applied sciences

MDPI

Article

MoLIP-Hasse: Flexible Multimodal Pain Recognition via Möbius Interaction Atoms and Explicit Hasse Cover-Relation Matrix Completion

Bowen Ju * and Chenglizhao Chen

Qingdao Institute of Software, College of Computer Science and Technology, China University of Petroleum (East China), Qingdao 266580, China
* Correspondence: z24070115@s.upc.edu.cn

Abstract

Multimodal pain recognition usually assumes a fixed sensor set, yet, in clinical practice, video, electrocardiography (ECG), electromyography (EMG), and galvanic skin response (GSR) channels may become unavailable, removing modality-specific evidence and higher-order cross-modal interactions. We propose MoLIP-Hasse (Möbius-lattice interaction prediction with boundary-clamped Hasse completion), which treats missing modalities as structured completion of interaction atoms on a Boolean subset lattice. A shared Transformer subset encoder and vector-valued Möbius inversion decompose representations into atoms with fixed subset identities. Atoms computable from the observed modalities are kept as immutable boundaries; unknown atoms are completed by bidirectional gated propagation over modality-typed Hasse cover-relation matrices with layer-wise re-clamping, and a fixed Zeta transform reconstructs the complete representation. In 87-fold leave-one-subject-out evaluation on BioVid Part A (no pain vs. tolerance-level pain), a single model handles all 15 non-empty modality combinations with 80.69% average accuracy and 80.05% Macro-F1, reaching 90.25% accuracy with video and GSR. Under identical encoders and protocol, its average accuracy exceeds the best of six missing-modality baselines by 3.74 percentage points. Removing Hasse propagation lowers average accuracy by 2.87 percentage points, and removing clamping raises the known-atom error from below 10^{-6} to 0.081. Structured interaction completion enables unified pain recognition under dynamic modality availability.

Keywords: pain recognition; missing-modality learning; Möbius inversion; interaction atoms; Hasse diagram; multimodal learning; BioVid

Check for updates

Academic Editor: Alexander N. Pisarchik

Received: 23 October 2026
Revised: 19 November 2026
Accepted: 24 November 2026
Published: 26 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Pain assessment directly affects diagnosis and intervention, but patient self-report is unavailable or unreliable for infants, patients with impaired consciousness, and others with limited ability to communicate, and clinical observation is subjective. Machine learning and deep learning have advanced objective pain recognition [1–3], and physiological signals such as the electrocardiogram (ECG), electromyogram (EMG), and galvanic skin response (GSR; also known as electrodermal activity, EDA) provide a data basis for pain modeling [4].

The BioVid Heat Pain Database provides a standardized benchmark for research on video and physiological signals [5]. Building on it, multimodal pain assessment has progressed from comprehensive analyses [6] to physiological time-series modeling, video–heart-rate fusion, Transformer-based fusion, and robust learning under sensor failure [7–11].

However, most methods still rely on a predefined, fixed combination of modalities [7–9]. In real clinical acquisition, sensors may be unavailable because of cost, comfort, privacy, acquisition conditions, or data loss [12], so fixed-input models adapt poorly to dynamic modality configurations.

Existing missing-modality methods mainly rely on modality-dropout training [13], prompt adaptation [14,15], generative recovery [16], or latent-representation compensation [17]. Prompt-based methods are parameter-efficient, but it is usually difficult to verify whether they compensate for modality-specific information or for higher-order complementary information, and, for combination-level prompts, the number of prompts to be maintained (one per missing combination) grows rapidly with the number of modalities.

Generative methods recover the missing data directly, but the differing difficulty of generating each modality can cause bias and training imbalance [18]. Instead, we complete the structured interaction gap in the classification representation: we identify which unimodal and higher-order interactions remain unknown while keeping the observable evidence unchanged.

To this end, we propose MoLIP-Hasse. The power set of {Video, ECG, EMG, GSR} is organized as a Boolean subset lattice, a shared subset encoder defines a vector-valued subset function, and Möbius inversion decomposes this function into interaction atoms. For the currently available set S, atoms with $A \subseteq S$ can be computed exactly, whereas atoms with $A \not\subseteq S$ are completed over explicit Hasse cover-relation matrices: fixed binary matrices restrict propagation to legal cover edges, known atoms are re-clamped after every propagation layer, and a Zeta transform finally synthesizes the complete representation.

The main contributions of this work are as follows:

(1) We propose a vector-valued Möbius interaction-atom decomposition that maps all 16 subsets into a unified representation space and explicitly separates first-order modality-specific information from higher-order joint components, so that every compensated unit has a fixed subset identity.
(2) We propose a boundary-clamped Hasse completion operator driven by explicit cover-relation matrices. Modality-typed bidirectional matrices restrict messages to legal cover edges, gates update only the unknown region, and observable atoms are restored exactly after every layer.
(3) We introduce a strict-superset path-consistency constraint that requires the same completed atoms to reconstruct every strict superset of the available set, so that the atoms remain consistent along all relevant Zeta paths and a single model controlled only by the modality mask handles all 15 non-empty combinations of four modalities.
(4) Under identical modality encoders and an 87-fold leave-one-subject-out (LOSO) protocol, we compare MoLIP-Hasse with six missing-modality learning methods covering modality dropout, generative completion, shared–specific representation, prompt adaptation, and parameter-efficient adaptation. We further evaluate multi-level pain recognition and non-random missingness and analyze how the computational complexity grows with the number of modalities.

The remainder of this paper is organized as follows. Section 2 reviews related work, Section 3 describes MoLIP-Hasse, Section 4 presents the experimental setup, Section 5 reports the main results, and Section 6 presents ablation studies. Section 7 discusses trade-offs, scalability, and limitations, and Section 8 concludes the paper.

2. Related Work

2.1. Facial Expression-Based Pain Assessment

Facial images and videos are important sources of information for objective pain analysis, and the corresponding methods have evolved from hand-crafted features to deep networks [19]. Reviews and meta-analyses have summarized the main methods, scales, datasets, and metrics and have confirmed the potential of multi-level assessment [20,21]. However, visual performance is sensitive to head pose, occlusion, individual expressiveness, and psychological state, and static images cannot describe how expressions evolve over time.

To exploit dynamic information, Patania et al. [22] built graphs from facial keypoints and their motion features and modeled spatiotemporal relations with a deep graph convolutional neural network (DGCNN), and Pittara et al. [23] compared a convolutional neural network (CNN), VGG16, the convolutional vision Transformer (CvT), and MobileViT on the BioVid binary task. The former relies on keypoints and hand-designed graph construction, and the latter selects only representative frames; neither learns the temporal dynamics directly from the full video clip.

Self-supervised video models learn temporally consistent representations through prediction in latent space. V-JEPA 2, a video joint-embedding predictive architecture (V-JEPA), jointly models spatiotemporal video information and outputs dense tokens [24], and V-JEPA 2.1 further improves the quality of dense features [25]. We adopt a frozen V-JEPA 2.1 ViT-B/16, i.e., a base-size Vision Transformer (ViT) with 16 × 16 patches, with the official pretrained weights (https://github.com/facebookresearch/vjepa2, accessed on 16 April 2026) and fuse its temporal tokens and their first- and second-order differences with a lightweight motion adapter.

Overall, a single visual modality is unstable under weak expressions, low image quality, or camera unavailability, which calls for combining it with physiological signals and accounting for missing modalities.

2.2. Multimodal Pain Assessment

Multimodal pain assessment combines visual cues with ECG, EMG, GSR, and other signals to reduce dependence on any single cue. Kächele et al. [26] compared several fusion strategies for hand-crafted features in bio-visual fusion and later fused facial geometry, dynamic texture, and physiological statistics with random forests for person-independent classification and continuous intensity estimation [27], confirming the complementarity of visual and physiological information.

These methods depend on hand-crafted features and fixed fusion rules, so different modality combinations usually require redesign or retraining, and their continuous estimation distinguishes low pain levels poorly [27].

Recent studies have turned to deep temporal and attention models. MMAPA fuses video with functional near-infrared spectroscopy (fNIRS) and GSR using MARLIN, the electroencephalography (EEG) Conformer, and bidirectional attention [28]; Twins-PainViT processes video and fNIRS representations with a shared PainViT [29]; and Farmani et al. [30] fused video, GSR, and ECG with a two-stream network combining a CNN and long short-term memory (LSTM).

These methods strengthen temporal and cross-modal modeling, but their input combinations are still mostly predefined, and arbitrary missing modalities are not handled explicitly. Some experiments also show that adding a weak modality can degrade performance because of noise or redundancy [28–30].

The key problem has therefore shifted from fusing more information to modeling consistently and robustly when the set of available modalities changes. Accordingly, we cover all non-empty combinations of video, ECG, EMG, and GSR.

2.3. Missing-Modality Learning

Existing missing-modality methods fall mainly into training strategies, generative completion, representation compensation, and prompt-based or parameter-efficient adaptation; among these, prompt learning adapts to different missing patterns with a small number of learnable parameters.

Beyond prompt learning, modality-dropout training randomly masks modalities during training so that a single network adapts directly to missing inputs [13]; the missing modality imagination network (MMIN) imagines missing modalities in representation space with cascade residual autoencoders and constrains the reconstruction with cycle consistency [16]; and ShaSpec decomposes each modality representation into shared and specific components and compensates for missing modalities with the shared components of the available ones [17]. These three methods represent training-strategy, generative-completion, and representation-compensation approaches, respectively; we re-implement them under the same encoders, together with prompt-based methods [31,32] and a parameter-efficient adaptation method [33], and compare them with MoLIP-Hasse in Section 5.3.

Lee et al. [31] configured missing-aware prompts for the complete, image-only, and text-only states with small parameter overhead, but samples in the same missing state share a fixed prompt, and the number of combinations grows with the number of modalities.

Jang et al. [32] replaced combination-level prompts with composable modality-specific prompts so that the number of parameters grows approximately linearly, and used an orthogonality constraint to encourage diversity; however, the prompts are still determined mainly by modality identity and lack sample-level adaptivity.

Zhao et al. [34] generated sample-dependent prompts from a modality memory bank and used shared prompts to establish cross-modal links. This approach is more adaptive, but it is limited by prompt length and memory coverage, and its prompts lack an explicit combination identity and verifiable interaction semantics.

Reza et al. [33] adapted to different missing combinations with lightweight scale-and-shift parameters, reducing the cost of training a complete network, but combination-level states still have to be maintained. In contrast, MoLIP-Hasse shares all parameters and uses the fixed Boolean-lattice topology to restrict information flow to adjacent Hasse nodes.

Generating missing data is susceptible to generation error, differences in modality scale, and training imbalance, whereas predicting free compensation vectors in a fusion space makes it difficult to separate unimodal information from higher-order interactions. Consequently, existing methods usually do not specify which components need to be completed and which observed evidence must be preserved.

2.4. Set Functions and Möbius Interactions

Li et al. [35] characterized the label-relevant information of a modality given the remaining modalities through conditional mutual information, showing that the information lost under different missing combinations is not equivalent; however, they did not decompose combination contributions into sample-level representation units that can be supervised and reconstructed.

Set Transformer provides a permutation-invariant shared modeling basis for unordered inputs [36]. Rota established the theory of Möbius inversion on finite partially ordered sets [37], and recent work has used Möbius coefficients to characterize higher-order interactions with set identities [38]. These works have neither used vector-valued interaction atoms for missing-modality completion nor treated observable atoms as fixed boundaries on a Hasse diagram.

In summary, MoLIP-Hasse extracts the known Möbius atoms exactly from the lower sets of the current available set, completes the unknown atoms with these atoms as boundaries,

and synthesizes the target representation with the Zeta transform, so that the identity, number, and reconstruction target of the compensated units are unified on the same Boolean subset lattice.

3. Methodology

3.1. Overview

The end-to-end pipeline of MoLIP-Hasse is shown in Figure 1. Given any set of available modalities, the video encoder and three physiological encoders first produce modality tokens of a common dimension. A shared Transformer subset encoder then computes the direct representations of the observable lower sets, and Möbius inversion decomposes these cumulative subset representations into interaction atoms with fixed subset identities. Atoms that can be computed exactly serve as the known boundary, while the unknown atoms are completed by three layers of bidirectional gated propagation over explicit Hasse cover-relation matrices. Finally, the completed atoms are summed by a fixed Zeta transform into an estimate of the complete-modality representation, from which the pain class is predicted.

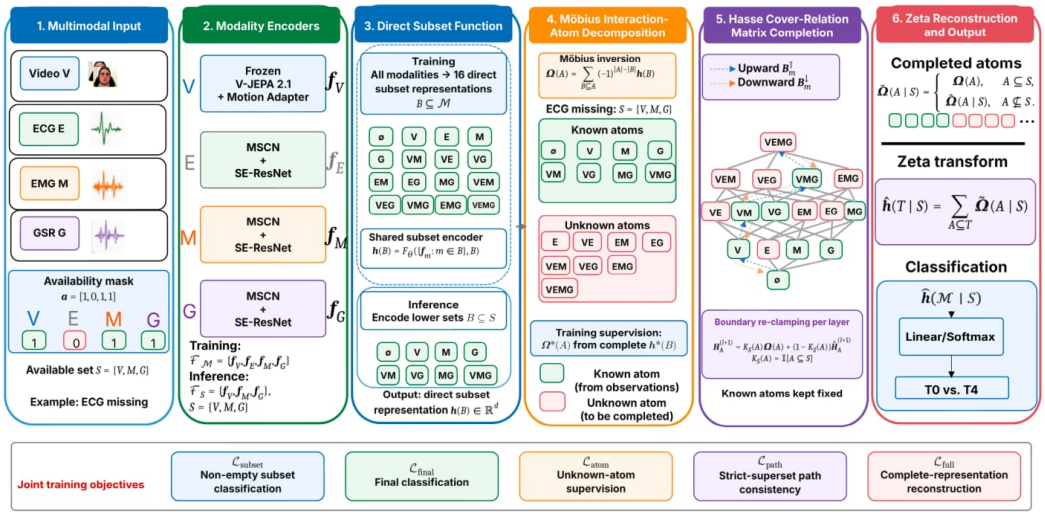

Figure 1. The overall architecture of MoLIP-Hasse, illustrated with ECG missing. The five training objectives are shown at the bottom.

The remainder of this section follows the data flow in Figure 1. Sections 3.2–3.4 describe the modality encoders; Section 3.5 defines the direct subset representations; Section 3.6 decomposes them into interaction atoms; Section 3.7 separates the known boundary from the unknown atoms; Section 3.8 completes the unknown atoms over the Hasse cover-relation matrices; and Sections 3.9 and 3.10 describe Zeta synthesis, classification, and the training objectives. The layer-level equations of Section 3.8 (Equations (18)–(27)) can be skipped at first reading without losing the main argument, and the main symbols are summarized in Table A1 (Appendix A).

Let the modality universe be $\mathcal{M} = \{V,E,M,G\}$, corresponding to video, ECG, EMG, and GSR. For the b-th sample, we define the availability mask $\boldsymbol{a}^{(b)} \in \{0,1\}^4$, whose entry $a_m^{(b)} = 1$ indicates that modality m is available, and obtain the available set $S^{(b)}$ (Equation (1)). The four modalities yield $2^4 - 1 = 15$ non-empty combinations.

$$\mathcal{M} = \{V,E,M,G\}, \quad S^{(b)} = \left\{ m \in \mathcal{M} \,\middle|\, a_m^{(b)} = 1 \right\} \tag{1}$$

The video branch extracts spatiotemporal tokens with a frozen V-JEPA 2.1 ViT-B/16. Following PainAttnNet [7], the physiological branch extracts ECG, EMG, and GSR repre-

sentations with a multiscale convolutional network (MSCN) and a squeeze-and-excitation residual network (SE-ResNet). Each of the four encoders outputs one modality token of dimension d. The back end consists of the shared subset encoder, Möbius interaction decomposition, extraction of the observable boundary, completion over explicit Hasse cover-relation matrices, Zeta reconstruction, and classification.

