A Pre-Standardization Multimodal ICT Framework for Agricultural Field-Note Visualization Using CLIP and AI Sketching

Hui Li[1,*] and Yu Chen[2]

[1]*Shandong Management University, Jinan, Shandong 250357, China*
[2]*Shandong Women's University, Jinan, Shandong 250357, China*
E-mail: leehui-1@163.com; 14438119920195@sdmu.edu.cn
[*]*Corresponding Author*

Received 15 May 2026; Accepted 09 August 2026

Abstract

Free-text agricultural field notes are difficult to exchange across farm information systems because crop, symptom, operation, and field-condition descriptions are heterogeneous and weakly visualized. This study develops a pre-standardization multimodal ICT framework that converts field notes into interoperable sketch-style visual diaries through Contrastive Language-Image Pre-training (CLIP)-guided semantic alignment and controllable diffusion generation. Agricultural entities and state descriptions are encoded by a CLIP text encoder, injected into the latent generation space through cross-attention, and constrained by edge-guided denoising, layout metadata, and semantic verification. Experiments on public agricultural image-text datasets show Fréchet Inception Distance (FID) below 32, Structural Similarity Index (SSIM) above 0.82, CLIP Score of 0.85 for 200-word notes, top-five retrieval

Journal of ICT Standardization, Vol. 14_4, 603–628.
doi: 10.13052/jicts2245-800X.1444
© 2026 River Publishers

recall (Recall@5) of 84%, and Learned Perceptual Image Patch Similarity (LPIPS) of 0.082. Results suggest that semantic metadata, interface parameters, quality metrics, and sample visual outputs can support interoperable AI-assisted agricultural documentation.

Keywords: ICT standardization, multimodal AI, agricultural field notes, interoperability, semantic metadata, visual diary.

1 Introduction

Agricultural practice increasingly depends on continuous observation, digital records, and timely exchange of field information. Farmers' daily notes remain an important source of first-hand knowledge about crop growth, disease symptoms, water conditions, management actions, and local environmental changes (Basir et al., 2024; Ferreira, 2026; Rabbi et al., 2023; Singh et al., 2024; Zhao, 2024). However, these notes are often recorded as short, informal, and highly contextual free text. When such records are transferred across farm management systems, extension-service platforms, mobile applications, and AI-based decision-support tools, their meaning may be weakened because the text lacks standardized semantic tags, visual references, and machine-readable metadata. Transforming field notes into visual diaries can improve readability and sharing, but such transformation must also consider interoperability, traceability, and evaluation consistency rather than image generation quality alone (Dhal et al., 2024; Namyenya et al., 2022, 2024; Utomo et al., 2025).

This problem is aligned with the broader agenda of ICT standardization. As digital agriculture converges with artificial intelligence, remote sensing, mobile services, and platform-based extension systems (Garg and Alam, 2023), a field-note visualization method should not be designed as an isolated text-to-image tool. It should provide a pre-standardization technical framework that clarifies input elements, semantic entities, model-interface requirements, output metadata, and quality verification indicators. Such a framework can support cross-platform exchange of agricultural records while preserving farmers' localized knowledge. In this context, multimodal AI is useful not only as a generative technology but also as a mechanism for structuring heterogeneous agricultural observations into reusable ICT artifacts.

This standardization framing also follows prior JICTS discussions on AI, Big Data, Internet of Things (IoT), and sustainable development,

which emphasize interoperable architectures, reusable data structures, and cross-sector ICT governance (Sivabalan et al., 2013; Walshe et al., 2020).

Farmers' field notes are difficult to standardize because they are abstract, personalized, and regionally variable. Expressions such as "yellowing leaves," "insufficient soil moisture," and "poor pollination" may describe different agronomic states depending on crop type, growth stage, climate, or local terminology (Bowen and Morris, 2024; Rodríguez et al., 2025; Sarku and Ayamga, 2025). A general visual generation system may interpret yellowing leaves as severe disease or complete wilting, although the intended observation may refer to early nitrogen deficiency. Similarly, descriptions of moderate spacing, flooded ground, or weak flowering may be visually exaggerated or structurally distorted when implicit agronomic constraints are not represented (Aryal et al., 2026; Dzreke, 2025; Mulungu et al., 2025; Wulandari et al., 2025). These limitations reduce the practical value of visual diaries as agricultural records and hinder the creation of reusable field-note data models.

Traditional agricultural visualization relies on photographs, manual sketches, or template-based diagrams. Photographs are intuitive but cannot fully represent non-visual information embedded in field notes, such as perceived soil texture, temperature trends, or causal explanations. Hand-drawn sketches can simplify visual complexity but require drawing skills and domain expertise (Pearce, 2021; Tessier et al., 2021). Earlier computational approaches based on rule parsing and image stitching improved automation but remained dependent on predefined templates, limiting their ability to represent abnormal conditions, rare varieties, or local expressions (Nismi Mol and Santosh Kumar, 2023; Wang and Wang, 2022). Generative adversarial networks and diffusion-based image synthesis have advanced visual realism, yet agricultural applications require stronger control over semantic alignment, structural consistency, and metadata compatibility (Croitoru et al., 2023; Rizvi et al., 2021).

Recent vision-language models provide a stronger basis for semantic alignment. Contrastive Language-Image Pre-training (CLIP) can project text and images into a shared representation space and has been applied to cross-modal retrieval, visual-language matching, and quality assessment (Peng, 2025; Shin et al., 2022; Xing et al., 2024; Zeng et al., 2023). Diffusion models, when combined with geometric priors such as edges or depth maps, offer controllable generation capacity (Barbu, 2023; Liu et al., 2024; Xu et al., 2022). Nevertheless, many existing methods treat semantic alignment and

structural control as separate modules. In agricultural scenarios, the generated visual content must answer both what is described and how it should be structurally represented. A field-note visualization framework should therefore unify semantic encoding, latent-space condition injection, edge-guided generation, post-processing, layout integration, and quality verification within a coherent and potentially standardizable workflow.

