Deep Multiple Clustering

From Foundations to Context-Aware Approaches

KDD'26 Lecture-Style Tutorial | August 10, 2026 | 9:00 AM–12:00 PM (KST)

International Convention Center Jeju (ICC Jeju)

224 Jungmungwangwang-ro, Seogwipo-si, Jeju-do, Republic of Korea

representation · multi-view · context · evaluation

Bangyu Zou | Jiawei Yao | Huiji Gao | Qi Qian | Jian Pei | Juhua Hu

Organizers

Dr. Juhua Hu
Associate Professor
University of Washington

Dr. Jian Pei
Arthur S. Pearse Distinguished Professor
Duke University

Dr. Qi Qian
AI Research Scientist
Meta

Dr. Huiji Gao
Senior Engineering Manager
Airbnb

Dr. Jiawei Yao
Machine Learning Engineer
Airbnb

Mr. Bangyu Zou
M.S. Student
University of California, Irvine

Roadmap

Part I **Motivation and Problem Formulation**
why multiple valid clusterings exist

Part II **Deep Representation-Based Methods**
facets, dimensions, attention heads, and disentangled factors
Coffee Break 9:30–10:00

Part III **Multi-View / Multi-Source Methods**
shared semantics, private cues, and incomplete views

Part IV **Context-Aware / User-Guided Methods**
prompts, proxies, language criteria, and agents

Part V **Comparisons, Applications, and Open Questions**
evaluation, method choice, and future directions

Part I

Motivation & Problem Formulation

Why one dataset can support several valid clusterings.

Traditional Single Clustering

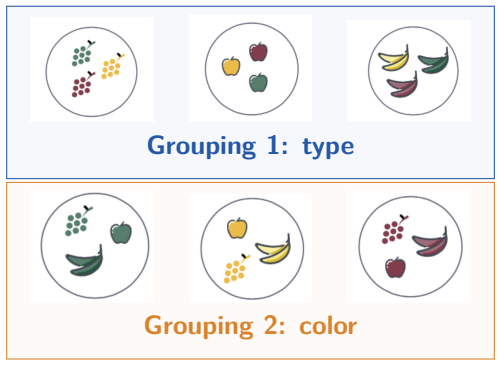
images $\mathcal{X}$

one partition $\mathcal{C}$

$$\min_{\mathcal{C}} \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|_2^2$$

NP-hard objectives are usually optimized approximately, so single clustering gives the user one local optimum.

Multiple Structures Hidden in the Data: Vision



Same fruit collection $\mathcal{X}$

One set of objects; change the similarity criterion, and the partition changes.

Multiple Structures Hidden in the Data: Text

The same four documents

D1	D2	D3	D4
team wins	team loses	patient recovers	outbreak worsens

Grouping 1: topic

Sports:	D1	D2
Health:	D3	D4

Grouping 2: tone

Positive:	D1	D3
Negative:	D2	D4

Topic words and sentiment words define two different notions of textual similarity.

Multiple Structures Hidden in the Data: Biology

Same samples, different informative gene programs

S1 T / baseline | **S2** T / inflamed | **S3** B / baseline | **S4** B / inflamed

schematic feature activity

Bright cells indicate an active gene program.

Grouping 1: cell identity

T: S1, S2 | B: S3, S4

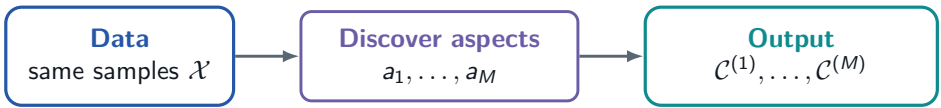
Grouping 2: biological state

Base: S1, S3 | Inflamed: S2, S4

Visual logic adapted from the phenotype-finding example in Hu et al., *Finding Multiple Stable Clusterings*, ICDM 2015 tutorial slides.

From One Dataset to Multiple Partitions

Input: one unlabeled dataset $\mathcal{X} = \{x_1, \ldots, x_N\}$



One partition per aspect

Non-redundant outputs

$\mathcal{C}^{(p)} = \{C_1^{(p)}, \ldots, C_{K_p}^{(p)}\}, \qquad p = 1, \ldots, M. \qquad \mathcal{C}^{(p)} \not\equiv \mathcal{C}^{(q)} \qquad \text{for } p \neq q.$

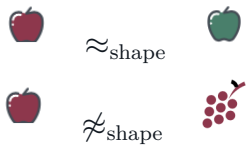
Goal: every partition should be meaningful for $\mathcal{X}$ and reveal a different organization of the samples.

Main Challenge: How Do We Discover Different Aspects?

$$a \longmapsto s_a(x_i, x_j) \longmapsto \mathcal{C}^{(a)}$$

Aspect: shape

Object identity defines similarity.



Aspect: color

Visual color defines similarity.

choose or learn criterion a → build similarity $s_a(x_i, x_j)$ → keep it if quality high → reject if redundant

The hard part is finding criteria that are meaningful, useful, and non-redundant.

Multiple Clustering Objective

$$\mathcal{S} = \{\mathcal{C}^{(1)}, \mathcal{C}^{(2)}, \ldots, \mathcal{C}^{(M)}\}, \qquad \mathcal{C}^{(p)} = \{C_1^{(p)}, \ldots, C_{K_p}^{(p)}\}$$

$$\max_{\mathcal{S}} \underbrace{\sum_{p=1}^{M} Q\left(\mathcal{C}^{(p)}\right)}_{\text{quality}} - \lambda \underbrace{\sum_{1 \le p < q \le M} R\left(\mathcal{C}^{(p)}, \mathcal{C}^{(q)}\right)}_{\text{redundancy penalty}}$$

$Q(\mathcal{C})$: quality
compact, well-separated, or likelihood-based clustering structure.

$R(\mathcal{C}^{(p)}, \mathcal{C}^{(q)})$: redundancy
similarity between two returned partitions; high values mean duplicate answers.

λ: trade-off
controls whether the method prioritizes cluster quality or diversity.

Each $\mathcal{C}^{(p)}$ should cover $\mathcal{X}$, use disjoint clusters, and reveal a different valid aspect.

From Traditional to Deep Multiple Clustering (DMC)

COALA / alternative clustering
constrain a new clustering to be dissimilar to a reference result.

Bae & Bailey, ICDM 2006.

Side-information methods
penalize similarity to a known grouping or user-provided constraints.

Gondek & Hofmann, 2004.

Meta clustering
generate many candidate partitions, then group the partitions.

Caruana et al., ICDM 2006.

Orthogonal subspace search
look for another projection where a different partition appears.

Jain, 2010.

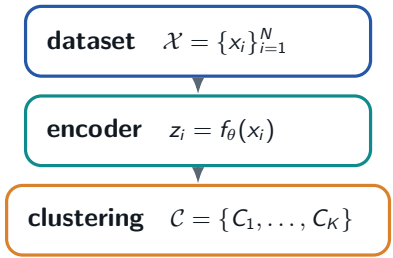
Summary: traditional methods search for alternatives in a fixed feature space; DMC learns the representation where alternatives become more visible.

Why Deep Representations Matter

Evidence from deep single clustering

Deep clustering learns $z_i = f_\theta(x_i)$, making cluster structure more visible than in raw or hand-crafted features.

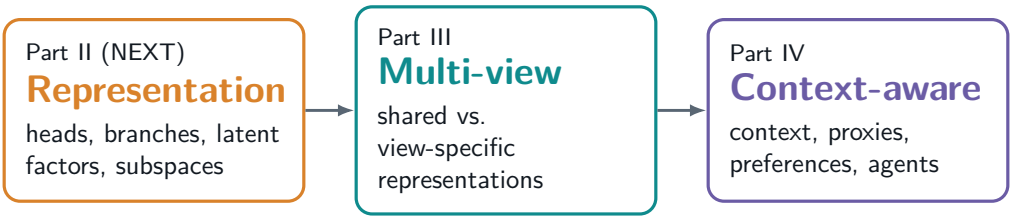
Next step: expose several aspects, not just one partition.



Deep representations expose latent structure that raw features can obscure.

Representative deep clustering evidence: DEC, Xie et al., ICML 2016; IDEC, Guo et al., IJCAI 2017; DeepCluster, Caron et al., ECCV 2018.

Handoff: three routes to useful alternatives



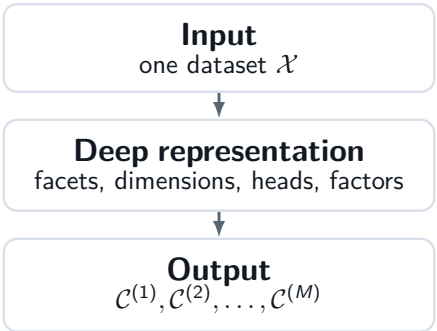
Q&A ?

Part II

Deep Representation-Based Multiple Clustering

Where does diversity live inside a deep model?

Start With The Right Problem



Not generic deep clustering

A standard deep clustering model learns one representation for one partition.

Deep multiple clustering

The model must encode diversity and prevent duplicate partitions.

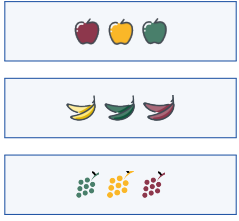
Key design question: **where can the diversity be encoded?**

One Dataset, Several Latent Factors

same samples
different grouping
criteria

Color

Shape



Nutrition and more

.
.
.

Training Objective: Representation + Clustering + Diversity

$$\mathcal{L} = \mathcal{L}_{\text{rep}} + \sum_{m=1}^{M} \mathcal{L}_{\text{clust}}^{(m)} + \lambda \mathcal{L}_{\text{div}}$$

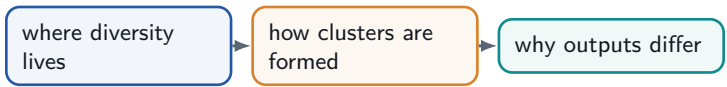
Representation
encode x into more informative features z

Clustering
produce $\mathcal{C}^{(1)}, \ldots, \mathcal{C}^{(M)}$

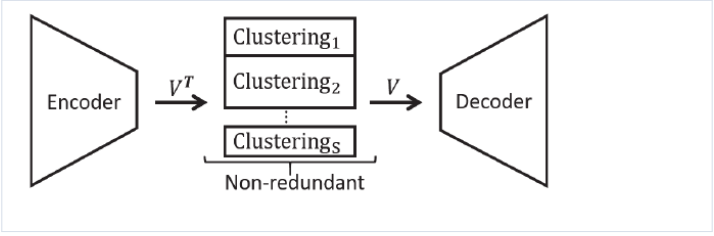
Redundancy
force solutions to reflect different aspects

Methods: What Changes?

Method	Diversity carrier	Clusters formed by	Key mechanism
ENRC [Miklautz et al., 2020]	rotated coordinates with solution weights β_s	nearest center under $\|V^\top z - \mu_{s,k}\|^2_{\beta_s}$	softmax over solutions assigns each latent dimension mostly to one clustering
MFCVAE [Falck et al., 2021]	latent facets z_j with separate mixture priors	$\arg\max_c q(c_j = c \mid x)$ inside each facet	reconstruction keeps facets useful; mixture priors make every facet clusterable
iMClusts [Ren et al., 2023]	attention head outputs O_h	k-means on each trained head output O_h	HSIC reduces head overlap; optional constraints guide heads
DDMC [Yao and Hu, 2024]	augmented paths X^k and disentangled latents Z^k	EM M-step assignment s_i^k by nearest center	E-step learns factors; M-step feeds cluster labels back into representation



ENRC: From Pixels to S Partitions



$$x_i \xrightarrow{\text{encode } f_\theta} z_i \xrightarrow{\text{align } V^\top} y_i \xrightarrow{\text{distance } (\beta_s,\mu_{s,k})} D_s(i,k) \xrightarrow{\arg\min_k} c_s(x_i), \quad s = 1,\dots,S$$

$z_i = f_\theta(x_i)$
shared latent evidence

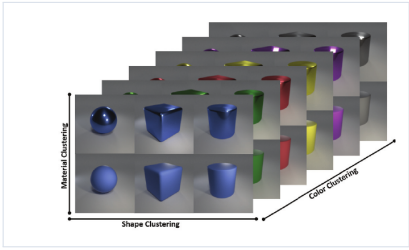
$y_i = V^\top z_i$
rotated semantic axes

$D_s(i,k)$
β_s reads different axes

$c_s(x_i)$
nearest center = label

Miklautz et al., *Deep Embedded Non-Redundant Clustering*, AAAI 2020, Fig. 2 and Sec. 2.

