scMAF: Single-Cell Multi-Omics Clustering via Adaptive Modality Fusion

Jun Fu[1], **Yiding Lu**[1], **Ruohong Yang**[1], **Yunfan Li**[1,3*], **Xi Peng**[2,3]
[1]College of Computer Science, Sichuan University, Chengdu, China
[2]School of Artificial Intelligence, Sichuan University, Chengdu, China
[3]Tianfu Jincheng Laboratory, Chengdu, China
{fujun.scu, yidinglu.gm, ruohong.yrh, yunfanli.gm, pengx.gm}@gmail.com

Abstract

Single-cell multi-omics sequencing has greatly advanced the characterization of cellular heterogeneity by jointly profiling multiple molecular modalities. However, since different modalities exhibit varying discriminative signals across cell types, existing integration strategies may dilute or even obscure the signals carried by more informative modalities when fused with less informative ones, thereby hindering accurate multi-omics clustering. To address this limitation, this work presents a novel adaptive fusion framework for unsupervised clustering of single-cell multi-omics data. The core of our approach is an attention-based fusion module that dynamically assigns modality-specific weights within a shared latent space, allowing the model to naturally focus on the most informative signals from each modality. Extensive experiments on eight real-world multi-omics datasets encompassing gene expression, chromatin accessibility, and protein modalities demonstrate that our method achieves the state-of-the-art single-cell clustering performance over fourteen competitive baselines. The code is available at https://github.com/XLearning-SCU/2026-NeurIPS-scMAF.

1 Introduction

Single-cell sequencing technology enables comprehensive molecular profiling at the resolution of individual cells [34, 25, 48, 22]. Early single-cell analyses typically focused on a single omic layer, such as transcriptomic data, which provide only a partial view of cellular states [24, 26]. However, cellular phenotypes are jointly determined by multiple molecular layers that interact through complex regulatory mechanisms [52]. With the continuous advancement of sequencing technologies, single-cell studies have evolved from single-omic measurements to multi-omics profiling [60]. Commonly used modalities include transcriptome (RNA), which reflects gene expression activity [69]; chromatin accessibility (ATAC), which characterizes the regulatory landscape and epigenetic openness [6]; and surface protein abundance (ADT), which captures phenotypic information at the protein level [53]. By integrating these complementary modalities, single-cell multi-omics data offer a more comprehensive and fine-grained view of cellular states and regulatory mechanisms [55, 31]. Nevertheless, accurately identifying cell subpopulations from such complex and heterogeneous data remains challenging. In practice, unsupervised clustering is widely employed to partition cells into coherent groups prior to downstream interpretation and annotation [30, 17, 71, 66].

Traditional clustering methods, such as k-means [3] and spectral clustering [43], struggle with the high dimensionality, sparsity, and nonlinear structures inherent in single-cell data. Deep clustering approaches alleviate these issues by learning low-dimensional latent representations where cellular structures become more separable [28]. Under the multi-omics scenario, existing methods typically

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

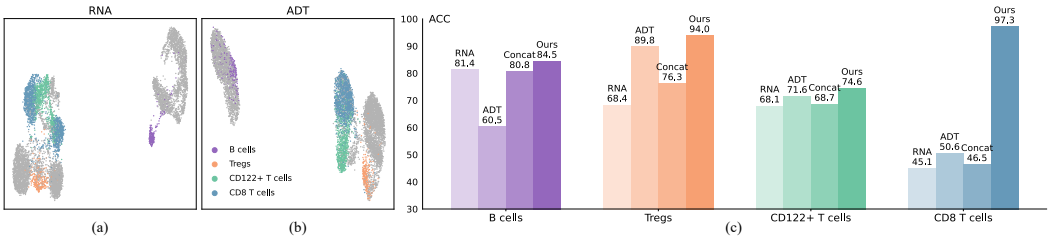


Figure 1: Our key observation on a subset of the *Spleen* dataset [35]. UMAP of RNA features (a) and ADT features (b) show that different cell clusters are better separable in different modalities, demonstrating the modality-dependent discriminative power. However, as shown in the clustering accuracy on four representative cell types (c), naive multi-omics fusion strategies, such as concatenation, would degrade performance because weaker or less-informative modalities can interfere with stronger ones. In contrast, our adaptive fusion effectively extracts the complementary signals across modalities, yielding a synergistic "1+1>2" improvement.

rely on shallow fusion strategies—early fusion through feature concatenation, as in scMDC [35]; parallel encoders with cross-modal alignment, as in TotalVI [18] and scMVAE [70]; or late fusion via fixed linear weighting, as in MoClust [66]. However, these fusion schemes remain modality-agnostic and fail to account for the heterogeneous discriminative power across modalities. As shown in Fig. 1, different modalities provide uneven, cell-type specific signals: RNA may resolve B-cell subpopulations more clearly, whereas ADT may better distinguish T-cell subtypes. When modalities are fused without considering such heterogeneity, weak or noisy modalities can override informative ones, degrading the learned representation and the resulting clustering performance. These limitations underscore the need for more advanced multi-omics fusion mechanisms that selectively extract and integrate discriminative information across different modalities.

To address these challenges, we propose scMAF, a single-cell multi-omics clustering framework designed to perform adaptive and non-linear fusion in the latent space. Motivated by the observation in Fig. 1 that different modalities exhibit heterogeneous and cell-type–dependent discriminative power, scMAF aims to integrate modalities in a way that preserves strong modality-specific signals while avoiding interference from weak or noisy ones. The core idea of scMAF is to model cross-modal interactions at a fine-grained level. Instead of merging modalities through shallow concatenation or fixed weighting, scMAF decomposes each modality's latent representation into token sequences and employs an attention-based fusion module to dynamically modulate their contributions. The attention weights are jointly optimized with the models through a reconstruction-driven objective that requires the fused representation to recover all input modalities. This objective encourages the model to selectively retain discriminative information while suppressing modality-specific noise. Such an adaptive fusion mechanism yields a unified latent representation that captures both shared and complementary information, ultimately improving multi-omics clustering performance.

The main contributions of this work could be summarized as follows:

- We reveal that existing multi-omics clustering methods are limited by shallow and modality-agnostic fusion strategies. As shown in Fig. 1, different omics provide different levels of discriminative information for different cell types, and simple fusion strategies such as concatenation would allow weak or noisy modalities to interfere with stronger ones.

- We propose scMAF, an adaptive fusion framework that enables the model to automatically extract informative signals from each modality. The proposed attention–based fusion module learns to emphasize useful modality cues while suppressing noisy or irrelevant information.

- Comprehensive experiments on eight real-world multi-omics datasets show that scMAF learns representations that capture both shared and complementary biological signals, leading to state-of-the-art clustering performance over fourteen existing baselines.

2 Related Work

In this section, we briefly review recent advances in two related fields, namely, single-cell multi-omics clustering and cross-modal fusion.

2.1 Single-Cell Multi-Omics Clustering

Existing single-cell multi-omics clustering methods fall into three main categories. *i)* Statistical modeling approaches assume explicit probabilistic relationships across modalities [13, 2, 58]. Examples include MOFA+ [2], which factorizes shared and modality-specific variation, and BREM-SC [58], which applies a hierarchical Bayesian mixture model. These methods are interpretable but limited in capturing nonlinear dependencies. *ii)* Deep generative models learn shared latent embeddings for clustering [4, 35], such as scMVAE [70] and scMDC [35]. While they better model complex structures, they often rely on a global shared embedding, reducing flexibility in handling modality heterogeneity. *iii)* Multi-modal representation learning methods align modalities through correlated latent spaces [65, 37, 57]. For instance, scCTClust [65] uses independent encoders with Deep CCA to enforce cross-modal agreement, but still adopts fixed fusion strategies.

Despite the differences in method design, existing works share a limitation that the fusion mechanisms are largely static and modality-agnostic, assuming equal modality contributions for all cells. This makes them vulnerable when different cell types are better distinguished by different modalities. In contrast, scMAF performs adaptive, cell-wise fusion by learning modality-specific embeddings and dynamically integrating them through self-attention, allowing the model to emphasize informative signals and suppress noise for each cell.

