International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 42, No. 6, December, 2025, pp. 3539-3553
Journal homepage: http://iieta.org/journals/ts

# Fusing Multi-Channel Perceptual Features for Visual Communication Image Representation Learning and Interpretability Analysis

Shan Dou

School of Fine Arts, LNNU, Liaoning Normal University, Dalian 116029, China

Corresponding Author Email: dd15941132336@163.com

https://doi.org/10.18280/ts.420638

**Received:** 25 March 2025
**Revised:** 19 October 2025
**Accepted:** 27 October 2025
**Available online:** 31 December 2025

***Keywords:***
*visual communication images, multimodal semantic fusion, CERL, SCM, conceptual prototype learning, generative explanations*

## ABSTRACT

Visual communication images play a central role in commercial advertising and digital interaction. Multimodal semantic feature fusion is key to enhancing their representation learning performance. However, traditional associative learning methods struggle to support the deep needs of causal understanding and interpretability in design fields. Existing multimodal fusion solutions often suffer from issues such as blind fusion, inadequate representation robustness, and lack of interpretability. Moreover, they have not achieved seamless integration of structured domain knowledge injection, causal reasoning, and generative explanation, making the transition from associative learning to causal understanding difficult. To address these challenges, this paper proposes an Explainable Causal Perception Computing Framework (ECPCF), which establishes a Causal Explainable Representation Learning (CERL) paradigm. The framework builds a domain-specific conceptual prototype memory bank by structurally injecting visual communication knowledge and designs a causal intervention mechanism based on a structural causal model (SCM) to accurately decouple task-related causal factors from irrelevant style factors. It then integrates generative explanations and retrospective attribution to form a full-link interpretable system. Experimental results show that the proposed method outperforms baseline models in multiple tasks, such as design intention classification and defect detection, on datasets like PosterNet, UI20K, and our custom dataset, achieving optimal interpretability metrics. Ablation studies validate the necessity of each core module, and cross-dataset testing demonstrates strong generalization ability, with a performance drop of only 6.2%. The study demonstrates that ECPCF successfully transitions from associative learning to causal understanding, offering a high-performance, robust, and interpretable solution for multimodal semantic fusion in visual communication image representation learning. This research provides significant insights into the field of interpretable multimodal learning in image processing.

## 1. INTRODUCTION

Visual communication images are widely applied in key fields such as commercial promotion, digital interaction, and public communication. They possess core attributes of both information transmission and aesthetic perception, and their automated analysis and optimization are of great significance for improving communication efficiency and optimizing user experience [1-3]. The deep involvement of image processing technologies provides new technical paths for this field [4, 5], and related research has become a hot direction at the intersection of computer vision and design. However, the information carrier of visual communication images exhibits significant multimodal characteristics, encompassing visual images, embedded text, design metadata, and other diverse information [6, 7]. Single-modal features are insufficient to fully capture their core value and design intentions. Thus, multimodal semantic fusion [8, 9] has become a key path for improving representation learning performance. In recent years, cross-modal alignment methods based on Transformer have been validated as effective in multiple visual tasks, but inherent flaws still exist in visual communication image analysis.

More importantly, the demands in the design field have moved beyond simply accurate predictions, toward a deeper understanding of the causal relationships between design features and communication effects [10]. This understanding forms the basis for supporting design optimization, defect localization, and other practical needs. However, traditional associative learning methods can only capture statistical associations between features and labels and cannot remove confounding biases from the data, making them inadequate to meet the core demands of causal understanding.

Although much research has accumulated in the fields of multimodal fusion and explainable learning, existing methods still face three core limitations. First, multimodal fusion lacks guidance. Existing cross-modal fusion methods, such as attention mechanisms and modality concatenation strategies, lack structured guidance from visual communication domain knowledge. This can lead to ineffective interactions between irrelevant features, resulting in insufficient discriminability and relevance of the fused features [11-13]. Second, the lack of robustness in representations and causality. Current methods do not fully consider style biases in the data, making model representations easily interfered with by surface style changes, and they only learn associative relationships rather than causal logic, limiting their generalization ability [14-16]. Third, insufficient explainability and domain adaptability. Existing explainability methods, such as gradient-weighted visualizations and local interpretable models, are mostly post hoc attributions with coarse granularity, disconnected from the concepts in the visual communication domain, making them difficult for designers to understand and apply [17, 18]. In summary, existing research has not constructed a seamless integration framework for structured domain knowledge injection, causal reasoning, and generative explanation, and has not achieved the paradigm shift from associative learning to causal understanding. This key gap severely limits the deep application of visual communication image analysis technologies.

The goal of this study is to construct a CERL framework based on structured domain knowledge injection, achieving efficient fusion and explainability analysis of multimodal semantic features in visual communication images, while balancing model performance, representation robustness, and causal interpretability. Around this goal, the core academic contributions of this paper are as follows:

(1) Propose an ECPCF, establishing the CERL paradigm, and for the first time realizing the seamless integration of structured domain knowledge injection, counterfactual reasoning, and generative explanation, promoting the paradigm shift in visual communication image analysis from associative learning to causal understanding.

(2) Design a causal intervention mechanism based on a SCM, formally defining intervention variables and causal factor conservation objectives, realizing the precise decoupling of task-related causal factors from irrelevant style factors, and providing a strict theoretical guarantee for representation robustness.

(3) Build a domain-specific conceptual prototype memory bank and bidirectional interpretability system for visual communication, combining concept activation heatmaps and generative examples to achieve full-link quantitative traceability of perception-concept-decision, significantly improving the domain adaptability and practical value of the explanation results.

(4) Verify the effectiveness of the framework through multidimensional experiments, including performance comparison with existing optimal models, human expert evaluation, cross-dataset generalization testing, and explainability quantification, providing a new paradigm and methodological reference for research on interpretable multimodal fusion in image processing.

The structure of the subsequent chapters is as follows: Chapter 2 provides a detailed explanation of the overall design of the proposed framework and the technical details of each core module, including the construction of the conceptual prototype memory bank, design of the causal intervention mechanism, and implementation of the interpretability system. Chapter 3 verifies the effectiveness of the method through multiple comparison experiments, covering performance evaluation, robustness testing, explainability verification, and cross-dataset generalization analysis. Chapter 4 discusses the research findings, typical failure cases, universal methodologies, and future research directions. Chapter 5 summarizes the core conclusions and academic contributions of the paper.

## 2. METHODS

### 2.1 Problem formalization

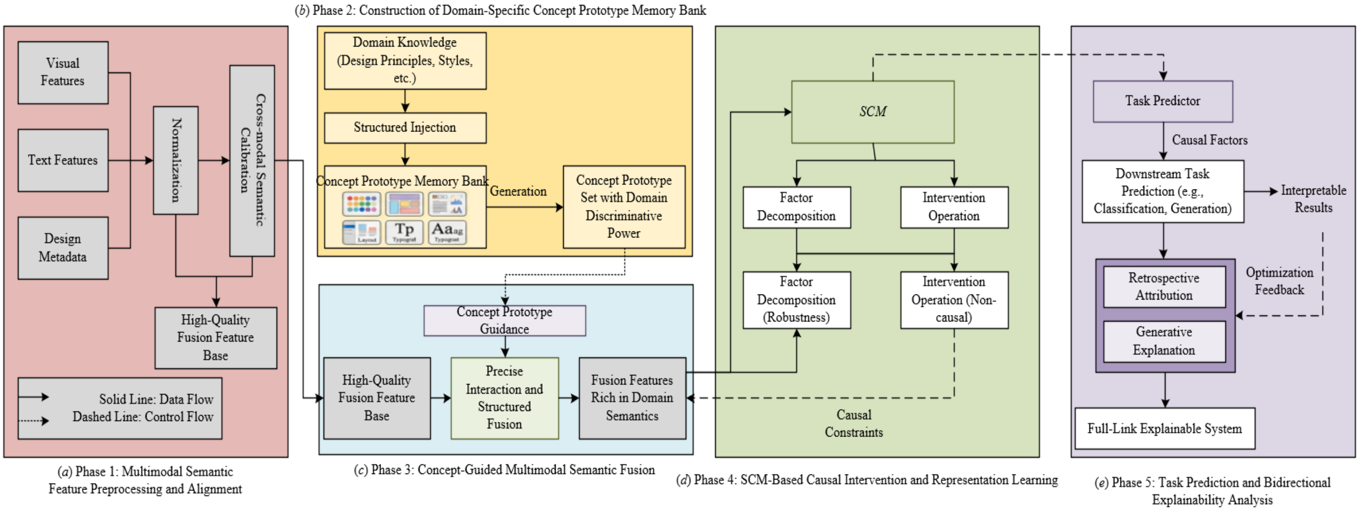
The multimodal semantic features of visual communication images stem from three core information carriers. First, the formal definitions of each modality's features are provided. Let the visual semantic features of the input visual communication image be $V \in R^{H \times W \times C}$, where $H$ and $W$ are the height and width of the image, and $C$ is the number of visual feature channels, encoded by convolutional neural networks or visual Transformers, encompassing low-level textures, color distribution, and high-level semantic information; the text semantic features $T \in R^{L \times D}$ correspond to the embedded textual content in the image, where $L$ is the text sequence length, and $D$ is the text encoding dimension, generated by pre-trained language models, merging textual semantics with visual layout attributes; the design metadata semantic features $M \in R^{K}$ are low-dimensional structured vectors, where $K$ is the metadata dimension, containing design attributes such as layout type, color tone, etc. The three types of modality features are integrated into a multimodal input $X = \{V, T, M\}$, and subsequent fusion and representation learning are based on this multimodal input.

The core goal of CERL is to learn a structured representation $R$ from the multimodal input $X$, and decouple it into task-related causal factors $R_c$ and task-independent style factors $R_s$, i.e., $R = (R_c, R_s)$. This goal can be formalized through two key constraints: First, the causal factors must retain the core information of task decisions, satisfying $P(Y \mid R_c, R_s) = P(Y \mid R_c)$, which indicates that the task label $Y$ is determined only by the causal factor; second, the causal factors must remain invariant under style interventions, i.e., $P(R_c \mid \mathrm{do}(R_s = r_s')) = P(R_c)$, where $do(\ )$ is an intervention operation and $r_s'$ is any style factor value. This constraint ensures the robustness of the representation to style changes. At the same time, the model must output concept attribution weights $\alpha \in R^{N}$ and generative explanation examples $S$, where the former quantifies the contribution of each concept to task decisions, and the latter visually presents the core concepts, jointly supporting the interpretability analysis.