ENRC Example: One Object, Three Labels



x_i = blue metal sphere

One image contains color, material, and shape.

$$x_i \xrightarrow{\text{encoder } f_\theta} z_i \xrightarrow{\text{rotation } V^\top} y_i$$

y_i: the same evidence, arranged along learned axes

$y_i =$ | color axes | material axes | shape axes | other |

β_{color} : color axes $\Rightarrow$ blue

β_{material} : material axes $\Rightarrow$ metal

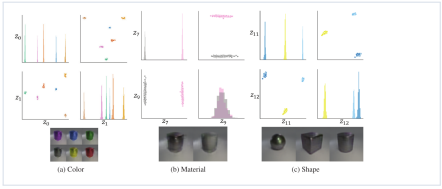
β_{shape} : shape axes $\Rightarrow$ sphere

$D_s(i,k)$ = distance from sample i to center k when solution s reads y_i

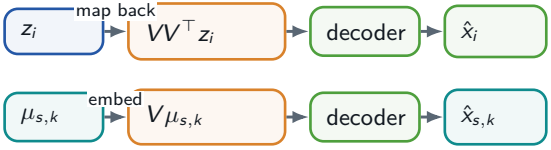
$$D_s(i,k) = \sum_d \beta_s[d]\Big(y_i[d] - \mu_{s,k}[d]\Big)^2, \qquad c_s(x_i) = \arg\min_k D_s(i,k)$$

Miklautz et al., *Deep Embedded Non-Redundant Clustering*, AAAI 2020, Figs. 1 and 3; weighted assignment: Eqs. (2)–(5).

ENRC: Decode Centers and Train Both Goals



Decoded centers reveal which factor a partition captured.



Upper path: reconstruct sample x_i. Lower path: visualize center $\mu_{s,k}$.

$$\min \mathcal{L} = \underbrace{\mathcal{L}_{\text{comp}}}_{\text{tight clusters}} + \lambda \underbrace{\mathcal{L}_{\text{rec}}}_{\text{retain input}}$$

Compact each solution

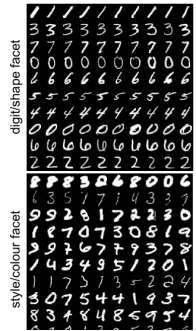
$$\frac{1}{S}\sum_{s,k}\frac{1}{|C_{s,k}|}\sum_{i\in C_{s,k}} D_s(i,k)$$

Reconstruct each sample

$$\sum_i \|x_i - \hat{x}_i\|_2^2$$

Miklautz et al., *Deep Embedded Non-Redundant Clustering*, AAAI 2020, Figs. 3–4 and Eqs. (1)–(9).

MFCVAE: One Image, Several Meaningful Facets



Facet 1: digit / shape
Images are grouped by identity: zeros with zeros, threes with threes, and so on.

Facet 2: style / stroke
The same images are regrouped by how they are written: thickness, slant, and shape style.

Intuition: the decoder needs both facets to reconstruct an image, while each mixture prior turns one factor into a partition.

Falck et al., *Multi-Facet Clustering Variational Autoencoders*, NeurIPS 2021.

MFCVAE: Each Facet Has Its Own Mixture

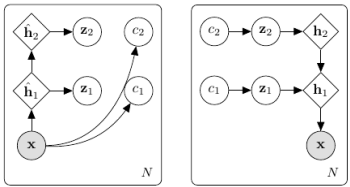


image x_i → ladder encoder → facet posteriors $q(z_j \mid x_i)$ → mixture responsibilities $q(c_j \mid x_i)$ → partitions $\mathcal{C}^{(1)}, \ldots, \mathcal{C}^{(J)}$

Representation A hierarchy of continuous latents models several generative factors.

Clustering Each facet has a categorical variable and a Gaussian-mixture prior.

Output Every facet directly produces one clustering of the same samples.

Falck et al., *Multi-Facet Clustering Variational Autoencoders*, NeurIPS 2021, Fig. 3.

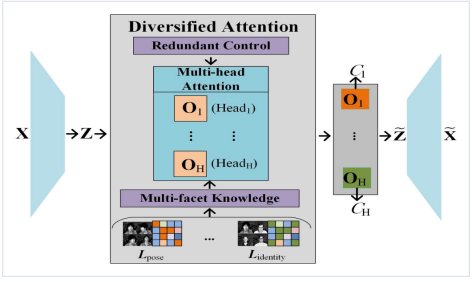
MFCVAE: The Posterior Gives Each Partition

Joint training objective

$$\mathcal{L}^{\text{MFCVAE}}(\mathcal{D};\theta,\phi) = \mathbb{E}_{x\sim\mathcal{D}}\Bigg[\underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|x)}\log p_\theta(x|\mathbf{z})}_{\text{reconstruct } x} - \sum_{j=1}^{J}\Bigg(\underbrace{\mathbb{E}_{q_\phi(c_j|x)}\mathrm{KL}\left(q_\phi(z_j|x)\parallel p_\theta(z_j|c_j)\right)}_{\text{fit facet } j \text{ to mixture}} + \underbrace{\mathrm{KL}\left(q_\phi(c_j|x)\parallel p(c_j)\right)}_{\text{regularize cluster assignments}}\Bigg)\Bigg]$$

Falck et al., *Multi-Facet Clustering Variational Autoencoders*, NeurIPS 2021.

iMClusts: From One Head to One Clustering



$$X \xrightarrow{\text{autoencoder } f_u} Z \xrightarrow{ZW_h^Q,\ ZW_h^K,\ ZW_h^V} (Q_h, K_h, V_h) \xrightarrow{\text{softmax}(Q_h K_h^\top/\sqrt{m})} A_h$$

$$(A_h, V_h) \xrightarrow{\text{aggregate } A_h V_h} O_h \xrightarrow[\text{after training}]{k\text{-means}} C_h$$

Z shared features A_h $N \times N$ neighbor weights O_h head representation C_h cluster labels

Ren et al., *A Diversified Attention Model for Interpretable Multiple Clusterings*, IEEE TKDE 2023, Fig. 1 and Algorithm 1

iMClusts Example: Identity Head vs. Pose Head

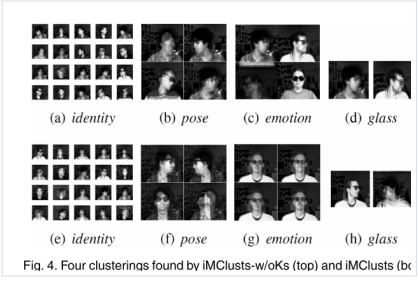


Fig. 4. Four clusterings found by iMClusts-w/oKs (top) and iMClusts (bo

CMUface varies identity, pose, expression, and glasses.

Every head uses the same computation

$$o_i^h = \sum_j \underbrace{A_h[i,j]}_{\text{neighbor weight}} v_j^h, \qquad O_h = [o_1^h; \ldots; o_N^h]$$

O_h stacks the new representation of every sample.
O_{id} and O_{pose} are two members of $\{O_1, \ldots, O_H\}$.

$$Z \xrightarrow{A_{\text{id}} V_{\text{id}}} O_{\text{id}} \xrightarrow{k\text{-means}} C_{\text{id}} = \text{person},$$

$$Z \xrightarrow{A_{\text{pose}} V_{\text{pose}}} O_{\text{pose}} \xrightarrow{k\text{-means}} C_{\text{pose}} = \text{pose}.$$

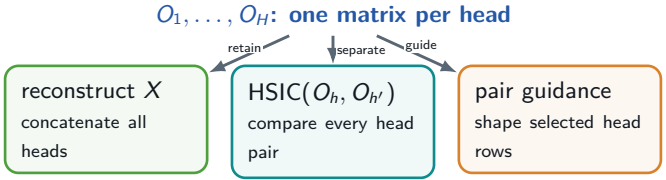
Identity head
A_{id} links the same person across poses.

Pose head
A_{pose} links the same pose across people.

Why they differ
HSIC penalizes duplicated O_h.

Ren et al., *A Diversified Attention Model for Interpretable Multiple Clusterings*, IEEE TKDE 2023, Fig. 4 and Eq. (7).

iMClusts: Joint Training



$$\min_{\Theta} \mathcal{J} = \underbrace{J_{\text{rec}}}_{\text{useful}} + \alpha \underbrace{J_R}_{\text{different}} + \beta \underbrace{J_K}_{\text{guided}}$$

Reconstruction

$$\|f_{\vec{w}}([O_1; \cdots; O_H]) - X\|_F^2$$

Redundancy

$$\sum_{h \neq h'} \text{HSIC}(O_h, O_{h'})$$

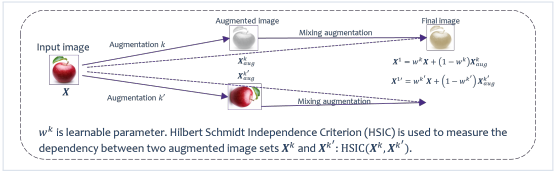
Knowledge

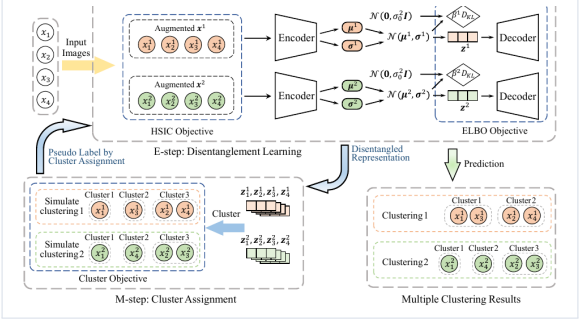
$$\frac{1}{2} \sum_{h,i,j} S_{ij}^h \|o_i^h - o_j^h\|_2^2$$

$+1$ pull; -1 push.

Ren et al., *A Diversified Attention Model for Interpretable Multiple Clusterings*, IEEE TKDE 2023, Eqs. (12)–(15).

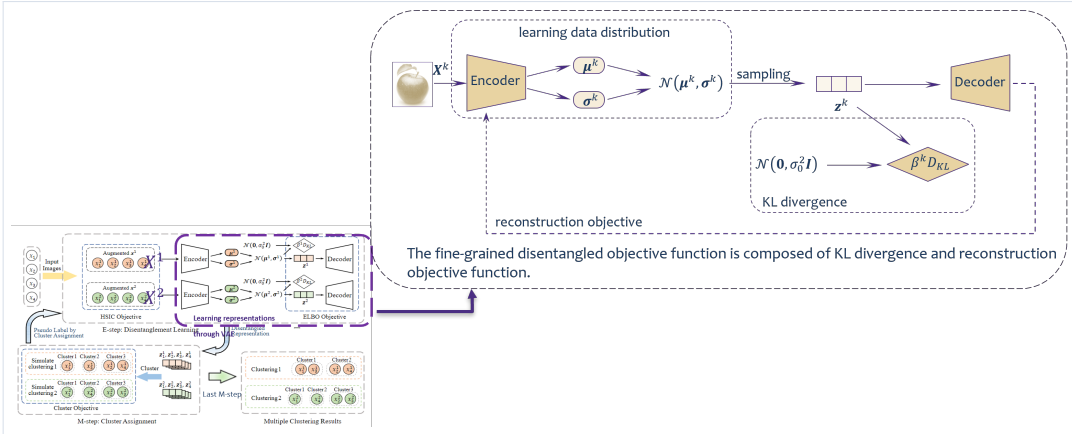
DDMC: Dual-disentangled Deep Multiple Clustering





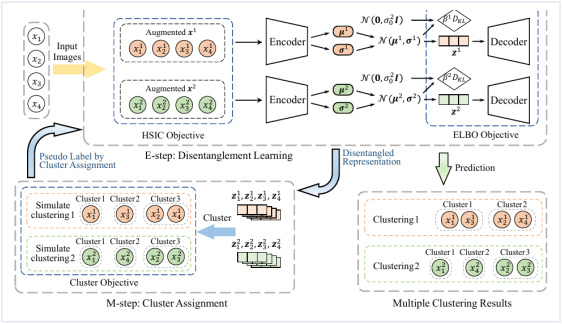
Yao and Hu, *Dual-disentangled Deep Multiple Clustering*, SDM 2024.

DDMC: Dual-disentangled Deep Multiple Clustering



Yao and Hu, *Dual-disentangled Deep Multiple Clustering*, SDM 2024, Eq. (3.7).