2.2 Cross-Modal Fusion

Cross-modal fusion is a central challenge in multi-modal learning [33, 44, 50, 32, 63, 10, 38, 39]. Early fusion approaches often rely on shallow operations such as feature concatenation, summation, or averaging [5, 51, 56, 16, 68]. These methods implicitly assume that all modalities contribute equally and lack the ability to adjust modality importance based on sample-specific context. As a result, noise or weak signals from one modality may be propagated into the fused representation. Another line of work learns shared latent spaces through approaches such as deep canonical correlation analysis [1, 46] and contrastive alignment methods [12, 40, 11]. These models enforce global agreement across modalities by maximizing correlation or mutual information, but they primarily focus on alignment rather than selective integration. Consequently, they cannot modulate modality contributions or filter out modality-specific noise at the level required for effective fusion.

Overall, existing fusion strategies provide limited adaptability when modality quality is imbalanced. Shallow fusion mixes modalities indiscriminately, while shared-space alignment enforces consistency without controlling which modality should dominate for each sample. In contrast, scMAF performs adaptive, cell-specific fusion, enabling the model to automatically emphasize informative modalities and suppress less reliable ones, leading to more robust and discriminative multi-omics representations.

3 Method

In this section, we present the scMAF method. As illustrated in Fig. 2, scMAF is composed of two main modules: *i)* the modality-specific feature extraction module, which learns compact representations for each modality while preserving modality-specific and complementary information; and *ii)* the adaptive multi-omics feature fusion module, which identifies the most discriminative signals across omics and adaptively aggregates them through a self-attention mechanism to produce a unified representation for clustering.

3.1 Modality-Specific Feature Extraction

Before performing multi-omics fusion, scMAF first learns a robust low-dimensional representation for each modality. Single-cell measurements are inherently sparse and noisy, and different omics exhibit distinct statistical characteristics. To account for these properties, we build a denoising autoencoder for each modality and adopt reconstruction losses tailored to its generative distribution. Formally,

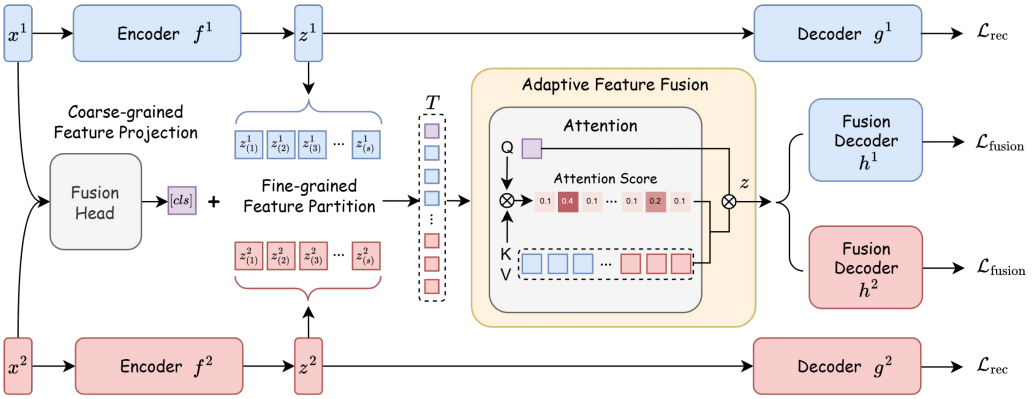


Figure 2: Overview of the proposed scMAF method with $V = 2$ input modalities as an example. Each modality is first encoded by a modality-specific autoencoder to obtain a low-dimensional latent representation. The input features are concatenated and passed through a fusion head to generate a coarse fused [cls] token. Along with chunked modality-specific features, the [cls] token is fed into a self-attention block, which adaptively aggregates informative signals across modalities and yields a deeply fused embedding for clustering. To preserve both modality-specific and fused information, scMAF includes two sets of decoders: modality-specific decoders that reconstruct each modality from its own latent representation, and fusion decoders that reconstruct each modality from the fused embedding.

let the multi-omics dataset with n cells profiled across V omics as $\mathbf{X} = \{X^1, X^2, \ldots, X^v\}_{v=1}^V$, where $X^v \in \mathbb{R}^{n \times m^v}$ refers to the v-th omics data and m^v is its feature number. For a cell x, a modality-specific encoder f^v maps corrupted data into a latent embedding by

$$z^v = f^v(x^v + \epsilon), \quad v \in \{1, 2, \cdots V\}, \tag{1}$$

where $\epsilon \sim \mathcal{N}(0, \sigma^v I)$ injects Gaussian noise and $z^v \in \mathbb{R}^d$ denotes the latent embedding of omics v. Then, a corresponding decoder g^v is introduced to reconstruct the original input x^v from z^v. To preserve biological signal while mitigating technical artifacts, each decoder outputs the parameters of a modality-specific likelihood that reflects the statistical nature of the data.

Specifically, RNA and ADT modalities are often characterized by over-dispersion. We therefore adopt the zero-inflated negative binomial (ZINB) distribution to model reconstruction, namely,

$$\begin{aligned} \mathcal{L}_{\text{ZINB}}(x \mid \pi, r, p) &= -\log \text{ZINB}(x \mid \pi, r, p) \\ &= -\log \left(\pi\, \delta_0(x) + (1-\pi) \frac{\Gamma(x+r)}{x!\,\Gamma(r)}\, p^r (1-p)^x \right), \end{aligned} \tag{2}$$

where π is the zero-inflation probability, r is the dispersion parameter, p is the success probability of the negative binomial (NB) distribution, and $\delta_0(x)$ is an indicator function which equals 1 iff $x = 0$. Parameters $\pi, r, p \in \mathbb{R}^{m^v}$ are estimated by the modality-specific decoder g^v, *i.e.*, $\pi, r, p = g^v(z^v)$. This formulation effectively captures both biological variability and technical dropout noise.

For ATAC data, which is sparse but generally less over-dispersed, we model them using the Poisson distribution, which provides a stable and tractable approximation by

$$\mathcal{L}_{\text{Pois}}(x \mid \lambda) = -\log \text{Pois}(x \mid \lambda) = -\log \left(\frac{e^{-\lambda} \lambda^x}{x!} \right), \tag{3}$$

where $\lambda > 0$ denotes the expected count estimated by the decoder g^v. The complete derivations of the ZINB loss and the Poisson loss, along with their parameterizations, are deferred to Appendices A.1 and A.2, respectively.

To briefly summarize, the modality-specific reconstruction loss aggregates all modalities and lies in the form of

$$\mathcal{L}_{\text{rec}} = \sum_{v=1}^{V} \begin{cases} \mathcal{L}_{\text{ZINB}}(x^v \mid g^v(z^v)), & \text{if } v \text{ is RNA or ADT,} \\ \mathcal{L}_{\text{Pois}}(x^v \mid g^v(z^v)), & \text{if } v \text{ is ATAC.} \end{cases} \tag{4}$$

This likelihood-guided reconstruction ensures that each modality encoder learns compact and biologically meaningful features, forming the foundation for the subsequent adaptive multi-omics fusion.

3.2 Adaptive Multi-Omics Feature Fusion

As mentioned earlier, different cell types may be better resolved in different omics, and weak or noisy modalities can easily distort the information extracted from stronger ones. To address this challenge, we introduce an adaptive modality fusion module that selectively emphasizes the most discriminative signals from different modalities, thereby enabling a deeper and more flexible integration of multi-omics features. The adaptive multi-omics feature fusion module consists of three key steps, including *i)* coarse-grained feature projection, *ii)* fine-grained feature partition, and *iii)* adaptive feature fusion.

Coarse-grained Feature Projection. Given the data of V omics, we first generate a coarse unified cell descriptor by concatenating multi-omics data and feeding them into a lightweight fusion head, *i.e.*,

$$[\text{cls}] = \text{FusionHead}\big([x^1, x^2, \ldots, x^v]\big), \tag{5}$$

where [cls] serves as a holistic but potentially noisy summary, providing an initial anchor for subsequent adaptive fusion.

