framework in molecule captioning task.

A.3 Experimental Settings

Stage 1 Molecule-Text Retrieval Pretraining. Stage 1 serves to effectively transform molecular representations into query tokens interpretable in textual space. Using the PubChem324k pretraining subset with shorter textual descriptions, that is less informative but easier to align, MQ-former is trained for 35 epochs. A total of 301,658 molecules generated valid 2D graphs and 3D conformers, and thereby was used for pretraining. The goal of this stage was to optimize MQ-Former's universal query generation by multi-objective training (molecule-text contrasting, molecule-text contrasting, and molecule captioning). Pretraining was conducted for 35 epochs using 3 NVIDIA A6000 GPUs with a batch size of 99. Learnable query tokens of each view was set to 12 tokens and were randomly initialized. Both the Uni-Mol and MAT graph encoders were frozen throughout the pipeline to prevent the model from focusing too much on modifying the graph encoders, ensuring the training prioritized aligning representations with the textual space. To put emphasis on the decoding ability given the molecule tokens, we assigned a weight of 2 to the captioning loss. Maximum text length was configured to 256. We used an optimizer with a warmup step of 200 and a learning rate scheduler with a decay rate of 0.9. Gradient accumulation was set to 1 batch per step.

Stage 1 Molecule-Text Retrieval Finetuning. After 35 epochs of pretraining, we loaded the checkpoint and fine-tuned MQ-Former for an additional 10 epochs on PubChem's train, validation and test datasets, consisting of 12,000, 1,000, and 2,000 molecules respectively. Training is conducted using our modified multi-token contrastive loss. This serves to raise alignment capability given longer and more complex textual descriptions. The optimizer, learning rate scheduler, batch size and text length settings are identical to the previous phase.

Stage 2 Molecule Captioning Pretraining. Stage 2 serves to further refine the universal tokens in a manner suited to a specific language model, LLaMA2 (Touvron et al., 2023) available at `https://huggingface.co/baffo32/decapoda-research-llama-7B-hf`. Using the trained model checkpoint from Stage 1 training stage, we conducted 4 epochs of pretraining on the PubChem dataset. The universal query generated by MQ-Former, along with the 1D SMILES string and an instruction prompt were given as input to the language model to generate textual descriptions for the molecules.

To fine-tune LLaMA2 efficiently, we employed LoRA (Hu et al., 2021) with a configuration of r=8, α=32, and a 0.1 dropout rate. These settings were applied to the [$k_{proj}, v_{proj}, q_{proj}, o_{proj}, gate_{proj}, up_{proj}, down_{proj}$] modules, adding 19 million trainable parameters, which constituted 0.29% of the total parameters in the LLaMA2-7B model. Unlike Stage 1, we used batch size of 30 with a maximum text length of 320 considering the prompt size. Token length for generation was set to range between 128 and 320. Gradient accumulation was set to 2. The training was carried out using 3 NVIDIA A6000 GPUs.

Stage 2 Molecule Captioning Fine-tuning. Stage 2 pre-training checkpoint was further fine-tuned on the train dataset for additional 10 epochs. Experimental settings are same as stage 2 pre-training phase, and validated using valid, test datasets.

A.4 Effectiveness of MQ-Former

In this section, we provide the detailed explanations and figures of Section 5.3. We illustrate the underlying mechanism for MQ-Former, which aligns two representations by providing (1) generated captions with ground truth, (2) caption comparison with Q-former based single-view alignment, and (3) attention map visualization.

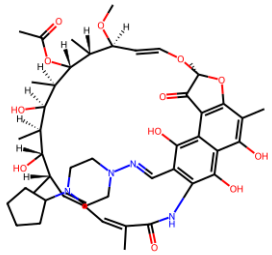
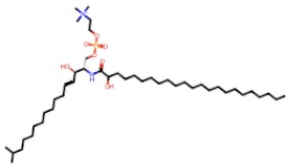
A.4.1 Comparison of MV-CLAM Captions with Ground Truth

Appendix Table 5 provides caption examples within the test dataset as specified in Section 5.2. MV-CLAM not only correctly generates IUPAC and generic names but also additional information unavailable in ground truth labels.

A.4.2 Single-View Alignment Captions

Appendix Figure 5 highlights the differences in captioning results between the uni-modal Q-Former ablation models and ours. This demonstrates that the multi-view approach generates richer and more precise molecular descriptions.