DDMC: Dual-disentangled Learning Objective



$$\mathcal{L} = \underbrace{\mathcal{L}_{\text{HSIC}}}_{\text{different inputs}} + \underbrace{\mathcal{L}_{\text{ELBO}}}_{\text{ELBO reconstruction}} + \underbrace{\mathcal{L}_{\text{cluster}}}_{\text{compact labels}}$$

Cluster feedback

$$-\max_{\mathbf{s}} \frac{1}{N} \sum_{i=1}^{N} \left\| z_i^k - W^k s_i^k \right\|_2^2$$

Same-Dataset Evidence: Methods We Just Read

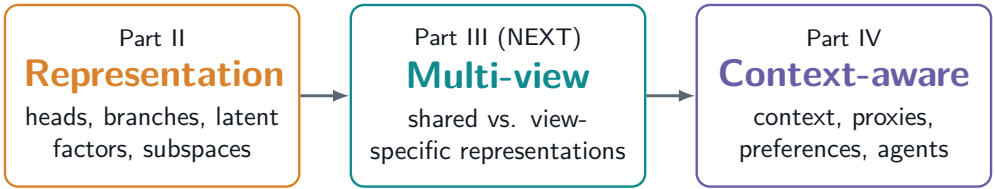
NMI on shared visual multiple-clustering benchmarks

Dataset	Aspect	ENRC	iMClusts	DDMC
Fruit	color	.7103	.7351	**.8973**
	species	.3187	.3029	**.3764**
Fruit360	color	.4264	.4097	**.4981**
	species	.4142	.3861	**.5292**
Card	order	.1225	.1144	**.1563**
	suits	.0676	.0716	**.0933**
ALOI	shape	.9732	.9963	**1.0000**
	color	.9833	**1.0000**	**1.0000**

Read this as a same-benchmark comparison: all three Part II methods are evaluated on the same datasets and metrics reported in the DDMC table.

Yao and Hu, *Dual-disentangled Deep Multiple Clustering*, SDM, 2024, Table 2. NMI; higher is better.

Handoff: three routes to useful alternatives



Q&A ?

Part III

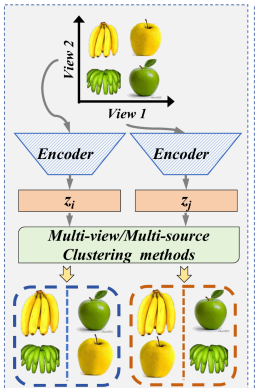
Multi-View / Multi-Source Methods

When diversity comes from several observations of the same object.

Part III Method Map: Many Views Of The Same Objects

Method	Input	How clusters are produced	What creates alternatives
DMClusts [Wei et al., 2019]	complete multi-view matrices	deep matrix factors H_m	different layers become different partitions
DEMVC [Xu et al., 2020]	complete views	collaborative target distribution	mainly consensus, useful contrast case
DiMVMC [Wei et al., 2020]	incomplete views	shared subspaces and decoders	multiple subspaces with HSIC separation
DIMVC [Xu et al., 2022]	incomplete views	complementarity mining	consistency plus complementary cluster cues
MFLVC [Xu et al., 2022]	complete views	semantic branch clustering	low-level detail separated from semantics
SCMC [Fu et al., 2022]	heterogeneous views	graph/self-expression clustering	contrastive subspaces

Problem Formulation: Same Objects, Many Observations



$$\mathcal{X} = \{X^{(v)}\}_{v=1}^{V} \quad \Longrightarrow \quad \{\mathcal{C}^{(m)}\}_{m=1}^{M}$$

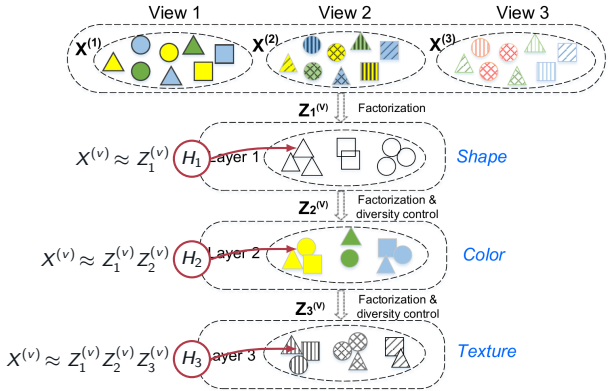
encode each view

align shared structure

keep view-specific cues

output several partitions

DMClusts Flow: One Matrix Factor Layer, One Clustering



Wei et al., *Multi-View Multiple Clusterings using Deep Matrix Factorization*, arXiv:1911.11396, 2019.

Input
same objects, several views

Representation
$X^{(v)} \approx Z_1^{(v)} \cdots Z_M^{(v)} H_M$

Clustering
layer H_m produces $\mathcal{C}^{(m)}$

Diversity
penalize repeated H_m

DMClusts Formula: Quality Term Plus Redundancy Term

$$\min_{\{Z_\ell^{(v)}, H_m, \alpha_m^{(v)}\}} \sum_{m=1}^{M} \sum_{v=1}^{V} \left(\alpha_m^{(v)}\right)^r \left\| X^{(v)} - Z_1^{(v)} \cdots Z_m^{(v)} H_m \right\|_F^2 + \lambda \sum_{1 \le m < m' \le M} \widetilde{R}(H_m, H_{m'})$$

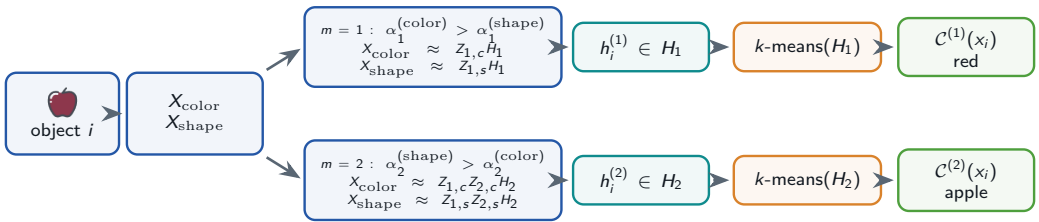
Soft co-association
$S_m = H_m^\top H_m$ measures pairwise proximity; J denotes the all-ones matrix.

Repeated near pairs
$R_+ = \mathrm{tr}(S_m S_{m'})$ penalizes pairs that stay close in both layers.

Balanced redundancy
$R_- = \mathrm{tr}[(J - S_m)(J - S_{m'})]$
$\widetilde{R} = \beta R_+ + (1 - \beta) R_-.$

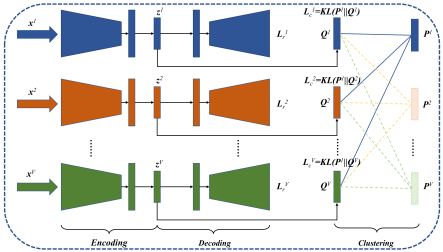
View weights $\alpha_m^{(v)}$ let different clusterings listen to different views.

DMClusts Example: Layers Reweight Views Differently

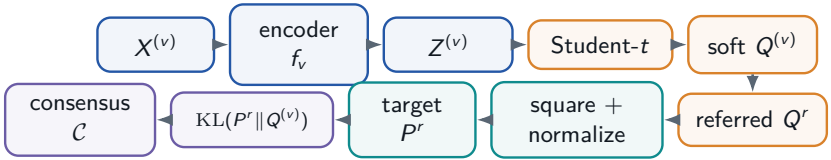


Input $\rightarrow$ **output:** $\widetilde{R}(H_1, H_2) \uparrow \implies \mathcal{C}^{(1)} \neq \mathcal{C}^{(2)}$: repeated factor-row neighborhoods are penalized.

DEMVC Flow: Collaborative Training Aligns Views



Xu et al., *Deep Embedded Multi-View Clustering with Collaborative Training*, arXiv:2007.13067, 2020.



DEMVC Formula: A Referred View Pulls The Others

$$\mathcal{L} = \sum_{v=1}^{V} \mathcal{L}_r^{(v)} + \gamma \sum_{v=1}^{V} \mathcal{L}_c^{(v)}, \qquad \mathcal{L}_c^{(v)} = \mathrm{KL}(P^r \| Q^{(v)})$$

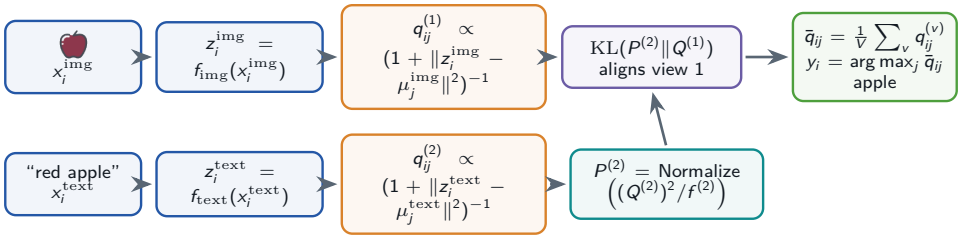
$$q_{ij}^{(v)} = \frac{(1 + \|z_i^{(v)} - \mu_j^{(v)}\|^2)^{-1}}{\sum_j (1 + \|z_i^{(v)} - \mu_j^{(v)}\|^2)^{-1}}$$

Representation
Each view learns an embedding $z_i^{(v)}$ with its own autoencoder.

Clustering
P^r is computed from the referred view and shared across views.

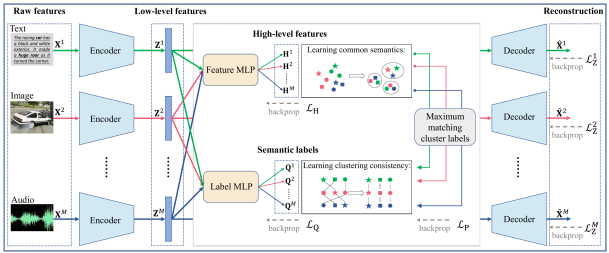
Role in this part
This is a contrast case: it aligns views into one clustering, not many outputs.

DEMVC Example: One Referred View Corrects Another

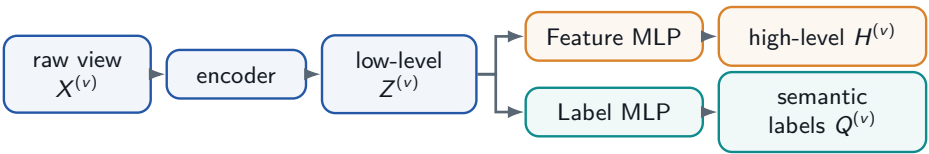


Input $\rightarrow$ **output:** $Q^{(2)} \xrightarrow{\text{square + frequency normalization}} P^{(2)} \xrightarrow{\mathrm{KL}} Q^{(1)}$: one referred view produces one consensus clustering.

MFLVC Flow: Put Reconstruction And Semantics In Different Levels



Xu et al., *Multi-Level Feature Learning for Contrastive Multi-View Clustering*, CVPR 2022.



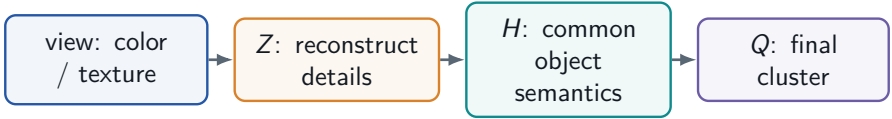
MFLVC Formula: Three Losses At Three Levels

$$\mathcal{L} = \mathcal{L}_Z + \mathcal{L}_H + \mathcal{L}_Q$$

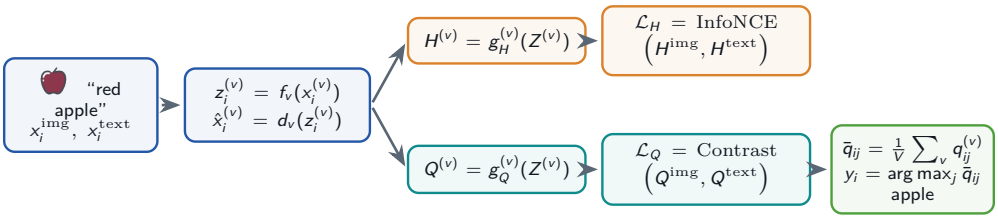
$\mathcal{L}_Z$: reconstruction
Low-level Z keeps view-private details out of the semantic space.

$\mathcal{L}_H$: feature contrast
High-level H pulls common semantics together across views.

$\mathcal{L}_Q$: label contrast
Semantic labels Q become the clustering output.