This study proposes a pre-standardization multimodal ICT framework for transforming agricultural field notes into sketch-style visual diaries. The technical core remains a CLIP-guided semantic alignment and AI sketching method, but the contribution is reframed for ICT standardization by emphasizing semantic data elements, interoperable processing layers, model-interface consistency, and output-verification indicators. The study makes three contributions. First, it defines a CLIP-guided semantic injection mechanism that encodes agricultural entities, state descriptions, and contextual relationships as conditional vectors for diffusion generation. Second, it designs an edge-guided denoising strategy that preserves crop morphology, field layout, and line-style consistency. Third, it integrates post-processing, layout standardization, and CLIP-based consistency verification to form a visual diary output suitable for cross-platform agricultural documentation. The objective is not to propose a formal standard, but to provide a technical basis for future standardization of AI-assisted agricultural field-note visualization.

2 Pre-Standardization Multimodal Architecture for Field-Note Visualization

Figure 1 presents the proposed architecture as a layered ICT workflow. The field-note semantic parsing layer normalizes agricultural text, extracts entities and state descriptions, and produces a global semantic vector through the CLIP encoder. The cross-modal conditioning layer injects semantic vectors into the latent space of the generation model using cross-attention and multi-head alignment. The controllable sketch generation layer incorporates edge-structure guidance during diffusion denoising to maintain line logic and crop morphology. The post-processing and layout integration layer performs line smoothing, contrast enhancement, semantic consistency checking, and visual diary arrangement. From a standardization perspective, these layers correspond to input normalization, semantic metadata generation, model-interface specification, output-format control, and quality verification.

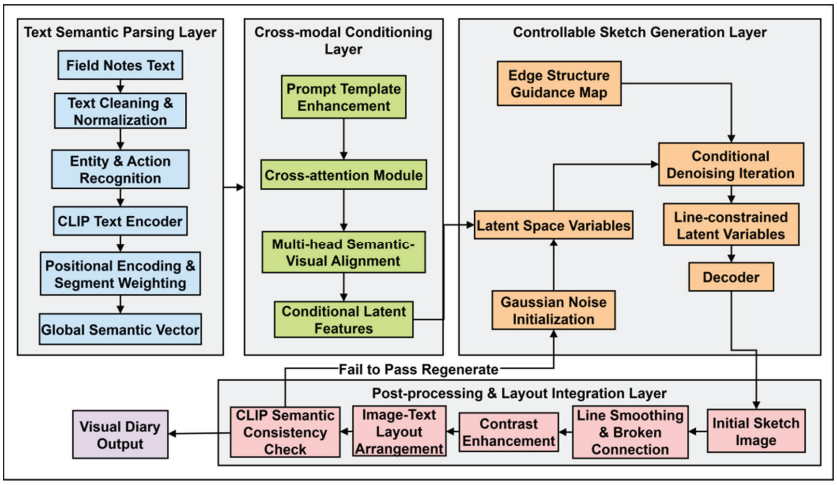


Figure 1 Semantic alignment and sketch generation architecture from agricultural text to visual diary.

2.1 CLIP-Based Semantic Embedding and Agricultural Field-Note Normalization

The field-note text is first cleaned and normalized to remove unrelated characters, repeated symbols, and incomplete fragments. The text sequence is segmented into fixed-length units while retaining chronological order and contextual continuity. Agricultural entities and action descriptions are then identified using a combination of named-entity recognition, a domain dictionary, and rule matching. The extracted entities include crop names, disease symptoms, growth stages, environmental conditions, agricultural operations, and spatial descriptions. These annotations are positionally encoded with the original text to support semantic traceability and to reduce ambiguity when the same term is used differently across regional contexts.

Each text segment is input into a pre-trained CLIP text encoder, which maps discrete tokens into a high-dimensional shared semantic space. Vector normalization is used to stabilize similarity calculations across segments. The encoded sequence is processed by a Transformer structure that captures cross-word dependencies and contextual relations. The hidden-state sequence is then aggregated to obtain segment-level semantic vectors representing crop type, growth status, operation context, and field environment. This process can be interpreted as a candidate semantic metadata layer for future agricultural ICT specifications.

To improve sensitivity to agricultural actions and state changes, a conditional weighting function reweights the encoder output according to entity category and action type. The conditional weighting formula is defined as follows:

$$\tilde{s} = \sum_{i=1}^{L} \alpha_i h_i, \quad \alpha_i = \text{softmax}(g(e_i, a_i)) \tag{1}$$

where $\tilde{s}$ is the conditionally weighted semantic vector, h_i is the i-th hidden-state output by the encoder, L is the text-segment length, and α_i is the normalized weight generated from the mapping of entity category e_i and action type a_i. The Softmax operation preserves the integrity of segment information while assigning greater weight to high-value agricultural terms.