During training, two data streams are constructed from the same complete sample: the complete-supervision branch computes all 16 direct subset representations and the reference Möbius atoms in a batch, while the simulated-missingness main branch samples the current available set S uniformly from the 15 non-empty sets, keeps only the modality tokens in S, and completes the unknown atoms. At inference there is no complete-supervision branch; the subset encoder runs only on the observable lower sets $B \subseteq S$, which reduces the computation from 16 subsets to $2^{|S|}$ subsets, and no actually missing modality is used.

To avoid ambiguity, we adopt the following notation throughout (see also Table A1 in Appendix A): $\mathcal{M} = \{V,E,M,G\}$ denotes the modality universe and $n = |\mathcal{M}| = 4$ the number of modalities; $S \subseteq \mathcal{M}$ denotes the current available set and $\overline{S} = \mathcal{M} \setminus S$ the current missing set; A, B, and R denote general modality subsets; and $T \supseteq S$ is used for the target subset in Zeta synthesis. B_s is the original batch size, N_{pair} the number of sample–subset pairs after parallel expansion, $d = 64$ the common feature dimension, and l the index of the Hasse propagation layer. Symbols used only within one subsection, such as the sizes of the video and physiological features, are defined where they appear. The superscript "*" denotes quantities from the complete-supervision branch, the hat "^" denotes network predictions, and the tilde "~" denotes completed quantities that merge the known boundary with the predicted unknown atoms.

3.2. Motion-Aware Video Encoder

Each 5.5 s clip is first converted into 16 aligned face frames of 384 × 384 pixels (Section 4.2). The video tensor is fed into the frozen V-JEPA 2.1 ViT-B/16 backbone, which uses 16 × 16 spatial patches. With a tubelet size of 2, the 16 frames are divided into 8 temporal tubelets and a 24 × 24 spatial patch grid, giving 4608 spatiotemporal tokens of dimension 768 [24,25]. The backbone remains frozen in all LOSO folds; only the motion adapter and the subsequent modules are trained end to end.

$$\boldsymbol{Z}_V = E_{\text{V-JEPA}}(\boldsymbol{X}_V) \in \mathbb{R}^{B_s \times N_V \times D_V} \tag{2}$$

where $N_t = 8$ is the number of temporal tubelets, $N_h = N_w = 24$ are the height and width of the spatial patch grid, $N_V = N_t N_h N_w = 4608$ is the number of spatiotemporal tokens, and $D_V = 768$ is the token dimension of the backbone; $t \in \{1,2,\ldots,N_t\}$ indexes the temporal tubelets.

The second dimension of the V-JEPA output is a flattened sequence of spatiotemporal tokens and cannot be treated directly as a temporal sequence. To characterize facial changes strictly along time, the output is first rearranged in the original spatiotemporal grid order into $B_s \times 8 \times 24 \times 24 \times 768$, and the 24 × 24 spatial tokens within each temporal tubelet are average-pooled to obtain a temporal representation of size $B_s \times 8 \times 768$.

$$\overline{\boldsymbol{z}}_{V,t} = \frac{1}{N_h N_w} \sum_{h=1}^{N_h} \sum_{w=1}^{N_w} \boldsymbol{z}^{\text{grid}}_{V,t,h,w}, \qquad t = 1,2,\ldots,N_t. \tag{3}$$

First- and second-order adjacent differences are then computed along the 8 temporal tubelets only. To keep the original representation and the two difference sequences the same length, the first position of each difference sequence is padded with a zero vector; the second-order difference is obtained by applying the adjacent difference again to the

first-order difference. This prevents adjacent spatial patches within the same tubelet from being misinterpreted as temporal changes.

$$\begin{aligned} \Delta\bar{z}_{V,1} &= \mathbf{0}, \qquad \Delta\bar{z}_{V,t} = \bar{z}_{V,t} - \bar{z}_{V,t-1}, \\ \Delta^2\bar{z}_{V,1} &= \mathbf{0}, \qquad \Delta^2\bar{z}_{V,t} = \Delta\bar{z}_{V,t} - \Delta\bar{z}_{V,t-1}, \qquad t \geq 2. \end{aligned} \tag{4}$$

The original representation and its first- and second-order differences at each time step are concatenated along the feature dimension to form a motion-enhanced sequence of size $B_s \times 8 \times 2304$. A lightweight motion adapter consisting of LayerNorm, two linear layers, and a Gaussian error linear unit (GELU) activation maps the 2304-dimensional features to the common dimension $d = 64$; the result is averaged over the 8 temporal tubelets and layer-normalized to obtain a single video token $\boldsymbol{f}_V$. Using a single token still retains local temporal dynamics, because the local states and their first- and second-order temporal changes are explicitly encoded before pooling.

$$\boldsymbol{f}_V = \mathrm{LN}\left(\frac{1}{N_t} \sum_{t=1}^{N_t} \mathrm{MLP}\left(\left[\bar{z}_{V,t}; \Delta\bar{z}_{V,t}; \Delta^2\bar{z}_{V,t} \right] \right) \right), \boldsymbol{f}_V \in \mathbb{R}^{B_s \times d}. \tag{5}$$

3.3. Multiscale Convolutional Network (MSCN)

Physiological signals are non-stationary and exhibit local variations at multiple scales. The MSCN, adopted from PainAttnNet [7], uses parallel branches with different kernel sizes to extract local waveform patterns under different short receptive fields; after several convolution and pooling layers, each branch develops effective receptive fields of different extents.

The MSCN consists of three parallel convolutional branches with kernel sizes $k \in \{3,9,27\}$, corresponding to receptive fields of about 5.9, 17.6, and 52.7 ms at a sampling rate of $f_s = 512$ Hz. Each branch contains three convolutional blocks (Conv1D + BatchNorm + GELU [39]), with max pooling after the first two blocks for downsampling. The outputs of the three branches are aligned to a common temporal length $L_{\mathrm{phy}} = 75$ by adaptive average pooling and concatenated along the channel dimension, yielding the multiscale feature $\boldsymbol{F}_{\mathrm{MSCN}} \in \mathbb{R}^{B_s \times 192 \times L_{\mathrm{phy}}}$, where 192 corresponds to the 64 channels of each of the three branches.

3.4. Squeeze-and-Excitation Residual Network (SE-ResNet)

SE-ResNet combines squeeze-and-excitation (SE) channel attention [40] with residual connections [41] and recalibrates features by modeling nonlinear dependencies between channels. For the 192-channel multiscale input, the main branch first maps the channels to 30 through two 1×1 convolutional layers, while the residual branch projects the input from 192 to 30 channels with an independent 1×1 convolution so that the channel dimensions match in the residual addition. The squeeze operation then compresses the temporal dimension of each main-branch channel into a scalar descriptor through global average pooling, and the excitation operation uses a bottleneck of two fully connected layers with reduction ratio $r = 4$, followed by a sigmoid that produces channel weights in $(0,1)$. The channel weights are multiplied channel-wise with the main-branch features and then added to the projected residual features, recalibrating the channels while preserving the input information. SE-ResNet finally outputs the 30-channel feature $\boldsymbol{F}_{\mathrm{SE}} \in \mathbb{R}^{B_s \times 30 \times L_{\mathrm{phy}}}$; global average pooling over time and a linear projection then map it to the d-dimensional token of each physiological modality.

3.5. Shared Subset Encoder

We use a parameter-shared Transformer encoder [42]. Here, shared means that all subsets of the four-modality power set use the same encoder parameters and the same

classification head, so that the direct representations of different modality combinations lie in the same vector space and can serve as values of a single set function for the subsequent Möbius inversion.

For any subset $B \subseteq \mathcal{M}$, the encoder constructs a token sequence of length 5:

$$\boldsymbol{X}_B = \left[\boldsymbol{q}_{\text{pool}} + \boldsymbol{e}_B,\ \boldsymbol{f}_V + \boldsymbol{e}_V,\ \boldsymbol{f}_E + \boldsymbol{e}_E,\ \boldsymbol{f}_M + \boldsymbol{e}_M,\ \boldsymbol{f}_G + \boldsymbol{e}_G\right] \in \mathbb{R}^{B_s \times 5 \times d} \tag{6}$$

where q_{pool} is a learnable pooling token, $\boldsymbol{e}_B$ is the subset-identity embedding, and $\boldsymbol{e}_m$ are the modality-type embeddings; positions of modalities not in B are fully masked by the padding mask and therefore do not participate in self-attention. The two-layer Transformer encoder preserves the sequence shape and models cross-modal relations within the current subset with four-head self-attention, and the output at the pooling position is taken as the direct subset representation $\boldsymbol{h}(B) \in \mathbb{R}^{B_s \times d}$. When multiple subsets are flattened and computed in parallel, the total number of pairs is denoted N_{pair}, and the shape after scattering back to the fixed node table is again $B_s \times 16 \times d$.

For example, when $B = \{V,G\}$, the pooling token can attend to the video and GSR tokens while the ECG and EMG positions are masked, and the output $\boldsymbol{h}(VG)$ is the cumulative representation of the sample under joint observation of video and GSR. During training, all 16 subsets are expanded at once; at inference, only the $2^{|S|}$ observable lower sets satisfying $B \subseteq S$ are encoded and scattered back to the 16-node table by fixed indices. The role of this encoder is not to guess the missing modalities directly but to define the sample-level vector-valued set function $\boldsymbol{h}{:}2^{\mathcal{M}} \to \mathbb{R}^d$, which provides comparable inputs for the interaction decomposition.

$$\boldsymbol{h}(B) = F_\theta\left(\{\boldsymbol{f}_m : m \in B\}, B\right) \in \mathbb{R}^{B_s \times d} \tag{7}$$

The four modalities yield 16 subsets, including the empty set $\emptyset$, whose baseline representation is produced only from the pooling token and the empty-set identity embedding.

$$\boldsymbol{H} = \left[\boldsymbol{h}(\emptyset), \boldsymbol{h}(V), \boldsymbol{h}(E), \ldots, \boldsymbol{h}(VEMG)\right] \in \mathbb{R}^{B_s \times 16 \times d} \tag{8}$$

During training, the shared linear classification head $g_{\text{cls}}{:}\mathbb{R}^d \to \mathbb{R}^2$ computes an auxiliary classification loss on the 15 non-empty direct subsets; the empty set is not used for classification supervision.

$$\mathcal{L}_{\text{subset}} = \frac{1}{15}\sum_{B \neq \emptyset} \text{CE}\left(g_{\text{cls}}\left(\boldsymbol{h}^*(B)\right), y\right) \tag{9}$$

The shared subset encoder defines the direct subset function $\boldsymbol{h}(B)$ using standard Transformer components. The core contribution of this work lies not in the encoder itself but in the exact interaction decomposition, the explicit Hasse cover topology, boundary clamping, and Zeta synthesis.

Note that the subset-identity embedding $\boldsymbol{e}_B$ contributes to the direct representations; therefore, a Möbius atom represents a structured increment in the model's representation space that is jointly determined by the observed modalities and the subset identity and is assigned to that subset identity.

3.6. Möbius Interaction-Atom Decomposition

On the Boolean lattice $2^{\mathcal{M}}$, we define for every interaction set $A \subseteq \mathcal{M}$ a vector-valued Möbius interaction atom $\boldsymbol{\Omega}(A)$. Under the subset partial order, $\boldsymbol{\Omega}(A)$ is uniquely determined from the representations of all lower sets of A by inclusion–exclusion [37,38]:

$$\boldsymbol{\Omega}(A) = \sum_{B \subseteq A} (-1)^{|A|-|B|}\boldsymbol{h}(B) \tag{10}$$

The direct subset representation $\boldsymbol{h}(B)$ contains the unimodal evidence of each modality in B together with lower-order interactions and higher-order joint information; therefore, $\boldsymbol{h}(B)$ cannot be interpreted directly as an independent or causal contribution of subset B. Möbius inversion successively subtracts the components already explained by all proper subsets, converting cumulative representations into interaction atoms that are uniquely assigned to fixed subset identities.

Taking first- and second-order interactions as examples, $\boldsymbol{\Omega}(V) = \boldsymbol{h}(V) - \boldsymbol{h}(\emptyset)$ and $\boldsymbol{\Omega}(VG) = \boldsymbol{h}(VG) - \boldsymbol{h}(V) - \boldsymbol{h}(G) + \boldsymbol{h}(\emptyset)$. Here $\boldsymbol{\Omega}(V)$ retains the component that video adds relative to the empty-set baseline, and $\boldsymbol{\Omega}(VG)$ retains the residual joint component that appears only when video and GSR are observed together and cannot be explained by the empty-set baseline plus the two first-order atoms. Third- and fourth-order atoms are obtained recursively by the same inclusion–exclusion rule.

The Möbius transform is exactly invertible: the representation of any subset can be recovered from all of its interaction atoms by Zeta summation:

$$\boldsymbol{h}(S) = \sum_{A \subseteq S} \boldsymbol{\Omega}(A) \tag{11}$$

The Zeta transform is therefore a fixed operation rather than another generative network to be learned. Once the atom values are given, the representation of any target subset T is obtained by directly summing the atoms of all its lower sets; when all atoms take their complete-observation reference values, the reconstruction coincides algebraically with the original $\boldsymbol{h}(T)$. The model can thus localize what is missing precisely as the atoms with $A \not\subseteq S$ and restrict the completion target to these unknown interactions without modifying the observable evidence $A \subseteq S$.

Apart from the empty-set baseline $\boldsymbol{\Omega}(\emptyset)$, the four modalities yield 15 non-empty atoms: four first-order atoms represent the increments of single modalities relative to the empty-set baseline, six second-order atoms represent increments under joint observation of two modalities, and four third-order atoms and one fourth-order atom represent higher-order joint increments.

Given the current available set S, the difference between the complete and current representations can be written exactly as the sum of all interaction atoms not contained in S:

$$\boldsymbol{h}(M) - \boldsymbol{h}(S) = \sum_{A \not\subseteq S} \boldsymbol{\Omega}(A) \tag{12}$$

For instance, when ECG is missing and $S = \{V,M,G\}$, the known non-empty atoms are V, M, G, VM, VG, MG, and VMG (seven in total), and the unknown atoms are the eight interactions containing E. When only GSR is available, only $\boldsymbol{\Omega}(G)$ is a known non-empty atom, and the remaining 14 non-empty interactions must be completed.

3.7. Known Boundary and Initialization of Unknown Atoms

For the current set S, all atoms satisfying $A \subseteq S$ depend only on the lower-set representations of the available modalities and can therefore be computed exactly without using any missing modality. We denote the observable boundary by $\mathcal{B}(S) = \{\boldsymbol{\Omega}(A) : A \subseteq S\}$ and the corresponding known-node mask by $K_S(A) = \mathrm{II}[A \subseteq S]$. The remaining nodes with $A \not\subseteq S$ form the unknown region.