Fine-grained Feature Partition. To provide a more expressive representation for adaptive fusion, each modality embedding z^v is evenly partitioned into s smaller feature slices as

$$z^v = [z^v_{(1)}, z^v_{(2)}, \ldots, z^v_{(s)}], \;\; z^v_{(s)} \in \mathbb{R}^c, \;\; c = \frac{d}{s}. \tag{6}$$

These slice tokens allow the fusion module to operate at a more fine-grained level than merging entire modality embeddings, enabling the attention mechanism to assign different weights to different subcomponents of each modality.

Adaptive Feature Fusion. We then construct the input token sequence by combining the [cls] token with all slice tokens, namely,

$$T = \left\{ [\text{cls}], \underbrace{z^1_{(1)}, \ldots, z^1_{(s)}}_{z^1}, \ldots, \underbrace{z^v_{(1)}, \ldots, z^v_{(s)}}_{z^v} \right\}. \tag{7}$$

The sequence T provides a structured view of multi-omics information, where the [cls] token encodes coarse holistic context, and the modality-specific slices carry fine-grained biological signals.

Next, we process T using a self-attention block. Let Attn denote the attention block. We extract the output corresponding to the [cls] token position as the final fused embedding, *i.e.*,

$$z = \text{Attn}(T)_{[\text{cls}]}, \quad z \in \mathbb{R}^c. \tag{8}$$

Such a cell-specific and modality-aware fusion process allows the [cls] token to retain both shared and complementary biological signals that carry stronger discriminative cues, while down-weighting noisy or irrelevant components, thus facilitating the downstream clustering.

To ensure that the fused embedding preserves modality-relevant information, scMAF introduces a set of fusion decoders h^v that reconstruct the original observations using the fused embedding through:

$$\mathcal{L}_{\text{fusion}} = \sum_{v=1}^{V} \begin{cases} \mathcal{L}_{\text{ZINB}}(x^v \mid h^v(z)), & \text{if } v \text{ is RNA or ADT,} \\ \mathcal{L}_{\text{Pois}}(x^v \mid h^v(z)), & \text{if } v \text{ is ATAC.} \end{cases} \tag{9}$$

The fusion decoders adopt the same architecture and likelihood formulations as the modality-specific decoders. Together with the modality-specific reconstruction loss $\mathcal{L}_{\text{rec}}$, we arrive at the overall training objective as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{fusion}}. \tag{10}$$

After training, we adopt the fused embeddings as the final cell representations and perform k-means clustering on z to derive the cell-type assignments.

4 Experiments

In this section, we evaluate scMAF by comparing its clustering performance against fourteen baselines on eight single-cell multi-omics datasets. We then conduct parameter analyses and ablation studies to demonstrate its robustness. Furthermore, we validate its generalizability on two image-text multi-view datasets and visualize the attention weights to demonstrate its interpretability.

4.1 Datasets

To comprehensively evaluate the proposed framework, we conduct experiments on both single-cell multi-omics datasets and image-text multi-view datasets. Specifically, we utilize eight real-world single-cell datasets, each containing paired transcriptomic and either chromatin accessibility or protein expression modalities. These datasets span various biological systems and experimental protocols, including human and mouse samples from diverse tissues such as blood, spleen, skin, and bone marrow. Furthermore, to demonstrate the model's generalizability to broader multimodal representation learning tasks, we incorporate two widely used image-text multi-view datasets. Detailed statistics for the datasets are summarized in Table 1.

Table 1: Summary of eight single-cell multi-omics datasets and two image-text multi-view datasets used in our experiments. "View 1" denotes the number of gene features for RNA data and image features for image data, while "View 2" denotes the number of protein features for ADT data, chromatin accessibility features for ATAC data, and text features for text data.

Modality	Dataset	Samples	View1	View2	Classes
RNA+ADT	Inhouse [49]	1,372	33,538	10	7
	Spector [35]	3,762	33,538	49	16
	Malt10x [49]	8,412	33,538	17	11
	Spleen [35]	16,828	13,553	112	35
	BMNC [35]	30,672	17,009	25	27
RNA+ATAC	PBMC [8]	3,009	36,601	81,156	8
	Ma2020A [67]	5,692	21,478	340,341	22
	Ma2020B [67]	5,927	21,478	340,341	22
Image+Text	NUS-WIDE [23]	9,000	4,096	300	10
	XMediaNet [21]	40,000	4,096	300	200

4.2 Implementation Details

To preprocess the data, we first select the top 2,000 highly variable genes or peaks for the RNA and ATAC modalities, while retaining all protein features for the ADT modality. Then, we normalize the total count of each cell to the median for sequencing depth correction, log-transform the count matrix, and scale each feature to zero mean and unit variance.

The modality-specific autoencoders are constructed using fully connected networks. For each modality with m selected highly variable features, the encoder is structured as $m \to 256 \to 128$, and the decoder is structured as $128 \to 256 \to m$. We apply a fusion head with dimensions $p \to 256 \to c$ to integrate input features, where p denotes the sum of feature dimensions from the two modalities, and c is the slice embedding dimension. To enhance the feature extraction ability, Gaussian noise with $\mu = 0$, $\sigma^1 = 3.5$, and $\sigma^2 = 2$ is added to the input data. The model is trained with a batch size of 256 using the Adam optimizer for 200 epochs, with all hyperparameters fixed across datasets. All experiments are conducted on a machine equipped with an NVIDIA RTX 4090 GPU.

We adopt three widely used clustering metrics for evaluation: normalized mutual information (NMI), adjusted Rand index (ARI), and clustering accuracy (ACC). Higher values of these metrics indicate better clustering performance. All results are averaged over 20 runs of k-means with different random seeds to reduce stochasticity.

4.3 Comparison with State-of-the-Art Methods

We compare scMAF with fourteen state-of-the-art clustering algorithms, including two single-omics methods: scDSC [17] and scTAG [64]; and twelve multi-omics methods: scMVP [27], scMCs [46], MoClust [66], scMVAE [70], GLUE [7], TotalVI [18], Liam [45], Seurat [20], sCIN [14], DCCA [71], scDMSC [59], and scMDC [35]. All baselines are evaluated using their officially released code or packages with default parameter configurations. To ensure fair comparison, we select the same set of features for different methods. For methods requiring gene activity matrices on ATAC data, we transform peak features into gene-level representations using the signac package [54]. Note that for scMDC [35], we report the results before its fine-tuning stage to avoid changes of target cluster numbers. As shown in Table 2, the proposed scMAF achieves state-of-the-art clustering performance

Table 2: Clustering performance on eight single-cell multi-omics datasets. The last column reports the average performance across all datasets within each row. Best and second-best results are highlighted in **bold** and underline, respectively. "-" indicates results not available due to incompatibility or memory overflow.

Method	Ma2020B			PBMC			Spleen			Spector			Average
	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	
scDSC[17]	3.89	4.06	23.76	36.47	24.18	40.48	42.81	24.14	32.96	64.32	47.34	62.36	33.90
scTAG[64]	58.01	36.38	50.97	64.51	53.19	60.65	54.75	26.03	41.48	66.85	51.67	67.38	52.66
scMVP[27]	11.16	3.32	16.36	70.81	57.60	68.37	61.02	26.39	41.61	46.59	28.57	42.48	39.52
scMCs[46]	55.57	29.60	38.75	68.10	55.26	66.30	42.74	13.36	22.27	56.29	34.65	46.70	44.13
MoClust[66]	58.15	35.02	48.34	67.99	53.80	70.74	47.93	27.60	34.06	63.45	47.88	54.77	50.81
scMVAE[70]	55.81	28.04	42.38	65.34	48.89	65.10	58.67	22.41	34.40	69.56	56.88	67.68	51.26
GLUE[7]	68.39	46.77	59.12	68.87	57.88	73.42	-	-	-	-	-	-	-
TotalVI[18]	-	-	-	-	-	-	60.09	26.51	38.84	67.98	54.24	66.35	-
Liam[45]	57.23	30.76	41.02	67.15	52.20	64.01	57.55	23.13	35.07	64.70	45.39	58.96	49.76
Seurat[20]	65.04	40.51	54.85	59.98	47.70	66.70	54.88	18.32	29.91	69.35	53.92	64.89	52.17
sCIN[14]	60.68	39.37	49.18	66.65	56.40	72.13	54.24	17.62	30.22	68.74	53.95	64.22	52.78
DCCA[71]	68.29	45.53	59.26	70.99	58.27	68.17	57.96	24.57	36.83	66.53	50.26	62.23	55.74
scDMSC[59]	67.57	50.37	57.52	63.97	47.99	56.23	59.00	36.32	45.41	67.78	45.94	60.15	54.85
scMDC[35]	61.32	42.29	53.89	66.28	54.56	68.39	59.86	29.0	40.02	71.97	59.13	69.06	56.31
scMAF (Ours)	**71.68**	**51.53**	**62.57**	**71.59**	**58.91**	**75.23**	**63.71**	**39.57**	**49.44**	**73.08**	**59.53**	**69.83**	**62.22**