Table 5: Comparison of ground truth and MV-CLAM descriptions. Matching keywords are highlighted in bold, while additional details provided by MV-CLAM are marked in red.

Molecule	Ground Truth	MV-CLAM
	**Rifapentine** is a **rifamycin antibiotic** that is similar in structure and activity to rifampin and rifabutin and that is used in combination with other agents as **therapy of tuberculosis**, particularly in once or twice weekly regimens. Rifapentine is associated with transient and asymptomatic elevations in serum aminotransferase and is a likely cause of **clinically apparent acute liver injury**.	**Rifapentine** is a **rifamycin antibiotic** that is similar in structure and activity to rifampin and rifabutin and that is used in combination with other agents as **therapy of tuberculosis**, particularly in once or twice weekly regimens. Rifapentine is associated with transient and asymptomatic elevations in serum aminotransferase and is a likely cause of **clinically apparent acute liver injury**. Rifapentine is a **long-acting, cyclopentyl-substituted derivative of rifamycin.**
	**N-(2-hydroxytricosanoyl)-15-methylhexadecasphing-4-enine-1-phosphocholine** is an N-acyl-15-methylhexadecasphing-4-enine-1-phosphocholine in which the **acyl group has 23 carbons and 0 double bonds and is 2-hydroxylated**. It is functionally related to a 15-methylhexadecasphing-4-enine.	**N-(2-hydroxytricosanoyl)-15-methylhexadecasphing-4-enine-1-phosphocholine** is an N-acyl-15-methylhexadecasphing-4-enine-1-phosphocholine in which the **acyl group has 23 carbons and 0 double bonds and is 2-hydroxylated**. It is functionally related to a 15-methylhexadecasphing-4-enine.

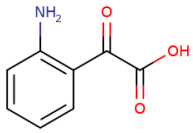
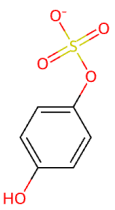
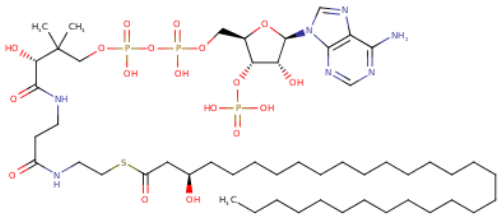
2D Only	3D Only	Original	Ground Truth
isatinic acid is a member of the class of 4-aminobenzoic acids that is anthranilic acid substituted by a hydroxy group at C-5. It has a role as a bacterial metabolite. It is a monohydroxybenzoic acid and a member of 4-aminobenzoic acids. It is functionally related to an anthranilic acid. It is a conjugate acid of an isatinate. Anthraniloic acid is a metabolite found in or produced by Escherichia coli (stra	**4-hydroxyphenyl sulfate(1-)** is a phenyl sulfate oxoanion that is the conjugate base of 4-hydroxyphenyl hydrogen sulfate, obtained by deprotonation of the sulfate group; major species at pH 7. 3. It has a role as a human metabolite. It is a conjugate base of a 4-hydroxyphenyl hydrogen sulfate. Phenyl hydrogen sulfate is a metabolite found in or produced by Escherichia col	**(R)-3-hydroxytriacontanoyl-CoA** is a 3-hydroxy fatty acyl-CoA that results from the formal condensation of the thiol group of coenzyme A with the carboxy group of (R)-3-hydroxytriacontanoic acid. It is a (R)-3-hydroxyacyl-CoA, a 3-hydroxy fatty acyl-CoA and an ultra-long-chain fatty acyl-CoA. It is a conjugate acid	**(R)-3-hydroxytriacontanoyl-CoA** is a 3-hydroxy fatty acyl-CoA that results from the formal condensation of the thiol group of coenzyme A with the carboxy group of (R)-3-hydroxytriacontanoic acid [(R)-3-hydroxymelissic acid]. It is a (R)-3-hydroxyacyl-CoA, a 3-hydroxy fatty acyl-CoA and an ultra-long-chain fatty acyl-CoA. It is functionally related to a triacontanoic acid. It is a conjugate acid of a (R)-3-hydroxytriacontanoyl-CoA(4-)
			