During training, the reference atoms $\boldsymbol{\Omega}^*(A)$ are obtained from the 16 direct subset representations of the complete sample, and the boundary values with $A \subseteq S$ are selected from them; at inference, only $B \subseteq S$ are enumerated and a local Möbius inversion is performed. To keep the algebraic structure fixed, all 16 lattice nodes are retained, and the known mask in Equation (13) indicates which of them form the boundary.

$$K_S(A) = \mathrm{II}[A \subseteq S], \qquad U_S(A) = 1 - K_S(A) \tag{13}$$

For unknown nodes, propagation does not start from all-zero vectors; instead, a structured initialization is constructed. Let $\boldsymbol{r}_A \in \{0,1\}^4$ be the set identity of node A, which is mapped by a two-layer network F_{id} to the node-identity code $\mathbf{e}_A = F_{\mathrm{id}}(\boldsymbol{r}_A) \in \mathbb{R}^d$. The local context of each node is then computed from the available modality tokens:

$$\boldsymbol{c}_A^S = \begin{cases} \frac{1}{|A \cap S|} \sum_{m \in A \cap S} \boldsymbol{f}_m, A \cap S \neq \emptyset \\ 0, A \cap S = \emptyset \end{cases} \tag{14}$$

where $\boldsymbol{c}_A^S$ is the average of the available modality tokens contained in $A \cap S$; when $A \cap S = \emptyset$, $\boldsymbol{c}_A^S$ is defined as the zero vector to avoid division by zero. $\boldsymbol{c}_S$ is the global average of all available modality tokens; in addition, $\mathcal{F}_S = \{\boldsymbol{f}_m : m \in S\}$ denotes the set of currently available modality tokens.

$$\boldsymbol{H}_A^{(0)} = \begin{cases} \boldsymbol{\Omega}(A), & A \subseteq S, \\ I_\theta(A, \mathcal{F}_S, S), & A \not\subseteq S \end{cases} \tag{15}$$

$$I_\theta(A, \mathcal{F}_S, S) = \mathrm{LN}\left(\mathrm{MLP}\left(\left[\boldsymbol{q}_A; \boldsymbol{e}_A; \boldsymbol{c}_A^S; \boldsymbol{c}_S; \boldsymbol{e}_S + \boldsymbol{e}_{A \cap \overline{S}}\right]\right)\right) \tag{16}$$

Specifically, $\boldsymbol{q}_A \in \mathbb{R}^d$ is the learnable query of node A. The current available set is encoded as $\boldsymbol{e}_S = F_{\mathrm{set}}(\boldsymbol{s}) \in \mathbb{R}^d$, where $\boldsymbol{s}$ denotes the availability mask $\boldsymbol{a}^{(b)}$ of Section 3.1 with the sample index b omitted, and the missing-modality signature of the node is $\boldsymbol{e}_{A \cap \overline{S}} = F_{\mathrm{miss}}(\boldsymbol{r}_A \odot (\boldsymbol{1} - \boldsymbol{s})) \in \mathbb{R}^d$. The initializer concatenates $\left[\boldsymbol{q}_A; \boldsymbol{e}_A; \boldsymbol{c}_A^S; \boldsymbol{c}_S; \boldsymbol{e}_S + \boldsymbol{e}_{A \cap \overline{S}}\right] \in \mathbb{R}^{5d}$ and passes the result through a two-layer multilayer perceptron (MLP) of $5d \to 192 \to d$, GELU, and LayerNorm to obtain the initial states of the unknown nodes.

3.8. Bidirectional Propagation over Hasse Cover Matrices

Figure 2 illustrates the construction of the explicit Hasse cover-relation matrices, the node initialization, and one propagation layer. The left panel lists the fixed binary indices of the 16 subset nodes and the observable lower sets when ECG is missing; the middle panel shows the modality-typed upward cover-relation matrices defined by adding one modality and their transposed downward matrices, and distinguishes the known boundary obtained exactly by Möbius inversion from the unknown nodes produced by the initializer; the right panel expands the data flow of one layer: modality transform, source gating, cover-matrix routing, target gating, summation over the four modalities, candidate update, and boundary re-clamping. This layer is applied three times, corresponding to $L_H = 3$. For any current available set $S \neq \emptyset$ and unknown atom $A \not\subseteq S$, node $A \cap S$ belongs to the exactly computable known boundary. Going from $A \cap S$ to A only requires adding the missing modalities in $A \cap \overline{S}$ one at a time, so the shortest Hasse path length satisfies

$$\mathrm{dist}(A \cap S, A) = \left|A \cap \overline{S}\right| \leq \left|\overline{S}\right| \leq 3 \tag{17}$$

Hence, for non-empty four-modality inputs, three propagation layers cover the maximum structural distance between an unknown atom and its known intersection boundary, and we therefore fixed $L_H = 3$ before the formal experiments.

The Hasse cover relation of the Boolean subset lattice connects only $A \leftrightarrow A \cup \{m\}$; that is, each edge adds or removes exactly one modality. For modality m, we construct the upward binary cover-relation matrix $\boldsymbol{B}_m^{\uparrow} \in \{0,1\}^{16 \times 16}$, whose rows index target nodes A_i and whose columns index source nodes A_j in the fixed 16-node index table:

$$\boldsymbol{B}_m^{\uparrow}[i,j] = \begin{cases} 1, & A_i = A_j \cup \{m\},\ m \notin A_j, \\ 0, & \text{otherwise}, \end{cases} \qquad \boldsymbol{B}_m^{\downarrow} = \left(\boldsymbol{B}_m^{\uparrow}\right)^{\top} \tag{18}$$

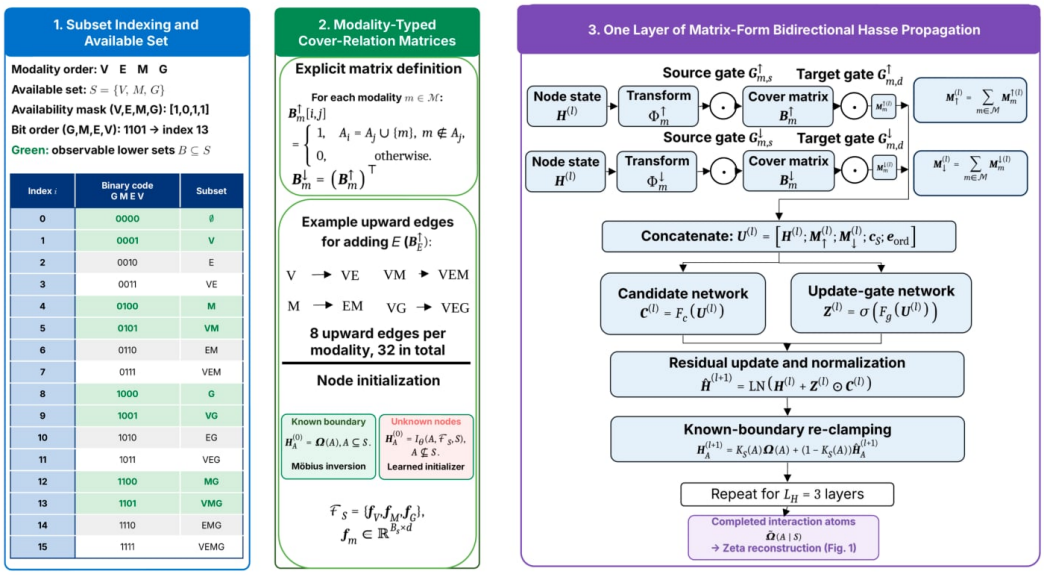

Figure 2. Hasse completion driven by explicit cover-relation matrices, illustrated with ECG missing.

With four modalities, each upward cover-relation matrix $\boldsymbol{B}_m^{\uparrow}$ contains exactly eight non-zero cover edges, giving 32 upward edges in total. The fixed cover-relation matrices encode only legal Hasse cover edges; the source gate $\boldsymbol{G}_{m,s}^{\mathrm{dir},(l)}$ and target gate $\boldsymbol{G}_{m,d}^{\mathrm{dir},(l)}$ reweight legal messages feature-wise and are both generated from the current node states by modality- and direction-specific $d \to d$ linear maps followed by a sigmoid. Each gate tensor therefore has shape $B_s \times 16 \times d$ rather than being a single scalar gate.

$\Phi_m^{\uparrow}$ and $\Phi_m^{\downarrow}$ are modality- and direction-specific linear transforms without bias, i.e., $\Phi_m^{\mathrm{dir}}(\boldsymbol{H}) = \boldsymbol{H}\boldsymbol{W}_m^{\mathrm{dir}}$ and $\boldsymbol{W}_m^{\mathrm{dir}} \in \mathbb{R}^{d \times d}$; the gate parameters are $\boldsymbol{W}_{m,s}^{\mathrm{dir}}$, $\boldsymbol{W}_{m,d}^{\mathrm{dir}} \in \mathbb{R}^{d \times d}$ with biases $\boldsymbol{b}_{m,s}^{\mathrm{dir}}$, $\boldsymbol{b}_{m,d}^{\mathrm{dir}} \in \mathbb{R}^d$. The available-set context $\boldsymbol{c}_S \in \mathbb{R}^d$ is broadcast to all 16 nodes. The interaction-order embedding is defined as $\boldsymbol{e}_{\mathrm{ord}}(A) = \mathrm{Emb}_{\mathrm{ord}}(|A|) \in \mathbb{R}^d$, which becomes $B_s \times 16 \times d$ after batch expansion. The candidate network F_c and the update-gate network F_g are both two-layer MLPs of $5d \to 192 \to d$; F_c uses GELU, and F_g ends with a sigmoid. Node states, feature-wise gates, and aggregated upward and downward messages all have shape $B_s \times 16 \times d$.

$$\boldsymbol{G}_{m,s}^{\mathrm{dir},(l)} = \sigma\left(\boldsymbol{W}_{m,s}^{\mathrm{dir}}\boldsymbol{H}^{(l)} + \boldsymbol{b}_{m,s}^{\mathrm{dir}}\right) \tag{19}$$

$$\boldsymbol{G}_{m,d}^{\mathrm{dir},(l)} = \sigma\left(\boldsymbol{W}_{m,d}^{\mathrm{dir}}\boldsymbol{H}^{(l)} + \boldsymbol{b}_{m,d}^{\mathrm{dir}}\right) \tag{20}$$

$$\boldsymbol{M}_{\uparrow}^{(l)} = \sum_{m \in \mathcal{M}} \boldsymbol{G}_{m,d}^{\uparrow(l)} \odot \left[\boldsymbol{B}_m^{\uparrow}\left(\boldsymbol{G}_{m,s}^{\uparrow(l)} \odot \Phi_m^{\uparrow}\left(\boldsymbol{H}^{(l)}\right)\right)\right] \tag{21}$$

$$\boldsymbol{M}_{\downarrow}^{(l)} = \sum_{m \in \mathcal{M}} \boldsymbol{G}_{m,d}^{\downarrow(l)} \odot \left[\boldsymbol{B}_m^{\downarrow}\left(\boldsymbol{G}_{m,s}^{\downarrow(l)} \odot \Phi_m^{\downarrow}\left(\boldsymbol{H}^{(l)}\right)\right)\right] \tag{22}$$

For each modality $m \in \mathcal{M}$, the upward branch first computes $\Phi_m^{\uparrow}\left(\boldsymbol{H}^{(l)}\right)$, multiplies it by the feature-wise source gate, routes it through $\boldsymbol{B}_m^{\uparrow}$, and scales the received messages by the feature-wise target gate; the downward branch performs the symmetric operation with $\boldsymbol{B}_m^{\downarrow}$. The results of the four modalities are summed to obtain $\boldsymbol{M}_{\uparrow}^{(l)}$ and $\boldsymbol{M}_{\downarrow}^{(l)}$. We then concatenate the node states, the aggregated messages, the available-set context, and the order embedding (Equation (23)); generate candidate content and a feature-wise update gate (Equations (24) and (25)); and finally obtain the temporary node states through a residual update and LayerNorm (Equation (26)), after which the known atoms are re-clamped to their exact values (Equation (27)).

$$\boldsymbol{U}^{(l)} = \left[\boldsymbol{H}^{(l)}; \boldsymbol{M}_{\uparrow}^{(l)}; \boldsymbol{M}_{\downarrow}^{(l)}; \boldsymbol{c}_S; \boldsymbol{e}_{\mathrm{ord}}\right] \tag{23}$$

$$\boldsymbol{C}^{(l)} = F_c\left(\boldsymbol{U}^{(l)}\right) \tag{24}$$

$$\boldsymbol{Z}^{(l)} = \sigma\left(F_g\left(\boldsymbol{U}^{(l)}\right)\right) \tag{25}$$

$$\hat{\boldsymbol{H}}^{(l+1)} = \mathrm{LN}\left(\boldsymbol{H}^{(l)} + \boldsymbol{Z}^{(l)} \odot \boldsymbol{C}^{(l)}\right) \tag{26}$$

$$\boldsymbol{H}_A^{(l+1)} = K_S(A)\,\boldsymbol{\Omega}(A) + (1 - K_S(A))\hat{\boldsymbol{H}}_A^{(l+1)} \tag{27}$$

This clamping is applied after every layer. The error of the clamped known atoms with respect to reference atoms recomputed from the complete test modalities (for evaluation only) is zero in exact arithmetic; the measured error, caused only by floating-point round-off, was below 10^{-6}. The constraint acts on node states rather than on shared network parameters; the parameters of the gates, modality transforms, and node-update networks are still optimized end to end by the joint loss.

After three layers, the final states of all 16 nodes are obtained. The completed atom table contains both the immutable known boundary and the unknown-atom predictions $\hat{\boldsymbol{\Omega}}(A \mid S) = \boldsymbol{H}_A^{(L_H)}$ obtained by three layers of Hasse propagation, and it is the sole input to the subsequent Zeta reconstruction.

$$\tilde{\boldsymbol{\Omega}}(A \mid S) = \begin{cases} \boldsymbol{\Omega}(A), & A \subseteq S, \\ \hat{\boldsymbol{\Omega}}(A \mid S), & A \not\subseteq S. \end{cases} \tag{28}$$

3.9. Zeta Synthesis and Superset-Path Constraint

The completed atoms are summed by the fixed Zeta transform to synthesize an estimate of the representation of any target subset $T \supseteq S$. In the missing case, however, the atoms with $A \subseteq S$ are network predictions, so $\hat{\boldsymbol{h}}(T \mid S)$ is an estimate of the target representation and in general differs from the complete-observation representation.

$$\hat{\boldsymbol{h}}(T \mid S) = \sum_{A \subseteq T} \tilde{\boldsymbol{\Omega}}(A \mid S),\ T \supseteq S \tag{29}$$

During training, we not only supervise the reconstruction of the complete set $\mathcal{M}$ but also require the same completed atoms to reconstruct all strict supersets $T \supset S$ of the current set S. This path constraint discourages erroneous atoms from canceling one another only in the complete-node sum and encourages every atom to be consistent along all relevant Zeta paths.

$$\mathcal{L}_{\text{path}} = \mathrm{Avg}_{T \supset S}\mathrm{SmoothL1}\left(\hat{\boldsymbol{h}}(T \mid S), \mathrm{sg}\left[\boldsymbol{h}^*(T)\right]\right) \tag{30}$$

3.10. Prediction and Training Objectives

Classification is performed on the complete representation reconstructed by the Zeta transform. The final complete representation $\hat{\boldsymbol{h}}(\mathcal{M} \mid S)$ is fed into the shared classification head g_{cls} to obtain the predicted probabilities of T0 and T4:

$$\hat{\boldsymbol{p}} = \mathrm{Softmax}\left(g_{\text{cls}}\left(\hat{\boldsymbol{h}}(\mathcal{M} \mid S)\right)\right) \tag{31}$$

$$\mathcal{L}_{\text{final}} = \mathrm{CE}\left(g_{\text{cls}}\left(\hat{\boldsymbol{h}}(\mathcal{M} \mid S)\right), y\right) \tag{32}$$

The training objective consists of the non-empty subset classification loss $\mathcal{L}_{\text{subset}}$, the final classification loss $\mathcal{L}_{\text{final}}$, the unknown-atom supervision loss $\mathcal{L}_{\text{atom}}$, the strict-superset path loss $\mathcal{L}_{\text{path}}$, and the cosine reconstruction loss of the complete representation $\mathcal{L}_{\text{full}}$:

$$\mathcal{L}_{\text{atom}} = \mathrm{Avg}_{A \not\subseteq S}\mathrm{SmoothL1}\left(\hat{\boldsymbol{\Omega}}(A \mid S), \mathrm{sg}\left[\boldsymbol{\Omega}^*(A)\right]\right) \tag{33}$$

$$\mathcal{L}_{\text{full}} = 1 - \cos\left(\hat{\boldsymbol{h}}(\mathcal{M} \mid S), \mathrm{sg}\left[\boldsymbol{h}^*(\mathcal{M})\right]\right) \tag{34}$$

$$\mathcal{L} = \lambda_{\text{subset}}\mathcal{L}_{\text{subset}} + \lambda_{\text{final}}\mathcal{L}_{\text{final}} + \lambda_{\text{atom}}\mathcal{L}_{\text{atom}} + \lambda_{\text{path}}\mathcal{L}_{\text{path}} + \lambda_{\text{full}}\mathcal{L}_{\text{full}} \tag{35}$$

where sg[·] denotes the stop-gradient operation and Avg denotes averaging first over the valid atom or strict-superset indices of each sample and then over the samples in the batch; a loss term is set to 0 if its valid index set is empty. $\boldsymbol{\Omega}^*(A)$ and $\boldsymbol{h}^*(T)$ are the reference atoms and reference subset representations produced by the complete-supervision branch, and $\hat{\boldsymbol{\Omega}}(A \mid S)$ and $\hat{\boldsymbol{h}}(T \mid S)$ are the predicted unknown atoms and the Zeta-synthesized representations, respectively.