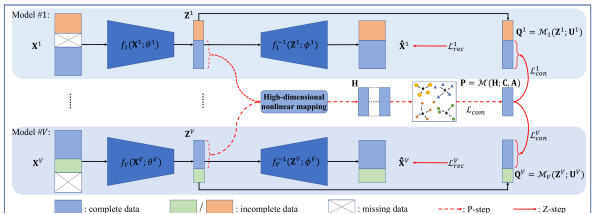


MFLVC Example: Keep Details Low, Cluster Semantics High

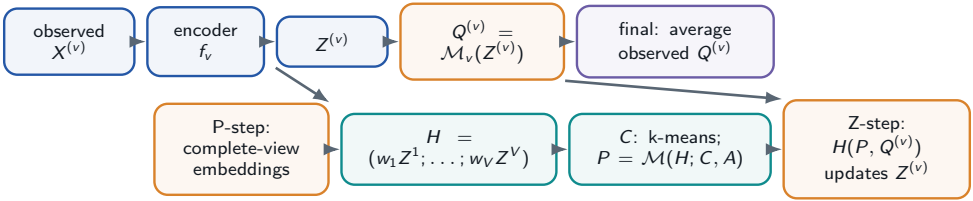


Input $\rightarrow$ **output:** Z keeps reconstructable details; H aligns object-level features; Q alone supplies the semantic cluster label.

DIMVC Flow: Mine Complementarity Without Filling Missing Views



Xu et al., *Deep Incomplete Multi-View Clustering via Mining Cluster Complementarity*, AAAI 2022.



DIMVC Formula: Complementarity And Consistency Alternate

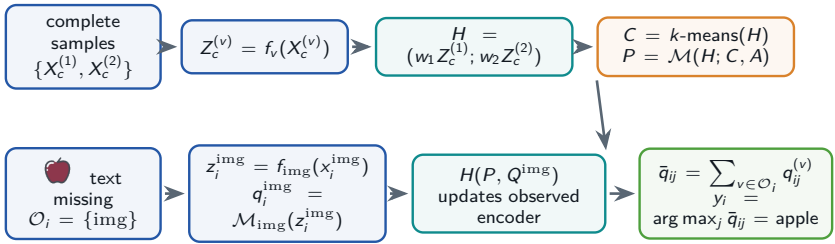
$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{com}} + \mathcal{L}_{\text{con}} = \sum_{v=1}^{V} \|X^{(v)} - \hat{X}^{(v)}\|_F^2 + \sum_i \sum_j \|h_i - c_j\|_2^2 + \sum_{v=1}^{V} H(P, Q^{(v)})$$

Representation
$Z^{(v)} = f_v(X^{(v)})$ is learned only from available samples.

P-step
H is mapped to confident pseudo labels P to mine complementarity.

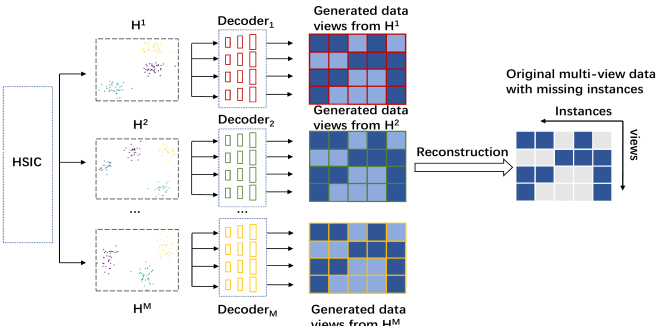
Z-step
$H(P, Q^{(v)})$ pushes each view toward the shared clustering signal.

DIMVC Example: No Imputation, Use Confident Pseudo Labels



Input → output: P is mined from complete training samples; the missing text view of sample i is never synthesized.

DiMVMC Flow: Multiple Subspaces With Missing Views



Wei et al., *Deep Incomplete Multi-View Multiple Clusterings*, arXiv:2010.02024, 2020.

Input
some views are missing

Representation
$H^1, \ldots, H^M$ shared subspaces

Clustering
one partition per subspace

Diversity
HSIC separates shared subspaces

DiMVMC Formula: Complete Views And Separate Clusterings Together

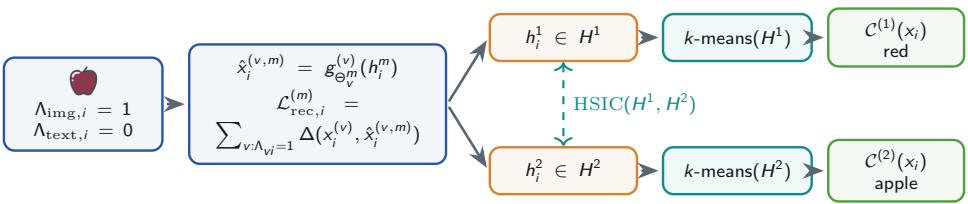
$$\min_{\{H^m, \Theta_v^m\}} \Phi \sum_{m=1}^{M} \sum_{v=1}^{V} \sum_{i=1}^{N} \Lambda_{vi} \, \Delta\left(x_i^{(v)}, g_{\Theta_v^m}^{(v)}(h_i^m)\right) + \lambda \sum_{\substack{m,m'=1 \\ m \neq m'}}^{M} \mathrm{HSIC}(H^m, H^{m'})$$

Representation
H^m is shared; Λ_{vi} selects observed views for reconstruction.

Clustering
Each H^m is clustered separately to produce $\mathcal{C}^{(m)}$.

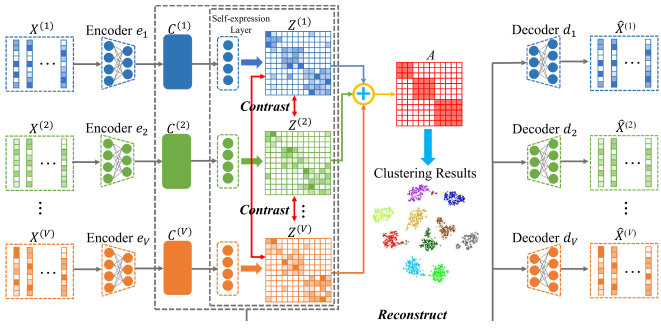
Diversity
HSIC reduces dependence between different shared subspaces.

DiMVMC Example: Incomplete Views, Separate Shared Subspaces



Input $\rightarrow$ **output:** Λ masks unavailable reconstruction terms; HSIC separates H^1, H^2 before each shared subspace is clustered.

SCMC Flow: Contrast Subspaces Before Fusing The Graph



Fu et al., *Subspace-Contrastive Multi-View Clustering*, arXiv:2210.06795, 2022.

Input
heterogeneous views

Encoder
$C^{(v)} = e_v(X^{(v)})$

Self-expression
$C^{(v)\top} \approx C^{(v)\top} Z^{(v)}$

Contrast + fuse
$A = \sum_{v=1}^{V} \omega_v Z^{(v)}$

Output
spectral clustering on $(A + A^\top)/2$

SCMC Formula: Reconstruct, Self-Express, Contrast, Fuse

$$\mathcal{L} = \mathcal{L}_{\mathrm{Re}} + \gamma_1 \mathcal{L}_{\mathrm{Sub}} + \gamma_2 \mathcal{L}_{\mathrm{Con}} + \gamma_3 \mathcal{L}_{\mathrm{Fu}}$$

$\mathcal{L}_{\mathrm{Re}}$
AEs capture nonlinear view structure.

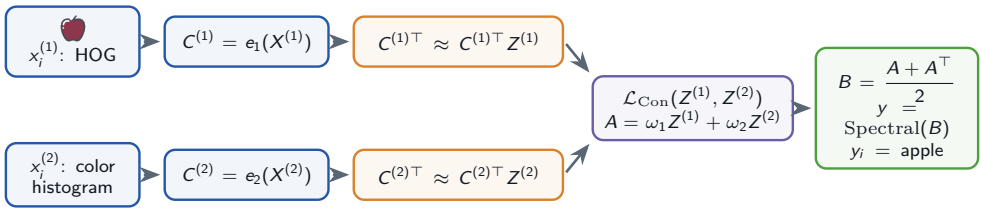
$\mathcal{L}_{\mathrm{Sub}}$
Self-expression builds subspace graphs.

$\mathcal{L}_{\mathrm{Con}}$
Cross-view positives align; negatives separate.

$\mathcal{L}_{\mathrm{Fu}}$
Graph regularization produces affinity A.

SCMC is useful here because it shows how heterogeneous views can be fused before Part IV adds user criteria.

SCMC Example: Contrast Subspaces, Then Cluster The Graph



Input → output: $Z^{(v)}$ is a sample-to-sample subspace graph; contrast precedes weighted graph fusion and spectral clustering.

Across Papers I: Complete Views And Shared Semantics

DMClusts [1]
multi-view matrices $\rightarrow$ several partitions; layers H_m are separated by redundancy control.

DEMVC [2]
several views $\rightarrow$ one consensus partition; a referred target P^r aligns all views.

MFLVC [3]
raw views $\rightarrow$ semantic clusters; Z reconstructs while H, Q align semantics.

[1] Wei et al., *Multi-View Multiple Clusterings using Deep Matrix Factorization*, arXiv:1911.11396, 2019.

[2] Xu et al., *Deep Embedded Multi-View Clustering with Collaborative Training*, arXiv:2007.13067, 2020.

[3] Xu et al., *Multi-Level Feature Learning for Contrastive Multi-View Clustering*, CVPR 2022.

Across Papers II: Missing Or Heterogeneous Views

DiMVMC [4]
incomplete views → multiple alternatives; decoders complete views while HSIC separates H^m.

DIMVC [5]
incomplete views → one clustering; P-step mines complementarity, Z-step enforces consistency.

SCMC [6]
heterogeneous views → fused affinity; self-expression and contrastive learning bridge modalities.

[4] Wei et al., *Deep Incomplete Multi-View Multiple Clusterings*, arXiv:2010.02024, 2020.

[5] Xu et al., *Deep Incomplete Multi-View Clustering via Mining Cluster Complementarity*, AAAI 2022.

[6] Fu et al., *Subspace-Contrastive Multi-View Clustering*, arXiv:2210.06795, 2022.

Same-Dataset Evidence: DMClusts vs DiMVMC

Complete multi-view datasets: diversity improves, quality can trade off

Dataset	DI ↑		Redundancy NMI ↓	
	DMClusts	DiMVMC	DMClusts	DiMVMC
Caltech20	.115	**.265**	.065	**.024**
Handwritten	.173	**.604**	.061	**.006**
Reuters	.030	**.434**	.007	**.001**
Mirflickr	.076	**.536**	.043	**.005**

DMClusts
Deep matrix factor layers produce several H_m; it is a good baseline for complete multi-view MC.

DiMVMC
Separate decoders and HSIC lower redundancy strongly, especially when views may be incomplete.

Wei et al., *Deep Incomplete Multi-View Multiple Clusterings*, arXiv:2010.02024, 2020, Table II. DI: diversity index; NMI: pairwise redundancy among generated clusterings.

When Does This Family Work Best?

aligned complete views

multiple semantic criteria

missing or unreliable views

heterogeneous modalities

MFLVC / SCMC

align common semantics after view encoding.

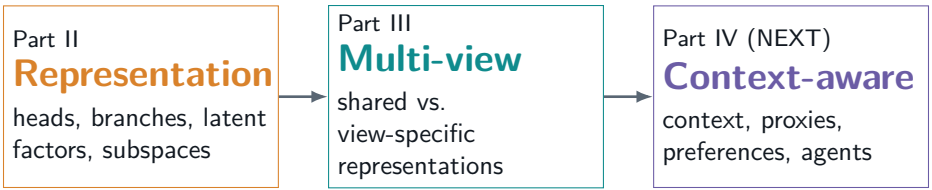
DMClusts / DiMVMC

assign different subspaces to different criteria.

DIMVC / DiMVMC

avoid letting imputed or noisy views dominate.

Handoff: three routes to useful alternatives



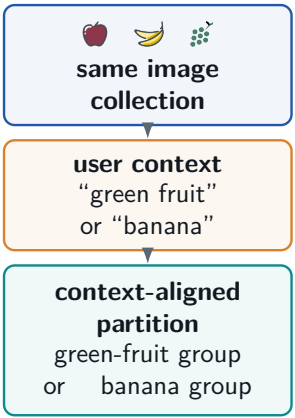
Q&A ?

Part IV

Context-Aware and User-Guided Methods

Preferences, prompts, proxy learning, open-ended semantics, and agents.