Method	Ma2020A			Malt10x			Inhouse			BMNC			Average
	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	
scDSC[17]	7.57	2.20	2.16	56.61	57.04	59.83	46.02	41.97	62.32	64.51	62.41	64.56	43.93
scTAG[64]	65.59	45.45	60.47	51.86	45.95	54.34	48.95	43.04	66.40	-	-	-	-
scMVP[27]	13.62	4.54	17.84	60.14	57.60	59.57	55.91	45.57	65.45	71.43	45.82	50.99	45.71
scMCs[46]	58.92	35.82	43.39	55.58	39.82	47.99	48.69	41.72	60.13	63.01	39.40	44.19	48.22
MoClust[66]	60.32	39.82	49.41	58.08	54.21	54.03	53.03	38.64	56.80	64.38	39.02	45.19	51.08
scMVAE[70]	64.24	48.33	58.36	55.50	38.53	50.04	56.97	42.46	60.50	77.38	51.00	56.23	54.96
GLUE[7]	75.53	58.90	68.42	-	-	-	-	-	-	-	-	-	-
TotalVI[18]	-	-	-	59.05	51.09	56.81	58.23	46.29	64.94	75.13	50.98	57.91	-
Liam[45]	61.45	39.30	47.89	60.19	45.74	54.42	66.11	59.34	70.92	76.12	50.64	54.24	57.20
Seurat[20]	74.06	56.88	67.87	52.87	42.95	46.24	70.82	55.53	60.35	77.93	48.28	51.52	58.78
sCIN[14]	64.54	42.98	55.04	56.72	45.07	50.28	75.03	62.52	73.79	77.13	53.86	59.87	59.74
DCCA[71]	74.86	58.29	67.11	58.39	46.68	**67.49**	56.20	47.80	56.34	72.45	46.21	54.62	58.87
scDMSC[59]	74.03	54.31	60.30	**62.28**	57.32	58.17	57.20	47.79	65.74	76.52	54.55	57.47	60.47
scMDC[35]	65.12	45.18	58.96	58.99	57.44	56.73	75.87	79.27	77.55	80.17	67.48	68.16	65.91
scMAF (Ours)	**78.77**	**63.28**	**70.72**	60.77	**66.57**	62.66	**76.51**	**80.82**	**80.83**	**84.45**	**79.87**	**79.73**	**73.75**

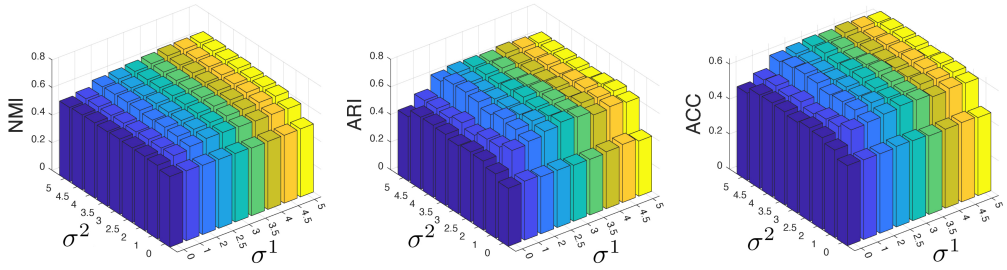


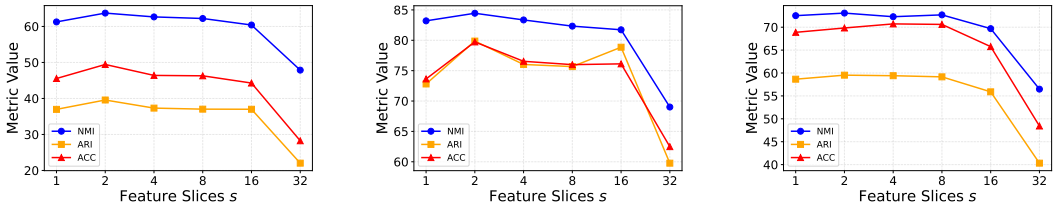
Figure 3: Clustering performance on the *Malt10x* dataset under different noise levels, σ^1 and σ^2 denote the standard deviations of Gaussian noise added to the RNA and ADT modalities, respectively.

on eight datasets. On the *BMNC* dataset, for example, scMAF improves the ARI by 12.39% over the most competitive baseline scMDC. Such a substantial gain highlights the effectiveness of our adaptive fusion mechanism, which enables the model to selectively extract informative signals from each modality and avoid interference from weaker ones.

4.4 Parameter Analysis

To investigate the robustness of the proposed scMAF, we perform parameter analysis on its two hyperparameters, namely, the strength of Gaussian noise σ^1, σ^2 and the number of feature slices s.

Strength of Gaussian Noise. Fig. 3 illustrates the clustering performance of scMAF with varying standard deviations of Gaussian noise σ^1 (on the RNA modality) and σ^2 (on the ADT modality). Specifically, we test noise levels $\sigma^{1,2} \in \{0,\ 1.0,\ 2.0,\ 2.5,\ 3.0,\ 3.5,\ 4.0,\ 4.5,\ 5.0\}$ on the *Malt10x*



(a) Spleen (b) BMNC (c) Spector

Figure 4: Clustering performance on three datasets under different numbers of feature slices s.

dataset and evaluate clustering performance. As shown, moderate noise contamination to the input data significantly improves clustering performance. Notably, even at higher noise levels (*e.g.*, $\sigma^{1,2} = 5.0$), the clustering performance remains stable, demonstrating the robustness of our model to input perturbations.

Number of Feature Slices. Fig. 4 examines how the number of feature slices s influences clustering performance. We evaluate $s \in \{1,\ 2,\ 4,\ 8,\ 16,\ 32\}$ on the *Spleen*, *BMNC*, and *Spector* datasets. The case $s = 1$ corresponds to using each modality's entire embedding as a single token without slicing. Across all three datasets, moderate slicing (*i.e.*, $s = 2 \sim 8$) consistently improves performance, suggesting that exposing multiple feature slices provides the attention module with a richer structure from which useful modality-specific cues can be extracted. However, excessive slicing (*i.e.*, $s > 16$) degrades performance, probably because overly small slice dimensions limit information capacity and weaken the effectiveness of the [cls] token in aggregating meaningful signals. Overall, these results indicate that scMAF is robust to a broad range of slice numbers.

4.5 Ablation Studies

To investigate the effectiveness of each module of scMAF, we conduct ablation studies on the *i)* multi-omics feature fusion strategy, *ii)* attention optimization strategy, and *iii)* objective functions.

Multi-Omics Feature Fusion. To demonstrate the superiority of the proposed attention-based multi-omics feature fusion module, *i.e.*, $z = \mathrm{Attn}(T)_{[\mathrm{cls}]}$, we compare it with five alternative fusion strategies, including *i)* element-wise addition, *i.e.*, $z = z_1 + z_2$; *ii)* concatenation, *i.e.*, $z = \mathrm{Concat}(z_1, z_2)$; *iii)* using the [cls] token itself without attention mechanism, *i.e.*, $z =$ [cls]; *iv)* modality gating, *i.e.*, $z = \alpha_1 z_1 + \alpha_2 z_2$; and *v)* modality-specific attention, *i.e.*, $z = \mathrm{Concat}(\mathrm{Attn}(T_1)_{[\mathrm{cls}]}, \mathrm{Attn}(T_2)_{[\mathrm{cls}]})$. As shown in Table 3, the proposed shared attention strategy achieves the best clustering performance across all metrics. Compared with modality gating and modality-specific attention, shared attention better captures fine-grained cross-modal interactions and enables more effective adaptive feature fusion. Moreover, removing the attention module leads to a noticeable performance drop, highlighting the critical role of the attention in modeling complex cross-modal complementarity.