The contextual dependency of the text sequence is enhanced through positional encoding. Sine and cosine functions introduce temporal and spatial offsets to each word vector, allowing the semantic representation to retain relations among agricultural events, entities, and positions.

$$\text{PE}(\text{pos}, 2i) = \sin\left(\frac{\text{pos}}{10000^{2i/d_{\text{model}}}}\right) \tag{2}$$

$$\text{PE}(\text{pos}, 2i+1) = \cos\left(\frac{\text{pos}}{10000^{2i/d_{\text{model}}}}\right) \tag{3}$$

where pos denotes the token position, i denotes the embedding dimension index, and d_{model} denotes the embedding dimension. After segment-level vectors are generated, vector reduction and feature fusion are performed to produce a global representation of the field note. Weighted pooling is used to emphasize segments with high entity coverage or action frequency:

$$S = \sum_{j=1}^{m} \beta_j \tilde{s}_j \tag{4}$$

where S is the global semantic vector, m is the number of segments, $\tilde{s}_j$ is the weighted semantic vector of the jth segment, and β_j is the segment-importance coefficient. This global vector preserves multi-level agricultural information while reducing redundant textual noise. In an ICT-standardization context, S functions as a compact semantic representation that can be associated with metadata fields such as crop entity, growth stage, symptom type, field condition, timestamp, and observation source.

Figure 2 illustrates the hidden-state matrix generated by the CLIP text encoder for a representative note describing corn leaves, slight yellowing,

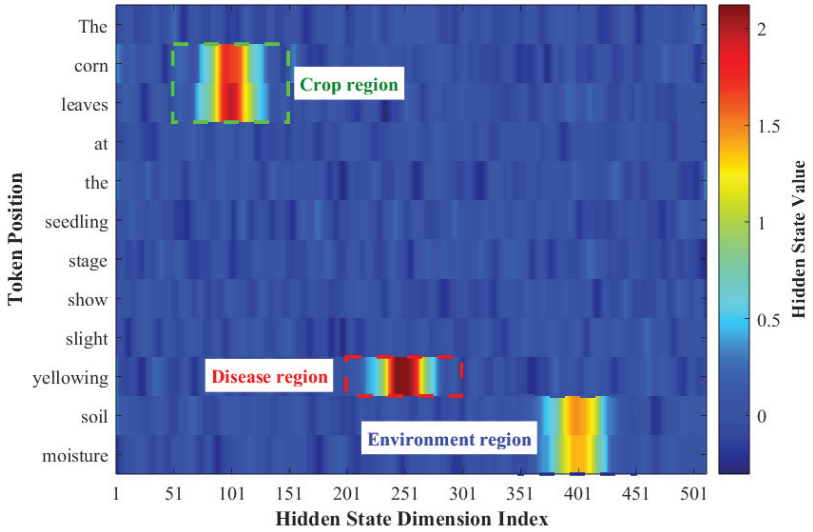


Figure 2 Schematic diagram of the hidden-state extraction process of the CLIP text encoder.

and insufficient soil moisture. The activation pattern shows that crop entities, disease-related terms, and environmental descriptions form distinguishable semantic regions. The disease-related expression obtains a stronger response than the crop and environment regions, indicating that the encoder can capture high-salience agricultural conditions. This feature separation provides a basis for conditional weighting and supports the construction of structured semantic metadata from free-text field notes.

2.2 Cross-Modal Semantic Alignment and Latent-Space Condition Injection

The text feature vectors are introduced into a cross-attention module and fused with latent visual features of the image generation model. The latent space is represented by a high-dimensional tensor in which each position carries local visual semantic information. Cross-attention adjusts the contribution of text semantics to latent visual features through query, key, and value mappings, enabling high-priority agricultural entities to guide spatial generation. The mapping matrices are updated during training according to attention scores, maintaining alignment between textual semantics and visual structures.

For each latent spatial location, a query vector is generated from visual features, while key and value vectors are generated from text features.

The scaled dot product between query and key is normalized by Softmax and then applied to the value vector. The operation is expressed as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{5}$$

where Q is the query matrix from latent visual features, K is the key matrix from text feature vectors, V is the value matrix, and d_k denotes the dimension of the key vectors. This computation distributes text semantics to corresponding positions in latent space and strengthens semantic-spatial consistency.

A prompt-template enhancement module is introduced to improve responsiveness to agricultural details. The module assigns additional weight to domain keywords by attaching adjustable prompt embeddings to the semantic vector. The enhanced text vector is calculated as follows:

$$T' = T + \gamma P \tag{6}$$

where T' is the enhanced text vector, T is the original text feature vector, P is the prompt embedding vector, and γ is the adjustment coefficient controlling prompt contribution. This formula increases the priority of key agricultural semantics in the latent space without changing the original field-note content.

The enhanced vector is further processed by multi-head cross-attention. Features from different latent subspaces are computed in parallel and projected back to a unified representation:

$$\begin{aligned}\text{MultiHead}(Q, K, V) &= \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O, \\ \text{head}_i &= \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)\end{aligned} \tag{7}$$

where W_i^Q, W_i^K, W_i^V, and W^O are learnable projection matrices, h is the number of attention heads, and head_i denotes the output of the i-th attention head. The multi-head structure supports semantic-visual alignment across crop morphology, disease symptoms, environmental cues, and layout features.