Why User Context Changes The Problem



Same data, different answers.

Analyst intent
maker, color, body type.

Domain context
lineage, state, response.

Prompt
“cluster by sleeve length”.

$$Q(\mathcal{C}) - \beta R(\mathcal{C}, \mathcal{C}') + \gamma A(\mathcal{C}, u)$$

Part IV Input-Output Pipeline

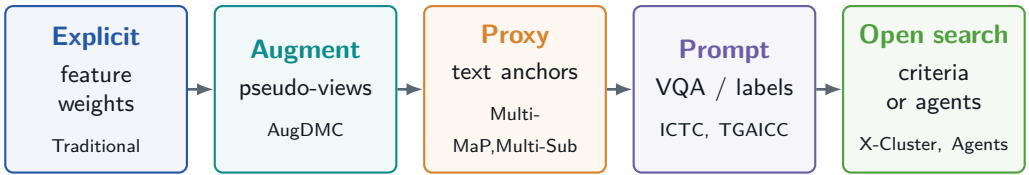
$$u \to g(u), \qquad z_i^{(u)} = f_\theta(x_i, g(u)), \qquad \mathcal{C}^{(u)} = \mathrm{Cluster}\big(\{z_i^{(u)}\}_{i=1}^n\big)$$

Input	Context
fruit images	color / type / shape
playing cards	suit / rank
car photos	maker / body style

Part IV Method Map: What Kind Of Context?

Method	Context signal	Bridge to clustering
AugDMC [Yao et al., 2023]	augmentation cue	pseudo-views from transformations
Multi-MaP [Yao et al., 2024]	keyword or preference	CLIP proxy
Multi-Sub [Yao et al., 2024]	text criterion	criterion-specific subspace
ICTC [Kwon et al., 2024]	text criterion	image-text conditioned embedding
TGAICC [Stephan et al., 2024]	text question	VQA answers as cluster signals
X-Cluster [Liu et al., 2024]	open language query	system proposes criteria
Agent-Centric [Chen et al., 2025]	user interest + MLLM agents	relation graph traversal and merge

A Spectrum Of Guidance Signals

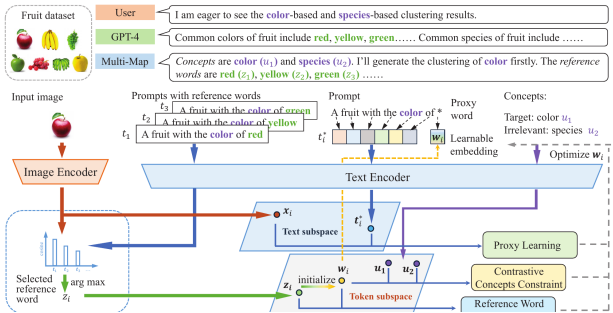


More manual
the user names dimensions or weights.

More semantic
VLMs and LLMs translate intent into anchors.

More autonomous
the system proposes or searches criteria.

Multi-MaP: From A User Keyword To A Personalized Partition



Yao et al., *Multi-Modal Proxy Learning Towards Personalized Visual Multiple Clustering*, CVPR 2024.

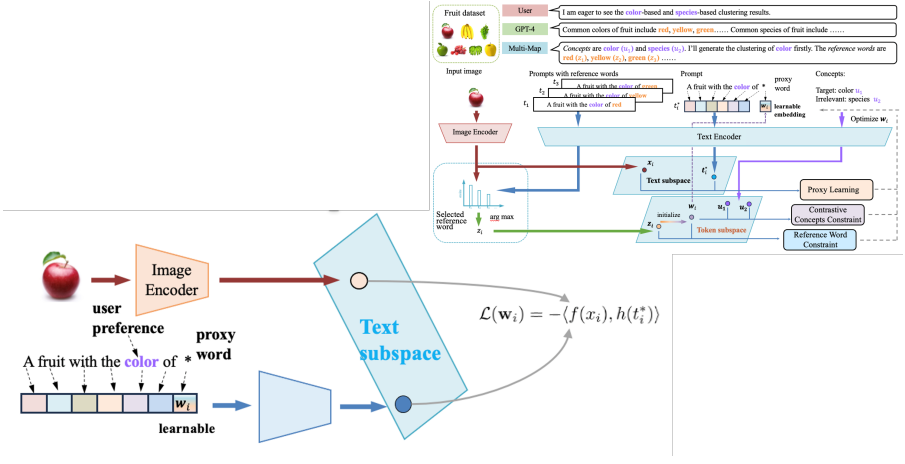
Input
images + keyword

→ **Representation**
CLIP proxy

→ **Clustering**
proxy-biased features

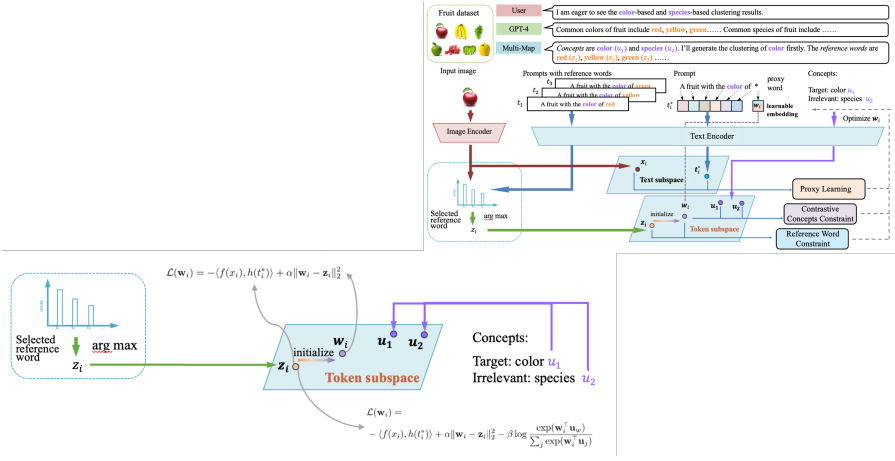
→ **Output**
personalized partition

Multi-MaP:Optimizing the proxy word embeddings to maximize similarity with image representations



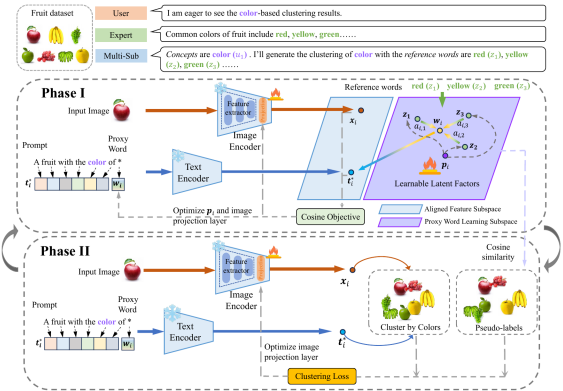
Yao and Hu, *Dual-disentangled Deep Multiple Clustering*, SDM 2024.

Multi-MaP:Concept-level Constraint and Constrained Optimization



Yao and Hu, *Dual-disentangled Deep Multiple Clustering*, SDM 2024.

Multi-Sub: Prompts Select Criterion-Specific Subspaces



Yao et al., *Customized Multiple Clustering via Multi-Modal Subspace Proxy Learning*, NeurIPS 2024.

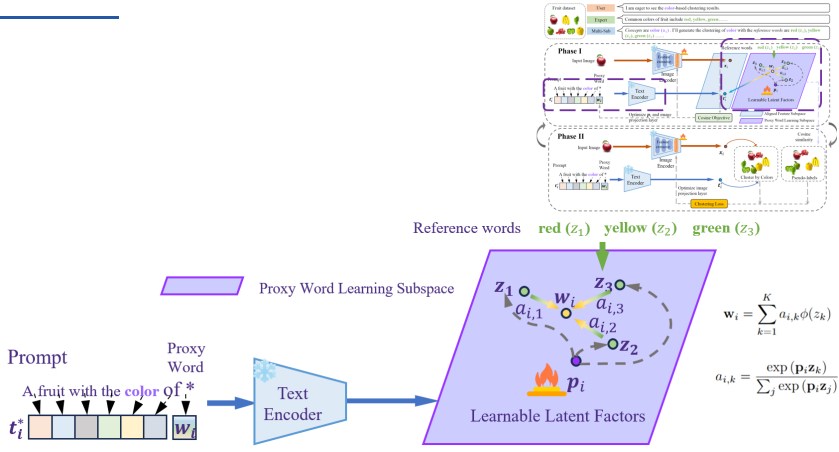
Input
images + prompt

Representation
proxy-word basis

Clustering
criterion subspace

Output
customized partition

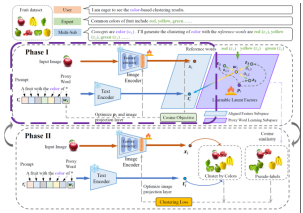
Multi-Sub: Subspace Proxy Word Representation

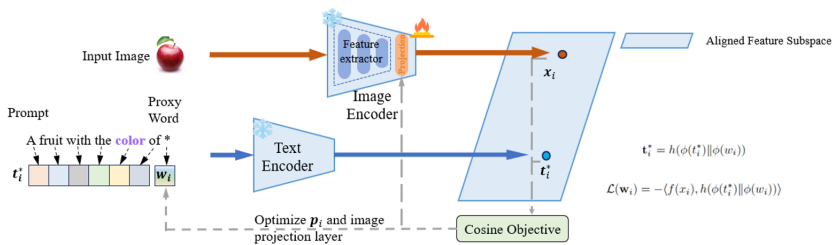


- Learn proxy word embeddings based on user-preferred aspects.
- Represent each proxy word as a linear combination of reference words.

Yao et al., *Customized Multiple Clustering via Multi-Modal Subspace Proxy Learning*, NeurIPS 2024.

Multi-Sub: Multi-Modal Subspace Proxy Learning

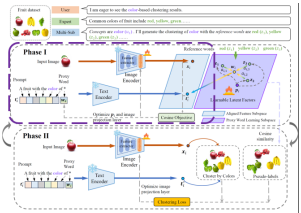


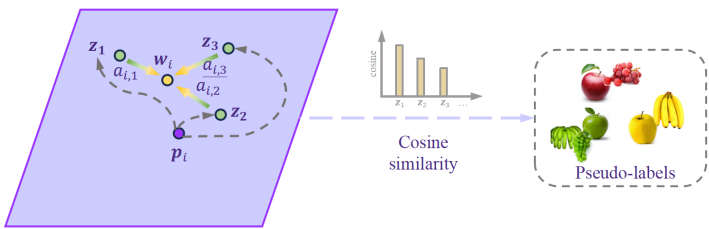


- Adjust trainable latent factors to align user-specific image representations with corresponding proxy word embeddings, optimizing the alignment through cosine similarity

Yao et al., *Customized Multiple Clustering via Multi-Modal Subspace Proxy Learning*, NeurIPS 2024.

Multi-Sub: Subspace Proxy Word Representation

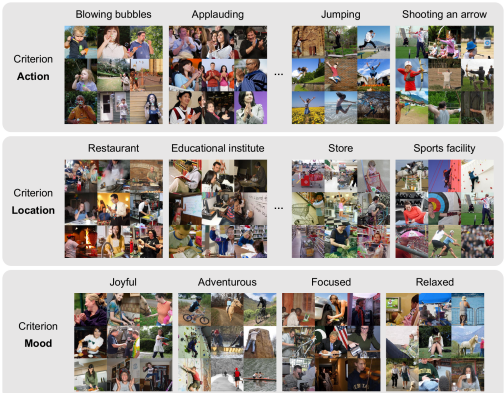




- Obtain the pseudo-labels using the highest cosine similarity between the currently learned proxy word embeddings and the reference word embeddings.

Yao et al., *Customized Multiple Clustering via Multi-Modal Subspace Proxy Learning*, NeurIPS 2024.

ICTC: Text Criteria Change The Image Clusters



Kwon et al., *Image Clustering Conditioned on Text Criteria*, ICLR 2024.

Input
images + criterion text

Representation
criterion-conditioned similarity

Clustering
group under one criterion

Output
named semantic clusters

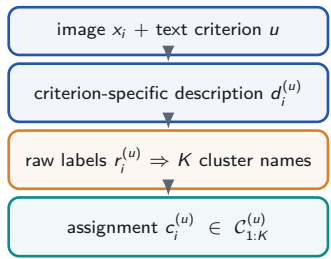
ICTC Mechanism: Describe, Name, Then Assign

Step 1: criterion-specific text
The VLM describes each image with respect to the user-specified criterion u.