Table 3: The clustering performance of scMAF with different feature fusion strategies for generating the representation z.

Fusion Strategy	Spleen			BMNC		
	NMI	ARI	ACC	NMI	ARI	ACC
Addition	60.34	29.28	40.80	79.42	63.28	64.04
Concatenation	62.79	37.68	45.65	79.12	60.77	64.28
[cls] only	60.55	30.97	43.64	80.36	68.07	68.76
Modality Gating	60.41	31.09	42.51	80.59	71.59	71.91
Modality-Specific Attention	62.62	34.12	45.52	80.64	70.97	71.87
Shared Attention (Ours)	**63.71**	**39.57**	**49.44**	**84.45**	**79.87**	**79.73**

Attention Optimization. We further evaluate two designs in our attention-based fusion module: *i)* whether to use a coarse-grained fusion head to generate the [cls] token instead of a learnable token initialized randomly; and *ii)* whether to share the decoder between the modality-specific and fusion branches. As shown in Table 4, replacing the fusion

Table 4: The clustering performance of scMAF with different attention optimizations.

Fusion Head [cls] Token	Individual Decoders	Spleen			BMNC		
		NMI	ARI	ACC	NMI	ARI	ACC
✗	✗	60.69	34.16	45.59	83.62	77.06	76.45
✗	✓	62.56	37.70	47.00	83.56	77.47	77.57
✓	✗	62.48	35.96	47.12	84.35	78.17	77.86
✓	✓	**63.71**	**39.57**	**49.44**	**84.45**	**79.87**	**79.73**

head with the randomly initialized [cls] token leads to consistent performance degradations. This indicates that providing the attention module with a shallow fused initialization offers a more informative and cell-specific starting point for learning cross-modal weighting. Furthermore, assigning independent decoders to the two reconstruction pathways yields additional gains, suggesting that dedicated reconstruction heads better preserve modality-specific structure and prevent interference between the branches.

Objective Functions. To evaluate the role of each loss term, we selectively enable *i)* the modality-specific reconstruction loss $\mathcal{L}_{rec}$, which encourages each omics to learn meaningful standalone representations; and *ii)* cross-modal fusion reconstruction loss $\mathcal{L}_{fusion}$, which ensures that the fused representation z effectively integrates multi-omics information. As Table 5 shows, using only $\mathcal{L}_{rec}$ leads to poor clustering, as no signal guides cross-modal interaction. Conversely, using only $\mathcal{L}_{fusion}$ yields substantially better results, since the fused embedding is directly optimized to capture shared and complementary information. When both loss terms are combined, scMAF achieves the best overall performance, highlighting the complementary roles of the two objectives.

Table 5: The performance of scMAF with different combinations of the loss terms.

$\mathcal{L}_{rec}$	$\mathcal{L}_{fusion}$	Spleen			BMNC		
		NMI	ARI	ACC	NMI	ARI	ACC
✔	✗	48.17	16.57	29.80	57.29	32.20	40.81
✗	✔	61.86	34.60	45.67	83.04	73.32	73.00
✔	✔	**63.71**	**39.57**	**49.44**	**84.45**	**79.87**	**79.73**

4.6 Generalization Analysis

While scMAF is primarily designed for multi-omics clustering, the proposed framework for fusing imbalanced modalities demonstrates a natural generalizability to broader multi-modal and multi-view pattern recognition scenarios. To validate this, we adapted our model to standard image-text multi-view datasets by substituting the original reconstruction objective with Mean Squared Error (MSE) losses. Benefiting from its superior capability in addressing modality imbalance, scMAF achieves state-of-the-art performance against recent multi-view clustering baselines, as detailed in Table 6. These results confirm its substantial value and potential as a highly adaptable framework for the broader multimodal representation learning field.

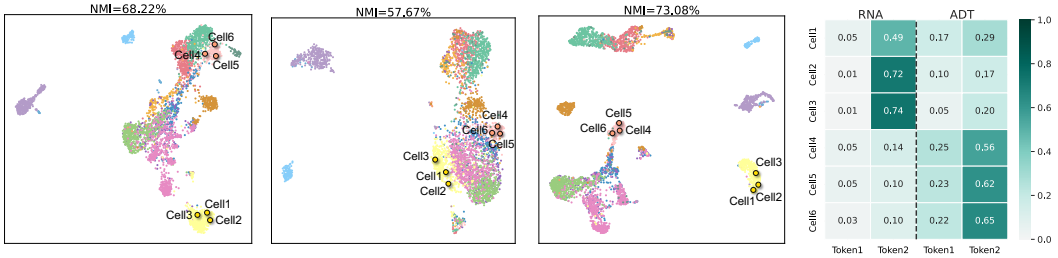
Table 6: Clustering performance on two image-text multi-view datasets.

Method	NUS-WIDE [23]			XMediaNet [21]		
	NMI	ARI	ACC	NMI	ARI	ACC
SURE [62]	46.01	38.60	56.30	82.92	25.29	51.39
DCP [36]	43.15	27.13	49.23	59.82	19.19	25.09
CVCL [9]	47.12	41.50	61.14	57.27	16.26	21.22
SCM-RE[41]	49.65	39.20	59.12	89.44	74.31	80.33
MVCAN[61]	49.24	40.34	63.03	89.36	74.20	79.62
scMAF(Ours)	**53.19**	**45.32**	**65.07**	**94.85**	**85.77**	**90.68**

4.7 Visualization

To understand how scMAF performs adaptive cross-modal fusion, we visualize the modality-specific features, the fused representation, and the corresponding attention weights on the *Spector* dataset in Fig. 5. Figs. 5a and 5b show the UMAP results derived separately from the RNA and ADT data. As can be seen, cells 1–3 of the same type are well gathered and separated from other types in the RNA modality, but suboptimally resolved in the ADT space, whereas cells 4–6 exhibit the opposite pattern. This highlights the complementary discriminative signals carried by the two modalities.

Fig. 5c presents the fused embedding produced by scMAF. Both groups of cells now form clear and coherent clusters, indicating that scMAF effectively integrates informative cues from each modality while suppressing modality-specific noise. To further interpret the behavior, Fig. 5d visualizes the attention weights from the [cls] token to all slice tokens. For cells 1-3, the [cls] token assigns higher weights to RNA slices, whereas for cells 4-6, the attention shifts toward ADT slices. This pattern is consistent with the separability observed in the modality-specific UMAPs and provides an interpretable explanation of how scMAF performs cell-wise, modality-aware fusion.



(a) RNA (b) ADT (c) scMAF (d) Attention Weights

Figure 5: Feature visualization on the *Spector* dataset, with the corresponding k-means clustering NMI annotated on the top. (a) UMAP of the RNA data; (b) UMAP of the ADT data; (c) UMAP of the fused features learned by scMAF; (d) Attention weights of the six highlighted cells.

5 Conclusion

In this work, we introduced scMAF, an adaptive fusion framework designed for unsupervised clustering of single-cell multi-omics data. scMAF learns modality-specific representations through distribution-aware autoencoders and integrates them using a tokenized self-attention mechanism that performs cell-wise, modality-aware fusion. This design enables the model to automatically extract informative signals from each omics while downweighting noisy or less discriminative components. Extensive experiments on diverse real-world datasets show that scMAF consistently outperforms state-of-the-art baselines, demonstrating its effectiveness in handling heterogeneous and complementary multi-omics signals. Moreover, the attention weights provide intuitive and interpretable insights into modality contributions at the cell level, helping reveal when and why certain modalities dominate the fusion process. In the future, we plan to explore incorporating biological priors into the fusion mechanism and extending scMAF to a wider range of downstream tasks, such as trajectory inference and multimodal integration across batches.