After conditional injection, the latent-space features are stabilized through normalization and residual connections:

$$Z' = \text{Norm}(Z + \text{Attention}(Z, T', T')) \tag{8}$$

where Z is the input latent feature tensor and Z' is the feature tensor after semantic injection. The residual design preserves local visual structure while

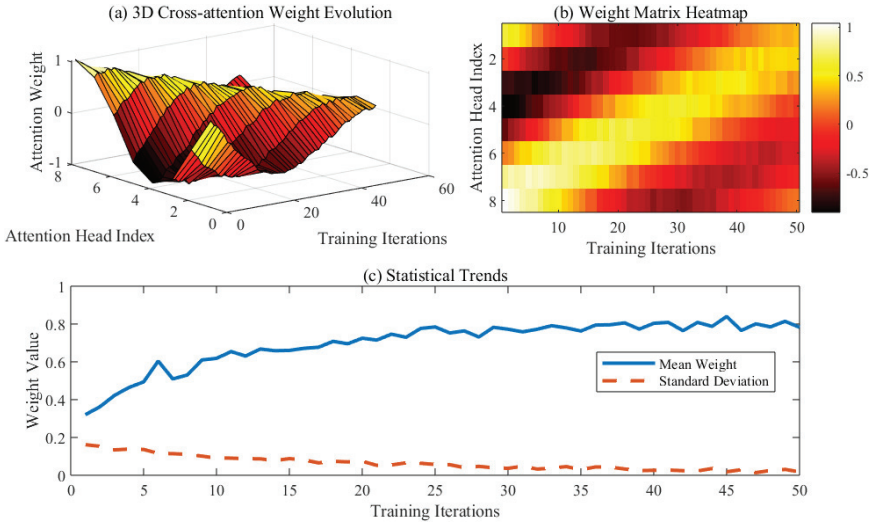


Figure 3 Dynamic evolution of cross-attention weights during training.

incorporating domain-specific semantic constraints. In a standardization-oriented interpretation, this layer functions as a model-interface component that specifies how semantic metadata are transmitted to generative modules.

Figure 3 shows the dynamic evolution of cross-attention weights. The three-dimensional trajectory indicates that the model adjusts attention across training iterations and attention heads. The heatmap reveals higher activation in selected stages and head combinations, and the statistical trend suggests that the mean attention weight increases while the standard deviation gradually decreases. This convergence pattern indicates that the model progressively concentrates on stable semantic cues. For ICT standardization, such weight behavior also suggests that semantic-alignment quality can be monitored through reproducible attention diagnostics.

2.3 Edge-Guided Denoising in the Controllable Sketch Diffusion Model

The aligned semantic vectors are input into the latent diffusion model, and latent variables are initialized with Gaussian noise (Wang et al., 2025; Zhang et al., 2025). The denoising process iteratively updates the latent variables using semantic conditions until they approximate the target image representation. At each denoising step, an edge-structure guidance map is introduced to preserve geometric boundaries and local contours. This design is important

for agricultural field-note visualization because crop morphology, disease location, row structure, and field layout must remain visually interpretable.

The latent-variable update is computed by the conditional denoising function:

$$z_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left[z_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(z_t, t, c) \right] + \sigma_t \eta \tag{9}$$

where z_t is the latent variable at step t, z_{t-1} is the updated latent variable, α_t and $\bar{\alpha}_t$ are noise attenuation coefficients, ϵ_θ is the conditional noise-estimation function, c is the semantic condition vector, σ_t is the noise scale, and η is sampled noise.

Edge guidance is injected into the denoising process to impose line constraints on latent variables:

$$z_t^e = z_t + \lambda_e E(z_t) \tag{10}$$

where z_t^e is the edge-guided latent variable, λ_e is the structural guidance strength coefficient, and $E(z_t)$ is the edge response calculated by the convolution operator. This residual adjustment guides the generation process toward clear line structures and preserves the visual logic of crops and field environments.

To balance semantic accuracy and line simplicity, controllable guidance is introduced during denoising:

$$\hat{\epsilon} = \epsilon_{\text{uncond}} + s(\epsilon_{\text{cond}} - \epsilon_{\text{uncond}}) \tag{11}$$

where $\hat{\epsilon}$ is the guided noise prediction, ϵ_{uncond} is the unconditional prediction, ϵ_{cond} is the conditional prediction, and s is the guidance scale. The formula strengthens semantic constraints while retaining visual diversity. After each iteration, latent variables are normalized and gradients are clipped to avoid extreme values that may cause broken lines or structural distortion. The final latent image is decoded to pixel space to generate an initial sketch aligned with the field-note semantics.

2.4 Post-Processing, Layout Standardization, and Consistency Verification

After latent decoding, the preliminary sketch image is refined through line smoothing, breakpoint connection, contrast enhancement, layout integration, and semantic consistency verification. These steps improve visual clarity and also introduce standardizable output-control rules. A visual diary intended

for cross-platform exchange should carry not only an image but also the associated text note, timestamp, location, semantic score, generation model version, and quality indicators.

Line smoothing is achieved through second-derivative convolution filtering, which reduces abrupt local variations while preserving important edges:

$$I_s(x,y) = I(x,y) + \sum_{u,v} w_{u,v} \left(\frac{\partial^2 I}{\partial x^2} + \frac{\partial^2 I}{\partial y^2} \right) \tag{12}$$

where I_s is the smoothed image, I is the original image, $w_{u,v}$ denotes the convolution-kernel weight, and the partial derivatives describe second-order changes in the horizontal and vertical directions. Breakpoint connection is then performed using a weighted shortest-path operation:

$$P^* = \arg\min_{P \in \Omega} \sum_{k} \omega_k d(p_k, p_{k+1}) \tag{13}$$

where P^* is the selected line-connection path, Ω is the candidate path set, ω_k is the position weight, $d(.)$ is the distance function between pixels, and p_k denotes the k-th point in a candidate path. Contrast enhancement is performed through local adaptive normalization:

$$I_{\text{enh}} = \kappa \frac{I - \mu_L}{\sigma_L + \epsilon} + b \tag{14}$$

where I_{enh} is the enhanced pixel value, μ_L and σ_L are the local mean and standard deviation, ϵ is a numerical stability term, κ is the contrast-gain coefficient, and b is the brightness offset. The enhanced image is then placed with the field-note text and metadata in a standardized visual diary layout. Text-box coordinates are generated from the image-boundary reference point:

$$C_{\text{text}} = B_{\text{img}} + \Delta_l \tag{15}$$

where C_{text} is the coordinate of the text box, B_{img} is the image-boundary reference coordinate, and Δ_l is the layout offset. Finally, CLIP-based semantic consistency verification is applied between the generated image and the original text vector:

$$\text{sim}(I,T) = (f_I(I)^T f_T(T))/(\|f_I(I)\| \|f_T(T)\|) \tag{16}$$

where $\text{sim}(I,T)$ is the similarity score, $f_I(I)$ is the image-encoding vector, and $f_T(T)$ is the text-encoding vector. If the score falls below the preset

threshold, regeneration is triggered. This verification step helps ensure that the final visual diary remains semantically consistent with the source note. From the perspective of ICT standardization, the similarity threshold, model parameters, layout metadata, and regeneration flag can serve as candidate elements for future interoperability specifications.

2.4.1 Step-by-step example of field-note-to-sketch conversion

To make the processing pipeline reproducible, this subsection gives a concrete example. Consider the field note: "Corn leaves at the seedling stage show slight yellowing, and the soil surface is dry after three days without irrigation." The normalization layer first removes irrelevant punctuation and segments the note into agricultural entities and states: crop = corn, growth stage = seedling, symptom = slight leaf yellowing, environment = dry soil surface, and operation context = irrigation delay. These fields form a structured metadata record that can be passed to farm management software or a mobile application.

The CLIP text encoder then maps the normalized note into hidden states. Entity- and action-aware weights assign larger attention to "corn," "seedling," "yellowing," and "dry soil," producing the weighted semantic vector in Equation (1). Positional encoding in Equations (2) and (3) preserves the relation between crop, symptom, and environment, while Equation (4) aggregates segment vectors into the global semantic representation S. In practical deployment, S can be stored together with the original note, timestamp, location, and observer ID as a semantic metadata unit.

In the generation stage, the global semantic vector is injected into latent visual features through cross-attention according to Equations (5)–(8). The diffusion module then generates a sketch under the edge-guided update in Equations (9)–(11). For the example note, the expected output is a line-style corn seedling with early leaf-yellowing marks and a lightly cracked soil surface rather than a photorealistic crop image. The post-processing layer applies line smoothing, breakpoint connection, contrast enhancement, and CLIP-based semantic verification. If the final similarity score is below 0.75, the image is regenerated and marked with a regeneration flag.

2.4.2 Real-world integration with farm management systems and mobile applications

The proposed framework can be embedded as a service layer in existing farm management systems. A mobile application collects the field note, optional location, time, crop type, and field identifier. These records are sent to the

semantic parsing layer, which returns standardized metadata and a generated sketch. The metadata can be stored in the farm database using common fields such as observation_id, crop_entity, growth_stage, symptom_type, environmental_condition, model_version, semantic_score, and regeneration_status.

At the application level, the generated visual diary can support extension-service consultation, field inspection, seasonal comparison, and farmer-to-advisor communication. The system does not replace expert diagnosis; rather, it provides a readable and machine-verifiable representation of field observations. The standardized output package also allows visual records to be shared across extension platforms, farm cooperatives, and AI advisory tools without losing the original text or quality-control information.

3 Experimental Validation and Interoperability-Oriented Quality Assessment

3.1 Experimental Data and Parameter Configuration

The experimental validation used two publicly available agricultural multimodal image-text datasets covering crop types, disease states, field layouts, and environmental conditions. The FarmSeg-VL dataset contains fine-grained image-text records for farmland segmentation across eight agricultural regions in China, covering approximately 4300 square kilometers and including temporal information for spring, summer, autumn, and winter. Each image is annotated with descriptions of farmland attributes, phenological characteristics, spatial distribution, topography, and surrounding environment. The VL-PAW dataset focuses on fruit-tree disease and weed identification, including pear tree diseases, apple surface defects, and weed categories, with 3900 image-text samples. These datasets differ in description length, semantic complexity, and visual scale, allowing the proposed framework to be evaluated from single-plant close-ups to field-level structures. Representative samples from the two datasets are shown in Figure 4.

Table 1 summarizes the core configuration of the proposed method. The prompt coefficient controls semantic-vector enhancement, structure guidance strength regulates the edge constraint, and the guidance scale determines the strength of conditional denoising. Attention-head number, key-vector dimension, denoising steps, maximum text length, and semantic consistency threshold define the main model-interface parameters. These parameters are retained as experimental settings and can also be understood as candidate configuration fields for a standardized multimodal agricultural visualization pipeline.



Figure 4 Representative samples from the agricultural multimodal image-text datasets.

Table 1 Core parameter configuration

Parameter	Setting	Description
Prompt Coefficient	Default: 0.5	Controls the enhancement strength of prompt embedding on semantic vectors
Structure Guidance Strength	Default: 0.6	Controls the influence of edge constraints on the generation process
Guidance Scale	Default: 3.0	Controls the constraint intensity of semantic conditions on denoising
Number of Attention Heads	8	Number of parallel subspaces for semanticvisual alignment
Key Vector Dimension	64	Dimension of key vectors in cross-attention computation
Denoising Steps	Training: 1000; Inference: 50	Iteration steps from noise to target latent representation
Max Text Length	75 tokens	Maximum input length of the CLIP text encoder
Semantic Consistency Threshold	0.75	Minimum similarity score to trigger image regeneration