2D only	**3D only**	**Original**	**Ground Truth**

Figure 5: Comparison of Uni-modal Q-Former Ablation and Ours

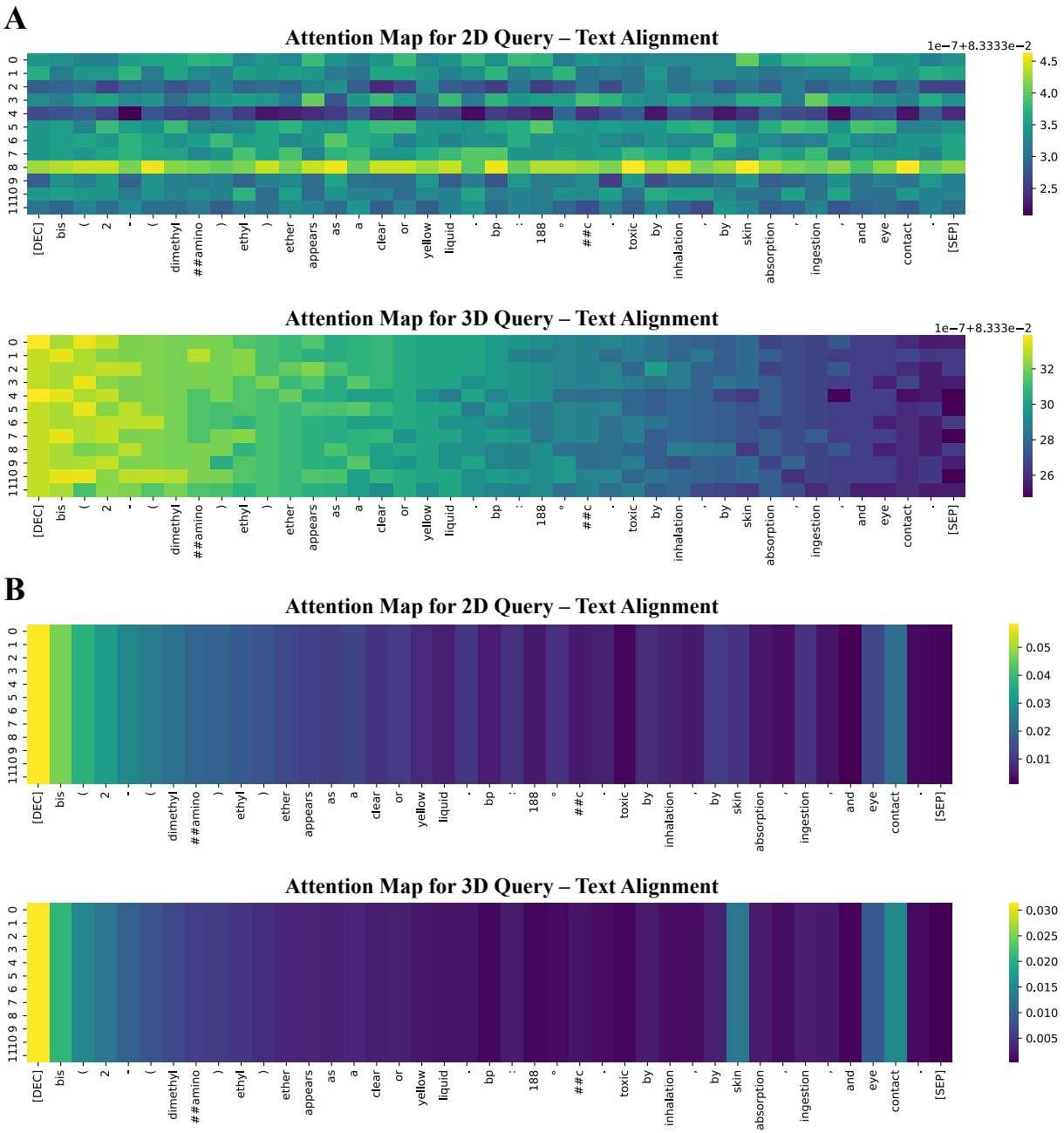


Figure 6: Comparison of attention map visualizations using different contrasting losses. The x-axis represents the word tokens in the sentence: *[DEC] bis (2 - (dimethylamino) ethyl) ether appears as a clear or yellow liquid . bp : 188 °c . toxic by inhalation , by skin absorption , ingestion , and eye contact . [SEP].*, while the y-axis corresponds to the query tokens representing the molecule. (A) shows each query exhibiting different attention weights across the textual descriptions. Additionally, 2D query tokens focus on chemical and material properties (e.g., boiling point, toxic, eye contact), while 3D query tokens capture structural information (e.g., bis(2-(dimethylamino)ethyl)). Comparatively in (B), all query tokens have consistent attention distributions for all text tokens and lack word specificity for each dimension.