Step 2: discover K names
2a. Produce one raw label per description. **2b.** Aggregate the labels into K interpretable cluster names.

Step 3: assign images
The LLM assigns each original description to one of the K discovered cluster names.

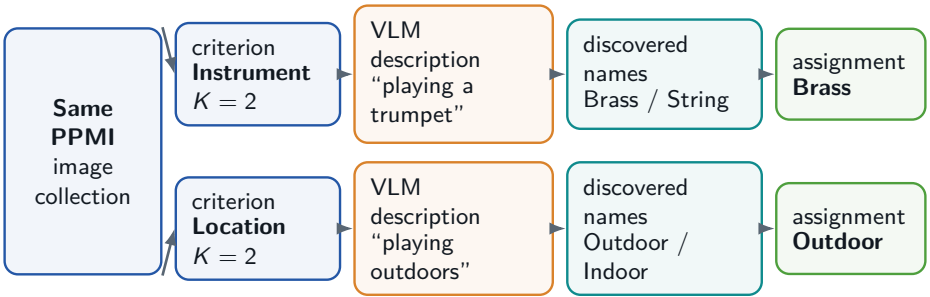
$$
\begin{aligned}
d_i^{(u)} &= \mathrm{VLM}(x_i, u), \\
r_i^{(u)} &= \mathrm{LLM}(d_i^{(u)}, u), \\
\mathcal{C}_{1:K}^{(u)} &= \mathrm{LLM}\left(\{r_i^{(u)}\}, K, u\right), \\
c_i^{(u)} &= \mathrm{Assign}\left(d_i^{(u)}, \mathcal{C}_{1:K}^{(u)}, u\right).
\end{aligned}
$$



Source: Kwon et al., *Image Clustering Conditioned on Text Criteria*, ICLR 2024, Fig. 2 and Secs. 3.1–3.3.

ICTC Example: Same Images, Different Criteria

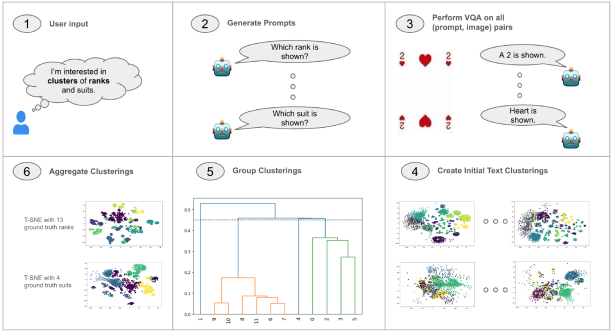
The same PPMI image collection produces different partitions when only the text criterion changes.



Input → output: **Key distinction:** "trumpet", "street", or "kitchen" may appear as criterion-conditioned descriptions or raw labels. The final assignment must instead be one of the discovered K cluster names, such as **Brass instruments** or **Outdoor**.

Source: Kwon et al., *Image Clustering Conditioned on Text Criteria*, ICLR 2024, Figs. 1, 2, and 4.

TGAICC: VQA Answers Become Alternative Clusters



Stephan et al., *Text-Guided Alternative Image Clustering*, arXiv:2406.18589, 2024.

Input
images + prompts

Representation
VQA answers

Clustering
answer-similarity partitions

Output
consensus alternatives

TGAICC Mechanism: Many Questions, Then Consensus

$$y_i^{(q)} = \mathrm{VQA}(x_i, q), \qquad P_q = \mathrm{Cluster}\big(\{y_i^{(q)}\}_{i=1}^{n}\big), \qquad \mathcal{C}^{(m)} = \mathrm{Consensus}\big(\{P_q : q \in G_m\}\big)$$

Representation
Each prompt asks for a text answer: card suit, rank, color, visible object.

Clustering
A prompt creates one candidate partition from answer similarity.

Diversity
Hierarchical grouping merges similar candidates and keeps distinct consensus clusters.

X-Cluster: From An Image Collection To Proposed Criteria

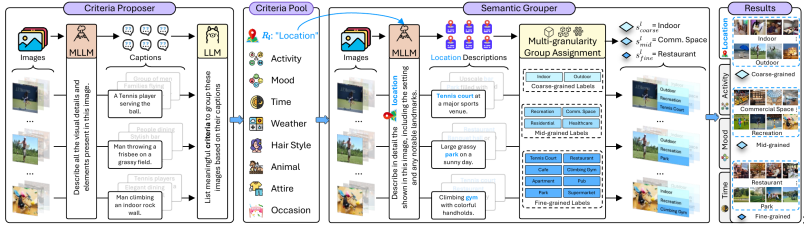


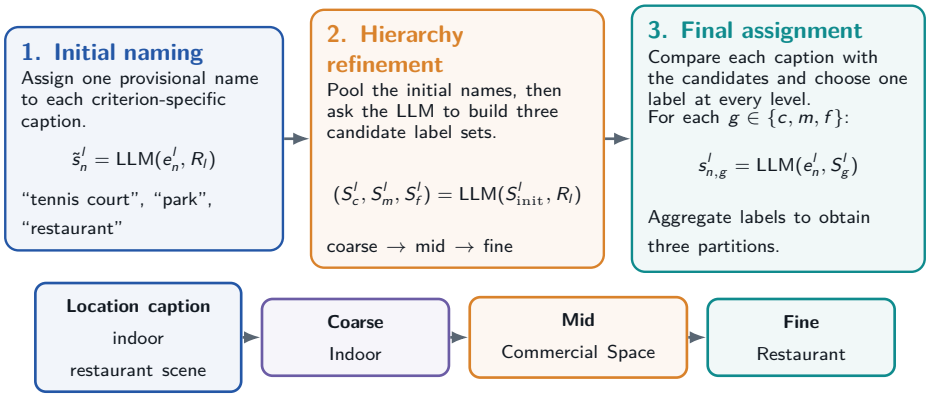
Figure 3. **$\mathcal{X}$-Cluster** consists of a *Criteria Proposer* and a *Semantic Grouper*. **(left)** Given a set of images, the Proposer discovers and outputs a pool of grouping criteria in natural language. **(right)** The Grouper subsequently extracts criterion-specific descriptions from images relevant to each criterion, discovers the underlying semantic clusters, and groups each image at three semantic granularity

Liu et al., *Organizing Unstructured Image Collections using Natural Language*, arXiv:2410.05217, 2024.

Input unstructured images → **Representation** language-proposed criteria → **Clustering** group under each axis → **Output** labeled alternatives

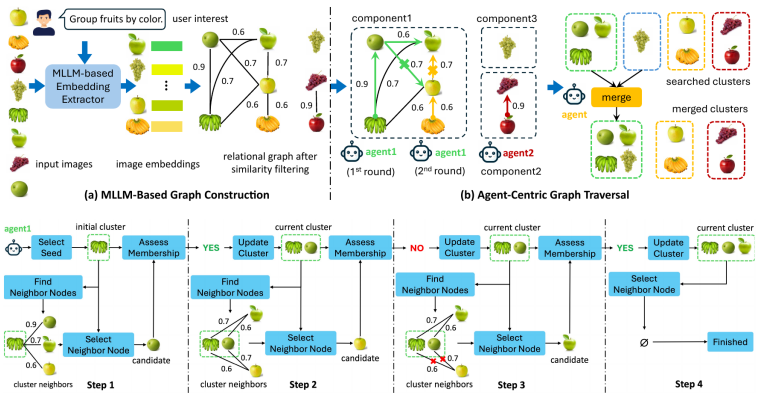
X-Cluster: Multi-Granularity Group Assignment

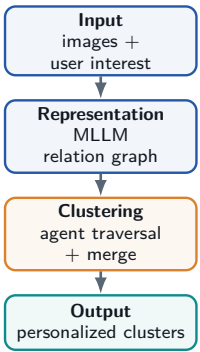
Input → output: Criterion-specific input: $e_n^l = \mathrm{MLLM}(x_n, R_l)$ for every image x_n and discovered criterion R_l.



Source: Liu et al., arXiv:2410.05217, Sec. 4.2; Supp. Tables 14–17.

Agent-Centric: Personalized Clustering via Relation Graphs





Chen et al., *Agent-Centric Personalized Multiple Clustering with Multi-Modal LLMs*, arXiv:2503.22241, 2025.

Agent-Centric: Assess Membership And Repair Local Edges

Input → output: At each step, choose the boundary node with strongest connectivity:
$v_j^* = \arg\max_{v \in N(S_i)} \sum_{u \in S_i} w(v, u)$.

1. Represent the cluster

Select the top-K weighted-degree nodes in S_i as visual representatives:

$$d_i(v) = \sum_{u \in N(v) \cap S_i} w(v, u),$$

$$R_i = \mathrm{TopK}_{v \in S_i}\, d_i(v).$$

2. Assess membership

Give the MLLM the representative images R_i, candidate v_j^*, and user interest T.

Output: reasoning plus a structured **YES/NO** conclusion.

3. Update cluster and graph

YES: $S_i \leftarrow S_i \cup \{v_j^*\}$; recompute R_i.

NO: remove every edge (u, v_j^*) with $u \in S_i$.

Repeat until $N(S_i) = \emptyset$.

Interest T
fruit color

Representatives
green fruits

Candidate
yellow-green apple

NO
cut local edges

Source: Chen et al., arXiv:2503.22241, Secs. 3.2 and 3.4; Algorithm 1; Fig. 4.

Agent-Centric: Merge Redundant Clusters Globally

Input → output: After parallel local traversal, the searched set $\mathcal{S} = \{S_1, \ldots, S_P\}$ may contain duplicate semantic clusters because the initial graph can miss useful edges.

1. Propose a cluster pair

Represent each cluster by its top-K nodes R_p, R_q.

Select a neighboring cluster using summed inter-cluster connectivity.

2. Merge assessment

Give the MLLM R_p, R_q, and interest T. Ask whether the two clusters express the same user-relevant concept.

Output: $\langle \text{CONCLUSION} \rangle$ YES/NO.

3. Apply and repeat

YES: $S_{pq} = S_p \cup S_q$; recompute representatives.

NO: remove edges between S_p and S_q.

Continue until $\mathcal{S}$ no longer changes.

YES repairs a missing edge

Two separately searched red-fruit clusters are semantically redundant under color, so the merge agent unifies them.

NO preserves a boundary

A red-fruit cluster and a green-fruit cluster remain separate; their cross-cluster edges are removed.

Source: Chen et al., arXiv:2503.22241, Algorithm 1; Appendix A.6 and A.8; Table 5.

Criterion Guidance Improves NMI On Shared Benchmarks

Dataset	Aspect	Part II: representation-driven			Part IV: context-guided		
		ENRC	iMClusts	DDMC	Multi-MaP	Multi-Sub	Agent-Centric
Fruit	color	.7103	.7351	.8973	.8619	.9693	**1.0000**
	species	.3187	.3029	.3764	**1.0000**	**1.0000**	**1.0000**
Fruit360	color	.4264	.4097	.4981	.6239	.6654	**.7214**
	species	.4142	.3861	.5292	.5284	.6123	**.6532**
Card	order	.1225	.1144	.1563	.3653	.3921	**.9667**
	suits	.0676	.0716	.0933	.2734	.3104	**.9481**
Stanford Cars	color	.2465	.2336	.6899	.7360	**.7533**	–
	type	.2063	.1963	.6045	.6355	**.6616**	–

Criterion control and performance

Agent-Centric is best or tied-best on all six reported pairs; Multi-Sub leads on Stanford Cars, where Agent-Centric reports no result.

[1] Yao et al., *Customized Multiple Clustering via Multi-Modal Subspace Proxy Learning*, NeurIPS 2024, Table 2.
[2] Chen et al., *Agent-Centric Personalized Multiple Clustering with Multi-Modal LLMs*, arXiv:2503.22241v3, Table 1.

Choosing A Method From The User Workflow

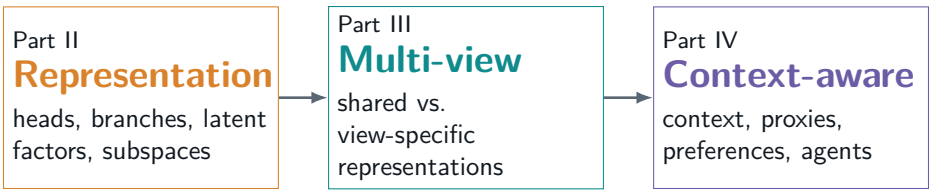
Known visual criterion
style, material, color
Multi-MaP / Multi-Sub / Agents

Unknown criterion
discover semantic axes
X-Cluster

Text question
rank, suit, action, mood
ICTC / TGAICC

Summary: read context-aware DMC from user intent to semantic bridge to labeled alternatives.