At inference, neither the complete modalities, the reference atoms, nor the auxiliary losses are needed. The model encodes only the currently available modalities, runs $2^{|S|}$ lower-set encodings, extracts the exact boundary atoms, performs three layers of propagation over the explicit Hasse cover-relation matrices, and then applies one fast Zeta transform and the shared linear classifier.

In terms of complexity, direct subset encoding costs $O(2^n C_{\text{set}})$ during training and $O(2^{|S|} C_{\text{set}})$ at inference, where C_{set} denotes the cost of one subset encoding, and the fast Möbius/Zeta transforms cost $O(n \cdot 2^n \cdot d)$. The four-modality lattice has $n \cdot 2^{n-1} = 32$ valid Hasse cover edges. For convenient batching, the current implementation uses 16×16 dense matrix multiplications, but all non-cover positions are strictly zeroed by the fixed binary cover-relation matrices. The complexity bound for a growing number of modalities is analyzed in Section 7.2.

The complete-supervision branch and the simulated-missingness main branch share all encoders and the classification head, and no teacher network is used; the stop-gradient operation only fixes the reference targets of the complete branch.

4. Experimental Setup

4.1. Dataset

Experiments were conducted on Part A of the BioVid Heat Pain Database [5], which contains facial videos and physiological signals recorded synchronously from 87 healthy participants (43 women and 44 men, aged 20–65 years) under controlled heat stimulation.

The data cover five levels, T0–T4: T0 is the 32 °C pain-free baseline, T1–T4 increase from each participant's pain threshold to tolerance, and T4 corresponds to the tolerance threshold. Each participant received 20 stimuli per level, with the peak temperature held for 4 s and random intervals of 8–12 s, giving 8700 samples in total.

Four synchronized modalities were used:

(a) Video: frontal RGB facial video at 25 fps and 1388 × 1038 pixels.
(b) ECG: 512 Hz standard-lead electrocardiogram, reflecting autonomic cardiovascular responses.
(c) EMG: 512 Hz electromyogram of the trapezius muscle, characterizing muscle tension and motor reflexes.
(d) GSR: 512 Hz skin conductance measured at the fingers, reflecting sympathetic activation.

The main task is the commonly used binary classification of T0 vs. T4, which covers the most distinct pain levels and enables fair comparison with existing methods; Section 5.7 further reports the binary tasks T0 vs. T1, T0 vs. T2, and T0 vs. T3, five-class classification, and ordinal regression.

4.2. Data Preprocessing

Physiological signals. The filtered recordings provided with the dataset were used directly, and the 5.5 s segment corresponding to each stimulus was extracted, giving 2816 time points per segment at 512 Hz.

Video preprocessing and feature extraction. Faces were detected and five facial landmarks were located with multi-task cascaded convolutional networks (MTCNNs) [43],

and the facial region was aligned by an affine transform based on the landmarks and cropped. Sixteen frames were sampled uniformly from each 5.5 s stimulus segment, resized to 384 × 384 pixels, and normalized according to the V-JEPA 2.1 input specification. The video token is then extracted as described in Section 3.2.

4.3. Evaluation Protocol

We adopted an outer 87-fold LOSO cross-validation protocol. In fold i, all T0 and T4 samples of participant i form an independent test set of 40 samples (20 per class). From the remaining 86 participants, a predefined subject-level split assigns 80 participants to training and 6 to validation, with 3200 and 240 samples, respectively. Training, validation, and test participants never overlap.

The validation set is used only for checkpoint selection and does not participate in training or in computing any statistics. After every epoch, the model is evaluated on all 15 non-empty combinations for the 6 validation participants, and the checkpoint with the highest subject–combination mean Macro-F1 is selected; the held-out test participant is then evaluated once. The per-fold training/validation assignments are fixed in advance and reused in all ablation and sensitivity experiments.

4.4. Evaluation Metrics

Accuracy, Macro-F1, reconstruction similarity, and interaction-atom reconstruction error were used for evaluation.

(1) Accuracy: the proportion of correctly classified samples,

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{36}$$

where TP, TN, FP, and FN are the numbers of true positives, true negatives, false positives, and false negatives, respectively.

(2) F1 score: the harmonic mean of precision and recall. For class k, precision, recall, and F1 are defined as

$$\text{Precision}_k = \frac{TP_k}{TP_k + FP_k}, \text{Recall}_k = \frac{TP_k}{TP_k + FN_k} \tag{37}$$

$$F1_k = \frac{2 \cdot \text{Precision}_k \cdot \text{Recall}_k}{\text{Precision}_k + \text{Recall}_k} \tag{38}$$

Macro-F1 is the arithmetic mean of the per-class F1 scores:

$$F1_{\text{macro}} = \frac{1}{C} \sum_{k=1}^{C} F1_k \tag{39}$$

where $C = 2$ corresponds to the classes T0 and T4 in the main binary task ($C = 5$ for the five-class task in Section 5.7). Because every participant has the same number of T0 and T4 samples, Macro-F1 mainly complements accuracy by reflecting how balanced the recognition of the two classes is, and it facilitates comparison with studies that report F1.

(3) Reconstruction similarity: the cosine similarity between the Zeta-synthesized representation and the complete-observation reference representation; larger values indicate reconstructions closer to the reference. Section 5.6 reports aggregate similarity statistics and improvement rates and reconstruction similarity distributions with their means and medians, Section 6.2 reports medians, and Section 6.3 reports the arithmetic mean of the medians for the VMG, VG, and V scenarios (S2, S8, and S14; Section 5.4).

(4) Interaction-atom reconstruction error: the normalized L_2 error, i.e., $\left\|\tilde{\boldsymbol{\Omega}}(A \mid S) - \boldsymbol{\Omega}^*(A)\right\|_2 / \left(\|\boldsymbol{\Omega}^*(A)\|_2 + 10^{-8}\right)$, computed separately for known and unknown atoms and averaged over valid atoms and samples. The known-atom error verifies whether layer-wise boundary clamping preserves the observable evidence, and the unknown-atom error evaluates how well Hasse completion recovers the missing interactions; lower is better for both.

4.5. Modality Sampling and Inference Protocol

Testing covers all 15 non-empty combinations of the four modalities, all inferred by the same model. During training, each complete four-modality sample is processed by the complete-supervision branch, and a main-branch available set S is sampled uniformly from the 15 non-empty sets, so that complete, single-missing, double-missing, and triple-missing scenarios all appear in training.

For each training sample, the four modality encoders are run only once. The complete-supervision branch evaluates the 16 direct subsets in a batch and produces the reference subset representations $\boldsymbol{h}^*(B)$ and reference interaction atoms $\boldsymbol{\Omega}^*(A)$; the simulated-missingness main branch uses only the modalities in S for unknown-node initialization and atom completion. At inference, to keep fixed-length tensors, excluded modality positions are filled with zero vectors and prevented by the padding mask from acting as keys or values in attention, so these positions contain no features extracted from missing modalities.

4.6. Implementation Details

The model was implemented in PyTorch (version 2.7.1) and trained with LOSO on a single NVIDIA RTX 3090 GPU (24 GB; NVIDIA Corporation, Santa Clara, CA, USA). The V-JEPA 2.1 ViT-B/16 backbone was frozen, while the motion adapter, the three physiological encoders, the shared subset encoder, and the explicit Hasse completion module were trained end to end. The common token dimension is $d = 64$; the subset Transformer encoder has 2 layers, 4 attention heads, a 128-dimensional feed-forward layer, and a dropout rate of 0.1. The source and target gates both use modality- and direction-specific $d \to d$ linear maps with sigmoid activation and output $d = 64$-dimensional feature-wise weights, and the hidden dimension of the node update is 192. The motion adapter has an input dimension of 2304 and a hidden dimension of 128; the other video-branch settings are given in Section 3.2.

The total loss is the five-term joint objective defined in Section 3.10, with $\lambda_{\text{subset}} = 0.25$, $\lambda_{\text{final}} = 1.00$, $\lambda_{\text{atom}} = 0.80$, $\lambda_{\text{path}} = 0.60$, and $\lambda_{\text{full}} = 0.50$. No additional order-sparsity regularization is used. These weights were set manually, with $\lambda_{\text{final}} = 1$ for the final classification loss and smaller values for the four auxiliary terms; they were not tuned systematically, and the same values were used in all folds and experiments. Their sensitivity was not analyzed (Section 7.3). During training, all loss terms, the maximum known-boundary error, the unknown-atom error, and the per-fold results for the 15 modality combinations were recorded.

AdamW [44] was used as the optimizer, with an initial learning rate of $\eta_{\text{enc}} = 1 \times 10^{-4}$ for the encoders and $\eta_{\text{core}} = 5 \times 10^{-5}$ for the MoLIP-Hasse core modules, a weight decay of $\lambda_{\text{wd}} = 1 \times 10^{-4}$, and a batch size of 16. Each fold was trained for at most 100 epochs; after every epoch, a checkpoint was selected on the validation participants according to the rule in Section 4.3, and after training the best checkpoint was loaded and evaluated once on the held-out test participant. The architecture, loss weights, learning rates, Hasse propagation depth (Section 3.8), and maximum number of epochs were all fixed before the outer LOSO evaluation, and neither validation nor test participants were used to select any hyperparameter.

5. Results and Analysis

5.1. Compared Methods

We compare against two groups of methods: (1) five representative pain recognition methods that report LOSO results for T0 vs. T4 on BioVid, whose results are taken from the original papers; and (2) six missing-modality learning methods re-implemented with the same modality encoders, data split, and training budget. The first group includes the following:

Gkikas et al. [8]: video representations extracted with two-stage pretraining are fused through a Transformer with heart-rate signals derived from the ECG.

CrossMod-Transformer [10]: GSR and ECG are fused with cross-modal attention.

MM-SCAAE [11]: GSR, ECG, and EMG are fused with contrastive pretraining and adversarial autoencoders.

TNT-based Transformer [45]: a two-level Transformer-in-Transformer (TNT) architecture models local facial regions and global temporal dynamics in video.

PainFormer [46]: a vision foundation model extracts video (RGB and synthetic thermal and depth videos derived from it) and GSR representations; it reports the best previously published result for this configuration.

The second group consists of missing-modality learning methods covering training strategies, generative completion, representation compensation, prompt adaptation, and parameter-efficient adaptation:

ModDrop [13]: modalities are randomly masked during training, and missing modalities are ignored through the padding mask at inference; the fused representation of the available modalities is classified directly, which serves as a no-completion reference.

MMIN [16]: cascade residual autoencoders imagine the missing modality tokens from the available ones, supervised by the complete modality tokens.

ShaSpec [17]: modality representations are decomposed into shared and specific components, and missing modalities are compensated by the mean shared component of the available modalities.

MAP [31]: a learnable missing-aware prompt (MAP) is configured for each missing combination and prepended to the input sequence of the fusion Transformer.

MSP [32]: each modality has a pair of present/missing modality-specific prompts (MSP), which are combined according to the current mask under an orthogonality constraint.

Reza et al. [33]: lightweight scale-and-shift parameters are learned for each missing combination on top of a frozen fusion backbone.

For fairness, all missing-modality baselines use the same V-JEPA 2.1 video encoder, MSCN/SE-ResNet physiological encoders, and $d = 64$ modality tokens as MoLIP-Hasse; only the missing-modality handling and fusion modules after the modality tokens are replaced, and the fusion network is uniformly a two-layer Transformer of the same size as our subset encoder. Training uses the same balanced sampling of the 15 combinations, optimizer, learning rates, maximum number of epochs, and validation-based checkpoint selection, and all methods are evaluated on the same 87-fold LOSO split. Because MAP and Reza et al. require a pretrained backbone, their fusion networks are first trained with complete modalities, after which the backbone is frozen and only the prompts or scale-and-shift parameters are trained.

5.2. Comparison with Pain Recognition Methods

Table 1 compares MoLIP-Hasse with dedicated models under the same modality configurations. The accuracy of MoLIP-Hasse is 0.84, 0.40, 1.09, and 1.17 percentage points higher than that of the corresponding methods for video, video + ECG, ECG + GSR, and video + GSR, respectively, but 2.99 percentage points lower than that of MM-SCAAE for

ECG + EMG + GSR. The compared methods are trained specifically for a fixed input combination, whereas MoLIP-Hasse covers all 15 combinations with one set of parameters; this gap may reflect the trade-off between a unified model and combination-specific models (Section 7.1).

Table 1. Comparison of T0 vs. T4 pain classification performance on BioVid Part A.

Method	Modalities	Acc (%)	Macro-F1 (%)
TNT-based Transformer [45]	Video	73.28	—
MoLIP-Hasse	Video	**74.12**	73.44
Gkikas et al. [8]	Video + ECG	82.74	—
MoLIP-Hasse	Video + ECG	**83.14**	82.96
CrossMod-Transformer [10]	ECG + GSR	87.52	87.15
MoLIP-Hasse	ECG + GSR	**88.61**	**88.27**
MM-SCAAE [11]	ECG + EMG + GSR	**84.22**	**83.72**
MoLIP-Hasse	ECG + EMG + GSR	81.23	80.68
PainFormer [46]	Video + GSR	89.08	88.87
MoLIP-Hasse	Video + GSR	**90.25**	**89.47**
MoLIP-Hasse	Video + ECG + EMG + GSR	84.56	84.02

Note: All results are for T0 vs. T4 under LOSO evaluation. The results of the compared methods are taken from the original papers, whose preprocessing, segment length, and video representation differ (e.g., the video branch of PainFormer uses RGB video together with synthetic thermal and depth videos derived from it), and are therefore for reference only. Each MoLIP-Hasse row is the same model evaluated with the corresponding available modalities. Bold indicates the better result for the same modality configuration; "—" indicates a metric not reported in the original paper.

5.3. Comparison with Missing-Modality Learning Methods

Table 2 compares MoLIP-Hasse with six missing-modality learning methods under identical encoders and evaluation protocol. Every method handles all 15 non-empty combinations with a single model. The table reports the complete-modality result; the group means of the single-, double-, and triple-missing scenarios; and the mean over all 15 combinations.