A.4.3 Attention Map Visualization

We provide images of the attention maps explained in Section 5.3 (Appendix Figure 6). The attention maps of the shared self-attention layers are visualized to compare the processing of 2D and 3D query tokens with and without the multi-token contrasting loss. With the proposed loss, query tokens exhibit diverse attention scores for each word in the captioning sentences while effectively distinguishing 2D- and 3D-related terms. Specifically, 2D query tokens focus on chemical and material properties (e.g., *boiling point, toxic, eye contact*), while 3D query tokens capture structural information (e.g., *bis(2-(dimethylamino)ethyl)*). In contrast, the original contrastive loss reduces query token diversity and weakens MQ-Former's ability to align with 2D- and 3D-specific terms. This demonstrates that MQ-Former with the revised contrastive loss not only effectively preserves modality-specific information from 2D and 3D while aligning seamlessly with textual semantics but also guarantees query token diversity.

A.5 Downstream Task 1. Question Answering

A.5.1 Dataset: 3D-MolT

A total of 18439K molecule-instruction text pairs are employed using the dataset split as given in the original paper (Li et al., 2024). The dataset consists of two types of molecular property prediction tasks: (1) Computed property prediction including 3D-dependent properties (e.g. HOMO) and (2) descriptive property prediction.

Table 6: Statistics of the PubChemQC and PubChem datasets across different subsets.

Subset	PubChemQC		PubChem		
	#Mol	#Comp. QA	#Mol	#Comp. QA	#Desc. QA
Pretrain	3,119,717	12,478,868	301,658	1,199,066	1,508,290
Train	623,944	2,495,776	12,000	46,680	60,000
Valid	77,993	311,972	1,000	3,898	5,000
Test	77,993	311,972	2,000	7,785	10,000

A.5.2 Experimental Settings

For the molecular question-answering task, we utilized the 3D-MolT (Li et al., 2024) dataset, which includes question-prompt and text-answer pairs derived from the same PubChem data we used in prior. Dataset statistics are in Appendix Table 6 The dataset consists of three distinct subsets: (1) Question-answering about non-3D properties, (2) Question-answering about 3D properties, and (3) Descriptive molecular properties.

For robust guidance into instruction tuning, the three sub-datasets of 3D-MolT (Li et al., 2024) were used in combination for training a single epoch. To ensure a fair comparison with single-view methods, we initialized the instruction-tuning process using the pretrained MV-CLAM checkpoints from the molecule captioning stage, employing the original loss function rather than the multi-token contrasting loss. Given the dataset size, the model was further fine-tuned for 5 epochs on non-3D, descriptive property tasks and 1 epoch on 3D property tasks. For computed property prediction, we evaluated performance using mean absolute error (MAE). For descriptive property prediction, we measured BLEU, ROUGE, and METEOR scores.

A.5.3 Results

For baselines, we reproduced results for 3D-MoLM and 2D-MoLM (with MAT (Maziarka et al., 2020) graph encoder). These baselines represent single-modal alignment using Q-Former, and provides a fair point of comparison to demonstrate the efficacy of our multi-view cross-modal alignment. Appendix Tables 7, 8 and 9 show that MV-CLAM consistently outperformed the single-modal models.

Table 7: Comparison of Descriptive Property Generation Performance

Model	B-2	B-4	R-1	R-2	R-L	M
2D-MoLM	31.24	25.13	39.30	25.16	34.11	49.88
3D-MoLM	29.22	22.82	37.38	22.54	31.47	27.29
MV-CLAM‡	**31.70**	**25.60**	**39.61**	**25.46**	**34.51**	**50.61**

Table 8: Q&A performance on 3D properties

Model	HOMO	LUMO	HOMO-LUMO	SCF Energy
2D-MoLM	0.78 (0.99)	0.47 (0.99)	0.39 (0.90)	0.98 (1.00)
3D-MoLM	0.42 (0.99)	0.44 (0.98)	1.26 (0.99)	1.22 (0.98)
MV-CLAM‡	**0.35 (0.98)**	**0.42 (0.93)**	**0.35 (0.99)**	**0.32 (0.99)**

Table 9: Q&A performance on non-3D properties. MW, TPSA denotes molecular weight and topological surface area.