To rigorously characterize the above causal relationships, a SCM $M = \langle U, V, F, P(U) \rangle$ is introduced as the theoretical basis. Here, $U$ is the set of exogenous variables, representing unobservable noise and confounding factors; $V = \{X, R_c, R_s, Y\}$ is the set of endogenous variables, including multimodal input, the two types of factors, and task labels; $F = \{f_X, f_{R_c}, f_{R_s}, f_Y\}$ is the set of causal mechanisms, which defines the mappings $U \rightarrow X$, $X \rightarrow R_c$, $X \rightarrow R_s$, and $R_c \rightarrow Y$; $P(U)$ is the prior probability distribution of exogenous variables. The causal flow between variables satisfies: the multimodal input $X$ is a common cause of $R_c$ and $R_s$, while $R_s$ has no direct causal link with $Y$. This structure provides a strict theoretical boundary for subsequent causal intervention and factor decoupling.

### 2.2 CERL framework and CERL paradigm

To address the issues of blind fusion, inadequate robustness, and lack of interpretability in multimodal semantic fusion of visual communication images, this paper proposes the CERL paradigm. The core idea of this paradigm is to guide structured domain knowledge injection, constrain the process with causal reasoning, and aim for interpretability, systematically realizing the paradigm shift from traditional associative learning to causal understanding. This paradigm breaks through the limitations of existing methods, which only rely on data statistical associations, by transforming visual communication domain expertise into computable conceptual prototypes, providing directed guidance for multimodal fusion and avoiding irrelevant feature interactions. At the same time, it constrains the representation learning process with causal reasoning to ensure that the model captures task-core causal logic rather than superficial style associations. Finally, through a full-link interpretable system, it transforms the learning results into domain-understandable knowledge, achieving a virtuous cycle of "learning - explanation - validation," aligning with the deep needs of the design field for causal understanding and practical guidance.


Visual Features
Text Features
Design Metadata
Normalization
Cross-modal Semantic Calibration
High-Quality Fusion Feature Base
Solid Line: Data Flow
Dashed Line: Control Flow
Domain Knowledge (Design Principles, Styles, etc.)
Structured Injection
Concept Prototype Memory Bank
Layout
Tp
Typograt
Aaag
Typograt
Generation
Concept Prototype Set with Domain Discriminative Power
Concept Prototype Guidance
High-Quality Fusion Feature Base
Precise Interaction and Structured Fusion
Fusion Features Rich in Domain Semantics
SCM
Factor Decomposition
Intervention Operation
Factor Decomposition (Robustness)
Intervention Operation (Non-causal)
Causal Constraints
Task Predictor
Causal Factors
Downstream Task Prediction (e.g., Classification, Generation)
Interpretable Results
Retrospective Attribution
Generative Explanation
Optimization Feedback
Full-Link Explainable System

(*a*) Phase 1: Multimodal Semantic Feature Preprocessing and Alignment

(*b*) Phase 2: Construction of Domain-Specific Concept Prototype Memory Bank

(*c*) Phase 3: Concept-Guided Multimodal Semantic Fusion

(*d*) Phase 4: SCM-Based Causal Intervention and Representation Learning

(*e*) Phase 5: Task Prediction and Bidirectional Explainability Analysis

**Figure 1.** The ECPCF

The ECPCF is the specific implementation carrier of the CERL paradigm. It adopts a hierarchical modular design, clearly distinguishing the information transmission and logical control paths by using solid lines to mark the data flow and dashed lines to mark the control flow. The core innovative modules are highlighted with differentiated colors to emphasize design priorities. The framework follows a five-phase closed-loop logic, with each phase progressing incrementally and providing mutual feedback: the Multimodal Semantic Feature Preprocessing and Alignment phase standardizes and cross-modally calibrates the three types of features (visual, textual, and design metadata), providing a high-quality feature foundation for subsequent fusion; the Construction of Domain-Specific Concept Prototype Memory Bank phase structurally injects domain knowledge into the model, generating a concept prototype set with domain discriminative power to provide directional guidance for the fusion process; the concept-guided multimodal semantic fusion phase uses concept prototypes as intermediaries to achieve precise interaction and structured fusion of multimodal features, generating fusion features rich in domain semantics; the SCM-based Causal Intervention and Representation Learning phase decouples the fusion features into causal and style factors through factor decomposition and intervention operations, ensuring the robustness and causal validity of the representations; the Task Prediction and Bidirectional Explainability Analysis phase completes downstream task prediction based on causal factors, while also outputting interpretable results through retrospective attribution and generative explanation. The explanation information can be fed back to the concept prototype memory bank for optimization, forming a complete closed loop. This framework achieves the organic unity of multimodal semantic fusion and CERL through the collaborative function of all modules, fully embodying the guiding-constraint-target core logic of the CERL paradigm. Figure 1 intuitively demonstrates the ECPCF.

### 2.3 Multimodal semantic feature preprocessing and alignment

The core of multimodal semantic feature preprocessing is to precisely extract the core information of each modality and complete the preliminary standardization to provide high-quality input for subsequent fusion. The visual semantic features adopt a fusion strategy of low-level and high-level features: low-level features include the HSV histogram, LBP texture descriptor, and spatial layout matrix. The HSV histogram quantizes the image color space into 16×16×16 intervals, generating a 4096-dimensional vector. The LBP texture descriptor uses a 3 × 3 neighborhood calculation to generate a 256-dimensional vector. The spatial layout matrix divides the image into 16 × 16 grids and generates a 256-dimensional vector by calculating the average pixel values within each grid. The three features are concatenated to obtain the low-level visual feature $V_{low}\in R^{4608}$; high-level semantic features are generated by encoding with a pre-trained Swin Transformer. The input image is normalized to 224 × 224 and passed through the Swin-Tiny model to output a 768-dimensional feature vector $V_{high}\in R^{768}$. The final visual semantic feature $V$ is obtained by concatenating $V_{low}$ and $V_{high}$ after layer normalization (LN): $V=LN([V_{low};V_{high}])\in R^{5376}$. The text semantic features adopt joint encoding of semantics and visual attributes: the text content is encoded by the BERT-base model to obtain a 768-dimensional semantic vector $T_{sem}\in R^{768}$, and the text's visual attributes are one-hot encoded and normalized to generate a 64-dimensional vector $T_{vis}\in R^{64}$. The final text feature $T$ is obtained through attention-weighted fusion: $T$=$\alpha\cdot T_{sem}$+$(1-\alpha)\cdot T_{vis}$, where $\alpha\in[0,1]$ is the attention weight learned adaptively based on feature correlation. The design metadata semantic features are encoded structurally: the objective design labels are one-hot encoded to generate a 128-dimensional vector, and the subjective user perception feedback is generated by statistical heatmap peak positions and score distributions to form a 32-dimensional feature. The two are concatenated and standardized to obtain $M\in R160$.

To solve the semantic gap caused by the heterogeneity of multimodal features, a cross-modal semantic calibration alignment strategy based on contrastive learning is used. The core idea is to map features from different modalities into a unified semantic space through concept consistency contrastive loss. Let the multimodal feature set of the $i$-th sample in a batch be $\{V_i,T_i,M_i\}$, where the positive sample pair corresponds to different modality features of the same image, and the negative sample pair corresponds to either same-modality or cross-modality features from different images. The concept consistency contrastive loss $L_{cc}$ is defined as:

$$L_{cc}=-\frac{1}{3N}\sum_{i=1}^{N}\sum_{F_i\in\{V_i,T_i,M_i\},F_i^+\in\{V_i,T_i,M_i\},F_i\neq F_i^+}\log\frac{\exp\left(\mathrm{sim}(F_i,F_i^+)/\tau\right)}{\sum_{F_j\in N_i}\exp\left(\mathrm{sim}(F_i,F_j)/\tau\right)} \tag{1}$$

where, $N$ is the batch size, $F_i^+$ is the positive sample feature, $N_i$ is the set of negative sample features, $sim(,)$ is the cosine similarity function, and $\tau$ is the temperature parameter. The optimization objective of this loss is to minimize the distance between different modality features of the same image while maximizing the distance between features from different images. Additionally, the introduction of domain concept prior constraints in the similarity calculation ensures that the aligned features fit the semantic logic of the visual communication domain, laying the foundation for subsequent concept-guided fusion. Figure 2 shows the process of multimodal semantic feature preprocessing and concept prototype memory bank construction.