Acknowledgments

This work was supported in part by Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China under Grant JYB2025XDXM610; in part by the Fundamental Research Funds for the Central Universities under Grant CJ202303, 20826041K4049; in part by Sichuan Science and Technology Planning Project under Grant 2024NSFTD0038; in part by China National Postdoctoral Program for Innovative Talents under Grant BX20250392.

References

[1] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International conference on machine learning*, pages 1247–1255. PMLR, 2013.

[2] Ricard Argelaguet, Damien Arnol, Danila Bredikhin, Yonatan Deloro, Britta Velten, John C Marioni, and Oliver Stegle. Mofa+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome biology*, 21(1):111, 2020.

[3] David Arthur, Sergei Vassilvitskii, et al. k-means++: The advantages of careful seeding. In *Soda*, volume 7, pages 1027–1035, 2007.

[4] Jenna L Ballard, Zexuan Wang, Wenrui Li, Li Shen, and Qi Long. Deep learning-based approaches for multi-omics data integration and analysis. *BioData Mining*, 17(1):38, 2024.

[5] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[6] Jason D Buenrostro, Beijing Wu, Ulrike M Litzenburger, Dave Ruff, Michael L Gonzales, Michael P Snyder, Howard Y Chang, and William J Greenleaf. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523(7561):486–490, 2015.

[7] Zhi-Jie Cao and Ge Gao. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nature biotechnology*, 40(10):1458–1466, 2022.

[8] Fuqun Chen, Guanhua Zou, Yongxian Wu, and Le Ou-Yang. Clustering single-cell multi-omics data via graph regularized multi-view ensemble learning. *Bioinformatics*, 40(4):btae169, 2024.

[9] Jie Chen, Hua Mao, Wai Lok Woo, and Xi Peng. Deep multiview clustering by contrasting cluster assignments. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16752–16761, 2023.

[10] Man-Sheng Chen, Chang-Dong Wang, and Jian-Huang Lai. Low-rank tensor based proximity learning for multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 35(5):5076–5090, 2022.

[11] Man-Sheng Chen, Xi-Ran Zhu, Jia-Qi Lin, and Chang-Dong Wang. Contrastive multiview attribute graph clustering with adaptive encoders. *IEEE Transactions on Neural Networks and Learning Systems*, 36(4):7184–7195, 2024.

[12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.

[13] Maria Colomé-Tatché and Fabian J Theis. Statistical single cell multi-omics integration. *Current Opinion in Systems Biology*, 7:54–59, 2018.

[14] Amir Ebrahimi, Alireza Fotuhi Siahpirani, and Hesam Montazeri. scin: a contrastive learning framework for single-cell multi-omics data integration. *Briefings in Bioinformatics*, 26(4):bbaf411, 2025.

[15] Gökcen Eraslan, Lukas M Simon, Maria Mircea, Nikola S Mueller, and Fabian J Theis. Single-cell rna-seq denoising using a deep count autoencoder. *Nature communications*, 10(1):390, 2019.

[16] Uno Fang, Man Li, Jianxin Li, Longxiang Gao, Tao Jia, and Yanchun Zhang. A comprehensive survey on multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 35(12):12350–12368, 2023.

[17] Yanglan Gan, Xingyu Huang, Guobing Zou, Shuigeng Zhou, and Jihong Guan. Deep structural clustering for single-cell rna-seq data jointly through autoencoder and graph neural network. *Briefings in Bioinformatics*, 23(2):bbac018, 2022.

[18] Adam Gayoso, Zoë Steier, Romain Lopez, Jeffrey Regier, Kristopher L Nazor, Aaron Streets, and Nir Yosef. Joint probabilistic modeling of single-cell multi-omic data with totalvi. *Nature methods*, 18(3):272–282, 2021.

[19] Christopher Heje Grønbech, Maximillian Fornitz Vording, Pascal N Timshel, Casper Kaae Sønderby, Tune H Pers, and Ole Winther. scvae: variational auto-encoders for single-cell gene expression data. *Bioinformatics*, 36(16):4415–4422, 2020.

[20] Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M Mauck, Shiwei Zheng, Andrew Butler, Maddie J Lee, Aaron J Wilk, Charlotte Darby, Michael Zager, et al. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587, 2021.

[21] Xiangteng He, Yuxin Peng, and Liu Xie. A new benchmark and approach for fine-grained cross-media retrieval. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1740–1748, 2019.

[22] Dayu Hu, Zhibin Dong, Ke Liang, Hao Yu, Siwei Wang, and Xinwang Liu. High-order topology for deep single-cell multiview fuzzy clustering. *IEEE Transactions on Fuzzy Systems*, 32(8):4448–4459, 2024.

[23] Peng Hu, Liangli Zhen, Dezhong Peng, and Pei Liu. Scalable deep multimodal learning for cross-modal retrieval. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 635–644, 2019.

[24] Dragomirka Jovic, Xue Liang, Hua Zeng, Lin Lin, Fengping Xu, and Yonglun Luo. Single-cell rna sequencing technologies and applications: a brief overview. *Clinical and translational medicine*, 12(3):e694, 2022.

[25] Yukie Kashima, Yoshitaka Sakamoto, Keiya Kaneko, Masahide Seki, Yutaka Suzuki, and Ayako Suzuki. Single-cell sequencing techniques from individual to multiomics analyses. *Experimental & Molecular Medicine*, 52(9):1419–1427, 2020.

[26] Aleksandra A Kolodziejczyk, Jong Kyoung Kim, Valentine Svensson, John C Marioni, and Sarah A Teichmann. The technology and biology of single-cell rna sequencing. *Molecular cell*, 58(4):610–620, 2015.

[27] Gaoyang Li, Shaliu Fu, Shuguang Wang, Chenyu Zhu, Bin Duan, Chen Tang, Xiaohan Chen, Guohui Chuai, Ping Wang, and Qi Liu. A deep generative model for multi-view profiling of single-cell rna-seq and atac-seq data. *Genome biology*, 23(1):20, 2022.

[28] Yunfan Li, Peng Hu, Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng. Contrastive clustering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 8547–8555, 2021.

[29] Yunfan Li, Hancong Li, Yijie Lin, Dan Zhang, Dezhong Peng, Xiting Liu, Jie Xie, Peng Hu, Lu Chen, Han Luo, et al. Metaq: fast, scalable and accurate metacell inference via single-cell quantization. *Nature Communications*, 16(1):1205, 2025.

[30] Yunfan Li, Yijie Lin, Peng Hu, Dezhong Peng, Han Luo, and Xi Peng. Single-cell rna-seq debiased clustering via batch effect disentanglement. *IEEE Transactions on Neural Networks and Learning Systems*, 35(8):11371–11381, 2023.

[31] Yunfan Li, Dan Zhang, Mouxing Yang, Dezhong Peng, Jun Yu, Yu Liu, Jiancheng Lv, Lu Chen, and Xi Peng. scbridge embraces cell heterogeneity in single-cell rna-seq and atac-seq data integration. *Nature communications*, 14(1):6045, 2023.

[32] Ke Liang, Lingyuan Meng, Hao Li, Meng Liu, Siwei Wang, Sihang Zhou, Xinwang Liu, and Kunlun He. Mgksite: Multi-modal knowledge-driven site selection via intra and inter-modal graph fusion. *IEEE Transactions on Multimedia*, 27:1722–1735, 2024.

[33] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM computing surveys*, 56(10):1–42, 2024.

[34] Jongsu Lim, Chanho Park, Minjae Kim, Hyukhee Kim, Junil Kim, and Dong-Sung Lee. Advances in single-cell omics and multiomics for high-resolution molecular profiling. *Experimental & molecular medicine*, 56(3):515–526, 2024.

[35] Xiang Lin, Tian Tian, Zhi Wei, and Hakon Hakonarson. Clustering of single-cell multi-omics data with a multimodal deep learning method. *Nature communications*, 13(1):7705, 2022.

[36] Yijie Lin, Yuanbiao Gou, Xiaotian Liu, Jinfeng Bai, Jiancheng Lv, and Xi Peng. Dual contrastive prediction for incomplete multi-view representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4447–4461, 2022.