Handoff: three routes to useful alternatives



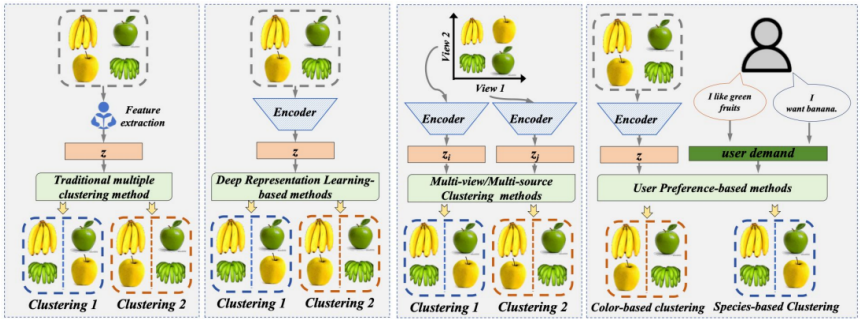
Q&A ?

Part V

Comparative Insights, Applications, and Open Questions

When each family works, how to evaluate it, and what remains open.

From Taxonomy To Method Choice



Method Selection: Start From Data And Task

One dataset, no criterion

images have latent factors: color, material, pose

ENRC / iMClusts / DDMC / MFCVAE

Named user criterion

"cluster cars by color", "cards by suit"

Multi-Sub / Multi-MaP / ICTC / TGAICC / Agent-Centric

Same objects, multiple views

image-text, RNA-ATAC, multi-sensor observations

DMClusts / DiMVMC / DIMVC / MFLVC

SCMC / scMCs (DEMVC: consensus)

Open-ended exploration

user has an interest but not a fixed axis

X-Cluster

Evaluation: Report Three Axes Together

$$S = Q(\mathcal{C}) - \beta\,\mathcal{R}(\mathcal{C}) + \gamma\,\mathcal{A}(\mathcal{C}, u)$$

quality: coherent clusters

diversity:
non-duplicate outputs

alignment: answers the task

Quality metrics
ACC, NMI, ARI, RI.

Diversity metrics
pairwise NMI, VI, DI, SI.

Alignment checks
human agreement, prompt score, labels.

Benchmark Anchors Across The Tutorial

Image collections

ALOI, StickFig, CMUface, NR-Objects, MNIST/C-MNIST, Fruit/Fruit360, Card, Cars, Flowers, CUB-200.[D1–D5]

Multi-view / multi-modal benchmarks

MNIST-USPS, BDGP, CCV, Fashion-MNIST views, Caltech-5V, Yale, ORL, Reuters.[D6–D7]

Single-cell multi-omics

CellMix and PBMC: paired scRNA-seq/scATAC views for cell type, state, and alternative structures.[D8]

Context is cross-cutting, not a dataset class

A criterion, text prompt, or user intent can guide any compatible collection above; image datasets are simply the best-developed examples.

Block A: Multi-view Clustering Benchmarks

Complete multi-view clustering; ACC under one reported protocol.

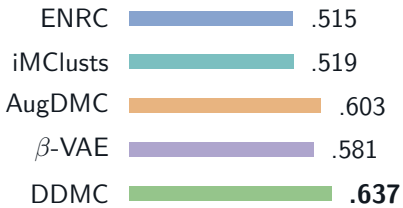
Dataset	IMVTSC-MVI	SiMVC	CoMVC	MFLVC
MNIST-USPS	.669	.981	.987	**.995**
BDGP	.981	.704	.802	**.989**
CCV	.117	.151	.296	**.312**
Fashion	.632	.825	.857	**.992**
Caltech-5V	.760	.719	.700	**.804**

Read: MFLVC has the highest ACC on all five shown datasets, including a large Fashion gain over CoMVC.

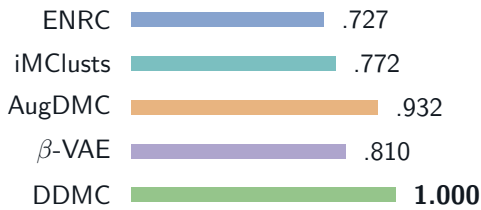
Source: Xu et al., *Multi-Level Feature Learning for Contrastive Multi-View Clustering*, CVPR 2022, Tables 2–3.

Block B: Automatic Deep Multiple Clustering

Fruit: mean NMI over color/species



C-MNIST: mean NMI over left/right digit



Conclusion

DDMC is not tied with ENRC in this comparison: across all 16 criterion-level NMI rows in Table 2, DDMC records 14 wins and 2 ties versus ENRC.

Source: Yao and Hu, *Dual-Disentangled Deep Multiple Clustering*, SIAM SDM 2024, Table 2.

Block C: Context-guided Visual Clustering

Same table and protocol; criterion-level NMI (higher is better).

Dataset	Criterion	DDMC	Multi-MaP	Multi-Sub	Agent-Centric
Fruit	color	.8973	.8619	.9693	**1.0000**
Fruit	species	.3764	**1.0000**	**1.0000**	**1.0000**
Fruit360	color	.4981	.6239	.6654	**.7214**
Fruit360	species	.5292	.5284	.6123	**.6532**
Card	order	.1563	.3653	.3921	**.9667**
Card	suits	.0933	.2734	.3104	**.9481**

Read: Agent-Centric wins five rows and ties one against Multi-Sub; the largest gains are on Card order and suits.

Source: Chen et al., *Agent-Centric Personalized Multiple Clustering with Multi-Modal LLMs*, arXiv:2503.22241v3, Table 1.

What Can We Conclude From Comparable Blocks?

Benchmark block	Best-supported choice	Why this family fits
One dataset, latent factors	DDMC (14 wins, 2 ties vs ENRC)	clustering-oriented disentanglement recovers hidden factors
Complete multi-view data	MFLVC (best ACC on 5/5 shown)	separates reconstruction from high-level semantics
Incomplete multi-view data	DIMVC / DiMVMC	mines complementarity while handling missing views
Named visual criterion	Agent-Centric (5 wins, 1 tie of 6)	MLLM reasoning follows the requested semantic axis
Open-ended criterion discovery	TeDeSC (X-Cluster)	proposes interpretable criteria before clustering

Main rule: compare methods only under a fixed dataset, criterion, metric, and protocol.

Evidence: DDMC Table 2; MFLVC Tables 2–3; Agent-Centric Table 1. Choices are block-specific, not universal rankings.

Application Domains By Method Family

Family	Concrete domain	Useful partition examples	Representative methods
Representation-based	rendered objects and face images	color/material/shape; identity/pose	ENRC, iMClusts, DDMC
Multi-view	multi-omics, image-text, multi-sensor data	cell type/state; visual/semantic groups; sensor-specific behavior	MFLVC, DIMVC, SCMC
Context-guided (cross-cutting)	any domain above; strongest evidence is visual	any named task-relevant criterion	Multi-MaP, Multi-Sub, ICTC, TGAICC, Agent-Centric
Open-ended discovery	personal image collections, browse/search systems	automatically proposed criteria and semantic substructures	TeDeSC (X-Cluster)

Application view
choose the method by the user's decision, not by the dataset name alone.

Deep Multiple Clustering in Industry

Huiji Gao

Search and Recommender Systems in Industry

Search and Recommendation are Everywhere

Search and Recommender Systems in Industry

Web Search

Search for specific information through the web search engine

*https://www.autosnyders.com/how-to/technology/top-10-search-engines-to-make-your-career-searching-for-information-in-cyberspace_742.php

Online Social Networks

Search and Recommend with social networks for items such as news, jobs, videos, as well as online advertising

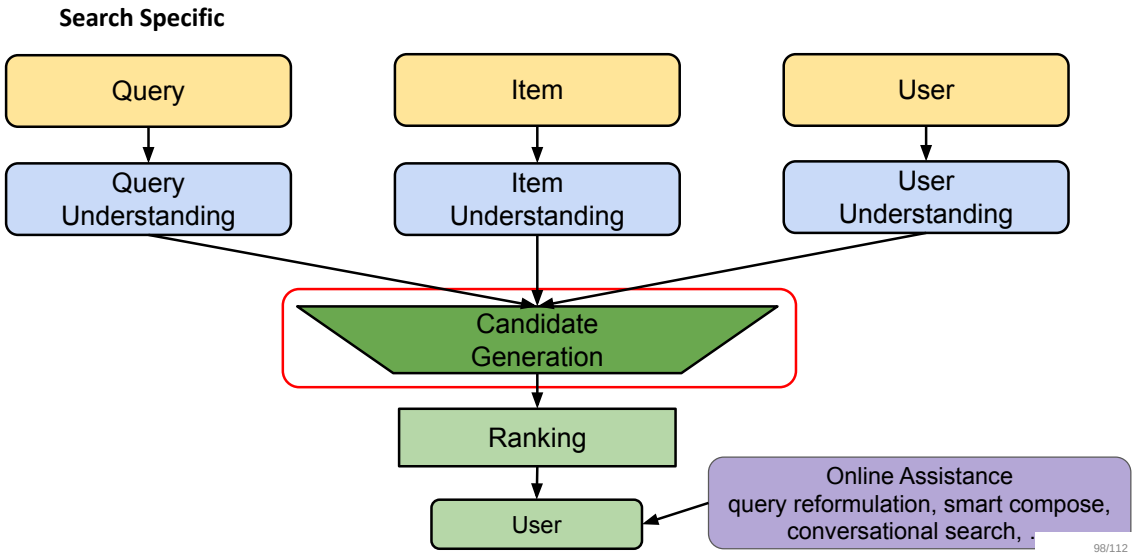
*https://60secondmarketer.com/2021/04/05/the-top-25-social-media-networks-you-should-know-in-2021/

Online E-commerce Marketplace

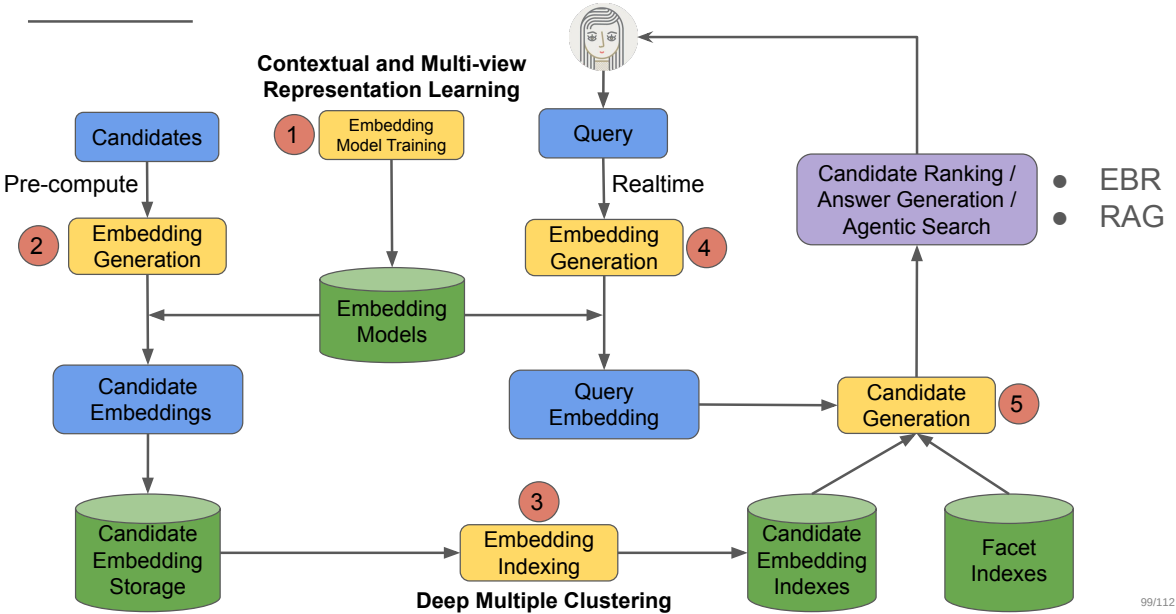
Buy & Sell Goods and / or Services

* https://www.danhock.com/posts/the-future-of-marketplaces
* https://internetdevels.com/blog/start-online-marketplace-website