Model	MW	LogP	Complexity	TPSA
2D-MoLM	47.51 (0.98)	0.89 (0.99)	110.78 (0.99)	16.65 (0.99)
3D-MoLM	42.76 (0.96)	1.25 (0.96)	105.03 (0.96)	20.97 (0.92)
MV-CLAM‡	**21.35 (0.92)**	**0.69 (0.94)**	**55.14 (0.91)**	**9.65 (0.91)**

A.6 Downstream Task 2: Zero-shot Molecule Editing

Unlike conventional natural languages, SMILES encode molecular topology and properties demanding a specialized understanding of its notation system. Thereby, previous efforts in text-based

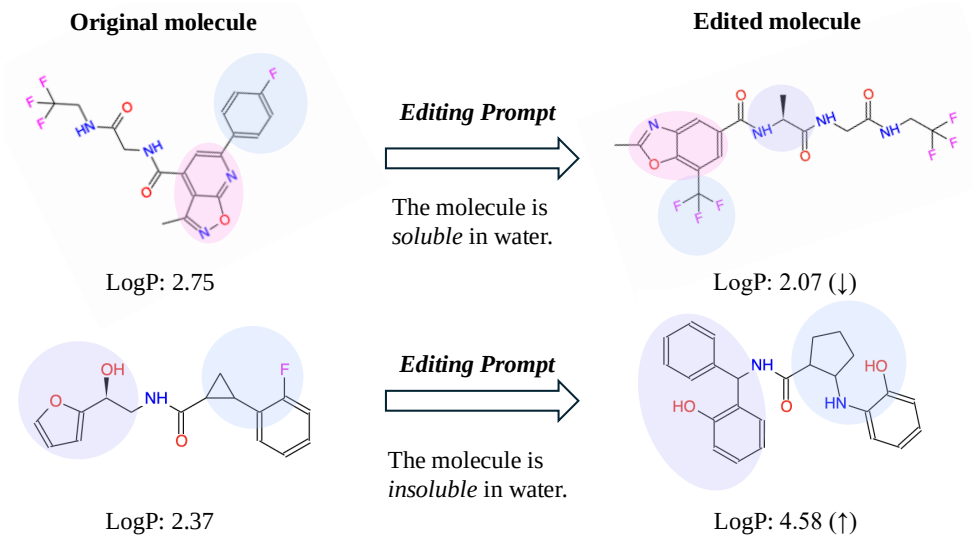


Figure 7: Zero-shot editing with chemical properties

de-novo molecule generation with large language models typically involves training or developing tokenizers that account for the unique grammar of SMILES (Edwards et al., 2022). By fine-tuning MV-CLAM, we enabled the model to output SMILES strings without additional tokenizer training.

A.6.1 Dataset: ZINC20

Following the experiment settings of (Liu et al., 2023a), 200 molecules randomly selected from the ZINC20 dataset are given 6 single-objective molecule editing instructions. The 200 molecules follow the property distribution of the entire dataset, and do not overlap with the PubChem324k training dataset in previous stages. The six instructions are the following. (1) The molecule is soluble in water. (2) The molecule is insoluble in water. (3) The molecule has high permeability. (4) The molecule has low permeability. (5) The molecule is like a drug. (6) The molecule is not like a drug. (7) The molecule has more hydrogen bond donors. (8) The molecule has more hydrogen bond acceptors.

A.6.2 Experimental Settings

Zero-shot molecule editing was conducted on the curated dataset presented in (Liu et al., 2023a) which consists of 200 randomly sampled molecules from the ZINC dataset. Each molecule was paired with molecule editing prompts (chemical instructions such as "*The molecule is more soluble in water*") and their corresponding SMILES. The dataset included molecular structures that were unseen during training. Starting with the original SMILES, the universal molecular token generated by the trained MQ-Former, and the editing prompt, we generated SMILES of the edited molecule. Using the pretrained MV-CLAM checkpoints from the molecule captioning stage, we conducted zero-shot molecule editing, utilizing the model's pre-existing multi-view molecular understanding from prior stages. The model was further fine-tuned for 4 epochs on the PubChem 324k pretraining and training datasets. This fine-tuning enabled MV-CLAM to directly generate SMILES from molecular universal tokens and was crucial to produce valid SMILES, considering the nature of LLaMA's general-purpose tokenizer which was not explicitly trained for SMILES generation. We evaluate the edited results by computing desired chemical properties using RDKit (Landrum et al., 2013), and classify whether the modification was valid shot.