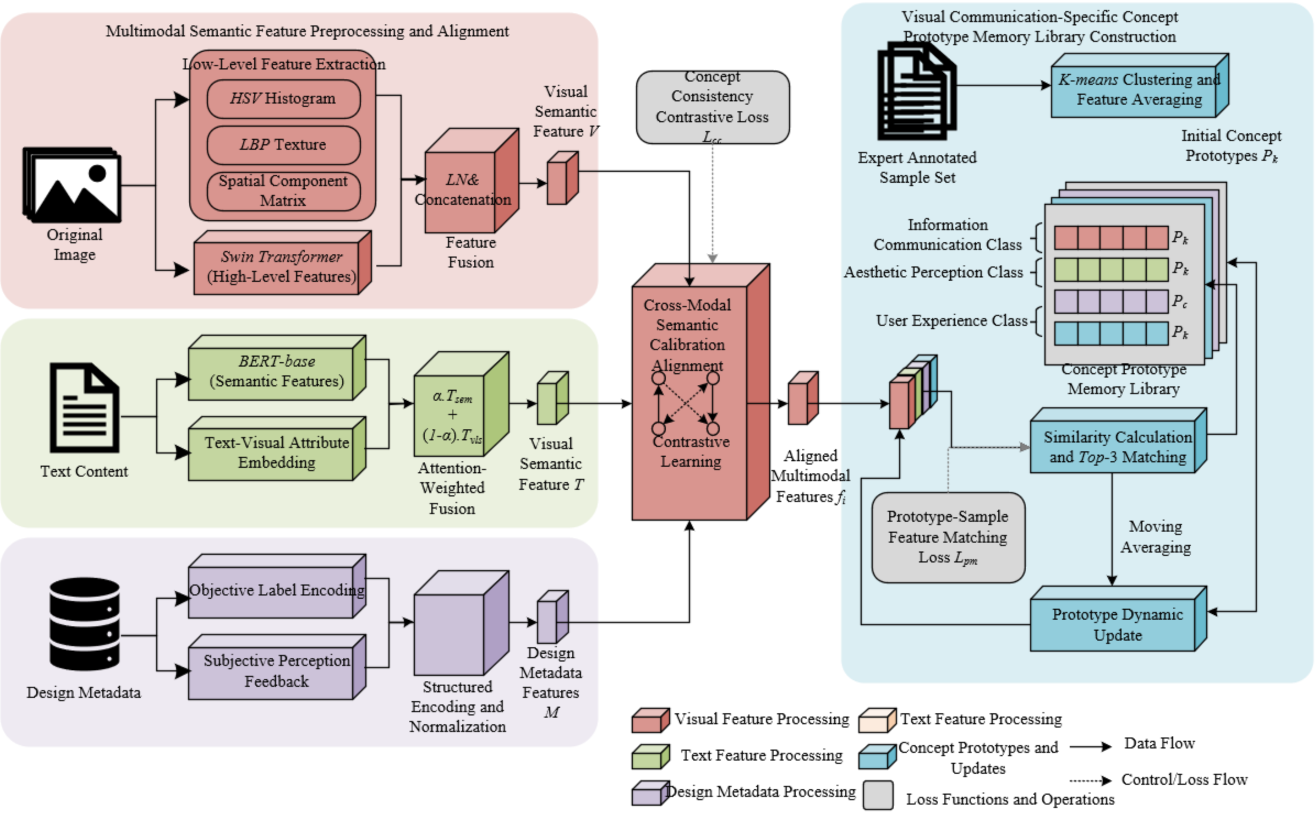

Multimodal Semantic Feature Preprocessing and Alignment
Low-Level Feature Extraction
HSV Histogram
LBP Texture
Spatial Component Matrix
Swin Transformer (High-Level Features)
Original Image
LN& Concatenation
Feature Fusion
Visual Semantic Feature V
BERT-base (Semantic Features)
Text-Visual Attribute Embedding
Text Content
α.T_sem + (1-α).T_vis
Attention-Weighted Fusion
Visual Semantic Feature T
Objective Label Encoding
Subjective Perception Feedback
Design Metadata
Structured Encoding and Normalization
Design Metadata Features M
Concept Consistency Contrastive Loss L_cc
Cross-Modal Semantic Calibration Alignment
Contrastive Learning
Aligned Multimodal Features f_i
Visual Communication-Specific Concept Prototype Memory Library Construction
Expert Annotated Sample Set
K-means Clustering and Feature Averaging
Initial Concept Prototypes P_k
Information Communication Class
Aesthetic Perception Class
User Experience Class
P_k
P_k
P_c
P_k
Concept Prototype Memory Library
Similarity Calculation and Top-3 Matching
Prototype-Sample Feature Matching Loss L_pm
Moving Averaging
Prototype Dynamic Update
Visual Feature Processing
Text Feature Processing
Text Feature Processing
Concept Prototypes and Updates
Design Metadata Processing
Loss Functions and Operations
Data Flow
Control/Loss Flow

**Figure 2.** Multimodal semantic feature preprocessing and concept prototype memory bank construction process

## 2.4 Visual communication domain-specific concept prototype memory bank construction

The core of constructing the visual communication domain-specific concept prototype memory bank is to build a structured concept system that fits domain needs and generate learnable concept prototypes. The concept set definition follows both the consensus of domain experts and the practical requirements of design, covering three core categories: information transmission, aesthetic perception, and user experience. Information transmission concepts focus on the core information delivery function of design, including information hierarchy clarity, theme prominence, and text readability; aesthetic perception concepts rely on visual design aesthetics principles, covering color harmony, layout balance, and style consistency; user experience concepts relate to audience perception feedback, including visual appeal, information retrieval efficiency, and emotional resonance. The concept screening process is completed collaboratively by three visual communication domain experts and two computer vision researchers, with multiple rounds of discussion to eliminate ambiguous concepts. Ultimately, 60 core concepts form the concept set, ensuring the domain adaptability, discriminative power, and operability of the concepts.

Concept prototype initialization is completed based on an expert-labeled sample set, using a strategy combining clustering and feature averaging. First, an expert-labeled sample set is constructed. For each concept $c_k$ ($k$ = 1,2,...,$K$, $K$

= 60), the expert selects 100 representative visual communication image samples to form the labeled set $S_{c_k}$; for each sample, the aligned multimodal semantic features are extracted. The K-means clustering algorithm is used to cluster the feature set corresponding to $S_{c_k}$, and the initial concept prototype $p_k$ is obtained by averaging the features. The initialization formula for the concept prototype vector set $P$=[$p_1$,$p_2$,...,$p_K$]∈$R^{D\times K}$ is:

$$p_k = \frac{1}{|S_{c_k}|} \sum_{x \in S_{c_k}} Feat(x) \tag{2}$$

where, $Feat(x)$ denotes the multimodal semantic features of sample $x$, and $|S_{c_k}|$ is the size of the concept $c_k$ labeled sample set. This initialization method ensures that the prototypes can accurately capture the core feature distribution of similar concepts.

To make the concept prototypes adaptive to data distribution and dynamically optimized, a prototype-sample feature matching loss is designed to enable dynamic updating of the memory bank. During training, for each sample's multimodal feature $f_i$ in the batch, the cosine similarity $sim(f_i,p_k)$ between the sample and each concept prototype $p_k$ is calculated. The top-3 concept prototypes most matching the sample are selected to form the matching prototype set. The prototype-sample feature matching loss $L_{pm}$ is defined as:

$$L_{pm} = \frac{1}{N} \sum_{i=1}^{N} (\,1 - \max_{k \in K_i} sim\,(f_i, p_k)) \tag{3}$$

where, $N$ is the batch size, and $K_i$ is the matching prototype set for sample $i$. This loss minimizes the distance between the sample features and matching prototypes, and during training, each iteration updates the concept prototypes based on the current batch features:

$$p_k \leftarrow \eta \cdot p_k + (1-\eta) \cdot \frac{1}{|B_k|} \sum_{f_i \in B_k} f_i \tag{4}$$

where, $\eta \in (0,1)$ is the update weight, and $B_k$ is the set of sample features in the batch that match prototype $p_k$, ensuring that the prototypes dynamically adapt to the data distribution during training and improve the precision of concept-guided fusion.

### 2.5 Concept-guided multimodal semantic fusion

Concept-guided multimodal semantic fusion uses domain-specific concept prototypes as intermediaries and achieves precise interaction of multimodal features through a cross-attention mechanism. The core idea is to avoid ineffective fusion by relying on domain concepts, enhancing the semantic relevance of features to the domain. First, feature-concept prototype matching is performed. For the preprocessed and aligned visual, textual, and design metadata features $V \in R^D$, $T \in R^D$, $M \in R^D$, the cosine similarity with each prototype in the concept prototype set $P \in R^{D\times K}$ is calculated, and the top-5 most similar concepts are selected to form the active concept set $P_{act} = [p_1^{act}, \ldots, p_5^{act}]$. The cosine similarity calculation formula is:

$$sim(f,p_k) = \frac{f \cdot p_k^{\perp}}{\|f\|2 \cdot \|p_k\|_2} \tag{5}$$

where, $f$ is any modality feature, and $p_k$ is the $k$-th concept prototype. This process ensures that only domain concepts related to the current sample's semantics are activated, providing directional guidance for subsequent fusion.

The concept-guided cross-attention mechanism is constructed based on the active concept set to achieve cross-modal feature interaction. The core idea is to calculate the attention weights between modalities through concept prototypes. First, the modality features are associated with the active concept prototypes, yielding modality-concept associated features $V_{rel}$=$V \cdot P_{act}$, $T_{rel}$=$T \cdot P_{act}$, $M_{rel}$=$M \cdot P_{act}$. Then, using modality-concept associated features as a bridge, the cross-modal attention weights between visual and text, visual and metadata, and text and metadata are calculated. Taking the visual-text attention weight calculation as an example:

$$A_{V\to T} = softmax\left(\frac{V_{rel} \cdot T_{rel}^{\mathrm{T}}}{\sqrt{D_{act}}}\right) \tag{6}$$

where, $D_{act}$=5 is the number of active concepts, and the attention weight $A_{V\to T}$ quantifies the guidance weight of visual features on text features. Similarly, other modality attention weights $A_{T\to V}$, $A_{V\to M}$, etc., can be obtained. Finally, the fused feature $F$ is obtained by concatenating the weighted features from each modality:

$$F = LN([A_{T\to V}T + A_{M\to V}M + V, A_{V\to T}V + A_{M\to T}M + T]) \tag{7}$$

where, LN ensures stable feature distribution. This fusion method achieves precise semantic alignment and interaction of multimodal features through concept mediation.

To ensure semantic consistency between the fused features and domain concepts, a concept consistency loss $L_{con}$ is defined to constrain the fusion process. The average cosine similarity between the fused feature $F$ and the active concept set $P_{act}$ is computed, and the loss function is defined as:

$$L_{\mathrm{con}} = 1 - \frac{1}{5} \sum_{k=1}^{5} \mathrm{sim}\,(F, p_k^{act}) \tag{8}$$

The optimization objective of this loss is to maximize the semantic similarity between the fused features and the active concepts, forcing the fused features to encode domain-related information and enhance their structure and explainability. During training, the concept consistency loss is jointly optimized with the prototype-sample matching loss and task loss to ensure that the fused features meet both task requirements and domain concept constraints. Figure 3 presents the complete process of concept-guided multimodal semantic fusion.

### 2.6 Causal intervention and representation learning based on SCM

Based on the structure causal model defined earlier, this section formalizes the dependencies between variables using causal graphs and constructs causal constraints based on intervention theory to provide strict theoretical support for representation learning. The causal graph clearly depicts the causal flow between exogenous variables, multimodal inputs, causal factors, style factors, and task labels: multimodal inputs directly drive the generation of causal and style factors, task labels are only determined by causal factors, and style factors have no direct causal relation with task labels. Exogenous variables provide noise disturbances to the endogenous variables. Based on this structure, the intervention operation $do(R_s{=}r_s')$ is defined, meaning fixing the causal factor $R_c$ while replacing the style factor $R_s$ with any value $r_s'$. Combining the stability assumption of the causal mechanism in SCM, the core theoretical guarantee under intervention is derived:

$$P(Y|\mathrm{do}(R_s{=}r'_s),R_c){=}P(Y|R_c) \tag{9}$$

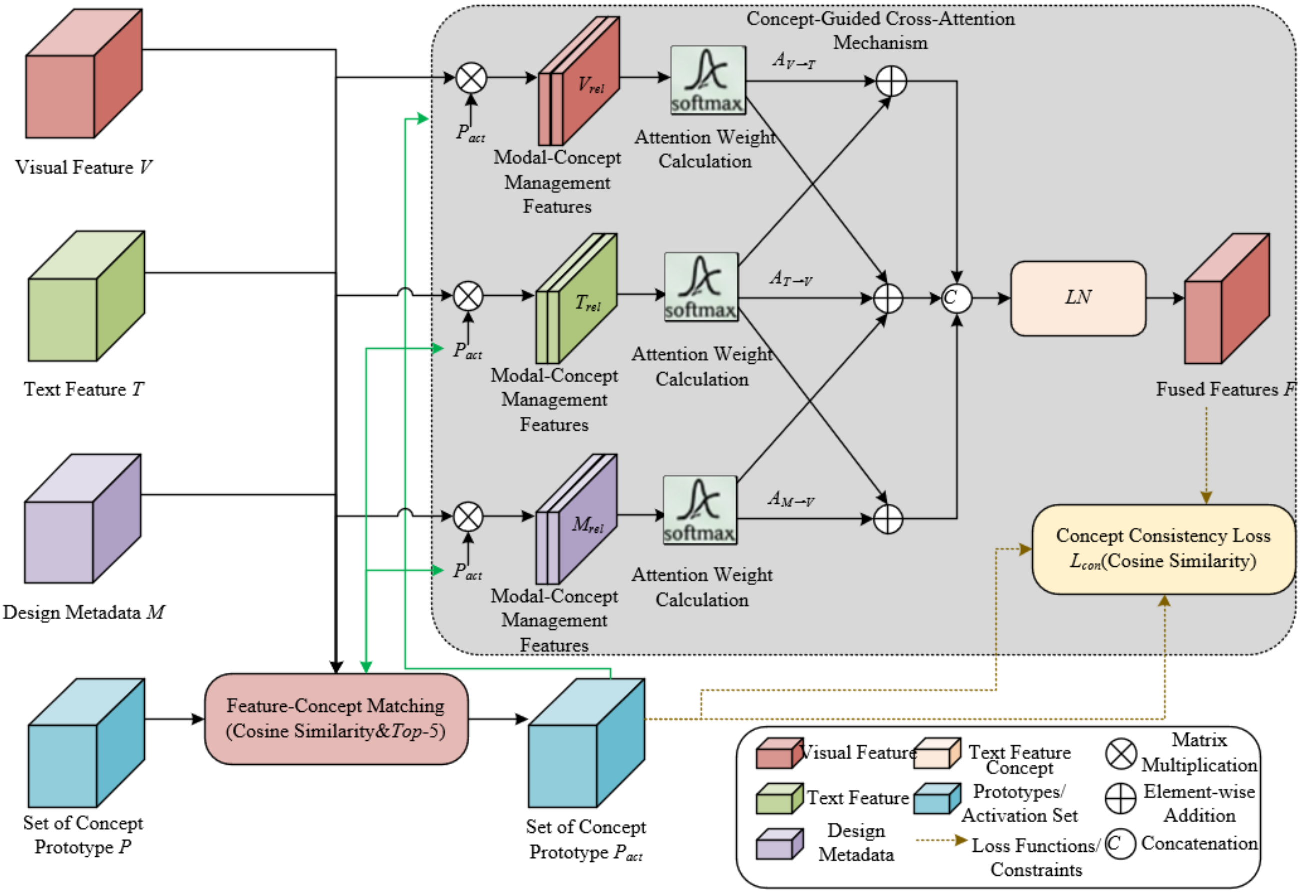

Concept-Guided Cross-Attention Mechanism
Visual Feature $V$
Text Feature $T$
Design Metadata $M$
$P_{act}$
$V_{rel}$
$T_{rel}$
$M_{rel}$
Modal-Concept Management Features
softmax
Attention Weight Calculation
$A_{V\to T}$
$A_{T\to V}$
$A_{M\to V}$
C
LN
Fused Features $F$
Concept Consistency Loss $L_{con}$(Cosine Similarity)
Set of Concept Prototype $P$
Feature-Concept Matching (Cosine Similarity&$Top$-5)
Set of Concept Prototype $P_{act}$
Visual Feature
Text Feature
Design Metadata
Text Feature Concept Prototypes/ Activation Set
Loss Functions/ Constraints
Matrix Multiplication
Element-wise Addition
Concatenation

**Figure 3.** Concept-guided multimodal semantic fusion process

The derivation process is as follows: since the causal mechanism of $Y$ is only determined by $R_c$, the intervention only changes the value of $R$ without disrupting the causal mechanism from $R_c$ to $Y$, so the conditional probability remains unchanged. This conclusion provides a theoretical basis for representation robustness, meaning that as long as the learned representation can accurately decouple $R_c$ and $R_s$, task predictions will remain stable during style changes.

Variational inference is used to perform factorization of the fused feature $F$, decoupling it into task-relevant causal factors $R_c\in\mathbb{R}^{Dc}$ and task-independent style factors $R_s\in\mathbb{R}^{Ds}$, where $D_c{+}D_s{=}D$ and $D$ is the dimension of the fused features. A variational encoder $q_\varphi(R_c,R_s|F)$ is introduced to approximate the posterior distribution, aiming to approach the true posterior $P(R_c,R_s|F)$. The prior distribution is defined as $p(R_c,R_s){=}p(R_c)p(R_s)$, assuming that $R_c$ and $R_s$ are independent. Based on the variational inference principle, the evidence lower bound (ELBO) is maximized to optimize the encoder parameters. The ELBO objective function is derived as:

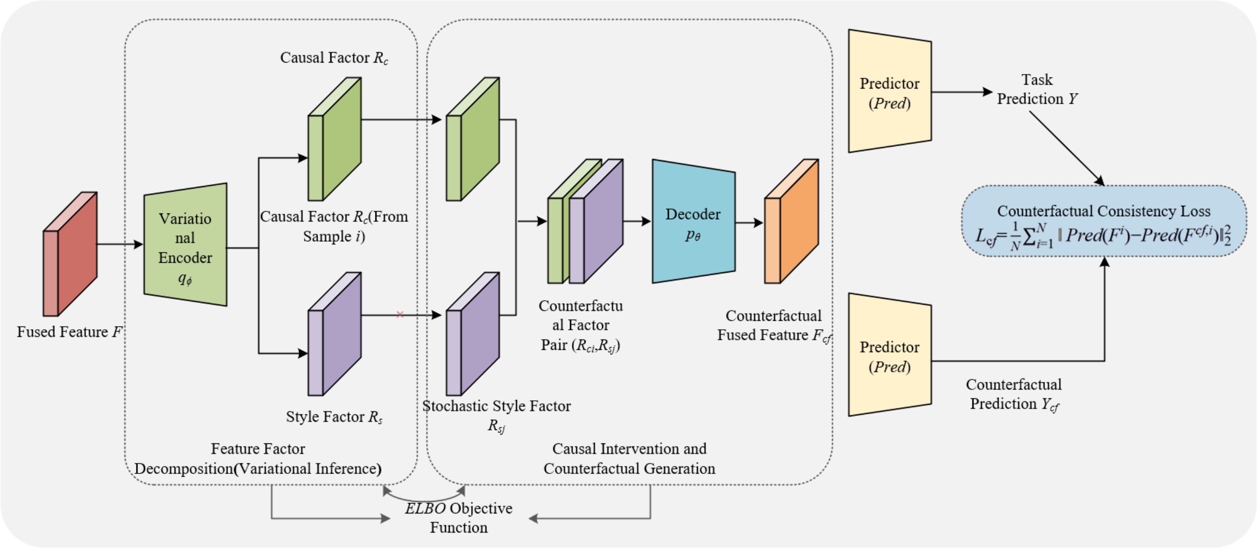
$$\begin{gathered} log\,P\,(F){\geq}E_{q_\phi(R_c,R_s|F)}[\,log\,P\,(F|R_c,R_s)] \\ \text{-}KL(q_\phi(R_c,R_s|F)\|p(R_c)p(R_s)) \end{gathered} \tag{10}$$

where, the first term is the reconstruction loss, which constrains the decoder to accurately reconstruct the fused features from the decoupled factors. The second term is the KL divergence, which ensures that the approximate posterior approximates the prior distribution and ensures the effectiveness of factor decoupling. A decoder $p_\theta(F|R_c,R_s)$ is introduced to implement feature reconstruction. The final factorization is completed by jointly optimizing the ELBO, resulting in decoupled representations $(R_c,R_s)$.

To strengthen the factor decoupling effect and verify the validity of the intervention theory, a counterfactual training strategy is designed, integrating the *do* operation into the training process. Counterfactual samples are generated based on the principle "causal factors remain unchanged, style factors are replaced": for the decoupled factors $(R^i_c,R^j_s)$ of the original sample, a style factor $R^j_s$ from another sample in the batch is randomly selected as a replacement, generating a counterfactual factor pair $(R^i_c,R^j_s)$, and then the counterfactual fused feature $F^{cf}{=}p_\theta(F|R^i_c,R^j_s)$ is generated through the decoder. The generation process must satisfy the constraint: the cosine similarity between the counterfactual factor pair $R^i_c$ and the original $R^i_c$ should not be lower than 0.95, ensuring that the causal factor is unaffected. To constrain the consistency of model predictions, a counterfactual consistency loss is defined:

$$L_{cf}{=}\frac{1}{N}\sum_{i=1}^{N}\|\,Pred(F^i)\text{-}Pred(F^{cf,i})\|_2^2 \tag{11}$$

where, *Pred*( ) represents the model's task prediction output, and $N$ is the batch size. This loss minimizes the prediction difference between the original and counterfactual samples, forcing the model to rely solely on causal factors for decision-making, further strengthening the task relevance of $R_c$ and the irrelevance of $R_s$, ultimately improving representation robustness and causal validity. During training, this loss is optimized together with the variational inference ELBO and concept consistency loss, forming a complete causal constraint system. Figure 4 shows the principles of causal intervention and representation learning based on SCM.



Causal Factor $R_c$
Variational Encoder $q_\phi$
Causal Factor $R_c$(From Sample $i$)
Fused Feature $F$
Style Factor $R_s$
Feature Factor Decomposition(Variational Inference)
Stochastic Style Factor $R_{sj}$
Counterfactual Factor Pair ($R_{ci}$,$R_{sj}$)
Decoder $p_\theta$
Counterfactual Fused Feature $F_{cf}$
Causal Intervention and Counterfactual Generation
$ELBO$ Objective Function
Predictor (*Pred*)
Task Prediction $Y$
Counterfactual Consistency Loss
$L_{cf}=\frac{1}{N}\sum_{i=1}^{N}\|Pred(F^i)-Pred(F^{cf,i})\|_2^2$
Predictor (*Pred*)
Counterfactual Prediction $Y_{cf}$

**Figure 4.** Causal intervention and representation learning based on SCM

## 2.7 Task prediction and bidirectional interpretability analysis

Task prediction is implemented through hierarchical decomposition of the causal purified representation, with the core idea of further decoupling the causal factor $R_c$ into three sub-representations to adapt to different downstream task requirements: core communication intention representation $R_{\mathrm{c}1}∈R^{D_{\mathrm{c}1}}$, aesthetic perception representation $\mathrm{R}_{c2}∈R^{D_2}$, and style-independent representation $\mathrm{R}_{c3}∈R^{D_{c3}}$ , such that $D_{c1}+D_{c2}+D_{c3}=D_c$, where $R_{c1}$ encodes the core information delivery target of the design, $R_{c2}$ characterizes aesthetic-related features, and $R_{c3}$ retains general structural information unrelated to style. To avoid overfitting, lightweight task heads are designed for different tasks: for the design intention classification task, a two-layer fully connected layer is used to build the task head, with the prediction formula:

$$\widehat{Y}_{cls}=softmax(W_2 \cdot ReLU(W_1 \cdot R_{c1})+b_2) \tag{12}$$

where, $W_1∈R^{256×D_{c1}}$, $W_2∈R^{256×D_{c2}}$ are learnable weights, $C$ is the number of categories, and $b_2∈R^{256}$ is the bias. For the defect detection task, a 3×3 convolutional layer and global average pooling are used to construct the task head, outputting the defect probability map:

$$\hat{Y}_{det}=\mathrm{s}igmoid(Conv(R_{c3})+b_{det}) \tag{13}$$

where, *Conv*( ) is a single-channel convolution operation that achieves pixel-level defect localization.