3.1.1 Metric definitions for agricultural visualization assessment

Fréchet Inception Distance (FID) measures the distributional distance between generated sketch outputs and reference agricultural sketch or image

distributions. In this study, a lower FID means that the generated visual diaries are closer to the expected agricultural visual style and are less likely to contain unrealistic structures. Structural Similarity Index (SSIM) evaluates local structural similarity, including edges, texture organization, and contrast. In the agricultural context, higher SSIM indicates better preservation of crop morphology, leaf boundaries, field rows, and disease-region structures. Learned Perceptual Image Patch Similarity (LPIPS) measures perceptual distance between generated images and reference visual patterns using deep visual features. A lower LPIPS indicates that the sketch is perceptually closer to the target agricultural drawing style and contains fewer distracting visual deviations. CLIP Score measures the cosine similarity between the original field-note text and the generated visual output in the CLIP semantic space. A higher CLIP Score indicates that the sketch better reflects the crop type, symptom, operation, and field condition described in the note. Recall@5 evaluates cross-modal retrieval consistency. In this study, it measures whether the correct field-note description can be retrieved within the top five text candidates when the generated sketch is used as the query. Higher Recall@5 suggests stronger semantic traceability between the sketch and the source note. The Non-Photorealistic Rendering Metric (NPRM) is defined here as an operational sketch-quality index that combines line continuity, edge-density appropriateness, and grayscale sparsity: NPRM $= 0.40$ LC $+ 0.35$ (1–$|\mathrm{ED} - \mathrm{ED}_{\mathrm{ref}}|$)+0.25 GS, where LC denotes normalized line-continuity score, ED denotes edge density, $\mathrm{ED}_{\mathrm{ref}}$ is the reference edge-density range derived from validation sketches, and GS denotes grayscale sparsity. Higher NPRM indicates clearer line drawings with fewer photorealistic textures.

3.2 Evaluation of Prompt Embedding Regulation

The effect of prompt embedding on semantic vectors was evaluated through a controlled experiment. A fixed base semantic vector was used, and learnable prompt increments with different intensity coefficients $\gamma = 0.1$, 0.3, 0.5, and 0.7 were introduced to generate enhanced semantic vectors. All vectors were computed within a unified vocabulary index range of 1–100. The semantic shift caused by prompt injection was quantified by comparing the baseline and enhanced vectors. This design verifies whether the prompt-template mechanism can strengthen agricultural keywords without destroying the original semantic structure.

Figure 5 shows that the deviation between the enhanced and baseline vectors increases with gamma. The trend indicates that prompt embedding

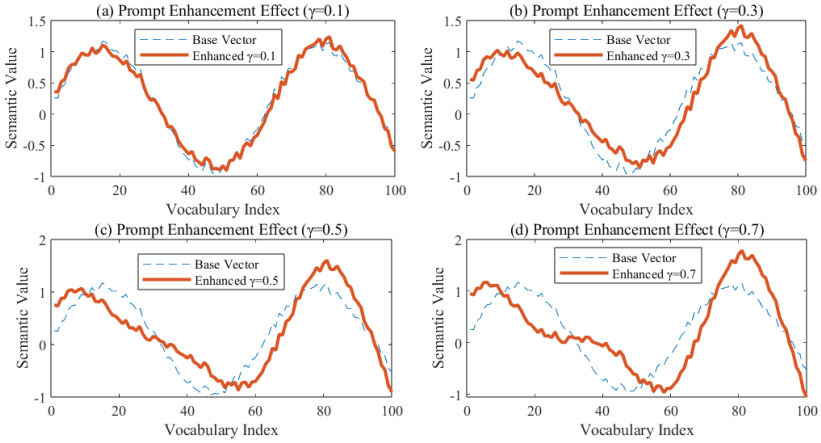


Figure 5 Effect of prompt embedding on agricultural semantic representation.

amplifies the representation strength of key agricultural terms while preserving the overall vector pattern. The mechanism therefore offers a controllable means of strengthening standardized entities such as crop, symptom, operation, and environment. This result supports the use of prompt coefficients as tunable semantic-regulation parameters in a pre-standardization ICT framework.

3.3 Image Generation Fidelity and Structural Integrity

Image generation fidelity and structural preservation were evaluated using FID and SSIM. The proposed method was compared with SDXL, DeepFloyd IF, CogView2, DALL•E 3, and Midjourney v6 under three agricultural scene types: simple single plants, complex fields, and disaster close-ups. FID measures the distributional distance between generated sketches and reference sketch data, while SSIM evaluates local structural similarity. These indicators are retained from the original experiment and are interpreted here as candidate quality metrics for multimodal field-note visualization.

Figure 6 shows that the proposed method maintains an FID below 32 across the three scenario types, indicating closer alignment with the reference sketch distribution. SSIM remains above 0.82, showing stable preservation of crop edges, leaf texture, and field structure. As scene complexity increases, comparison models show greater FID variation and lower structural consistency. The relative stability of the proposed method suggests that the integration of semantic conditions and edge constraints is effective

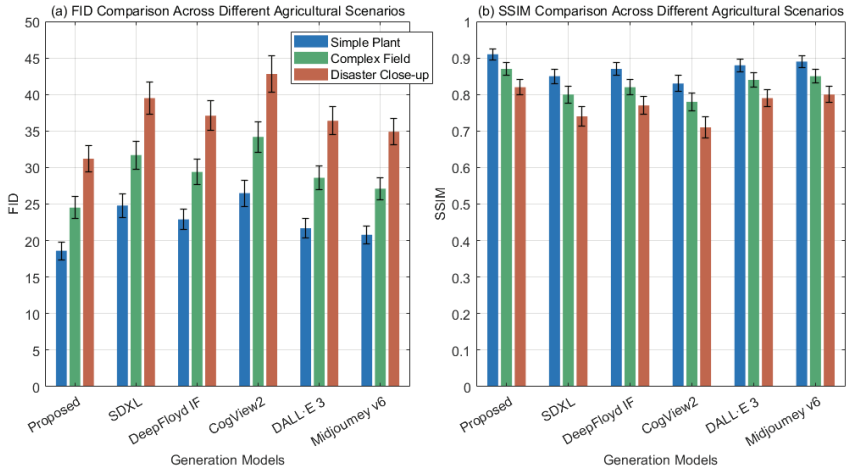


Figure 6 FID and SSIM assessment across different agricultural scenarios.

for generating visual outputs that remain interpretable across heterogeneous agricultural contexts.