[37] Meng Liu, Ke Liang, Hao Yu, Lingyuan Meng, Siwei Wang, Sihang Zhou, and Xinwang Liu. Multiview temporal graph clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

[38] Suyuan Liu, Siwei Wang, Ke Liang, Xueling Zhu, Jinyan Li, Xinwang Liu, and Xiaohong Chen. Propagating cross-view semantics for multi-view clustering: A unified anchor refinement paradigm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

[39] Suyuan Liu, Lijun Zhang, Siwei Wang, Miaomiao Li, Xueling Zhu, and Xinwang Liu. Towards scalable multi-view clustering without predefined clusters via inter-anchor graph learning. *IEEE Transactions on Knowledge and Data Engineering*, 2026.

[40] Yiding Lu, Yijie Lin, Mouxing Yang, Dezhong Peng, Peng Hu, and Xi Peng. Decoupled contrastive multi-view clustering with high-order random walks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 14193–14201, 2024.

[41] Caixuan Luo, Jie Xu, Yazhou Ren, Junbo Ma, and Xiaofeng Zhu. Simple contrastive multi-view clustering with data-level fusion. In *IJCAI*, pages 4697–4705, 2024.

[42] Laura D Martens, David S Fischer, Vicente A Yépez, Fabian J Theis, and Julien Gagneur. Modeling fragment counts improves single-cell atac-seq analysis. *Nature Methods*, 21(1):28–31, 2024.

[43] Andrew Ng, Michael Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 14, 2001.

[44] Anil Rahate, Rahee Walambe, Sheela Ramanna, and Ketan Kotecha. Multimodal co-learning: Challenges, applications with datasets, recent advances and future directions. *Information Fusion*, 81:203–239, 2022.

[45] Pia Rautenstrauch and Uwe Ohler. Liam tackles complex multimodal single-cell data integration challenges. *Nucleic Acids Research*, 52(12):e52–e52, 2024.

[46] Liangrui Ren, Jun Wang, Zhao Li, Qingzhong Li, and Guoxian Yu. scmcs: a framework for single-cell multi-omics data integration and multiple clusterings. *Bioinformatics*, 39(4):btad133, 2023.

[47] Martin Ridout, Clarice GB Demétrio, and John Hinde. Models for count data with many zeros. In *Proceedings of the XIXth international biometric conference*, volume 19, pages 179–192. International Biometric Society Invited Papers Cape Town, SA, 1998.

[48] Ehud Shapiro, Tamir Biezuner, and Sten Linnarsson. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews Genetics*, 14(9):618–630, 2013.

[49] Yinan Shi, Yanchi Su, Yue Cheng, Ka-Chun Wong, Yunhe Wang, and Xiangtao Li. Dynamic synthesis of multi-modal representations for cite-seq data integration and analysis. *Advanced Science*, 12(44):e09247, 2025.

[50] Binyang Song, Rui Zhou, and Faez Ahmed. Multi-modal machine learning in engineering design: A review and future directions. *Journal of Computing and Information Science in Engineering*, 24(1):010801, 2024.

[51] Sören Richard Stahlschmidt, Benjamin Ulfenborg, and Jane Synnergren. Multimodal deep learning for biomedical data fusion: a review. *Briefings in bioinformatics*, 23(2):bbab569, 2022.

[52] Stefan Stanojevic, Yijun Li, Aleksandar Ristivojevic, and Lana X Garmire. Computational methods for single-cell multi-omics integration and alignment. *Genomics, Proteomics & Bioinformatics*, 20(5):836–849, 2022.

[53] Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nature methods*, 14(9):865–868, 2017.

[54] Tim Stuart, Avi Srivastava, Shaista Madad, Caleb A Lareau, and Rahul Satija. Single-cell chromatin state analysis with signac. *Nature methods*, 18(11):1333–1341, 2021.

[55] Indhupriya Subramanian, Srikant Verma, Shiva Kumar, Abhay Jere, and Krishanpal Anamika. Multi-omics data integration, interpretation, and its application. *Bioinformatics and biology insights*, 14:1177932219899051, 2020.

[56] Shiliang Sun. A survey of multi-view machine learning. *Neural computing and applications*, 23(7):2031–2038, 2013.

[57] Wenxuan Tu, Sihang Zhou, Xinwang Liu, Xifeng Guo, Zhiping Cai, En Zhu, and Jieren Cheng. Deep fusion clustering network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 9978–9987, 2021.

[58] Xinjun Wang, Zhe Sun, Yanfu Zhang, Zhongli Xu, Hongyi Xin, Heng Huang, Richard H Duerr, Kong Chen, Ying Ding, and Wei Chen. Brem-sc: a bayesian random effects mixture model for joint clustering single cell multi-omics data. *Nucleic acids research*, 48(11):5814–5824, 2020.

[59] Zile Wang, Fengyu Lei, Xiaoping Shi, Jianping Zhao, and Junfeng Xia. scdmsc: deep multi-view subspace clustering for single-cell multi-omics data. *IEEE Journal of Biomedical and Health Informatics*, 2025.

[60] Xiangyu Wu, Xin Yang, Yunhan Dai, Zihan Zhao, Junmeng Zhu, Hongqian Guo, and Rong Yang. Single-cell sequencing to multi-omics: technologies and applications. *Biomarker research*, 12(1):110, 2024.

[61] Jie Xu, Yazhou Ren, Xiaolong Wang, Lei Feng, Zheng Zhang, Gang Niu, and Xiaofeng Zhu. Investigating and mitigating the side effects of noisy views for self-supervised clustering algorithms in practical multi-view scenarios. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22957–22966, 2024.

[62] Mouxing Yang, Yunfan Li, Peng Hu, Jinfeng Bai, Jiancheng Lv, and Xi Peng. Robust multi-view clustering with incomplete information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):1055–1069, 2022.

[63] Yan Yang and Hao Wang. Multi-view clustering: A survey. *Big data mining and analytics*, 1(2):83–107, 2018.

[64] Zhuohan Yu, Yifu Lu, Yunhe Wang, Fan Tang, Ka-Chun Wong, and Xiangtao Li. Zinb-based graph embedding autoencoder for single-cell rna-seq interpretations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 4671–4679, 2022.

[65] Musu Yuan, Liang Chen, and Minghua Deng. Clustering cite-seq data with a canonical correlation-based deep learning method. *Frontiers in Genetics*, 13:977968, 2022.

[66] Musu Yuan, Liang Chen, and Minghua Deng. Clustering single-cell multi-omics data with moclust. *Bioinformatics*, 39(1):btac736, 2023.

[67] Youlin Zhan, Jiahan Liu, and Le Ou-Yang. scmic: A deep multi-level information fusion framework for clustering single-cell multi-omics data. *IEEE Journal of Biomedical and Health Informatics*, 27(12):6121–6132, 2023.

[68] Changqing Zhang, Yajie Cui, Zongbo Han, Joey Tianyi Zhou, Huazhu Fu, and Qinghua Hu. Deep partial multi-view learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2402–2415, 2020.

[69] Grace XY Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, et al. Massively parallel digital transcriptional profiling of single cells. *Nature communications*, 8(1):14049, 2017.

[70] Chunman Zuo and Luonan Chen. Deep-joint-learning analysis model of single cell transcriptome and open chromatin accessibility data. *Briefings in Bioinformatics*, 22(4):bbaa287, 2021.

[71] Chunman Zuo, Hao Dai, and Luonan Chen. Deep cross-omics cycle attention model for joint analysis of single-cell multi-omics data. *Bioinformatics*, 37(22):4091–4099, 2021.

A Appendix

In this appendix, we first present the probabilistic reconstruction objectives adopted in scMAF, including the ZINB loss for RNA and ADT data and the Poisson loss for ATAC data. We then provide additional analysis of different reconstruction targets. Finally, we discuss the limitations of this work.

A.1 ZINB Distribution Loss

To derive the reconstruction objective used in our model, we begin by recalling the standard Negative Binomial (NB) distribution [19, 29, 15] and then extend it to the Zero-Inflated Negative Binomial (ZINB) [47, 19, 30, 35, 15] formulation commonly used for modeling over-dispersed count data with excessive zeros.