Under identical encoders and evaluation protocol, the 15-combination mean accuracy of MoLIP-Hasse is 80.69%, which is 3.74 percentage points higher than that of the strongest baseline, MSP, under the same 87-fold LOSO split (two-sided paired Wilcoxon signed-rank test over the 87 test participants, $p < 10^{-6}$). Measured against MSP, the margin is 2.41, 3.48, 3.73, and 4.35 percentage points in the complete, single-, double-, and triple-missing scenarios, respectively; it is therefore smallest in the complete case, where all atoms are known and no completion is required, and increases as fewer modalities remain available. In terms of degradation, the six baselines lose between 5.20 and 8.25 percentage points between the complete case and the 15-combination mean (MSP 5.20, MMIN 7.65, ShaSpec 7.80, MAP 7.80, Reza et al. 8.05, and ModDrop 8.25), whereas MoLIP-Hasse loses 3.87 percentage points. A within-model comparison shows where this advantage comes from: the "w/o Hasse propagation" configuration of the ablation study (Section 6.1; $L_H = 0$) keeps identical encoders, training objective, and classification head but completes the unknown atoms only with the one-shot structured initializer, and its 15-combination mean (77.82%) is only 0.87 percentage points above that of MSP; most of the 3.74-percentage-point margin (2.87 points) is therefore attributable to Hasse propagation rather than to the modality encoders or the classification head. In terms of parameter organization, MAP and Reza et al. store prompts or scale-and-shift parameters separately for each of the 15 modality combinations, so their parameter count grows with the number of combinations; like ModDrop, MMIN, ShaSpec, and MSP, MoLIP-Hasse shares all parameters, and it additionally guarantees that known atoms are not modified during completion.

Table 2. Comparison with missing-modality learning methods under identical encoders and protocol.

Method	Category	Per-Combination Parameters	Complete (S0)	Single Missing	Double Missing	Triple Missing	Mean (15)
ModDrop	Training strategy	No	80.40/79.56	77.24/76.34	74.73/73.87	61.14/60.27	72.15 ± 0.74/ 71.28
MMIN	Generative completion	No	81.70/80.77	79.02/78.03	76.62/75.67	63.31/62.35	74.05 ± 0.74/ 73.09
ShaSpec	Representation compensation	No	81.40/80.38	78.60/77.65	76.17/75.05	62.79/61.74	73.60 ± 0.72/ 72.55
MAP	Prompt adaptation	Yes	80.95/80.07	78.15/77.21	75.72/74.83	62.34/61.43	73.15 ± 0.74/ 72.24
Reza et al.	Parameter-efficient adaptation	Yes	80.80/80.05	77.80/76.84	75.32/74.39	61.83/60.85	72.75 ± 0.76/ 71.81
MSP (strongest baseline)	Prompt adaptation	No	82.15/81.18	81.45/80.42	79.50/78.52	67.33/66.33	76.95 ± 0.73/ 75.95
MoLIP-Hasse	Structured interaction completion	No	84.56/84.02	84.93/84.38	83.23/82.66	71.68/70.83	80.69 ± 0.66/ 80.05

Note: Values are Acc/Macro-F1 (%). The Mean (15) column is the arithmetic mean over all 15 non-empty modality combinations, and its accuracy is reported as mean ± standard error over the 87 test participants; the strongest baseline is the one with the highest 15-combination mean accuracy (MSP). "Per-combination parameters" indicates whether prompts or adaptation parameters must be stored separately for each missing combination. MoLIP-Hasse differs significantly from every baseline in 15-combination mean accuracy (two-sided paired Wilcoxon signed-rank test over the 87 test participants; $p < 10^{-6}$ for all six baselines).

5.4. Robustness to Missing Modalities

Table 3 reports the results of the same MoLIP-Hasse model under the 15 non-empty combinations. Only the availability mask changes between scenarios, without retraining, which tests the joint robustness of interaction decomposition, Hasse propagation, and boundary clamping.

Table 3. Classification performance of MoLIP-Hasse under different combinations of available modalities.

Missing Level	Scenario	Available Modalities	Acc (%)	Macro-F1 (%)
Complete	S0	Video + ECG + EMG + GSR	84.56	84.02
Single missing	S1	ECG + EMG + GSR	81.23	80.68
	S2	Video + EMG + GSR	87.41	86.93
	S3	Video + ECG + GSR	88.72	88.15
	S4	Video + ECG + EMG	82.35	81.74
Double missing	S5	EMG + GSR	79.64	78.92
	S6	ECG + GSR	88.61	88.27
	S7	ECG + EMG	76.18	75.43
	S8	Video + GSR	90.25	89.47
	S9	Video + EMG	81.56	80.91
	S10	Video + ECG	83.14	82.96
Triple missing	S11	GSR only	72.83	71.96
	S12	EMG only	68.47	67.52
	S13	ECG only	71.29	70.41
	S14	Video only	74.12	73.44
Mean (single missing)		S1–S4	84.93	84.38
Mean (double missing)		S5–S10	83.23	82.66
Mean (triple missing)		S11–S14	71.68	70.83

The mean accuracies of the single- and double-missing scenarios are 84.93% and 83.23%, i.e., +0.37 and −1.33 percentage points relative to the complete modalities, showing that, on average, performance remains stable as long as at least two modalities are available, although individual pairs such as ECG + EMG (76.18%) are clearly weaker.

With a single remaining modality, the mean accuracy drops to 71.68%, 12.88 percentage points below that with all modalities; although the number of unknown atoms increases substantially, every single-modality scenario remains well above the 50% chance level.

Performance does not vary monotonically with the number of available modalities: video + GSR (90.25%) outperforms the complete combination (84.56%), and among two-modality pairs, video + GSR and ECG + GSR reach 90.25% and 88.61%, whereas ECG + EMG reaches only 76.18%, indicating that specific complementary relations matter more than the number of modalities.

5.5. Modality Contribution Analysis

We further compare the stand-alone discriminative ability of each modality with the performance change caused by removing it from the complete combination (Table 4).

Table 4. Stand-alone discriminative ability of each modality and the effect of its absence.

Modality	Stand-Alone Acc (%)	Acc Change When Missing (pp)	Physiological Meaning
Video	74.12 (S14)	−3.33 (S0 → S1)	Facial expression, action units
ECG	71.29 (S13)	+2.85 (S0 → S2)	Heart-rate variability, autonomic regulation
EMG	68.47 (S12)	+4.16 (S0 → S3)	Muscle tension, pain reflex
GSR	72.83 (S11)	−2.21 (S0 → S4)	Skin conductance, sympathetic activation

Note: The accuracy change when a modality is missing is the accuracy of the corresponding single-missing scenario minus that of the complete scenario S0 (pp, percentage points). Positive values indicate that accuracy increases after removing the modality, and negative values indicate a decrease; the change reflects only the performance difference under the current experimental setting and does not by itself imply that the modality is harmful. S0 → S1 denotes the change from the complete combination S0 to S1, and similarly for the other rows.

Ranked by stand-alone accuracy, the modalities are video (74.12%), GSR (72.83%), ECG (71.29%), and EMG (68.47%): video is the strongest single modality, but every physiological modality alone performs well above chance.

Removing video or GSR from the complete combination lowers accuracy by 3.33 and 2.21 percentage points, respectively, whereas removing ECG or EMG raises it by 2.85 and 4.16 percentage points. The two modalities with the lowest stand-alone accuracy are also the two whose removal improves the complete combination, which suggests that their information is not consistently converted into accuracy gains in the current joint representation when video and GSR are available. This is an observation about the current experimental setting and does not show that ECG or EMG is generally ineffective.

Accuracy rises from 74.12% (video alone) and 72.83% (GSR alone) to 90.25% for video + GSR, the strongest complementarity observed; adding ECG, EMG, or both to this pair lowers accuracy to 88.72%, 87.41%, and 84.56%, respectively, suggesting that the additional information may be accompanied by redundancy and noise.

Overall, video and GSR play a prominent role in the current task, but the classification results cannot be interpreted directly as causal contributions.

5.6. Reconstruction and Atom Analysis

Figure 3 summarizes the effect of Hasse completion on representation similarity under representative missing combinations. Here, the observed representation is the direct representation of the available set produced by the shared subset encoder (equivalently, the Zeta sum of the known atoms only), and the reconstructed representation is the Zeta synthesis of the completed atom table. Cosine similarities are computed sample by sample in the 64-dimensional representation space and then summarized: panel (a) compares the mean similarities of the observed and reconstructed representations to the complete-observation reference, while panel (b) reports the percentage of samples for which reconstruction increases this similarity. VMG, VG, and V represent single-, double-, and triple-missing scenarios, respectively; the complete combination VEMG requires no completion.

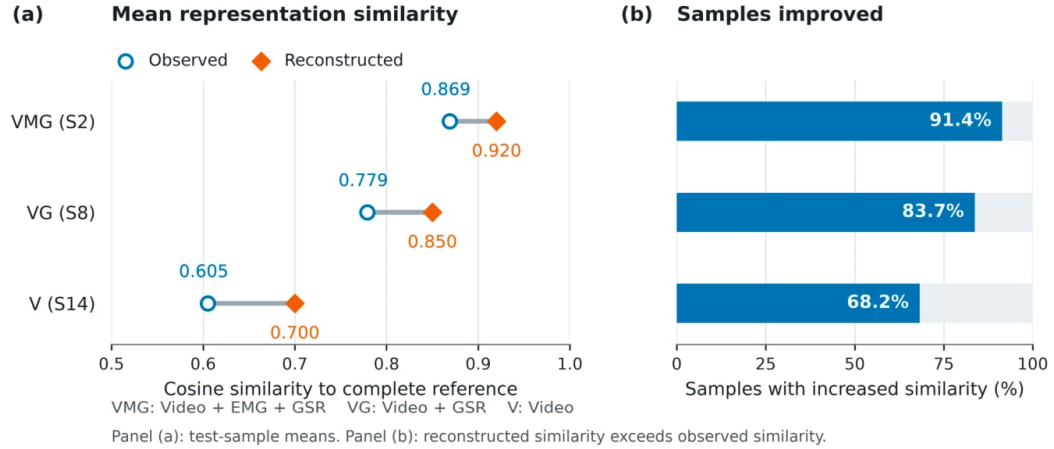

Figure 3. Representation reconstruction under representative missing-modality combinations (VMG, VG, and V). (**a**) Mean cosine similarities of the observed and reconstructed representations to the complete-observation reference. (**b**) Percentage of samples whose similarity to the reference increases after Hasse completion. Values summarize the test samples across the 87 LOSO folds.

Under VMG, VG, and V, the reconstructed representation is closer to the complete reference than the observed representation for 91.4%, 83.7%, and 68.2% of the samples, respectively (Figure 3b), and the mean cosine similarity increases from 0.869, 0.779, and 0.605 to 0.920, 0.850, and 0.700 (Figure 3a; the reconstructed means also appear as the green triangles in Figure 4). The corresponding medians are 0.928, 0.857, and 0.710 (Section 6.2); the mean is slightly below the median for all three combinations, which indicates a small fraction of samples with a markedly lower reconstruction similarity rather than a few outliers above it.

In Figure 4, the median reconstruction cosine similarities for VEMG, VMG, VG, and V are 1.000, 0.928, 0.857, and 0.710, respectively. As fewer modalities are available, the number of unknown non-empty atoms increases from 0 to 8, 12, and 14, and reconstruction becomes harder and more variable across samples.

The ranking of reconstruction similarity does not agree with that of classification accuracy: VG has lower reconstruction similarity than VEMG and VMG but the highest accuracy, showing that being close to the complete representation is not equivalent to preserving the most discriminative directions.

Figure 5 shows the normalized energy of each atom, defined as the ratio of the atom's L2 norm to the sum of the norms of the 15 non-empty completed atoms of the same sample, averaged over samples; this computation does not alter the classification input.

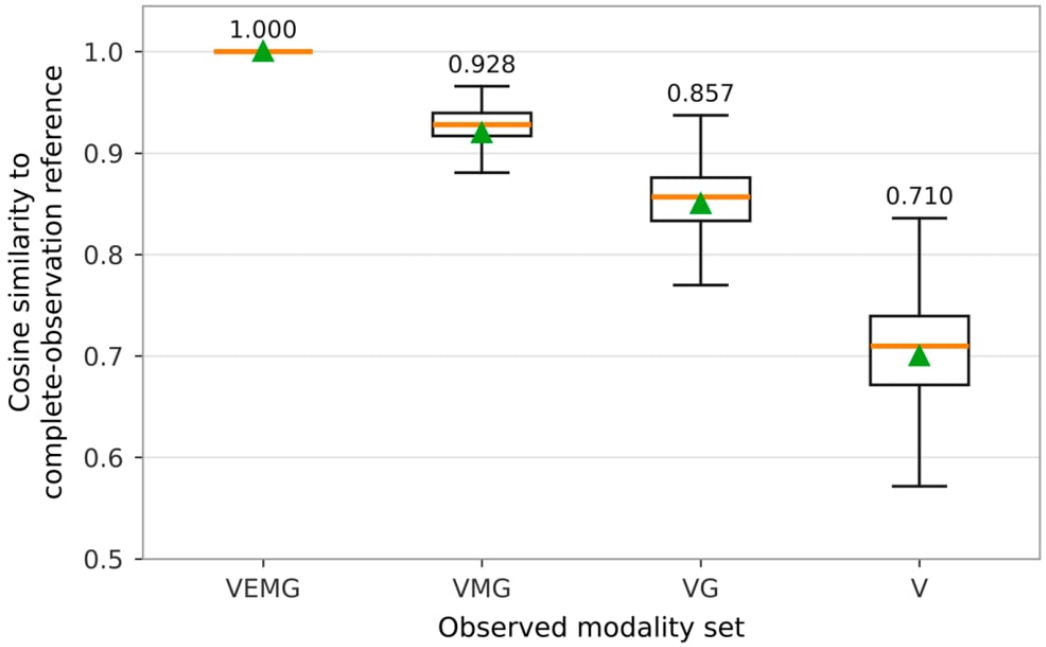

Figure 4. Cosine similarity between the reconstructed representations and the complete-observation reference under representative modality combinations. Each box summarizes the per-sample reconstruction similarity of the 87 test participants (87 folds × 40 samples); the box spans the interquartile range (IQR), the orange line is the median, the whiskers extend to 1.5 × IQR (outliers are not shown), and the green triangles mark the mean. The numbers above the boxes are the medians, the last three of which are also reported in Section 6.2.

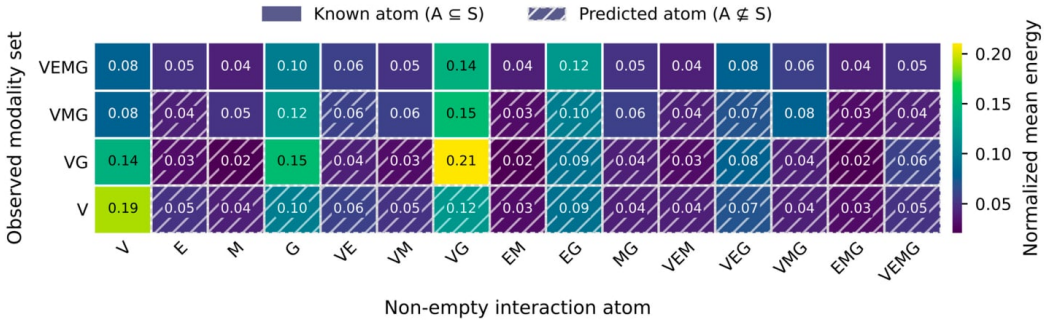

Figure 5. Normalized energy distribution of interaction atoms under representative modality combinations. Solid cells are known atoms ($A \subseteq S$); hatched cells with dashed borders are predicted atoms ($A \not\subseteq S$).