3.4 Text-Image Semantic Alignment Accuracy

Text-image semantic alignment was evaluated using CLIP Score and Image-Text Retrieval Recall@5. CLIP Score measures the cosine similarity between generated image features and text features in a shared semantic space. Recall@5 measures whether the correct text is retrieved within the top five candidates in cross-modal retrieval. The evaluation was conducted under text-description lengths of 20, 50, 100, and 200 words. These metrics directly reflect whether field-note semantics can be preserved during visual diary generation and thus are relevant for standardizable quality verification.

Figure 7 indicates that the proposed method maintains high semantic matching across all text lengths. Under 200-word descriptions, the CLIP Score remains at 0.85 and Recall@5 reaches 84%, suggesting that long field-note descriptions can be compressed into visual representations with limited semantic loss. The advantage is particularly visible when multiple agricultural entities and temporal cues are present. In contrast, comparison models show stronger degradation under longer descriptions, indicating weaker preservation of multi-entity relations. This result supports the role of CLIP-guided semantic injection as a stable mechanism for transforming free-text agricultural notes into machine-verifiable visual records.

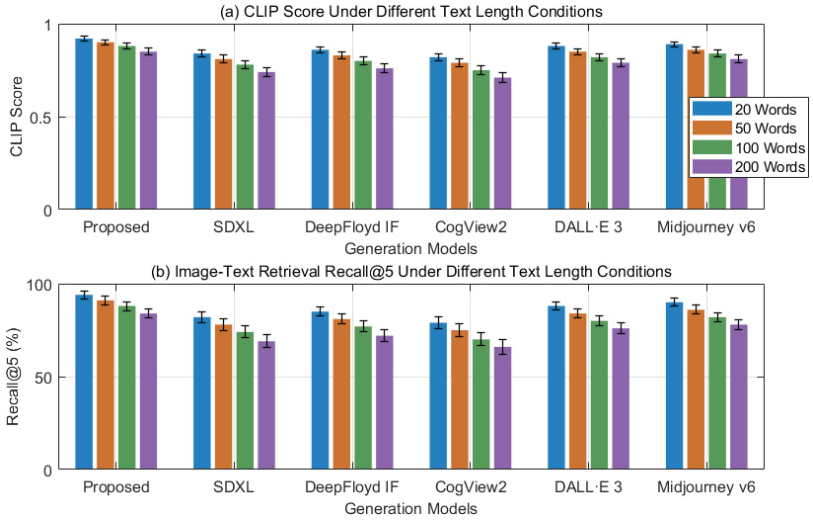


Figure 7 Text-image semantic alignment accuracy under different text-length conditions.

Table 2 Consistency of sketch style and artistry

Model	LPIPS ↓	NPRM ↑	p-value (LPIPS)	p-value (NPRM)
Proposed	0.082 ± 0.004	0.915 ± 0.006	1.20E-08	3.40E-09
SDXL	0.141 ± 0.006	0.842 ± 0.008	2.70E-06	5.10E-06
DeepFloyd IF	0.129 ± 0.005	0.856 ± 0.007	4.60E-07	7.80E-07
CogView2	0.152 ± 0.007	0.821 ± 0.009	6.30E-06	9.40E-06
DALL•E 3	0.118 ± 0.005	0.868 ± 0.006	1.90E-07	2.60E-07
Midjourney v6	0.105 ± 0.004	0.883 ± 0.005	8.70E-08	1.10E-07

3.5 Sketch-Style Consistency and Perceptual Evaluation

Sketch-style consistency was evaluated using LPIPS and the NPRM. LPIPS measures perceptual distance between generated images and the target sketch distribution. NPRM is defined in this study as a composite sketch-style indicator that combines line continuity, appropriate edge density, and grayscale sparsity, as described in the metric-definition subsection. Each model was generated ten times under the same data conditions, and paired t-tests were used to assess differences among methods.

Table 2 shows that the proposed method achieves the lowest LPIPS value of 0.082 ± 0.004 and the highest NPRM value of 0.915 ± 0.006. This indicates a lower perceptual distance from the target sketch distribution and stronger line-style purity. SDXL and CogView2 show higher LPIPS and lower NPRM

Table 3 User-perceived usefulness and comprehension efficiency

Model	Task Completion Time ↓ (s)	SUS Score ↑	p-value (time)	p-value (SUS)
Proposed	18.4 ± 1.6	86.7 ± 2.1	3.10E-09	2.60E-08
SDXL	27.9 ± 2.3	74.2 ± 2.8	6.40E-06	8.90E-06
DeepFloyd IF	25.6 ± 2.0	76.8 ± 2.5	1.20E-06	2.10E-06
CogView2	29.3 ± 2.6	72.5 ± 3.0	7.80E-06	9.60E-06
DALL•E 3	23.1 ± 1.9	86.4 ± 2.2	4.50E-07	6.30E-07
Midjourney v6	21.7 ± 1.8	86.6 ± 2.0	9.20E-08	1.40E-07

because their outputs tend to retain excessive texture details. DALL·E 3 and Midjourney v6 perform well in overall composition but show weaker non-photorealistic control in dense agricultural scenes. The statistically significant p-values indicate that the style-consistency improvement is stable rather than random. In a standardization context, LPIPS and NPRM can serve as complementary indicators for output-style verification.