Search and Recommender Systems in Industry



Overall Architecture on Candidate Generation



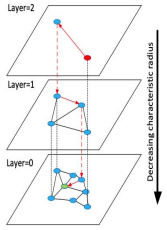
Building Embedding Index with Deep Clustering

Inverted Index (Embedding)

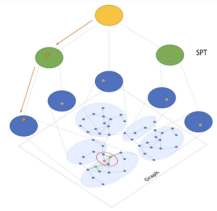
Cluster ID of Embeddings

	Candidates
	Candidates
	Candidates
	Candidates
.....	Candidates

HNSW



SPTAG



Inverted Index (Hybrid)

Keyword

Cluster ID of Embeddings

	Candidates
	Candidates
....	Candidates
	Candidates
	Candidates
	Candidates

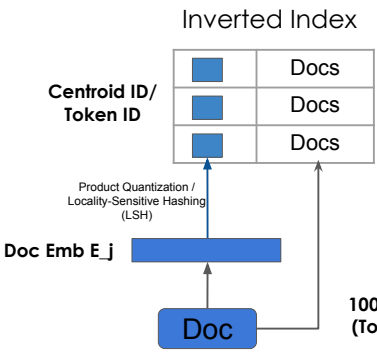
Hybrid Candidate Generation

Embedding-based Index (Inverted)

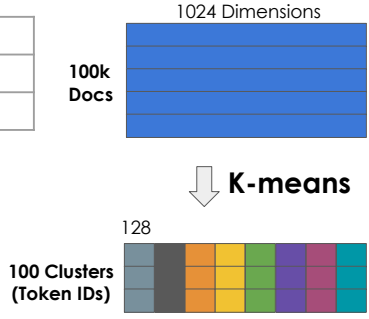
Keyword-based Index (Inverted or Forward)

Building Embedding Index with Deep Clustering

- **Compatible with Inverted Index Systems**



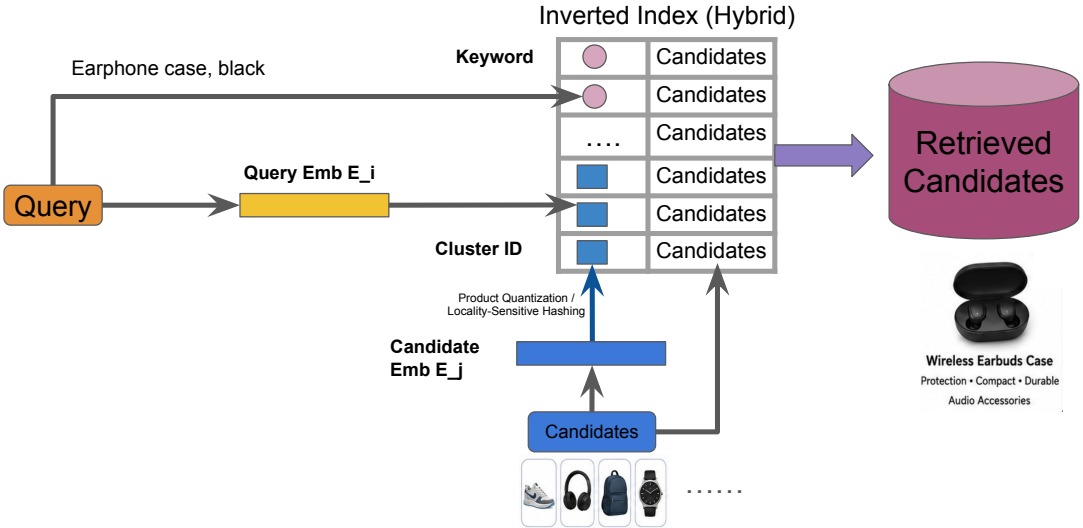
- Product Quantization



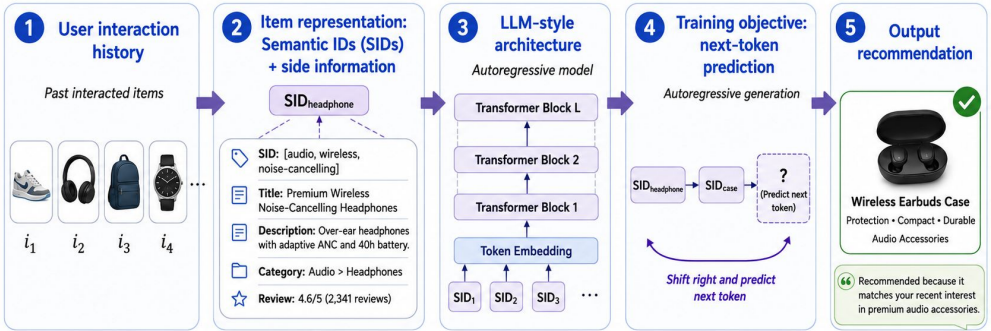
  - 100 x 8 centroids, each centroid is represented by a 128 dimension vector
  - Each embedding can be represented by 8 centroid embedding vectors, based on its distance to each centroid
  - The centroids can be used as keys in inverted index

- Locality-Sensitive Hashing (LSH)

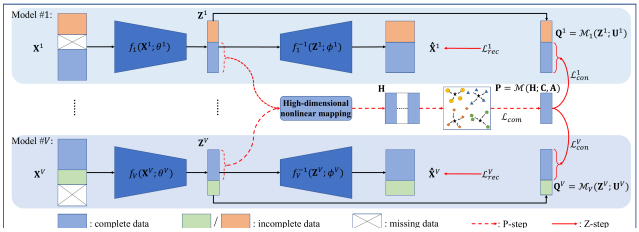
Candidate Retrieval with Embedding Clusters



Semantic ID and Generative Retrieval



Challenge: Imperfect Or Contradictory Views



Xu et al., *Deep Incomplete Multi-View Clustering via Mining Cluster Complementarity*, AAAI 2022.

Missing
not every sample has every view.

Noisy
one modality may dominate.

Contradictory
disagreement can be signal or noise.

Challenge: Scalability And Cost

Model cost
multiple encoders, heads, subspaces, decoders.

Semantic cost
LLM, VQA, CLIP proxy generation and search.

Evaluation cost
several outputs times several metrics and users.

Practical fixes: shared encoders, anchors, caching, early stopping.

Challenge: Explain, Audit, Protect

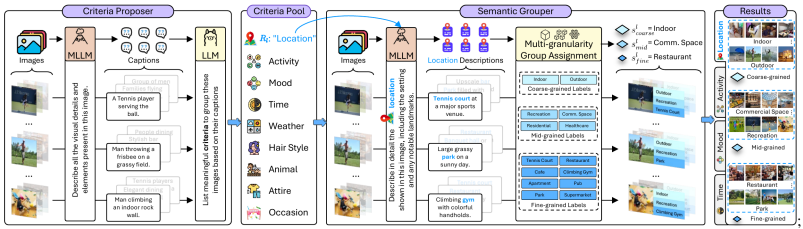


Figure 3. $\mathcal{X}$-**Cluster** consists of a *Criteria Proposer* and a *Semantic Grouper*. **(left)** Given a set of images, the Proposer discovers and outputs a pool of grouping criteria in natural language. **(right)** The Grouper subsequently extracts criterion-specific descriptions from images relevant to each criterion, discovers the underlying semantic clusters, and groups each image at three semantic granularity

Liu et al., *Organizing Unstructured Image Collections using Natural Language*, arXiv:2410.05217, 2024.

Explain
name the criterion and show examples.

Audit
prompts and proxies can encode bias.

Protect
user context may be sensitive.

Open Research Questions

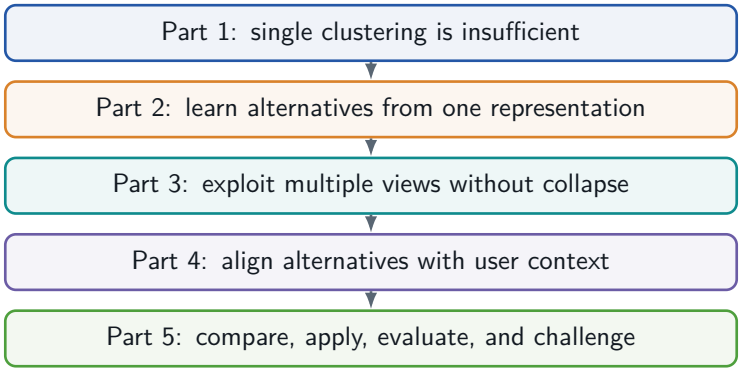
Robustness
complementarity or noise?

Evaluation
one benchmark for all three axes?

Interaction
specify, select, or steer?

Deployment
document sensitive criteria?

Tutorial Recap



Foundations / Views

Foundations / representation

- Miklautz et al., *Deep Embedded Non-Redundant Clustering*, AAAI 2020.
- Falck et al., *Multi-Facet Clustering Variational Autoencoders*, NeurIPS 2021.
- Ren et al., *A Diversified Attention Model for Interpretable Multiple Clusterings*, TKDE 35(9), 2023.
- Yao and Hu, *Dual-Disentangled Deep Multiple Clustering*, SIAM SDM 2024.

Multi-view / multi-source

- Wei et al., *Multi-View Multiple Clusterings Using Deep Matrix Factorization*, AAAI 2020.
- Xu et al., *Deep Embedded Multi-View Clustering with Collaborative Training*, Information Sciences 573, 2021.
- Wei et al., *Deep Incomplete Multi-View Multiple Clusterings*, arXiv:2010.02024, 2020.
- Xu et al., *Multi-Level Feature Learning for Contrastive Multi-View Clustering*, CVPR 2022.
- Xu et al., *Deep Incomplete Multi-View Clustering via Mining Cluster Complementarity*, AAAI 2022.
- Fu et al., *Subspace-Contrastive Multi-View Clustering*, arXiv:2210.06795, 2022.
- Ren et al., *scMCs: A Framework for Single-Cell Multi-Omics Data Integration and Multiple Clusterings*, Bioinformatics 39(4), 2023.

Context / Agents

Context-aware / user-guided

- Yao, Qian, and Hu, *Multi-Modal Proxy Learning Towards Personalized Visual Multiple Clustering*, CVPR 2024.
- Yao, Qian, and Hu, *Customized Multiple Clustering via Multi-Modal Subspace Proxy Learning*, NeurIPS 2024.
- Kwon et al., *Image Clustering Conditioned on Text Criteria*, ICLR 2024.
- Chen et al., *Agent-Centric Personalized Multiple Clustering with Multi-Modal LLMs*, arXiv:2503.22241v3, 2025.

Open-ended / agentic

- Stephan et al., *Text-Guided Alternative Image Clustering*, arXiv:2406.18589, 2024.
- Liu et al., *Organizing Unstructured Image Collections using Natural Language*, arXiv:2410.05217, 2024.

Dataset Provenance

- **D1** Geusebroek et al., *The Amsterdam Library of Object Images*, IJCV 61, 2005 (ALOI).
- **D2** Miklautz et al., *Deep Embedded Non-Redundant Clustering*, AAAI 2020 (NR-Objects); Günnemann et al., *SMVC*, KDD 2014 (CMUface / StickFig protocol).
- **D3** LeCun et al., *Gradient-Based Learning Applied to Document Recognition*, Proc. IEEE, 1998 (MNIST; C-MNIST construction in DDMC).
- **D4** Hu et al., *Finding Multiple Stable Clusterings*, KAIS 51, 2017 (Fruit); DDMC Table 1 (Fruit360 / Card benchmark definitions).
- **D5** Krause et al., *3D Object Representations for Fine-Grained Categorization*, ICCV Workshops 2013 (Stanford Cars); Nilsback and Zisserman, IJCV 2008 (Flowers); Welinder et al., Caltech TR 2010 (CUB-200).
- **D6** Xu et al., MFLVC, CVPR 2022, Table 1: MNIST-USPS, BDGP, CCV, Fashion-MNIST views, and Caltech-5V; original sources are cited therein.
- **D7** Wei et al., DMClusts, AAAI 2020: Yale and Reuters multi-view benchmark protocols; Samaria and Harter, 1994 (ORL).
- **D8** Ren et al., scMCs, Bioinformatics 39(4), 2023: CellMix (GEO GSE126074) and PBMC_3K (10x Genomics).

Benchmark labels and transformations follow the cited method papers; they are not all independent newly collected datasets.

Thank you

Questions and Discussion

Deep Multiple Clustering: From Foundations to Context-Aware Approaches

Useful diversity should be coherent, explainable, and aligned with context.