A.6.3 Results

In this section we show successful case studies of the language model generating valid SMILES strings with adequate property modifications. Compared to previous works which mostly generate mere modifications of a single functional group, MV-CLAM generates diversified chemical structure modifications that may not be immediately obvious. This ability to generate more complex modifications is particularly advantageous for domain experts, as simple functional group changes are typically easy to perform manually. We attribute this diversity to the model's robust understanding of molecules within the textual space. The alignment between molecules and text is achieved by focusing on distinct substructures and molecular properties through the multi-view approach.

(Appendix Figure 7, 8,9,10,11). The values presented indicate the predicted LogP (octanol-water partition coefficient), topological surface area (TPSA), quantitative estimate of drug-likeness (QED) and number of hydrogen bond and acceptors. Each figure showcases original molecules alongside their modified counterparts with numerical indicators representing the chemical properties before and after the zero-shot editing. LogP values reflect solubility in water, while topological surface area relates to molecular permeability. QED reflects drug likeliness. The modifications are aligned with targeted property-based editing prompt, demonstrating the flexibility and chemical expertise of MV-CLAM.

A.7 Ablation Studies for Stage 2. Specializing LLaMA2 for Molecule Captioning

1D Molecular Representations We conducted an ablation study to compare the use of SELFIES (Krenn et al., 2020) with SMILES as input representations (Appendix Table 10). Using the pretrained Stage 2 checkpoint, the model was further trained for captioning under identical settings. After 10 stages of training with SELFIES, SMILES consistently demonstrated superior performance across metrics such as BLEU, METEOR, and ROUGE, validating the effectiveness of our selection.

Table 10: Captioning performance comparison for 1D molecular representations

Model	B-2	B-4	R-1	R-2	R-L	M
SELFIES	28.39	20.89	33.25	37.58	22.49	31.37
SMILES	31.75	24.48	40.43	25.72	33.79	36.54

A.8 Failure Case Study

Appendix Table 11 showcases two instances where MV-CLAM fails to differentiate structurally similar molecules. First, the model misclassifies lactoyl-CoA as oleoyl-CoA despite the key difference being the length of the carbon chain. This indicates a limitation in the model's capacity to capture subtle variations in carbon chain lengths. Second, the model misidentifies Ajugaciliatin B as subtypes E and C, demonstrating that while it successfully recognizes the molecule's primary backbone, it struggles to distinguish the small functional groups that define each subtype. This suggests that the model is not sufficiently sensitive to minor structural modifications. Both errors appear to stem from the model's difficulty in perceiving refine differences in chemical properties and spatial structure between the ground truth and its predictions. This underscores a broader challenge in molecular captioning: capturing subtle yet critical molecular features that may not greatly impact the primary structure but are crucial contributors for property.

To overcome these limitations, we propose several future studies. First, expanding our MQ-Former to align additional views or modalities, along with finer-grained molecular or related biological embeddings, could offer complementary insights to enhance the model's ability to differentiate between similar molecules. This multi-view alignment could offer a more holistic understanding of the molecule's structure and properties. In addition, curating larger molecule datasets would enhance the model's capacity to generalize, ensuring it has sufficient exposure to a wide range of molecular variations during training. These developments will address the current shortcomings and pave the way for more accurate molecular identification in future iterations of the model.

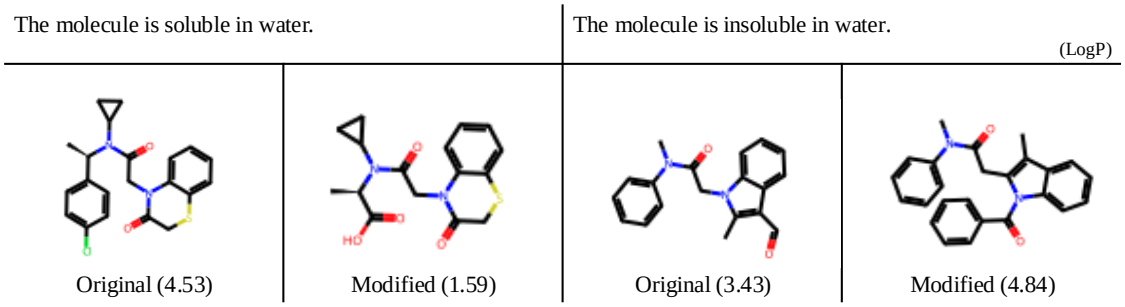


Figure 8: Editing Solubility (LogP Adjustments): Smaller LogP indicates higher solubility in water. Molecules were successfully modified given the prompt *"The molecule is soluble/insoluble in water"*.