The largest energies occur in the video and GSR atoms and in the video–GSR and ECG–GSR second-order interactions. In the VG scenario, $\mathbf{\Omega}(\mathrm{VG})$ is highest, followed by $\mathbf{\Omega}(\mathrm{G})$ and $\mathbf{\Omega}(\mathrm{V})$; with video only, $\mathbf{\Omega}(\mathrm{V})$ dominates, and the predicted GSR- and video–GSR-related atoms also carry appreciable energy.

The energy is thus concentrated in a small number of atoms; however, energy reflects only vector magnitude and should not be equated with classification contribution or causal importance.

5.7. Multi-Level Pain Recognition

To examine applicability to finer-grained pain intensity, we keep the architecture, training strategy, and LOSO split unchanged and add binary tasks T0 vs. T1, T0 vs. T2, and T0 vs. T3, as well as five-class classification of T0–T4, for which only the output dimension of the classification head is changed to 5. In addition, T0–T4 are treated as ordinal labels

0–4: the classification head is replaced by a single-output linear regression head, the cross-entropy in the subset auxiliary loss and the final classification loss is replaced by the mean squared error, and the mean absolute error (MAE) is reported. The results are given in Table 5.

Table 5. Results of multi-level pain recognition.

Task	Classes	Complete (S0)	Video + GSR (S8)	Video Only (S14)	Mean (15)
T0 vs. T1	2	71.86/71.02	74.83/74.04	67.20/66.41	71.45/70.68
T0 vs. T2	2	77.42/76.52	79.87/79.02	70.65/69.80	76.45/75.62
T0 vs. T3	2	81.34/80.40	83.26/82.37	73.78/72.89	79.95/79.08
T0 vs. T4 (main task)	2	84.56/84.02	90.25/89.47	74.12/73.44	80.69/80.05
T0–T4 five-class	5	46.28/44.46	51.24/49.47	40.72/38.86	42.10/40.20
T0–T4 ordinal regression (MAE ↓)	—	0.912	0.977	1.149	1.011

Note: Classification results are Acc/Macro-F1 (%); regression results are MAE in pain-level units (lower is better). S0, S8, and S14 are the complete, video + GSR, and video-only combinations of Table 3, and Mean (15) is the arithmetic mean over all 15 non-empty combinations for the same task. All tasks use the same outer 87-fold LOSO split. ↓: lower is better.

As the gap between the two pain levels narrows, classification performance decreases monotonically for all three reported combinations and for the 15-combination mean; the 15-combination mean accuracy for T0 vs. T1 is 71.45%, consistent with the known difficulty of separating low-intensity pain from the baseline [27]. Among the reported combinations, video + GSR (S8) still performs best in the five-class task, whereas in ordinal regression the complete combination (S0) yields the lowest MAE (0.912, against 1.149 with video only).

5.8. Non-Random Missingness

The 15-combination evaluation in Table 3 assumes that all combinations occur with equal probability and that missingness is independent of the sample content, i.e., missing completely at random (MCAR). In clinical practice, missingness is often related to signal quality or even to the pain state itself, i.e., missing not at random (MNAR) [47]. We therefore designed three test-time missingness protocols without retraining the model:

(1) Label-dependent missingness (P1): To simulate electrode detachment caused by pain-induced body movement, ECG, EMG, and GSR are independently missing with probability 0.7 for T4 samples and 0.4 for T0 samples, and video is missing with probability 0.25; samples with all four modalities missing are resampled. With this setting, the mean number of available modalities (2.20) is close to that of the MCAR protocol (2.13), so the protocol mainly probes the missingness mechanism rather than the amount of missing data, although the composition of the available modalities differs (e.g., video is available more often than under MCAR).
(2) Within-segment interruption (P2): On top of MCAR combination sampling, partial signal loss caused by poor contact is simulated by zeroing a random contiguous 1–3 s interval of each available physiological channel and replacing 3–6 consecutive frames (about 1–2 s) of the 16 sampled video frames with the previous valid frame. These modalities are still flagged as available, which tests the sensitivity of the model to local degradation that is not reflected in the mask.
(3) Quality-driven missingness (P3): On top of the MCAR combination sampling, additional missingness is determined from measured signal quality. Video is treated as missing when MTCNN face detection fails or its confidence is below 0.90; ECG is treated as missing when R-peak detection fails or the heart rate falls outside 40–180

beats per minute (bpm); and EMG or GSR is treated as missing when amplitude saturation or a flat segment longer than 1 s occurs.

Table 6 reports the results of MoLIP-Hasse and of the strongest missing-modality baseline in Table 2 under each protocol.

Table 6. Recognition results under non-random missingness protocols.

Protocol	Missingness Mechanism	Mean No. of Available Modalities	MoLIP-Hasse	Strongest Baseline (MSP)
MCAR (balanced 15 combinations)	MCAR	2.13	80.69/80.05	76.95/75.95
P1 label-dependent	MNAR	2.20	79.43/78.62	74.05/73.14
P2 within-segment interruption	MCAR + local degradation	2.13	78.94/78.15	74.86/73.89
P3 quality-driven	MCAR + measured signal quality	2.07	80.34/79.70	76.45/75.46

Note: P1–P3 are applied only at test time; the model and training settings are unchanged. Values are Acc/Macro-F1 (%).

Under label-dependent missingness, the accuracy of MoLIP-Hasse decreases by 1.26 percentage points relative to MCAR (80.69% to 79.43%) and remains 5.38 percentage points higher than that of the strongest baseline. Within-segment interruption reduces accuracy by 1.75 percentage points (80.69% to 78.94%), because the degraded modalities are still flagged as available and therefore enter the known boundary unchanged without being corrected by completion. Under the quality-driven protocol, the quality rules additionally mask 0.6%, 3.1%, 4.4%, and 5.2% of the still-available video, ECG, EMG, and GSR segments. Because a modality has to be retained by the MCAR sampling before these rules can apply, the rule-hit probability of each modality over the whole test set is $(8/15)$ times its measured rate. Summing over the 15 equally likely combinations, the rules fire for 6.9% of the test samples. In the 0.9% of the samples in which the flagged modality is the last remaining one, the rules are not applied, so the sample keeps its MCAR combination and the evaluated test set stays identical to that of the MCAR protocol. The expected number of modalities removed is therefore 0.06 per sample (0.07 before this safeguard), the mean number of available modalities falls from 2.13 to 2.07, and 6.0% of the test samples actually lose at least one modality. Accuracy decreases from 80.69% to 80.34%. Because quality failures are rare in BioVid Part A, quality-driven missingness has a much smaller effect than the label-dependent and within-segment mechanisms. These protocols nevertheless remain simulations of clinical missingness mechanisms (Section 7.3).

6. Ablation Studies

6.1. Core Components

Each ablation in Table 7 modifies only the corresponding component of the three-layer full model, which achieves a mean accuracy of 80.69% and a mean Macro-F1 of 80.05%. Removing Hasse propagation (i.e., $L_H = 0$; Section 6.3) lowers the two metrics by 2.87 and 3.01 percentage points and the S14 accuracy by 4.24 percentage points, showing that one-shot initialization is insufficient under severe missingness.

With only upward or only downward propagation, the mean accuracy is 79.34% and 78.96%, respectively, and removing boundary clamping lowers it by 2.08 percentage points. Merging the modality-typed edges, removing all gates, and replacing the Transformer subset encoder with mean pooling lower the mean accuracy by 1.42, 1.05, and 1.67 percentage points, respectively, suggesting that the components are complementary.

Table 7. Ablation study of the core components of MoLIP-Hasse (Acc/Macro-F1, %).

Method	S0 VEMG	S2 VMG	S8 VG	S14 V	Mean (15)
Full MoLIP-Hasse	84.56/84.02	87.41/86.93	90.25/89.47	74.12/73.44	80.69/80.05
w/o Hasse propagation	83.62/83.00	84.96/84.31	86.72/85.94	69.88/68.96	77.82/77.04
Upward propagation only	84.01/83.45	86.12/85.55	88.91/88.14	72.31/71.48	79.34/78.68
Downward propagation only	83.84/83.21	85.73/85.02	88.47/87.71	71.62/70.80	78.96/78.23
w/o boundary clamping	83.31/82.68	85.54/84.82	88.02/87.24	70.95/70.02	78.61/77.85
w/o modality-typed edges	84.05/83.48	86.03/85.36	88.62/87.80	72.44/71.60	79.27/78.59
w/o all gates	84.18/83.62	86.57/85.96	89.14/88.39	72.96/72.08	79.64/78.98
Transformer → mean pooling	83.76/83.11	85.88/85.19	88.34/87.55	71.90/71.05	79.02/78.32

Note: 'Transformer → mean pooling' denotes replacing the Transformer subset encoder with mean pooling.

6.2. Interaction-Atom Reconstruction Quality

In Table 8, the full model achieves similarities of 0.928, 0.857, and 0.710 for S2, S8, and S14 and the lowest unknown-atom error (0.243). Without Hasse propagation, the error increases to 0.401 and the S14 similarity drops to 0.612, indicating that severe missingness depends more on structural propagation.

Upward-only propagation outperforms downward-only propagation, suggesting that the main recovery path runs from known low-order subsets to unknown high-order subsets; bidirectional propagation remains best, indicating that downward feedback has a corrective effect. Without boundary clamping, the known-atom error rises from below 10^{-6} to 0.081, and removing the modality-typed edges or the gates also increases the unknown-atom error.

Table 8. Effects of structural components on interaction-atom reconstruction quality.

Method	S2 VMG Similarity ↑	S8 VG Similarity ↑	S14 V Similarity ↑	Known-Atom Error ↓	Unknown-Atom Error ↓
Full MoLIP-Hasse	0.928	0.857	0.710	$<10^{-6}$	0.243
w/o Hasse propagation	0.882	0.788	0.612	$<10^{-6}$	0.401
Upward propagation only	0.912	0.834	0.679	$<10^{-6}$	0.294
Downward propagation only	0.901	0.815	0.651	$<10^{-6}$	0.344
w/o boundary clamping	0.889	0.801	0.634	0.081	0.367
w/o modality-typed edges	0.907	0.827	0.669	$<10^{-6}$	0.315
w/o all gates	0.916	0.840	0.687	$<10^{-6}$	0.281

Note: Similarity is the median per-sample cosine similarity; atom error is the normalized L2 error averaged over valid atoms and samples. "Transformer → mean pooling" is evaluated only at the classification level (Table 7) and is therefore not listed here. ↑: higher is better; ↓: lower is better.

6.3. Sensitivity to Propagation Depth

To examine the sensitivity of the model to the Hasse propagation depth, a post hoc comparison over $L_H \in \{0,1,2,3,4\}$ was performed with the modality encoders, subset encoder, loss function, optimization settings, and data split unchanged. Table 9 is used only for analysis, not for model selection.

When $L_H = 0$, the model predicts the unknown atoms only with the structured initializer and performs no propagation over cover edges. As the number of propagation layers increases from 0 to 3, both the mean classification performance over the 15 combinations and the three-scenario reconstruction similarity improve steadily, indicating that multi-layer propagation over cover edges helps integrate observable information across interaction nodes of different orders. Increasing the depth to 4 layers brings no further gain in either classification or reconstruction, suggesting that propagation beyond the maximum structural

distance yields no stable benefit, possibly because of repeated message aggregation or accumulated prediction errors; the exact cause requires further study.

Table 9. Sensitivity analysis of the Hasse propagation depth.

LH	Mean Acc Over 15 Combinations (%)	Mean Macro-F1 Over 15 Combinations (%)	Mean Reconstruction Similarity (S2, S8, S14)
0	77.82	77.04	0.761
1	79.29	78.51	0.801
2	80.21	79.53	0.822
3	80.69	80.05	0.832
4	80.40	79.75	0.829

Note: The 15-combination mean accuracy and Macro-F1 are computed from fold-level results of the 87 outer test participants under all 15 non-empty combinations; the reconstruction similarity is the arithmetic mean of the median per-sample cosine similarities for S2, S8, and S14. The $L_H = 0$ configuration is the "w/o Hasse propagation" variant of Tables 7 and 8, so the values for $L_H = 0$ and $L_H = 3$ are the means of the corresponding medians in Table 8 (0.882/0.788/0.612 and 0.928/0.857/0.710, respectively).

7. Discussion

7.1. Trade-Offs of a Unified Model

As shown in Section 5.4, performance does not increase monotonically with the number of available modalities: the single-missing mean (84.93%) is slightly higher than the complete-modality accuracy (84.56%), and the choice of complementary modalities matters more than their number.

Notably, the accuracy with all four modalities (84.56%) is lower than that with video + GSR (90.25%), ECG + GSR (88.61%), and two of the four three-modality combinations (S2 and S3); moreover, with ECG + EMG + GSR, MoLIP-Hasse (81.23%) trails the dedicated MM-SCAAE (84.22%) by 2.99 percentage points (Table 1). Two factors may contribute. First, EMG and ECG have the two lowest stand-alone accuracies (68.47% and 71.29%). Adding either of them to the strongest pair degrades performance (87.41% and 88.72%, respectively, versus 90.25% for video + GSR; Section 5.5), and removing either of them from the complete combination increases accuracy (Table 4); their atoms may therefore introduce noise or redundancy into the shared representation. Second, a single model must accommodate 15 combinations, so its parameters represent a compromise among them. Future work could introduce gating that adapts modality weights to signal quality or, at deployment, select the best subset for frequently available combinations on validation data, to mitigate negative transfer from weak modalities.

To examine where this decrease arises, we compared, sample by sample, the test predictions of the same trained models under video + GSR and after adding ECG, EMG, or both (3480 test samples from the 87 LOSO folds). Adding ECG changed the predicted class of 3.6% of the samples (89 became incorrect and 36 became correct; exact McNemar test, $p < 0.001$), adding EMG changed 6.6% (165 vs. 66; $p < 0.001$), and adding both changed 12.1% (310 vs. 112; $p < 0.001$). With both modalities added, the per-participant accuracy decreased for 70 of the 87 participants, increased for 7, and was unchanged for 10 (41 and 57 participants showed a decrease when only ECG or only EMG was added); the decrease is therefore spread across participants rather than caused by a few of them. Of the samples that became incorrect, 68% were pain-free (T0) samples classified as pain (T4) and 32% were T4 samples classified as T0 (70% and 72% when only ECG or only EMG was added). The added ECG and EMG information thus mainly shifts pain-free samples toward the pain class; in a monitoring setting, these errors would appear as false pain alarms.

The encoders also differ in capacity and prior knowledge: the video token is produced by a frozen V-JEPA 2.1 backbone pre-trained on large video corpora, whereas the ECG, EMG, and GSR tokens come from compact MSCN/SE-ResNet encoders trained from scratch on about 3200 training samples per fold. This imbalance may partly explain why atoms involving video carry a large share of the normalized energy in Figure 5 (about 0.56 under complete observation) and why adding ECG or EMG does not improve on video + GSR. It is unlikely to be the only factor, however. GSR is encoded by the same type of from-scratch encoder as ECG and EMG, yet atoms involving GSR carry a comparable share (about 0.64), GSR alone is nearly as accurate as video alone (72.83% vs. 74.12%; Table 4), and video + GSR is the best combination, whereas the atoms involving only ECG and/or EMG (E, M, and EM) account for about 0.13. Although energy does not measure classification contribution (Section 5.6), this pattern suggests that the task relevance of each signal, and not only the capacity of its encoder, determines how much a modality contributes. Separating the two factors would require capacity-matched or pre-trained physiological encoders, which we leave for future work.