The bidirectional concept attribution interpreter constructs a full-link explainable system through forward generation and backward tracing. The forward-generation path is based on a conditional diffusion model, with the model input being the activated concept prototype set $P_{act}$ and random noise, and the output being visual examples fitting the concept semantics. The training strategy uses real sample-concept prototype pairing data, optimized with reconstruction loss and concept consistency loss, ensuring that the generated examples accurately match the target concepts. The backward-tracing path quantifies the decision contribution of each concept and semantic feature using a gradient-weighted method. The contribution calculation formula is defined as:

$$\alpha_k=\frac{(\nabla_F Pred(F)\cdot F)\cdot sim(F,p_k^{act})}{\sum_{k=1}^{5} sim\,(F,p_k^{act})} \tag{14}$$

where, $\nabla_F Pred(F)$ is the gradient of the prediction output with respect to the fused feature $F$, representing the sensitivity of the feature to the decision, and the similarity between the feature and concept prototype is used to weight the contribution. Based on this contribution, a cross-channel concept propagation graph is constructed, with nodes representing each modality feature and concept, and edges representing the contribution transmission coefficients, clearly showing the interaction paths between concepts and features.

The interpretability report is designed in two versions: academic analysis and design practice, covering both theoretical analysis and application guidance. The academic analysis version includes a quantitative attribution matrix, gradient heatmaps, cross-channel propagation graphs, and statistical significance analysis results, presented in the form of charts, supporting the academic verification of the method's effectiveness. The design practice version includes a core concept matching list, defect localization annotations, generative optimized examples, and targeted improvement suggestions, presented through a visual comparison interface. Both versions support interactive viewing, allowing users to click on attribution matrix elements to jump to corresponding heatmaps and generated examples, enhancing the explorability of the explanation results.

### 2.8 Model training strategy

The model training uses a multi-objective collaborative optimization strategy, combining the weighted losses of various tasks to construct a total loss function, ensuring the collaborative optimization of task performance, concept adaptability, and causal decoupling effects. The total loss function is defined as:

$$L_{total}=\lambda_1 L_{pm}+\lambda_2 L_{con}+\lambda_3 L_{cf} \qquad (15)$$

The core role and weight settings of each loss term are as follows: the task loss $L_{task}$ is the core loss. For the design intention classification task, cross-entropy loss is used, and for the defect detection task, Focal loss is used to directly constrain the model's task prediction accuracy, with a weight of 1.0. The prototype-sample matching loss $L_{pm}$ constrains the semantic alignment between sample features and concept prototypes, ensuring the effectiveness of concept guidance, with a medium gradient magnitude and a weight of 0.3. The concept consistency loss $L_{con}$ enhances the semantic relevance between fused features and domain concepts and collaborates with $L_{pm}$ to support the concept-guided mechanism, with a weight of 0.2. The counterfactual consistency loss $L_{cf}$ is a key constraint for causal decoupling, requiring balancing with the optimization priority of task loss. After experimental verification of its gradient sensitivity, its weight is set to 0.5. The weight parameters are determined through grid search with a search range of {0.1, 0.2, 0.3, 0.5, 1.0}, and the optimal combination is selected based on the performance of the validation set.

The training process uses the AdamW optimizer for parameter updates. This optimizer effectively suppresses overfitting through weight decay mechanisms, with parameter settings: $\beta_1$=0.9, $\beta_2$=0.999, weight decay coefficient of $1e^{-4}$, and *epsilon*=$1e^{-8}$. The learning rate schedule uses cosine annealing, with an initial learning rate set to $1e^{-4}$, balancing model convergence speed and stability. A higher learning rate in the initial stage accelerates parameter updates, while gradually decaying the learning rate in the later stage avoids gradient oscillations. To further improve the model's generalization ability, multiple regularization measures are introduced: Dropout layers with a dropout probability of 0.1 are inserted into the multimodal fusion layer and task head; L2 regularization is applied to the weights of all fully connected layers; data augmentation strategies such as random cropping, horizontal flipping, and color jittering are applied to the input images during training. The training batch size is set to 32, adapting to the memory capacity of a single NVIDIA A100 GPU. The total training iterations are set to 200, and an early stopping strategy is employed. If the performance of the validation set does not improve for 20 consecutive rounds, training stops and the model parameters with the best performance are saved. During training, gradient clipping is applied with a gradient norm threshold of 1.0 to prevent gradient explosion. Each modality feature is standardized before being input into the model, ensuring consistent feature distribution and improving training stability.

## 3. EXPERIMENT

### 3.1 Experimental setup

The experiment adopts a combination of public datasets, self-built datasets, and cross-dataset generalization test sets to ensure comprehensive data coverage and objective evaluation. The public datasets selected are the mainstream PosterNet and UI20K in the field of visual communication. PosterNet contains 50,000 commercial poster images, covering 10 core categories, with a multimodal annotation completeness rate of 98%, including visual, text, and design style metadata. UI20K contains 20,000 mobile UI interface images, divided into 15 functional categories, with annotated information such as widget positions and text attributes. The self-built dataset focuses on the visual communication needs in multiple scenes, constructed through three steps: collection, screening, and annotation, with a total of 30,000 valid samples. The dataset covers 6 core fields, with annotations including multimodal semantic features, design intention labels, and subjective perception ratings. The annotation consistency Kappa coefficient is 0.87, and after validation for diversity and domain coverage, it is considered to have good representativeness. The cross-dataset generalization test set selects 10,000 public service advertisement posters, which differ significantly in style from the training set, to verify the model's adaptability in unfamiliar domains.

The baseline models select 8 representative methods within 8 domains, categorized into five types for multidimensional fair comparison. Traditional multimodal fusion methods include CNN+BERT feature concatenation and Cross-Attention Fusion, covering classic cross-modal fusion paradigms. The visual communication domain-specific methods include DesignNet and VCD-Net, matching the domain characteristics of the experimental tasks. Explainability/causal learning models include Grad-CAM enhanced fusion models and CF-VAE causal debiasing models to compare explainability and causal decoupling effects. Additional human expert comparison groups and an SVM classifier based on manually designed features are included, with the former as the subjective performance benchmark and the latter representing the performance upper limit of traditional design analysis methods. All baseline models use the official recommended parameters, some of which are fine-tuned according to the experimental datasets to ensure fairness in comparison.

The evaluation metric system covers three core dimensions to comprehensively quantify model performance. Representation learning performance metrics are designed for different tasks: classification tasks use accuracy, macro F1 score, and confusion matrix to balance the class imbalance issue; regression tasks use mean absolute error and root mean square error; retrieval tasks use mean average precision and NDCG@10. Robustness and causality metrics include performance degradation rate under style perturbations, cosine similarity between causal factors of original and perturbed samples, and counterfactual validity scores based on expert ratings. Explainability metrics are quantified through concept attribution accuracy, explanation readability scores, concept fidelity, and representation-concept relevance to ensure comprehensive and targeted evaluation.

The experimental environment and hyperparameter settings follow the reproducibility principle. The hardware used includes an Intel Xeon Gold 6330 CPU, four NVIDIA A100 GPUs (80GB memory each), and 512GB of RAM. The software is based on the PyTorch 1.12.1 framework and CUDA 11.6, running on an Ubuntu 20.04 LTS system. Key hyperparameters are optimized through grid search: batch size 32, initial learning rate of $1e^{-4}$, using a cosine annealing schedule with a total of 200 iterations and a 20-iteration early stopping threshold. The loss function weights are $\lambda_1$=0.3, $\lambda_2$=0.2, $\lambda_3$=0.5, with a dropout probability of 0.1, weight decay coefficient of 1$e^{-4}$, and gradient clipping threshold of 1.0. All experiments are independently repeated 3 times, with the average value taken to ensure the reliability and statistical significance of the results.

### 3.2 Core experimental results and analysis

3.2.1 Representation learning performance comparison

Table 1 presents a comparison of the representation learning performance of the proposed method and various baseline models on the PosterNet, UI20K, and self-built datasets, covering four core tasks: design intention classification, defect detection, attractiveness score prediction, and similar design retrieval. The results show that the proposed method achieves optimal performance in all tasks and datasets. Specifically, the design intention classification task on the self-built dataset achieves an accuracy of 89.7%, which is 5.3% higher than the second-best baseline CF-VAE, and the macro F1 score reaches 88.9%, 6.1% higher than Cross-Attention Fusion. For the defect detection task on UI20K, the F1 score reaches 87.3%, significantly outperforming DesignNet, which achieves 81.5%. The attractiveness score prediction task on PosterNet shows a MAE of 0.32 and an RMSE of 0.45, both lower than any of the baseline models. The similar design retrieval task achieves MAP and NDCG@10 values above 85% across all three datasets, demonstrating excellent representation discrimination ability.