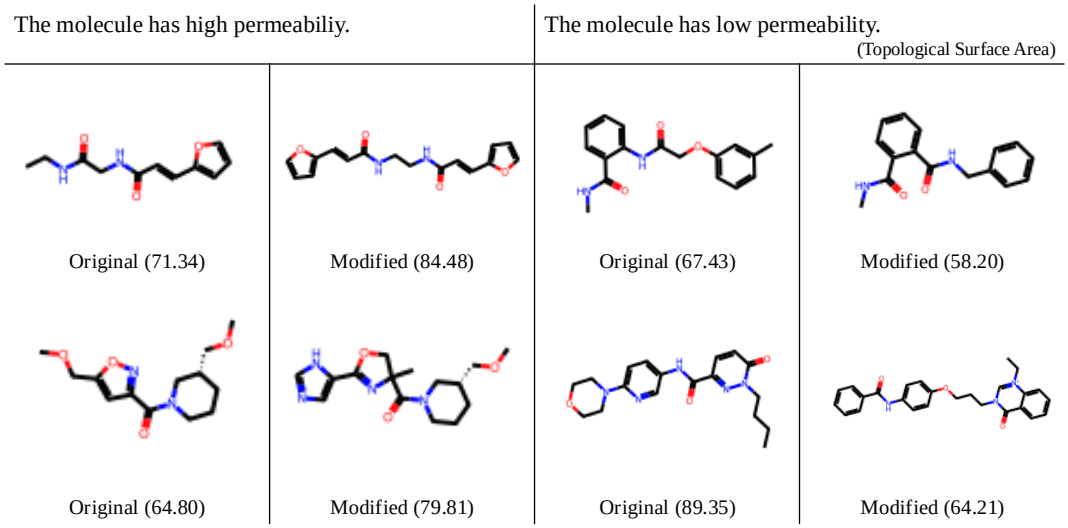


Figure 9: Editing Permeability (Topological Surface Area, TPSA Adjustments): A higher TPSA implies lower permeability, while a lower TPSA suggests higher permeability. Molecules were successfully modified given the prompt *"The molecule has high/low permeability"*.

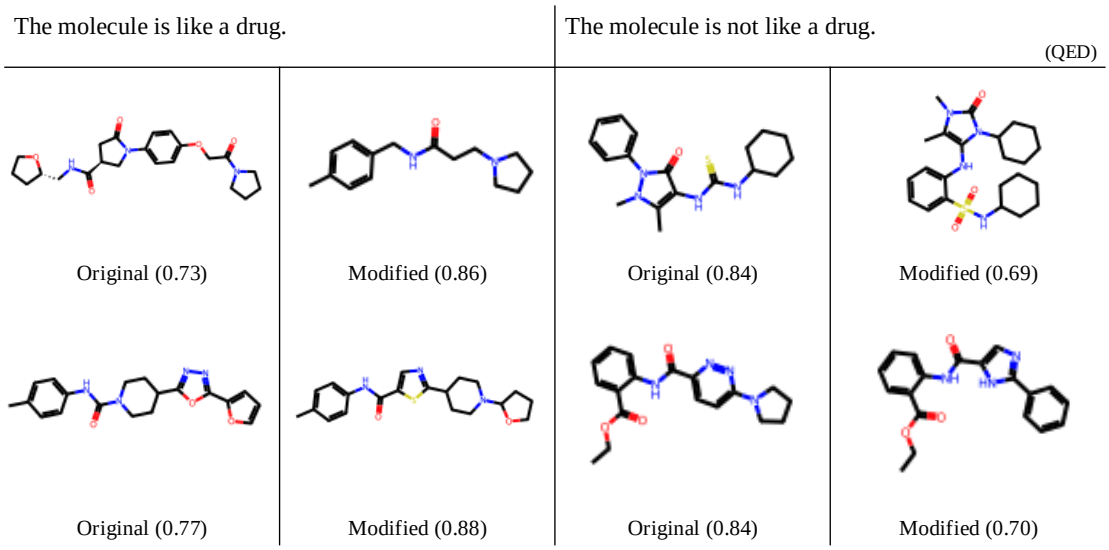


Figure 10: Editing Drug Likeliness (Quantitative Estimate of Drug-likeness, QED): A higher QED suggests a compound is more likely to possess favorable pharmacokinetic and ADMET (absorption, distribution, metabolism, excretion, and toxicity) properties, being more drug-likely. Molecules were successfully modified given the prompt *"The molecule is/is not like a drug"*.