A related question is how the reconstruction terms interact with classification. $\mathcal{L}_{\text{full}}$, $\mathcal{L}_{\text{atom}}$, and $\mathcal{L}_{\text{path}}$ pull the completed representation toward the complete-observation reference, whereas $\mathcal{L}_{\text{final}}$ only requires $\hat{\boldsymbol{h}}(\mathcal{M}|S)$ to be discriminative. Because the complete reference itself contains the weaker ECG and EMG components, strong reconstruction could pull a well-performing subset such as video + GSR toward a less discriminative region; consistent with this, video + GSR has a lower reconstruction similarity than video + EMG + GSR but a higher accuracy (Section 5.6). In the current model, this tension is limited in three ways. First, the known boundary is re-clamped after every layer, so the observed atoms of video + GSR, $\Omega(V)$, $\Omega(G)$, and $\Omega(VG)$, always enter $\hat{\boldsymbol{h}}(\mathcal{M}|S)$ with their exact values and cannot be overwritten by the propagation; the reconstruction losses can change them only indirectly, through the shared encoders. Second, $\mathcal{L}_{\text{subset}}$ requires every direct subset representation, including $\boldsymbol{h}(VG)$, to be discriminative on its own. Third, the classification loss has the largest weight ($\lambda_{\text{final}} = 1$, versus 0.50–0.80 for the three reconstruction terms). The balance is nevertheless set by fixed weights shared by all available sets, and the lower accuracy of the complete combination suggests that it is not optimal for every combination. Task-guided adaptive weighting is a natural extension: the contribution of each predicted atom to the Zeta synthesis, or the weight of each reconstruction term, could be modulated by its estimated benefit for classification, for example through atom-wise gates trained with $\mathcal{L}_{\text{final}}$ or through uncertainty-based loss weighting, so that predicted atoms that do not improve discrimination are down-weighted for subsets such as video + GSR.

7.2. Computational Complexity and Scalability

Let n be the number of modalities and d the token dimension. The complete Boolean lattice has 2^n nodes and $n{\cdot}2^{n-1}$ upward cover edges (with the same number of downward edges); for non-empty inputs, the maximum structural distance between an unknown atom and its known intersection boundary is $n-1$ (Equation (17)), so the propagation depth is set to $L_{\text{H}} = n-1$. Accordingly, the complete-supervision branch must encode 2^n subsets during training at a cost of $\text{O}(2^n{\cdot}n{\cdot}d^2)$; the fast Möbius/Zeta transforms cost $\text{O}(n{\cdot}2^n{\cdot}d)$; and each Hasse layer applies node-wise linear transforms and gates for every modality and direction at a cost of $\text{O}(n{\cdot}2^n{\cdot}d^2)$, i.e., $\text{O}(n^2{\cdot}2^n{\cdot}d^2)$ for L_{H} layers. If cover routing is implemented with dense $2^n \times 2^n$ matrices, as in the current implementation, the routing term costs $\text{O}(n{\cdot}4^n{\cdot}d)$ per layer, 2^{n+1} times that of a sparse edge-list implementation, $\text{O}(n{\cdot}2^n{\cdot}d)$. At inference, subset encoding reduces to $\text{O}(2^{|S|}{\cdot}n{\cdot}d^2)$, but propagation still runs over all 2^n nodes. Hence, even with a sparse implementation, the cost of the complete-lattice approach grows at least in proportion to $2^n{\cdot}d$, and exponential growth in the number of nodes is unavoidable.

Table 10 lists the structural size and estimated back-end computation for up to 12 modalities. For $n = 4$, the back end requires about 1.9×10^7 multiply–accumulate operations (MACs) per sample, more than four orders of magnitude fewer than the roughly 7.8×10^{11} MACs that the frozen video backbone needs for one clip; the current computational bottleneck is therefore video encoding rather than Hasse completion. For $n = 8$, the back end requires about 1.3×10^9 MACs and the dense upward cover matrices occupy only about 2 MB, which a single GPU handles easily; however, the number of non-empty combinations grows to 255, and with 3200 training samples per fold and balanced combination sampling, each combination is sampled only about 12.5 times per epoch, far fewer than the 213 times for $n = 4$. Thus, when the number of modalities reaches 7–8, the first bottleneck to appear is statistical, namely insufficient coverage of combinations. For $n \geq 10$, the computation and memory cost of dense routing rise rapidly: for $n = 12$, the back-end computation (3.1×10^{11} MACs) approaches that of the video backbone, and the dense cover matrices alone occupy about 0.75 GB in float32, increasing to about 14 GB for $n = 14$.

Table 10. Structural size and estimated back-end cost versus the number of modalities.

n	Subset Nodes	Non-Empty Combinations	Upward Cover Edges	Propagation Depth	Samples per Combination per Epoch	Back-End MACs/ Sample (Dense)	Back-End MACs/ Sample (Sparse)	Nodes with $\kappa = 2$
4	16	15	32	3	213.3	1.9×10^7	1.8×10^7	11
6	64	63	192	5	50.8	1.5×10^8	1.3×10^8	22
8	256	255	1024	7	12.5	1.3×10^9	7.9×10^8	37
10	1024	1023	5120	9	3.1	1.6×10^{10}	4.5×10^9	56
12	4096	4095	24,576	11	0.8	3.1×10^{11}	2.4×10^{10}	79

Note: Values are combinatorial counts and theoretical estimates, not measurements. Samples per combination per epoch assume 3200 training samples per fold with balanced sampling over the non-empty combinations. Back-end MACs assume $d = 64$, a feed-forward dimension of 128, node-update MLPs of $5d \to 192 \to d$, and encoding of all subsets during training, excluding the modality encoders; for reference, the frozen V-JEPA 2.1 ViT-B/16 requires about 7.8×10^{11} MACs per clip (4608 tokens). $\kappa = 2$ truncation keeps only atoms of order at most 2.

To complement these estimates with measurements, we timed inference on the same NVIDIA RTX 3090 GPU (PyTorch, float32, batch size 1, excluding data loading and face preprocessing; 20 warm-up and 100 timed runs, median latency). With all four modalities, end-to-end inference took 214.6 ms per sample, of which video encoding took 186.4 ms (86.9%), the three physiological encoders 3.6 ms, and the MoLIP-Hasse back end 1.9 ms; the remaining 22.7 ms was spent in the parts of the forward pass outside these three modules, namely token assembly, the classification head, and scheduling overhead. The peak GPU memory allocated by PyTorch was 2614 MB (excluding the CUDA context). With video + GSR, the latency (212.1 ms per sample) and peak memory (2559 MB) were almost the same, because video encoding dominates in both cases, in agreement with the MAC estimates in Table 10. Because each prediction covers a 5.5 s window, model inference takes about 3.9% of the window length, so the model itself is fast enough for continuous monitoring on this GPU (Section 7.3); face detection and alignment, which were not timed here, add to this in a deployed system.

The complexity for more modalities can be reduced in three ways. (1) Replacing dense matrix multiplication with sparse edge-list message passing reduces the routing cost to $O(n \cdot 2^n \cdot d)$, lowering the back-end computation for $n = 12$ to about 2.4×10^{10} MACs. (2) Limiting the maximum interaction order to κ and keeping only atoms with $|A| \leq \kappa$ reduces the number of nodes to $\sum_{k=0}^{\kappa} C(n,k) = O(n^\kappa)$ and the propagation depth to κ, which corresponds to the κ-order additive approximation in fuzzy-measure theory [48], but the Zeta reconstruction then becomes approximate. In Figure 5, first- and second-order

atoms account for about 0.73 of the normalized energy under complete observation and third- and fourth-order atoms for about 0.27, indicating that higher-order interactions are not negligible in this task and that the truncation order should be determined experimentally. (3) Grouping modalities by physiological system (e.g., treating multichannel EMG as one group) to build a group-level lattice, or sampling subsets for the complete-supervision branch and the superset-path loss during training, avoids encoding all 2^n subsets for every sample. Systematic validation of these schemes on pain datasets with more modalities, such as X-ITE [49], is left for future work.

For the present four-modality lattice, an order-2 truncation ($\kappa = 2$) would keep 11 of the 16 nodes—the empty set and the four first- and six second-order atoms—reduce the propagation depth from three to two, and discard the four third-order atoms and the fourth-order atom. Its effect would depend on the available set. For one- and two-modality inputs, including the best-performing video + GSR combination, every observable atom has order at most two, so the truncation would remove only predicted higher-order atoms and would leave the observed evidence unchanged. For three- and four-modality inputs, it would also discard observed atoms such as $\Omega(VEG)$ and $\Omega(VEMG)$, and the Zeta synthesis would no longer reproduce $\boldsymbol{h}(\mathcal{M})$ exactly even under complete observation. The discarded atoms carry about 0.27 of the normalized energy under complete observation and 0.23–0.26 in the three missing scenarios of Figure 5. We therefore expect the truncation to affect one- and two-modality inputs mainly through the removed predicted atoms, and three- and four-modality inputs also through the removed observed atoms. Because energy does not measure classification contribution, however, how much of the 80.69% mean accuracy would be retained cannot be inferred from Figure 5 alone; it requires retraining the truncated model under the same LOSO protocol, which we leave for future work.

7.3. Limitations

This study has several limitations. (1) The experiments use only laboratory-induced heat-pain data from healthy participants in BioVid Part A, and the main task is T0 vs. T4 classification; generalization across datasets and to clinical populations remains to be verified. (2) The non-random missingness protocols in Section 5.8 are still simulations. In clinical settings, the temporal structure and duration of sensor failures and their relation to patient state are more complex, and prospective validation on data with recorded real-world missingness, for example in postoperative or intensive care monitoring, is needed. Moreover, if missingness in the training data is itself label-dependent, the model may exploit missingness patterns as shortcut features, which requires dedicated debiasing strategies. (3) Interaction atoms are set increments in representation space and should not be regarded as physiological causal contributions. (4) The video branch relies on a frozen ViT-B/16 backbone, which dominates the inference cost (Section 7.2); bedside deployment will require a lighter video encoder. (5) The loss weights in Equation (35) were set manually and kept fixed, and their sensitivity was not analyzed; a systematic search on the validation participants, together with a sensitivity analysis, is needed to determine how strongly the results depend on them.

Moreover, availability is encoded by a binary mask, so a modality is either used in full or not at all. On BioVid Part A, the rules of the quality-driven protocol (P3) fire for only 6.9% of the test samples and lower the accuracy only from 80.69% to 80.34%, whereas within-segment interruption (P2), which leaves degraded channels flagged as available, lowers it to 78.94% (Section 5.8): the degraded atoms then enter the known boundary unchanged. In clinical deployment, where partial degradation is likely to be more frequent, the binary mask could be replaced by continuous quality scores $\rho_m \in [0,1]$ estimated from signal-quality indices, such as electrode contact, signal-to-noise ratio, or face-detection

confidence. Such scores could weight the modality tokens in the context vectors $\boldsymbol{c}_A^S$ and $\boldsymbol{c}_S$, modulate the source and target gates of the Hasse propagation, and soften the boundary clamping in Equation (27) by replacing $K_S(A)$ with a quality-weighted value, so that low-quality observed atoms are partly re-estimated from the lattice instead of being kept fixed. Because Equation (27) would then no longer preserve the observed evidence exactly, this extension trades part of the exactness guarantee for robustness to degraded signals and requires validation on data with recorded signal quality.

8. Conclusions

This paper formulated dynamic modality missingness as boundary-constrained completion of interaction atoms and proposed the MoLIP-Hasse framework. Vector-valued Möbius inversion assigns a fixed subset identity to every compensated unit, explicit Hasse cover-relation matrices and layer-wise boundary clamping complete the unknown interactions without modifying the observable evidence, and the Zeta transform synthesizes the complete representation. On BioVid Part A, a single model handles all 15 non-empty modality combinations with a 15-combination mean accuracy of 80.69% and reaches 90.25% with video and GSR; under identical encoders and protocol, its mean accuracy exceeds that of the strongest missing-modality learning baseline by 3.74 percentage points. The complexity analysis shows that the complete lattice can be used directly when the number of modalities is small (e.g., $n \leq 6$), whereas more modalities require sparse message passing and order truncation. Future work will validate the framework on datasets with more modalities and on clinical data with real-world missingness and will investigate quality-aware modality weighting and efficient sub-lattice approximations.

Author Contributions: Conceptualization, B.J. and C.C.; methodology, B.J.; software, B.J.; validation, B.J.; formal analysis, B.J.; investigation, B.J.; writing—original draft preparation, B.J.; writing—review and editing, C.C.; visualization, B.J.; supervision, C.C.; funding acquisition, C.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Science Foundation of Shandong Province for Excellent Young Scholars, grant number ZR2024YQ071, and the Youth Innovation and Technology Support Program in Colleges and Universities of Shandong Province, grant number 2021KJ062.

Institutional Review Board Statement: This study involved secondary analysis of existing data from the BioVid Heat Pain Database [5]. No new data were collected, and no interaction or intervention with participants took place. Ethical review was not required for the present study under Article 32, item (1), of the Measures for Ethical Review of Life Sciences and Medical Research Involving Humans (China, 2023), because the study used only lawfully obtained public data: the database is released by its providers for academic research under a usage agreement. The study caused no harm to the participants and involved no commercial interests; the participants are identified only by codes, and the authors had no access to their names or contact details and made no attempt to identify them. The exemption was confirmed in writing by the Qingdao Institute of Software, College of Computer Science and Technology, China University of Petroleum (East China), on 23 November 2026. The original data collection was approved by the Ethics Committee of Ulm University (approval no. 196/10-UBB/bal, 16 September 2010).

Informed Consent Statement: Informed consent was obtained from all participants by the original investigators of the BioVid Heat Pain Database [5]. The facial images in Figure 1 are taken from this database and show a participant who consented to the use of their images in scientific publications, in accordance with the terms of use of the database.

Data Availability Statement: The BioVid Heat Pain Database is available from its providers for non-commercial research upon request (https://www.nit.ovgu.de/nit/en/BioVid-p-2090.html, accessed on 18 November 2024). The code is available from the corresponding author upon reasonable request.

Acknowledgments: During the preparation of this manuscript, the authors used GPT-6 Astra (OpenAI) and Claude Opus 5.5 (Anthropic) for Chinese-to-English translation. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Table A1 summarizes the main notation used in Section 3. Symbols that are used only locally, such as the sizes of the encoder outputs and the embeddings inside the encoders and the initializer, are defined where they appear.

Table A1. Main notation of MoLIP-Hasse.