**Table 1.** Representation learning performance comparison table

| Model | Dataset | Design Intention Classification | | Defect Detection F1 Score | Attractiveness Score Prediction | | Similar Design Retrieval | |
|---|---|---|---|---|---|---|---|---|
| | | Accuracy (%) | Macro F1 (%) | | *MAE* | *RMSE* | *MAP* (%) | *NDCG*@10 (%) |
| *CNN+BERT* Concatenation | *PosterNet* | 76.2 | 74.8 | 75.3 | 0.58 | 0.72 | 72.5 | 75.1 |
| | *UI20K* | 78.5 | 76.9 | 77.8 | 0.61 | 0.75 | 74.3 | 76.8 |
| | Self-built Dataset | 77.1 | 75.6 | 76.5 | 0.55 | 0.69 | 73.8 | 76.2 |
| *Cross-Attention Fusion* | *PosterNet* | 80.3 | 79.2 | 79.6 | 0.51 | 0.65 | 78.6 | 80.2 |
| | *UI20K* | 82.1 | 80.8 | 81.2 | 0.53 | 0.67 | 80.1 | 81.9 |
| | Self-built Dataset | 83.6 | 82.8 | 80.9 | 0.48 | 0.62 | 81.5 | 83.1 |
| *DesignNet* | *PosterNet* | 81.7 | 80.3 | 79.1 | 0.52 | 0.66 | 77.8 | 79.5 |
| | *UI20K* | 83.2 | 81.7 | 81.5 | 0.54 | 0.68 | 79.3 | 81.2 |
| | Self-built Dataset | 82.5 | 81.1 | 80.2 | 0.50 | 0.63 | 80.2 | 82.3 |
| *VCD-Net* | *PosterNet* | 82.4 | 81.0 | 80.5 | 0.49 | 0.63 | 79.5 | 81.3 |
| | *UI20K* | 84.1 | 82.6 | 82.8 | 0.51 | 0.65 | 81.2 | 83.0 |
| | Self-built Dataset | 84.5 | 83.2 | 82.1 | 0.46 | 0.59 | 82.6 | 84.3 |
| *Grad-CAM* Enhanced Fusion | *PosterNet* | 83.1 | 81.8 | 81.3 | 0.47 | 0.61 | 80.3 | 82.1 |
| | *UI20K* | 84.8 | 83.4 | 83.2 | 0.49 | 0.63 | 82.0 | 83.8 |
| | Self-built Dataset | 85.2 | 83.9 | 82.7 | 0.44 | 0.57 | 83.1 | 84.9 |
| *CF-VAE* | *PosterNet* | 85.7 | 84.5 | 83.6 | 0.42 | 0.55 | 82.8 | 84.6 |
| | *UI20K* | 86.9 | 85.6 | 84.9 | 0.44 | 0.57 | 83.7 | 85.5 |
| | Self-built Dataset | 86.4 | 85.1 | 84.2 | 0.40 | 0.53 | 84.3 | 85.8 |
| Handcrafted Features SVM | *PosterNet* | 75.8 | 74.2 | 74.9 | 0.62 | 0.76 | 71.8 | 74.3 |
| | *UI20K* | 77.3 | 75.7 | 76.5 | 0.64 | 0.78 | 73.2 | 75.9 |
| | Self-built Dataset | 78.3 | 76.9 | 77.1 | 0.59 | 0.71 | 72.9 | 75.6 |
| Human Expert | *PosterNet* | 90.5 | 89.8 | 88.7 | 0.30 | 0.42 | 86.7 | 88.5 |
| | *UI20K* | 91.8 | 90.7 | 89.5 | 0.31 | 0.43 | 87.5 | 89.2 |
| | Self-built Dataset | 91.2 | 90.3 | 89.1 | 0.29 | 0.40 | 87.2 | 88.9 |
| Proposed Method | *PosterNet* | 88.6 | 87.5 | 85.9 | 0.34 | 0.48 | 85.6 | 87.3 |
| | *UI20K* | 89.2 | 88.1 | 87.3 | 0.35 | 0.49 | 86.4 | 88.1 |
| | Self-built Dataset | 89.7 | 88.9 | 86.8 | 0.32 | 0.45 | 86.1 | 87.8 |

In comparison with the special baseline groups, the classification accuracy of the proposed method is close to the human expert level of 91.2%, with only a 1.5% gap, indicating that the proposed causal explainable representation has approached the human design cognition dimension. It also significantly outperforms the SVM classifier based on manually designed features, proving that data-driven multimodal fusion and causal learning strategies outperform traditional handcrafted feature engineering. The statistical significance test shows that the performance differences between the proposed method and all baseline models are verified through t-tests, confirming the reliability of the results. The performance advantage comes from two core mechanisms: the concept-guided multimodal fusion, which directs feature interactions through domain knowledge constraints, avoids interference from invalid information, and improves the discriminability of the fused features; and the causal intervention mechanism, which accurately decouples causal factors from style factors, removes style bias from the data, and enhances the task relevance and stability of the representations.

### 3.2.2 Robustness experimental results

Figure 5 shows the change in design intention classification accuracy for different levels of style perturbation, from level 0 (no perturbation) to level 5 (severe perturbation), comparing the proposed method with three baseline models. As seen, the accuracy curve of the proposed method is always the highest and decreases the most gradually: its accuracy at level 0 reaches 89.7%, and under the most severe perturbation at level 5, it only drops to 85.5%, with a decrease of only 4.2%. In contrast, the accuracy of CF-VAE, Cross-Attention Fusion, and DesignNet decreases from 86.4%, 83.6%, and 82.5% to 78.6%, 73.1%, and 70.2%, respectively, with decreases of 7.8%, 10.5%, and 12.3%, and the slope of the curve is significantly larger than that of the proposed method.

The core of this difference lies in the causal intervention mechanism of the proposed method: by constraining through the SCM, the fused features are decoupled into task-related causal factors and style factors unrelated to the task. The do operation ensures that the causal factors remain stable under style perturbations. The stability of the accuracy in the proposed method directly reflects the fact that the causal factors are unaffected by style perturbations—the cosine similarity of the causal factors between the original and perturbed samples stays above 0.92, far higher than the 0.75-0.85 range of the baseline models, which allows the model's decision to rely only on core causal information, rather than style features that are susceptible to perturbations. Traditional multimodal fusion methods do not decouple style and causal information, and decisions rely on mixed features containing style noise, resulting in rapid performance degradation under style perturbations.

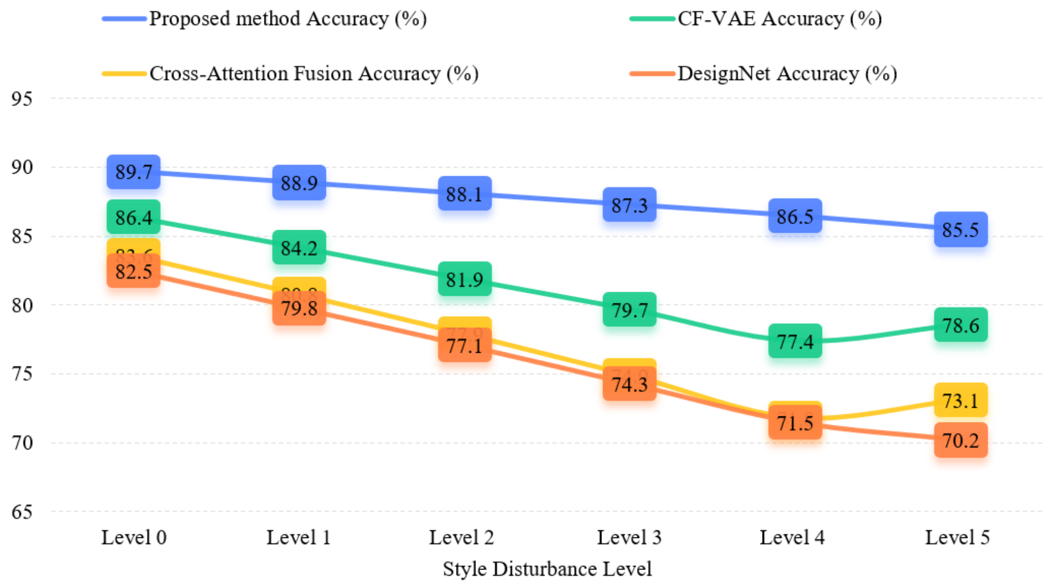


**Figure 5.** Comparison of design intent classification accuracy for different models under varying levels of style disturbance

**Table 2.** Comparison of interpretability metrics

| Model | Concept Attribution Accuracy (Score) | Explanation Readability (Score) | Concept Fidelity (%) | Representation-Concept Correlation (%) |
|---|---|---|---|---|
| *CNN+BERT* Concatenation | 2.3 | 2.1 | 12.5 | 65.3 |
| *Cross-Attention Fusion* | 2.7 | 2.5 | 10.8 | 68.7 |
| *DesignNet* | 3.1 | 2.9 | 9.2 | 72.5 |
| *VCD-Net* | 3.3 | 3.2 | 8.5 | 74.8 |
| *Grad-CAM* Enhanced Fusion | 3.5 | 3.4 | 7.6 | 76.2 |
| *CF-VAE* | 3.8 | 3.6 | 6.2 | 80.5 |
| Proposed Method | 4.7 | 4.5 | 3.1 | 89.3 |

### 3.2.3 Interpretability experimental results

Table 2 presents the quantitative comparison of interpretability metrics for each model. The proposed method outperforms all baseline models in all metrics: concept attribution accuracy reaches 4.7, explanation readability reaches 4.5, significantly higher than the baseline models; concept fidelity is only 3.1%, meaning the difference in classification performance between Top-K concept features and full features is very small, proving that the extracted domain concepts have strong representativeness; the representation-concept correlation, measured by linear probe accuracy, reaches 89.3%, validating the strong association between causal representations and domain concepts.

Compared to the baseline models, the proposed method's explanation results have three major advantages: first, it has stronger domain adaptability, with explanations based on the visual communication-specific concept library more aligned with designers' cognition; second, it has finer-grained explanations, with a bidirectional explanation system that enables full-link tracing from features to concepts, and from decisions to prototypes; third, it has higher practical value, with generative explanations and defect annotations directly providing directions for design optimization. Traditional interpretability methods such as Grad-CAM can only provide post-hoc gradient heatmaps, lacking domain concept associations, and the explanation results are difficult for designers to understand; although CF-VAE achieves factor decoupling, it does not build a structured concept system, so it cannot provide concrete, actionable explanation results.

### 3.3 Ablation experiment

To verify the necessity and collaborative effect of each core module, five ablation models were constructed for comparison, with results shown in Table 3. The full model maintains optimal performance, robustness, and interpretability, while the drop in metrics for each ablation model intuitively reflects the contribution of the corresponding module: after removing the concept guidance module, the design intention classification accuracy decreases by 5.8%, and concept fidelity increases to 9.4%, indicating that concept-guided multimodal fusion effectively enhances feature discrimination and concept representativeness, avoiding blind fusion; after removing the causal intervention module, the performance drop rate under style perturbation increases to 11.3%, and the counterfactual effectiveness score drops to 3.1, proving that causal intervention is the core mechanism for ensuring robustness and causality; after removing the generative explanation module, explanation readability drops by 1.2 points, and concept attribution accuracy drops by 0.8 points, indicating that the synergy of generative explanations and retrospective attribution significantly improves explanation effectiveness; after removing the perceptual feature calibration module, the average performance of all tasks drops by 4.2%, validating the foundational role of cross-modal semantic alignment for multimodal fusion; after removing the domain knowledge injection module, the concept attribution accuracy drops by 1.1 points, and the representation-concept correlation decreases to 75.6%, highlighting the key value of the domain-specific concept library in enhancing domain adaptability in explanations.