Definition 1. *The probability mass function of the Negative Binomial (NB) distribution is defined as*

$$\mathrm{NB}(x \mid r, p) = \binom{x+r-1}{r-1} p^r (1-p)^x, \tag{11}$$

where r denotes the target number of successes in a sequence of i.i.d. Bernoulli(p) trials, so that x represents the number of failures observed before the r-th success; p is the per-trial success probability.

To express the distribution in a form that is more amenable to continuous optimization, the combinatorial term can be rewritten using the Gamma function:

$$\binom{x+r-1}{r-1} = \frac{(x+r-1)!}{x!(r-1)!} = \frac{\Gamma(x+r)}{x!\,\Gamma(r)}, \tag{12}$$

where $\Gamma(n) = (n-1)!$ denotes the Gamma function for positive integers. Substituting this identity into the probability mass function yields the equivalent representation

$$\mathrm{NB}(x \mid r, p) = \frac{\Gamma(x+r)}{x!\,\Gamma(r)}\, p^r (1-p)^x. \tag{13}$$

Next, we compute the expectation of the NB distribution, which will later facilitate reparameterization in terms of the mean. Using the definition of expectation, we have

$$\begin{aligned}\mu &= \sum_{x=0}^{\infty} x \frac{\Gamma(x+r)}{x!\Gamma(r)} p^r (1-p)^x \\ &= r\frac{(1-p)}{p} \sum_{x=1}^{\infty} \frac{\Gamma(x-1+r+1)}{(x-1)!\Gamma(r+1)} p^{r+1}(1-p)^{x-1} \\ &= r\frac{(1-p)}{p}.\end{aligned} \tag{14}$$

This confirms that the NB mean is controlled by both r and p. Therefore, expressing the distribution using (μ, r) instead of (p, r) provides a more interpretable and numerically convenient parameterization.

Definition 2. *The Zero-Inflated Negative Binomial (ZINB) distribution introduces an additional zero-inflation parameter $\pi \in [0, 1]$ to explicitly model the probability of observing extra zeros:*

$$\mathrm{ZINB}(x \mid \pi, r, p) = \pi\, \delta_0(x) + (1-\pi)\, \mathrm{NB}(x \mid r, p), \tag{15}$$

where $\delta_0(x)$ is the Dirac delta function that places all probability mass at $x = 0$.

In practice, it is often convenient to reparameterize the NB component using the mean μ in Eq. 14, leading to

$$\begin{aligned}\mathrm{ZINB}(x_i \mid \pi_i, r_i, \mu_i) = \pi_i \delta_0(x_i) + (1-\pi_i)\frac{\Gamma(x_i + r_i)}{x_i!\Gamma(r_i)} \\ \times \left(\frac{r_i}{r_i + \mu_i}\right)^{r_i} \left(\frac{\mu_i}{r_i + \mu_i}\right)^{x_i}.\end{aligned} \tag{16}$$

The ZINB likelihood is then used as the reconstruction objective in the decoder. Given a set of observed counts $\{x_i\}_{i=1}^n$, the ZINB loss is defined as the average negative log-likelihood:

$$\mathcal{L}_{\mathrm{ZINB}}(x \mid \pi, r, \mu) = \frac{1}{n} \sum_{i=1}^{n} \Big[-\log \mathrm{ZINB}(x_i \mid \pi_i, r_i, \mu_i) \Big]. \tag{17}$$

Finally, to model the heterogeneity across individual cells, three independent fully connected layers W_π, W_r, and W_μ are applied to the feature map d_i output by the final hidden layer of the decoder to estimate

$$\pi_i = \mathrm{sigmoid}(W_\pi d_i), \tag{18}$$
$$r_i = \exp(W_r d_i), \tag{19}$$
$$\mu_i = \mathrm{diag}(s_i) \times \exp(W_\mu d_i), \tag{20}$$

where s_i denotes the size factor computed during data preprocessing.

A.2 Poisson Distribution Loss

We present the Poisson distribution and its corresponding negative log-likelihood, which is used as the reconstruction objective to model ATAC count data [42, 27, 29].

Definition 3. *The probability mass function of the Poisson distribution is defined as*

$$\mathrm{Pois}(x \mid \lambda) = \frac{e^{-\lambda}\lambda^x}{x!}, \tag{21}$$

where λ denotes the expected number of events in the interval, and x represents the actual number of events observed therein.

Given a set of observed counts $\{x_i\}_{i=1}^n$, the corresponding Poisson negative log-likelihood is defined as

$$\begin{aligned}\mathcal{L}_{\text{Pois}}(x \mid \lambda) &= \frac{1}{n}\sum_{i=1}^{n}\Big[-\log \text{Pois}(x_i \mid \lambda_i)\Big] \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[-\log \frac{e^{-\lambda_i}\lambda_i^{x_i}}{x_i!}\right].\end{aligned} \tag{22}$$

Finally, the Poisson mean λ_i is estimated by applying a fully connected layer W_λ to the latent representation z_i:

$$\lambda_i = \text{diag}(s_i) \times \exp(W_\lambda z_i), \tag{23}$$

where s_i denotes the size factor computed during data preprocessing. As shown in Table 7, modeling chromatin accessibility data with the Poisson distribution outperforms using the MSE or ZINB as the reconstruction target.

Table 7: Clustering performance of different ATAC reconstruction targets on three single-cell multi-omics datasets that contain ATAC modality. The last column reports the average performance across all datasets within each row.

ATAC Loss	Ma2020A			Ma2020B			PBMC			Average
	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	
MSE	59.85	42.94	57.27	48.28	39.77	55.20	67.49	52.19	63.86	54.09
ZINB	77.29	60.40	66.77	69.46	45.84	59.50	70.14	57.74	73.23	64.49
Poisson	**78.77**	**63.28**	**70.72**	**71.68**	**51.53**	**62.57**	**71.59**	**58.91**	**75.23**	**67.14**

A.3 Reconstruction Target

We further examine how different reconstruction losses affect clustering performance. For RNA and ADT, we compare the ZINB loss with the MSE loss, while for ATAC, we compare the Poisson loss with the MSE loss. We report the average performance of all eight datasets on three evaluation metrics. Table 8 reports the average performance across all eight datasets on three evaluation metrics. As shown, modeling RNA and ADT with ZINB consistently outperforms MSE, indicating that the ZINB distribution more accurately reflects the over-dispersed and zero-inflated characteristics of these modalities. For ATAC, the Poisson-based reconstruction achieves superior performance, likely due to its compatibility with the sparse count-based nature of chromatin accessibility measurements.

Table 8: Average clustering performance across all datasets using different reconstruction distributions per modality.

RNA	ADT		ATAC	
	ZINB	MSE	Poisson	MSE
ZINB	**68.22**	66.87	**66.91**	51.63
MSE	67.16	64.81	62.82	46.34

A.4 Limitations

Although scMAF achieves promising clustering performance on multiple commonly used single-cell multi-omics benchmark datasets, there are still several directions worth further exploration. First, this work mainly focuses on unsupervised clustering, and its applicability to broader downstream analyses, such as trajectory inference and multimodal integration across batches, remains to be investigated. Second, the effectiveness and scalability of scMAF on large-scale single-cell datasets have not yet been fully validated. In addition, incorporating biological priors into the fusion mechanism may further improve its interpretability and biological relevance. We leave these directions for future work.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the claims made.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations are discussed in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The assumptions and complete proofs for all theoretical results are provided in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The implementation and training details are clearly described for reproduction.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code and data will be released publicly upon acceptance.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The experiment settings are clearly presented in the paper and appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We report average performance over 20 k-means runs with different random seeds to reduce stochasticity, but do not include error bars or statistical significance tests in the current version.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The compute resources are reported in the implementation details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conform with the NeurIPS Code of Ethics in every respect.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Proper citations are provided throughout the document and the licenses will be included with the code when it is released.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The document will accompany the code upon its release.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method and experiments in this paper do not involve the use of LLMs.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.