Symbol	Description	Defined in
Sets, indices, and sizes		
$\mathcal{M} = \{V,E,M,G\};\ m \in \mathcal{M}$	Modality universe (video, ECG, EMG, and GSR); m denotes a single modality	Equation (1)
$n = \lvert\mathcal{M}\rvert$	Number of modalities ($n = 4$)	Section 3.1
$\boldsymbol{a}^{(b)}, a_m^{(b)}, \boldsymbol{s}$	Availability mask of the b-th sample, its entry for modality m, and the mask with the sample index omitted	Equation (1); Section 3.7
$S, \overline{S} = \mathcal{M} \setminus S$	Current available set and current missing set	Equation (1); Section 3.1
$A,\ B,\ R$	General modality subsets; A also indexes interaction atoms and lattice nodes	Section 3.1
$T \supseteq S$	Target subset in Zeta synthesis	Equation (29)
$B_s,\ N_{\text{pair}}$	Batch size and number of sample–subset pairs after parallel expansion	Sections 3.1 and 3.5
d	Common token dimension ($d = 64$)	Section 3.1
$l,\ L_H$	Index of the Hasse propagation layer and number of layers ($L_H = 3$)	Section 3.8; Equation (17)
Encoders, subset representations, and interaction atoms		
$\boldsymbol{f}_m$	Token of modality m produced by its encoder	Equation (5); Section 3.4
$\mathcal{F}_S$	Set of the tokens of the available modalities	Section 3.7
F_θ	Shared Transformer subset encoder	Equation (7)
$\boldsymbol{h}(B)$	Direct (cumulative) representation of subset B	Equation (7)
g_{cls}	Shared linear classification head	Section 3.5; Equation (31)
$\Omega(A)$	Möbius interaction atom of subset A	Equation (10)
$\mathcal{B}(S)$	Observable (known) boundary, i.e., the atoms with $A \subseteq S$	Section 3.7
$K_S(A),\ U_S(A)$	Known-node and unknown-node indicators	Equation (13)

Table A1. *Cont.*

Symbol	Description	Defined in
Initialization and Hasse propagation		
$\boldsymbol{c}_A^S, \boldsymbol{c}_S$	Local context of node A (mean of the available tokens in $A \cap S$) and global context of the available set	Equation (14)
I_θ	Learned initializer of the unknown atoms	Equations (15) and (16)
$\boldsymbol{H}^{(l)}, \boldsymbol{H}_A^{(l)}$	Node states after layer l (all nodes; node A)	Equations (15) and (27)
$\boldsymbol{B}_m^{\uparrow}, \boldsymbol{B}_m^{\downarrow}$	Upward and downward cover-relation matrices of modality m	Equation (18)
Φ_m^{dir}	Modality- and direction-specific linear transform, dir $\in \{\uparrow,\downarrow\}$	Section 3.8
$\boldsymbol{G}_{m,s}^{\mathrm{dir},(l)}, \boldsymbol{G}_{m,d}^{\mathrm{dir},(l)}$	Feature-wise source and target gates	Equations (19) and (20)
$\boldsymbol{M}_{\uparrow}^{(l)}, \boldsymbol{M}_{\downarrow}^{(l)}$	Aggregated upward and downward messages	Equations (21) and (22)
$\boldsymbol{U}^{(l)}$, $\boldsymbol{C}^{(l)}$, $\boldsymbol{Z}^{(l)}$	Concatenated layer input, candidate content, and update gate	Equations (23)–(25)
F_c, F_g	Candidate network and update-gate network	Equations (24) and (25)
$\hat{\boldsymbol{H}}^{(l+1)}$	Node states before boundary re-clamping	Equation (26)
Completion, synthesis, and training		
$\hat{\boldsymbol{\Omega}}(A\|S)$	Predicted unknown atom, equal to the final node state $\boldsymbol{H}_A^{(L_H)}$	Section 3.8
$\tilde{\boldsymbol{\Omega}}(A\|S)$	Completed atom: the known atom $\Omega(A)$ for $A \subseteq S$ and the predicted atom otherwise	Equation (28)
$\hat{\boldsymbol{h}}(T\|S), \hat{\boldsymbol{h}}(M\|S)$	Zeta-synthesized representation of the target subset T; estimated complete representation used for classification	Equations (29) and (31)
$\hat{\boldsymbol{p}}, y$	Predicted class probabilities (T0 vs. T4) and true label	Equations (9) and (31)
$\Omega^*(A)$, $\boldsymbol{h}^*(B)$	Reference atoms and subset representations from the complete-supervision branch	Section 3.7; Equation (9)
$\mathcal{L}_{\mathrm{subset}}$, $\mathcal{L}_{\mathrm{final}}$	Non-empty subset classification loss and final classification loss	Equations (9) and (32)
$\mathcal{L}_{\mathrm{atom}}$, $\mathcal{L}_{\mathrm{path}}$, $\mathcal{L}_{\mathrm{full}}$	Unknown-atom supervision, strict-superset path, and complete-representation reconstruction losses	Equations (30), (33), and (34)
$\lambda_{\mathrm{subset}}$, …, λ_{full}	Loss weights of the five terms (0.25, 1.00, 0.80, 0.60, and 0.50, in the order of Equation (35))	Equation (35); Section 4.6
sg[·], Avg	Stop-gradient operator; average over the valid indices of each sample and then over the samples	Section 3.10

Note: Bold symbols denote vectors and matrices. The superscript * marks quantities of the complete-supervision branch, the hat (^) network predictions, and the tilde (~) completed quantities that merge the known boundary with the predicted unknown atoms.

References

1. Werner, P.; Lopez-Martinez, D.; Walter, S.; Al-Hamadi, A.; Gruss, S.; Picard, R.W. Automatic recognition methods supporting pain assessment: A survey. *IEEE Trans. Affect. Comput.* **2022**, *13*, 530–552. [CrossRef]
2. El-Tallawy, S.N.; Pergolizzi, J.V.; Vasiliu-Feltes, I.; Ahmed, R.S.; LeQuang, J.K.; El-Tallawy, H.N.; Varrassi, G.; Nagiub, M.S. Incorporation of "artificial intelligence" for objective pain assessment: A comprehensive review. *Pain Ther.* **2024**, *13*, 293–317. [CrossRef] [PubMed]
3. Fang, R.; Hosseini, E.; Zhang, R.; Fang, C.; Rafatirad, S.; Homayoun, H. Survey on pain detection using machine learning models: Narrative review. *JMIR AI* **2025**, *4*, e53026. [CrossRef] [PubMed]
4. Camacho-Navas, C.; Li, L.; Poply, K.; Mehta, V.; Kyriacou, P. Pain assessment using physiological responses/markers in different types of pain: A scoping review. *npj Digit. Med.* **2026**, *9*, 64. [CrossRef] [PubMed]
5. Walter, S.; Gruss, S.; Ehleiter, H.; Tan, J.-W.; Traue, H.C.; Crawcour, S.C.; Werner, P.; Al-Hamadi, A.; Andrade, A.O. The BioVid heat pain database data for the advancement and systematic validation of an automated pain recognition system. In Proceedings of the 2013 IEEE International Conference on Cybernetics (CYBCO), Lausanne, Switzerland, 13–15 June 2013; pp. 128–131.
6. Benavent-Lledo, M.; Mulero-Pérez, D.; Ortiz-Perez, D.; Rodriguez-Juan, J.; Berenguer-Agullo, A.; Psarrou, A.; Garcia-Rodriguez, J. A comprehensive study on pain assessment from multimodal sensor data. *Sensors* **2023**, *23*, 9675. [CrossRef] [PubMed]
7. Lu, Z.; Ozek, B.; Kamarthi, S. Transformer encoder with multiscale deep learning for pain classification using physiological signals. *Front. Physiol.* **2023**, *14*, 1294577. [CrossRef] [PubMed]
8. Gkikas, S.; Tachos, N.S.; Andreadis, S.; Pezoulas, V.C.; Zaridis, D.; Gkois, G.; Matonaki, A.; Stavropoulos, T.G.; Fotiadis, D.I. Multimodal automatic assessment of acute pain through facial videos and heart rate signals utilizing transformer-based architectures. *Front. Pain Res.* **2024**, *5*, 1372814. [CrossRef] [PubMed]
9. Benavent-Lledo, M.; Lopez-Valle, M.D.; Ortiz-Perez, D.; Mulero-Pérez, D.; Garcia-Rodriguez, J.; Psarrou, A. Multimodal pain assessment with transformers. *Neurocomputing* **2026**, *664*, 132066. [CrossRef]
10. Farmani, J.; Bargshady, G.; Gkikas, S.; Tsiknakis, M.; Fernandez Rojas, R. A CrossMod-Transformer deep learning framework for multi-modal pain detection through EDA and ECG fusion. *Sci. Rep.* **2025**, *15*, 29467. [CrossRef] [PubMed]
11. Steur, N.A.K.; Schwenker, F. Multimodal pain recognition based on contrastive adversarial autoencoder pretraining. *Mach. Learn. Knowl. Extr.* **2025**, *7*, 165. [CrossRef]
12. Wu, R.; Wang, H.; Chen, H.-T.; Carneiro, G. Deep multimodal learning with missing modality: A survey. *Trans. Mach. Learn. Res.* **2026**, arXiv:2409.07825.
13. Neverova, N.; Wolf, C.; Taylor, G.; Nebout, F. ModDrop: Adaptive multi-modal gesture recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 1692–1706. [CrossRef]
14. Guo, Z.; Jin, T.; Zhao, Z. Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Bangkok, Thailand, 11–16 August 2024; pp. 1726–1736.
15. Hu, L.; Shi, T.; Feng, W.; Shang, F.; Wan, L. Deep correlated prompting for visual recognition with missing modalities. In Proceedings of the Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Vancouver, BC, Canada, 10–15 December 2024; pp. 67446–67466.
16. Zhao, J.; Li, R.; Jin, Q. Missing modality imagination network for emotion recognition with uncertain missing modalities. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Online, 1–6 August 2021; pp. 2608–2618.
17. Wang, H.; Chen, Y.; Ma, C.; Avery, J.; Hull, L.; Carneiro, G. Multi-modal learning with missing modality via shared-specific feature modelling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 15878–15887.
18. Dai, R.; Li, C.; Yan, Y.; Mo, L.; Qin, K.; He, T. Unbiased missing-modality multimodal learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Honolulu, HI, USA, 19–23 October 2025; pp. 24507–24517.
19. Gkikas, S.; Tsiknakis, M. Automatic assessment of pain based on deep learning methods: A systematic review. *Comput. Methods Programs Biomed.* **2023**, *231*, 107365. [CrossRef] [PubMed]
20. Ben Aoun, N. A review of automatic pain assessment from facial information using machine learning. *Technologies* **2024**, *12*, 92. [CrossRef]
21. Huo, J.; Yu, Y.; Lin, W.; Hu, A.; Wu, C. Application of AI in multilevel pain assessment using facial images: Systematic review and meta-analysis. *J. Med. Internet Res.* **2024**, *26*, e51250. [CrossRef] [PubMed]
22. Patania, S.; Boccignone, G.; Buršić, S.; D'Amelio, A.; Lanzarotti, R. Deep graph neural network for video-based facial pain expression assessment. In Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing (SAC '22), Online, 25–29 April 2022; pp. 585–591.
23. Pittara, M.; Anastasiou, A.; Andreou, K.; Matsangidou, M.; Panayides, A.; Petkov, N.; Pattichis, C.S. Machine learning methods on BioVid heat pain database for pain intensity estimation. *Sci. Rep.* **2025**, *15*, 38173. [CrossRef] [PubMed]

24. Assran, M.; Bardes, A.; Fan, D.; Garrido, Q.; Howes, R.; Komeili, M.; Muckley, M.; Rizvi, A.; Roberts, C.; Sinha, K.; et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv* **2025**, arXiv:2506.09985.
25. Mur-Labadia, L.; Muckley, M.; Bar, A.; Assran, M.; Sinha, K.; Rabbat, M.; LeCun, Y.; Ballas, N.; Bardes, A. V-JEPA 2.1: Unlocking dense features in video self-supervised learning. *arXiv* **2026**, arXiv:2603.14482.
26. Kächele, M.; Werner, P.; Al-Hamadi, A.; Palm, G.; Walter, S.; Schwenker, F. Bio-visual fusion for person-independent recognition of pain intensity. In Proceedings of the 12th International Workshop on Multiple Classifier Systems (MCS 2015), Günzburg, Germany, 29 June–1 July 2015; pp. 220–230.
27. Kächele, M.; Thiam, P.; Amirian, M.; Werner, P.; Walter, S.; Schwenker, F.; Palm, G. Multimodal data fusion for person-independent, continuous estimation of pain intensity. In Proceedings of the 16th International Conference on Engineering Applications of Neural Networks (EANN 2015), Rhodes, Greece, 25–28 September 2015; pp. 275–285.
28. Vianto, J.; Divakaran, A.; Yang, H.; Yeom, S.; Kim, S.; Kim, S.; Shin, J. Multimodal model for automated pain assessment: Leveraging video and fNIRS. *Appl. Sci.* **2025**, *15*, 5151. [CrossRef]
29. Gkikas, S.; Tsiknakis, M. Twins-PainViT: Towards a modality-agnostic vision transformer framework for multimodal automatic pain assessment using facial videos and fNIRS. In Proceedings of the 2024 12th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), Glasgow, UK, 15–18 September 2024; pp. 13–21.
30. Farmani, J.; Giuseppi, A.; Bargshady, G.; Fernandez Rojas, R. Multimodal automatic acute pain recognition using facial expressions and physiological signals. In Proceedings of the 31st International Conference on Neural Information Processing (ICONIP 2024), Auckland, New Zealand, 2–6 December 2024; pp. 49–62.
31. Lee, Y.-L.; Tsai, Y.-H.; Chiu, W.-C.; Lee, C.-Y. Multimodal prompting with missing modalities for visual recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 14943–14952.
32. Jang, J.; Wang, Y.; Kim, C. Towards robust multimodal prompting with missing modalities. In Proceedings of the 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024; pp. 8070–8074.
33. Reza, M.K.; Prater-Bennette, A.; Asif, M.S. Robust multimodal learning with missing modalities via parameter-efficient adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, *47*, 742–754. [CrossRef] [PubMed]
34. Zhao, Y.; Xi, W.; Fu, X.; Zhao, J. Enhancing multimodal model robustness under missing modalities via memory-driven prompt learning. In Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25), Montreal, QC, Canada, 16–22 August 2025; pp. 2458–2466.
35. Li, S.; Du, C.; Zhao, Y.; Huang, Y.; Zhao, H. What makes for robust multi-modal models in the face of missing modalities? *arXiv* **2023**, arXiv:2310.06383.
36. Lee, J.; Lee, Y.; Kim, J.; Kosiorek, A.; Choi, S.; Teh, Y.W. Set Transformer: A framework for attention-based permutation-invariant neural networks. In Proceedings of the 36th International Conference on Machine Learning (ICML), Long Beach, CA, USA, 9–15 June 2019; pp. 3744–3753.
37. Rota, G.-C. On the foundations of combinatorial theory I. Theory of Möbius functions. *Z. Wahrscheinlichkeitstheorie Verw. Geb.* **1964**, *2*, 340–368. [CrossRef]
38. Kang, J.S.; Erginbas, Y.E.; Butler, L.; Pedarsani, R.; Ramchandran, K. Learning to understand: Identifying interactions via the Möbius transform. In Proceedings of the Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Vancouver, BC, Canada, 10–15 December 2024; pp. 46160–46202.
39. Hendrycks, D.; Gimpel, K. Gaussian error linear units (GELUs). *arXiv* **2016**, arXiv:1606.08415.
40. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
41. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
42. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
43. Zhang, K.; Zhang, Z.; Li, Z.; Qiao, Y. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.* **2016**, *23*, 1499–1503. [CrossRef]
44. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. In Proceedings of the 7th International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.
45. Gkikas, S.; Tsiknakis, M. A full transformer-based framework for automatic pain estimation using videos. In Proceedings of the 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Sydney, Australia, 24–27 July 2023; pp. 1–6.

46. Gkikas, S.; Fernandez Rojas, R.; Tsiknakis, M. PainFormer: A vision foundation model for automatic pain assessment. *IEEE Trans. Affect. Comput.* **2025**, *16*, 3369–3386. [CrossRef]
47. Rubin, D.B. Inference and missing data. *Biometrika* **1976**, *63*, 581–592. [CrossRef]
48. Grabisch, M. k-order additive discrete fuzzy measures and their representation. *Fuzzy Sets Syst.* **1997**, *92*, 167–189. [CrossRef]
49. Gruss, S.; Geiger, M.; Werner, P.; Wilhelm, O.; Traue, H.C.; Al-Hamadi, A.; Walter, S. Multi-modal signals for analyzing pain responses to thermal and electrical stimuli. *J. Vis. Exp.* **2019**, *146*, e59057. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.