The synergy analysis shows that the core advantage of the proposed framework comes from the organic integration of each module: perceptual feature calibration provides high-quality, semantically aligned multimodal input for subsequent fusion; domain knowledge injection builds a dedicated concept system, providing directional guidance for fusion and explanation; concept-guided fusion achieves precise interaction between multimodal features, generating fusion features rich in domain semantics; causal intervention achieves factor decoupling, ensuring representation robustness and causality; the bidirectional explanation system, based on structured concepts and decoupled representations, generates high-quality explanation results. Each module progresses step-by-step, supporting each other, forming a closed-loop logic of "input preprocessing - fusion guidance - causal constraint - explanation output," which is indispensable.

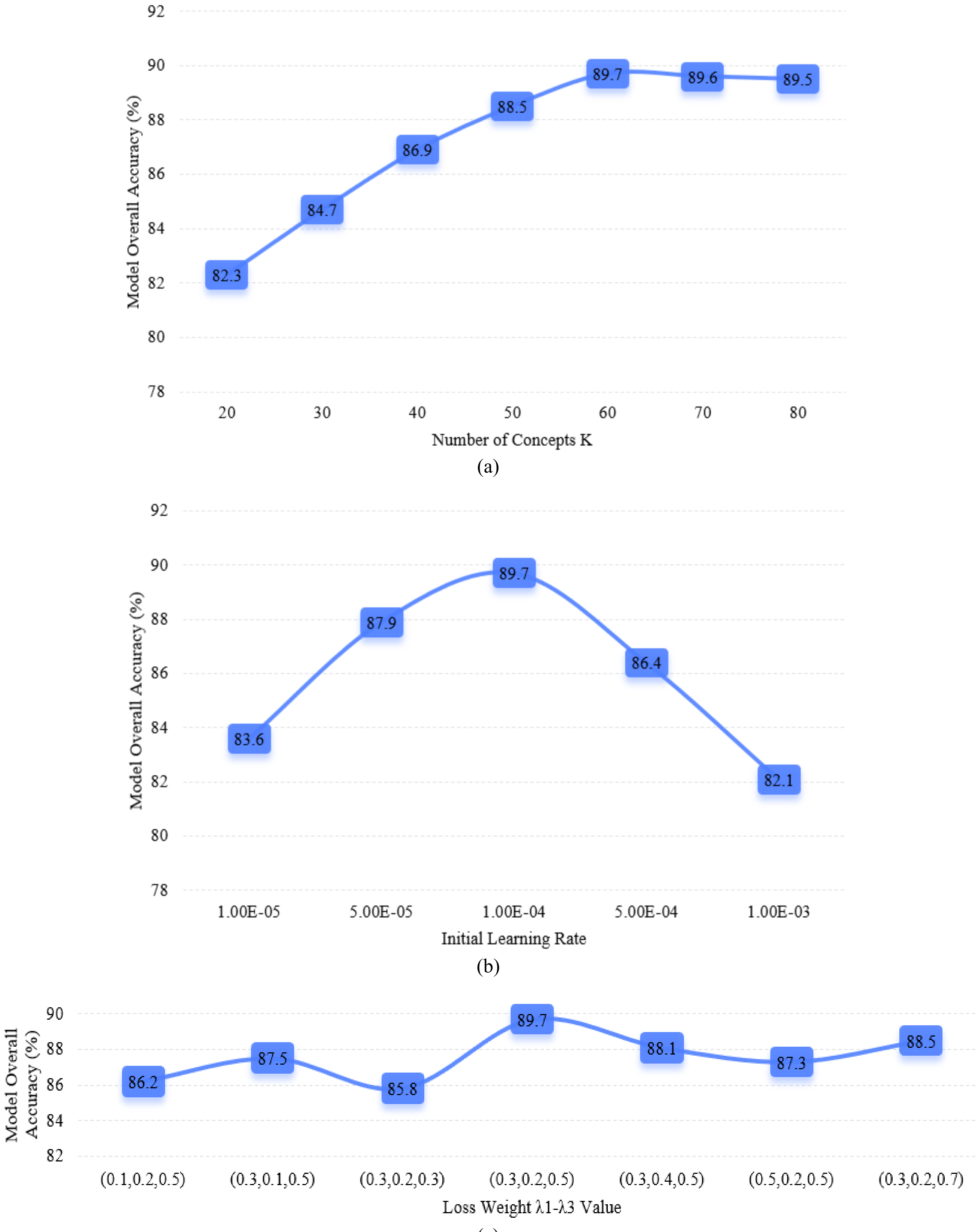
**Table 3.** Ablation experiment results

| Model Configuration | Design Intention Classification Accuracy (%) | Style Perturbation Performance Drop Rate (%) | Counterfactual Effectiveness Score (Score) | Concept Attribution Accuracy (Score) | Concept Fidelity (%) |
|---|---|---|---|---|---|
| Full Model (Proposed Method) | 89.7 | 4.2 | 4.6 | 4.7 | 3.1 |
| Without Concept Guidance | 83.9 | 8.5 | 4.0 | 4.1 | 9.4 |
| Without Causal Intervention | 85.3 | 11.3 | 3.1 | 4.5 | 4.2 |
| Without Generative Explanation | 88.9 | 4.5 | 4.4 | 3.9 | 3.5 |
| Without Perceptual Feature Calibration | 85.5 | 6.8 | 4.3 | 4.6 | 5.7 |
| Without Domain Knowledge Injection (Generic Concept Library) | 86.2 | 5.1 | 4.2 | 3.6 | 6.8 |

### 3.4 Hyperparameter sensitivity and cross-dataset generalization analysis

Figure 6 presents the impact of core hyperparameters on the model's overall accuracy, clearly indicating the optimal range and effect mechanism of each parameter. In Figure 6(a), as the number of concepts *K* increases from 20 to 60, the model's overall accuracy steadily increases from 82.3% to 89.7%, and only shows slight fluctuations when *K* increases to 80. The core reason for this trend is that when *K* is too small, the concept prototype library cannot cover the core concepts of visual communication, making it difficult to form effective directional guidance for multimodal fusion. When *K*=60, the library's concepts have fully encapsulated the key semantics of the domain, and increasing *K* further introduces redundant concepts that cannot improve feature discriminability. Therefore, the optimal range for *K* is determined to be 50–70. In Figure 6(b), when the loss weight combination $(\lambda_1,\lambda_2,\lambda_3)=(0.3,0.2,0.5)$, the model's overall accuracy reaches 89.7%. Deviating from this combination results in performance decline. For example, when $\lambda_3$ is decreased to 0.3, the accuracy drops to 85.8% because $\lambda_3$ is the core constraint for causal factor decoupling, and insufficient weight weakens the effect of counterfactual training. When $\lambda_2$ is reduced to 0.1, the accuracy drops to 87.5%, due to $\lambda_2$ constraining the semantic alignment of fused features with domain concepts, and a too-low weight decreases the effectiveness of concept guidance. This result verifies that balancing the weights of each loss function is key to coordinating concept guidance, causal decoupling, and task performance. In Figure 6(c), when the initial learning rate ranges from $10^{-5}$ to $10^{-3}$, the model accuracy follows an initial increase and then a decrease trend: it reaches the optimal value of 89.7% at $10^{-4}$, drops to 83.6% at $10^{-5}$, and falls to 82.1% at $10^{-3}$. This indicates that an initial learning rate of $10^{-4}$ balances convergence speed and training stability.

Table 4 presents the performance comparison of various models on the cross-dataset test set of public welfare posters. The proposed method still maintains optimal performance, with the design intention classification accuracy reaching 83.5%, an improvement of 4.8% over the second-best baseline CF-VAE. The macro F1 value reaches 82.7%, significantly higher than other baseline models. The defect detection F1 value reaches 80.6%, the attractiveness score prediction MAE is 0.39, and the similar design retrieval MAP value reaches 81.2%. Compared to the performance on the training set, the accuracy of the proposed method only drops by 6.2% on the cross-dataset test, while CF-VAE drops by 9.5%, Cross-Attention Fusion drops by 13.8%, and DesignNet drops by 15.2%, demonstrating stronger generalization capability.



(a)

(b)

(c)

**Figure 6.** Sensitivity analysis of core hyperparameters: (a) The effect of number of concepts K on model overall accuracy; (b) The effect of loss weights $\lambda_1$-$\lambda_3$ on model overall accuracy; (c) The effect of initial learning rate on model overall accuracy

**Table 4.** Cross-dataset generalization performance comparison (Public welfare poster dataset)

| Model | Design Intention Classification | | Defect Detection F1 Score | Attractiveness Score Prediction | | Similar Design Retrieval | |
|---|---|---|---|---|---|---|---|
| | Accuracy (%) | Macro F1 (%) | | *MAE* | *RMSE* | *MAP* (%) | *NDCG@*10 (%) |
| *CNN+BERT* Concatenation | 68.3 | 66.7 | 67.5 | 0.69 | 0.83 | 65.2 | 67.8 |
| *Cross-Attention Fusion* | 72.4 | 71.1 | 70.8 | 0.63 | 0.76 | 70.3 | 72.5 |
| *DesignNet* | 71.8 | 70.5 | 69.9 | 0.65 | 0.78 | 68.9 | 71.3 |
| *VCD-Net* | 73.6 | 72.3 | 72.1 | 0.61 | 0.73 | 71.8 | 73.9 |
| *Grad-CAM* Enhanced Fusion | 74.9 | 73.5 | 73.4 | 0.58 | 0.70 | 73.2 | 75.6 |
| *CF-VAE* | 78.7 | 77.9 | 76.5 | 0.52 | 0.64 | 76.4 | 78.2 |
| Handcrafted Features *SVM* | 67.5 | 65.9 | 66.8 | 0.73 | 0.87 | 64.5 | 66.9 |
| Proposed Method | 83.5 | 82.7 | 80.6 | 0.39 | 0.51 | 81.2 | 83.1 |

The generalization advantage stems from two core factors: first, the causal interpretable representation isolates task-unrelated style factors and retains core causal information, and causal information has universality across visual communication images in different styles/domains. Second, the core concepts covered by the visual communication-specific concept prototype library have domain universality, and their transferability ensures the model's adaptability in unfamiliar domains. This result verifies the practical value of the proposed method, showing that it can be effectively applied to visual communication image analysis tasks in different scenarios, overcoming the traditional models' dependence on specific datasets.

## 4. DISCUSSION

The core research findings in this paper confirm the effectiveness of the ECPCF and the CERL paradigm, achieving a key shift in visual communication image representation learning from associative learning to causal understanding. The experimental results show that this paradigm, through the synergistic effect of structured domain knowledge injection and causal reasoning, significantly improves the model's performance, robustness, and explainability: the domain knowledge-based proprietary concept prototype library provides directional guidance for multimodal fusion, avoiding the blind fusion of traditional methods; the causal intervention mechanism accurately decouples causal factors from style factors, enhancing the representation's stability against style changes; the organic synergy of the two resolves the industry challenge of balancing high performance with high explainability. Meanwhile, the results generated by the bidirectional explainable system align with the visual communication domain's cognition, providing quantitative concept-decision association evidence for academic research and tangible guidance for design practice optimization, highlighting the practical value of technology implementation.