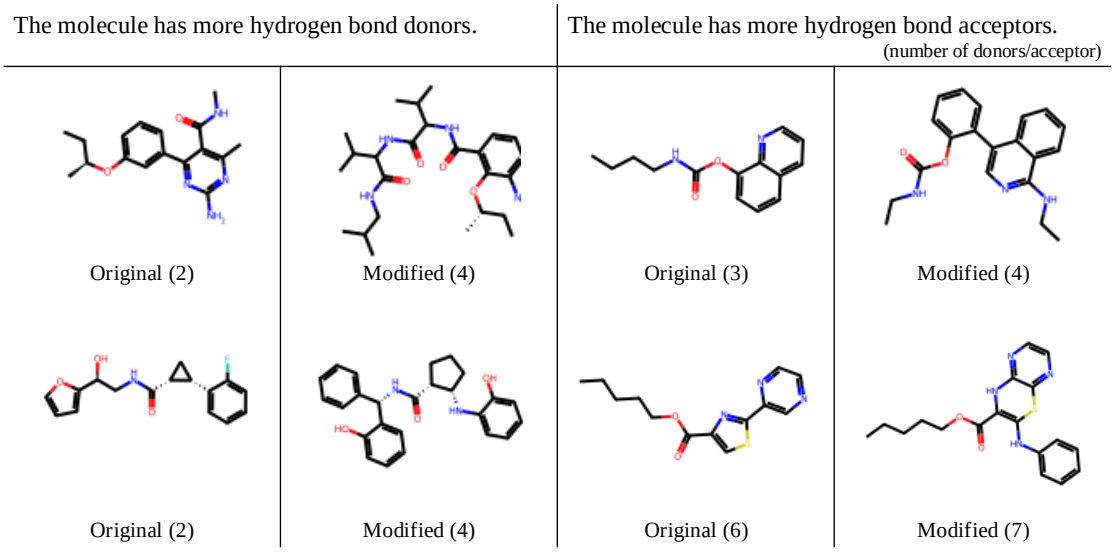
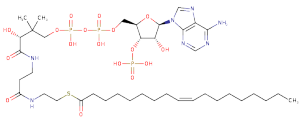
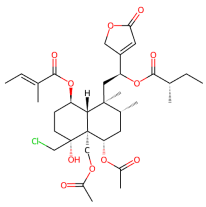


Figure 11: Editing Hydrogen Bond Acceptor/Donors: The number of hydrogen bond acceptors and donors in the molecule were given for evaluation. Molecules were successfully modified given the prompt *"The molecule has more hydrogen bond donors/acceptors"*.

Table 11: Failure Case Study: Comparison of Ground Truth and MV-CLAM Descriptions

Molecule	Ground Truth	MV-CLAM
	**Lactoyl-CoA(4-)** is an acyl-CoA(4-) that is the tetraanion of lactoyl-CoA arising from deprotonation of phosphate and diphosphate functions. It is a conjugate base of a lactoyl-CoA.	**oleoyl-CoA(4-)** is an acyl-CoA(4-) that is the tetraanion of oleoyl-CoA, arising from deprotonation of phosphate and diphosphate functions. It has a role as a human metabolite and a Saccharomyces cerevisiae metabolite. It is a saturated fatty acyl-CoA(4-) and a 3-substituted propionyl-CoA(4-). It is a conjugate base of an oleoyl-CoA
	**Ajugaciliatin B** is a diterpene lactone isolated from the whole plants of Ajuga ciliata and has been shown to exhibit neuroprotective activity. It has a role as a plant metabolite and a neuroprotective agent. It is a diterpene lactone, an acetate ester, a butenolide, a carbobicyclic compound, an organochlorine compound and a tertiary alcohol. Ajugaciliatin B is a natural product found in Ajuga ciliata with data available.	**ajugaciliatin C** is a natural product found in Ajuga ciliata with data available. **Ajugaciliatin E** is a diterpene lactone isolated from the whole plants of Ajuga ciliata. It has a role as a plant metabolite. It is a butenolide, an acetate ester, a diterpene lactone and a organochlorine compound. It is functionally related to a tiglic acid. **Ajugaciliatin E** is a natural product found in Ajuga ciliata