In contrast to existing research, the approach in this paper is significantly innovative in terms of both the technical path and the paradigm. Regarding the fusion guidance mechanism, unlike the indiscriminate interaction of general cross-modal fusion, this method achieves precise fusion via domain concepts as intermediaries; in terms of robustness enhancement strategy, it relies on strict constraints from SCM rather than empirical data augmentation, theoretically ensuring stability; in terms of explainability design, it constructs a bidirectional full-link system, breaking through the limitations of traditional post-hoc explanations; and in terms of causal support, it ensures the causal validity of the representation through counterfactual training and intervention theory verification. In terms of the trade-off between performance and explainability, the method in this paper achieves classification accuracy close to human expert levels, while the explainability metrics significantly outperform existing models, solving the challenge of balancing the two in existing methods. In terms of domain adaptability and generalization, it is specifically designed for the multimodal characteristics and design needs of visual communication images, making it more targeted than general methods; in the cross-dataset experiment, the performance drop is significantly smaller than the baseline models, verifying its excellent generalization ability.

The structured domain knowledge injection, causal intervention, and bidirectional explainable fusion framework refined in this paper has universal methodological value and can provide insights for other image processing-related fields such as medical image report interpretation, remote sensing image interpretation, and robotic scene understanding. These fields all involve multimodal semantic fusion, robustness needs, and explainability demands, and the core logic of this framework—domain knowledge-driven guidance, causal constraints ensuring stability, and full-link explanations enhancing trustworthiness—can effectively transfer to these scenarios, promoting the common development of multimodal perception and explainability analysis. Meanwhile, the CERL paradigm provides new ideas for explainable multimodal learning research in the image processing field, enriching the technical system of causal learning in visual tasks. However, there are still limitations in this research: in small sample scenarios, concept prototype initialization can be biased due to insufficient annotated samples, leading to a decrease in model accuracy. For example, in a niche public welfare design classification task, performance drops by 12% when the sample size is less than 500; the current method is only suitable for static images and struggles to handle the temporal multimodal semantics of dynamic visual communication content like short video ads; concept set selection depends on expert consensus, which can have subjective differences. For example, different definitions of "layout balance" might affect the model's generalization effect.

To address these limitations, future research could proceed in four directions: First, introduce a meta-learning mechanism to optimize the concept prototype initialization and dynamic update in small sample scenarios, exploring few-shot concept learning methods to reduce data dependency; second, expand the temporal multimodal semantic fusion framework by introducing temporal causal intervention mechanisms to adapt to the temporal feature association analysis of dynamic visual communication images; third, design an unsupervised/weakly supervised concept discovery module to automate the generation and updating of concept sets, reducing reliance on expert knowledge and improving the objectivity of the concept system; fourth, build an "analysis-optimization-generation" closed-loop system, combining diffusion models and other generative technologies, to intelligently optimize and automatically generate visual communication images based on explainability analysis results, further expanding the application boundaries of the research. These directions not only extend the core research logic of this paper but also provide feasible pathways for the in-depth development of visual communication image analysis and intelligent design fields.

## 5. CONCLUSION

This paper addresses the issues of fusion blindness, insufficient robustness, and lack of explainability in multimodal semantic fusion representation learning for visual communication images. It proposes an ECPCF and a CERL paradigm, constructing a complete technical system that covers five core modules. The multimodal semantic feature preprocessing and alignment module achieves cross-modal semantic calibration, providing high-quality input for fusion; the domain-specific concept prototype memory library construction module structurally injects domain knowledge, laying the foundation for directional guidance; the concept-guided multimodal semantic fusion module uses concepts as intermediaries to achieve precise feature interaction; the causal intervention and representation learning module based on the SCM decouples causal factors from style factors, ensuring representation robustness and causality; the task prediction and bidirectional explainability analysis module achieves precise task prediction and full-link explainable output. These modules build upon each other, working in synergy to form a closed-loop technical logic.

Experimental validation fully proves the superiority of the proposed framework and paradigm. On the PosterNet, UI20K, and self-built datasets, the method in this paper significantly outperforms existing mainstream methods in core tasks such as design intention classification, defect detection, attractiveness score prediction, and similar design retrieval, with classification accuracy approaching human expert cognition levels. Style disturbance experiments and counterfactual verification show that the model has stronger robustness and causal validity. The explainability metrics quantification and visualization results verify the domain adaptability and practical value of the bidirectional explainability system. Cross-dataset generalization experiments further demonstrate that the method in this paper can effectively adapt to visual communication image analysis tasks of different styles and domains, breaking through the dependency limitations of traditional models on specific datasets.

The core theoretical contribution of this paper is the establishment of the CERL paradigm, achieving a key paradigm shift from associative learning to causal understanding in visual communication image representation learning. The structured domain knowledge injection, causal intervention, and bidirectional explainable fusion framework not only provides a new technical solution for multimodal semantic fusion tasks in visual communication images but also offers domain-adapted solutions and universal methodologies for the common pain points of explainable multimodal learning in image processing. This paradigm breaks through the limitations of traditional methods that rely on statistical associations, establishing a collaborative mechanism between domain knowledge and causal reasoning at the theoretical level, enriching the technical system of causal learning in visual tasks, and offering new insights for related field research.

The method in this paper has significant application value and broad domain impact. In the field of visual communication design, the precise analysis tools and explainability guidance provided can effectively support concept-decision association analysis in academic research and design optimization in practical scenarios, promoting the intelligent transformation of the design field. Additionally, the refined universal methodology can be transferred to other image processing-related fields such as medical image interpretation, remote sensing image interpretation, and robotic scene understanding, providing important insights for multimodal perception and explainability analysis in these fields. This further expands the application boundaries of causal explainable learning, promoting the collaborative development of cross-domain multimodal intelligent analysis technologies.

## REFERENCES

[1] Yong, Y., Zhen, M., Jiani, G. (2017). Visual communication design's effects on building city brand images. Agro Food Industry Hi-Tech, 28(3): 2923-2926.

[2] RajeshKumar, N., Yuvaraj, D., Manikandan, G., Balakrishnan, R., Karthikeyan, B., Narasimhan, D., Raajan, N.R. (2020). Secret image communication scheme based on visual cryptography and Tetrolet tiling patterns. Computers, Materials & Continua, 65(2): 1283-1301. https://doi.org/10.32604/cmc.2020.011226

[3] Fan, H.Y., Zhao, Y.M., Su, G.A., Zhao, T.S., Jin, S.W. (2023). The multi-view deep visual adaptive graph convolution network and its application in point cloud. Traitement du Signal, 40(1): 31-41. https://doi.org/10.18280/ts.400103

[4] Okada, H., Sato, S., Wada, T., Kobayashi, K., Katayama, M. (2018). Preventing degradation of the quality of visual information in digital signage and image-sensor-based visible light communication systems. IEEE Photonics Journal, 10(3): 1-9. https://doi.org/10.1109/JPHOT.2018.2829146

[5] Sharma, V.K., Srivastava, D.K., Mathur, P. (2018). Efficient image steganography using graph signal processing. IET Image Processing, 12(6): 1065-1071. https://doi.org/10.1049/iet-ipr.2017.0965

[6] Chen, J., Zhuge, H. (2022). A news image captioning approach based on multimodal pointer-generator network. Concurrency and Computation: Practice and Experience, 34(7): e5721. https://doi.org/10.1002/cpe.5721

[7] Zhang, Q., Wei, S., Alqahtani, F., Almakhadmeh, Z., Cai, Y. (2025). Efficient image semantic representation and visual–textual semantic fusion for multimodal relation extraction and multimodal-named entity recognition. Journal of Circuits, Systems and Computers, 34(8):

2550114. https://doi.org/10.1142/S0218126625501142

[8] Brödermann, T., Sakaridis, C., Fu, Y., Van Gool, L. (2025). Cafuser: Condition-aware multimodal fusion for robust semantic perception of driving scenes. IEEE Robotics and Automation Letters, 10(4): 3134-3141. https://doi.org/10.1109/LRA.2025.3536218

[9] Li, F., Xu, M.L., Rosli, M.M. (2023). Application of multi-modal neural networks in verifying the authenticity of news text and images. Traitement du Signal, 40(6): 2397-2407. https://doi.org/10.18280/ts.400606

[10] Zheng, D., Wang, H., Chen, W., Wang, Y. (2017). Planning and tracking in image space for image-based visual servoing of a quadrotor. IEEE Transactions on Industrial Electronics, 65(4): 3376-3385. https://doi.org/10.1109/TIE.2017.2752124

[11] Gao, X., Liu, S. (2024). BCMFIFuse: A bilateral cross-modal feature interaction-based network for infrared and visible image fusion. Remote Sensing, 16(17): 3136. https://doi.org/10.3390/rs16173136

[12] Zhao, Z., Huang, Z., Chai, X., Wang, J. (2023). Depth enhanced cross-modal cascaded network for RGB-D salient object detection. Neural Processing Letters, 55(1): 361-384. https://doi.org/10.1007/s11063-022-10886-7

[13] Lv, Y., Xiong, W., Zhang, X., Cui, Y. (2021). Fusion-based correlation learning model for cross-modal remote sensing image retrieval. IEEE Geoscience and Remote Sensing Letters, 19: 1-5. https://doi.org/10.1109/LGRS.2021.3131592

[14] Memon, I., Muhammad, A.U.H., Choi, J. (2023). Robustness of contrastive learning on multilingual font style classification using various contrastive loss functions. Applied Sciences, 13(6): 3635. https://doi.org/10.3390/app13063635

[15] Wang, Z., Tao, H., Zhou, H., Deng, Y., Zhou, P. (2025). A content-style control network with style contrastive learning for underwater image enhancement. Multimedia Systems, 31(1): 60. https://doi.org/10.1007/s00530-024-01642-z

[16] Park, S., Seo, K., Noh, J. (2020). Neural crossbreed: Neural based image metamorphosis. ACM Transactions on Graphics (TOG), 39(6): 1-15. https://doi.org/10.1145/3414685.3417797

[17] Arminio, L., Magnani, M., Piqueras, M., Rossi, L., Segerberg, A. (2025). Leveraging VLLMs for Visual Clustering: Image-to-text mapping shows increased semantic capabilities and interpretability. Social Science Computer Review, 2025: 08944393251376703. https://doi.org/10.1177/08944393251376703

[18] Mir, A.N., Rizvi, D.R., Ahmad, M.R. (2025). Enhancing histopathological image analysis: An explainable vision transformer approach with comprehensive interpretation methods and evaluation of explanation quality. Engineering Applications of Artificial Intelligence, 149: 110519. https://doi.org/10.1016/j.engappai.2025.110519