3.6 User-Perceived Usefulness and Comprehension Efficiency

User-perceived usefulness and comprehension efficiency were evaluated through task-completion time and System Usability Scale (SUS) score. Task-completion time reflects the efficiency with which users locate and extract key agricultural information from a visual diary. SUS score reflects comprehensibility, learning cost, and perceived information clarity. The same comparative model set was evaluated to determine whether improvements in semantic alignment and structural consistency also translate into better information retrieval and usability.

Table 3 shows that the proposed method obtains the lowest average task-completion time of 18.4 ± 1.6 s and the highest SUS score of 86.7 ± 2.1. The results suggest that the generated visual diaries contain clearer visual partitions and more accurate representations of agricultural structures, reducing unnecessary visual search. Models with unstable boundaries or ambiguous crop conditions require longer interpretation time. The consistency between task-completion time and SUS score indicates that semantic-structural alignment contributes to both objective efficiency and subjective usability.

3.6.1 Representative visual sample outputs

Representative sketch-style visual diary outputs from the revised pipeline are shown in Figure 8. These examples show how agricultural field notes are converted into line-style visual records while retaining the original note,

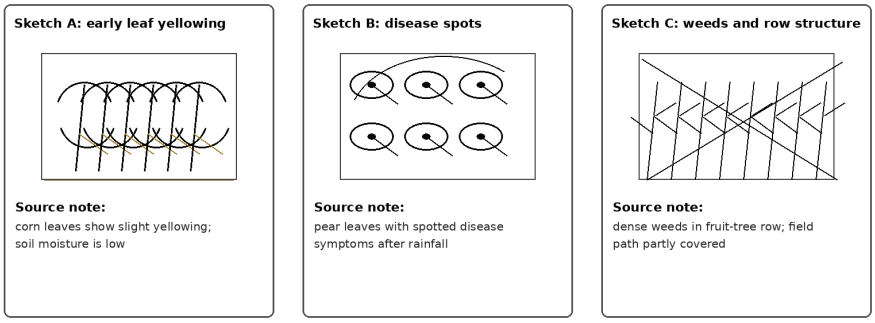


Figure 8 Representative field-note-to-sketch visual diary outputs generated by the revised pipeline.

dominant visual cue, and interpretable output structure. The examples are illustrative outputs from the revised pipeline and are intended to document output format, not to replace the quantitative evaluation reported above.

4 Discussion: Standardization Implications for Multimodal Agricultural ICT

The proposed method provides more than an image-generation pipeline; it offers a technical basis for pre-standardization of multimodal agricultural field-note visualization. The input side can be defined through minimum data elements, including field-note text, crop entity, symptom category, environmental condition, timestamp, location, observation source, and optional operational action. The processing side can be specified through semantic encoding, prompt weighting, cross-attention injection, edge-guided denoising, and quality verification. The output side can be represented through a visual diary package that includes the generated sketch, original note, metadata, semantic consistency score, model parameters, and regeneration status.

Such a layered structure is relevant to JICTS because it addresses interoperability, cross-sector convergence, and the transition from emerging AI workflows to reusable technical specifications. Agricultural platforms currently differ in how notes, images, diagnoses, and extension recommendations are stored and exchanged. Without a common structure, AI-generated visual outputs may become isolated artifacts that cannot be audited or reused. By associating each visual diary with semantic metadata and quality scores, the proposed framework improves traceability and supports future integration

with farm management systems, extension platforms, mobile recordkeeping tools, and AI advisory services.

The framework also raises governance issues. AI-generated agricultural images may hallucinate disease symptoms, exaggerate environmental stress, or misrepresent local terminology. These risks are not purely technical; they affect trust, accountability, and the interpretation of farmer knowledge. Therefore, candidate standards for visual diaries should include quality thresholds, human verification flags, prompt-version records, model-version records, and warnings for low-confidence outputs. The CLIP Score, Recall@5, SSIM, FID, LPIPS, NPRM, task-completion time, and SUS score used in this study can form an initial set of evaluation indicators, although future standardization should further examine robustness, bias, privacy, and cross-regional terminology adaptation.

Several limitations remain. First, the method continues to face difficulties in extremely complex field environments where multiple causal relations are implicit. Second, regional agricultural terminology may still create semantic ambiguity even after prompt enhancement. Third, although public datasets support controlled evaluation, broader validation is needed across crops, languages, and local farming systems. Fourth, the framework provides a pre-standardization technical path rather than a formal standard. Future work should develop explicit metadata schemas, application programming interfaces, version-control rules, and conformance-testing procedures for multimodal agricultural visual diaries.

5 Conclusion

This study developed a pre-standardization multimodal ICT framework for transforming agricultural field notes into sketch-style visual diaries through CLIP-guided semantic alignment and controllable AI sketching. The method uses CLIP-based semantic embedding to encode agricultural entities and state descriptions, cross-attention to inject semantic conditions into the latent generation space, edge-guided denoising to preserve crop morphology and field structure, and post-processing with consistency verification to improve output clarity and semantic fidelity. The experimental results show stable performance in semantic matching, structural preservation, sketch-style consistency, and user-oriented information extraction. Specifically, the proposed method maintains an FID below 32, SSIM above 0.82, CLIP Score of 0.85 for long descriptions, Recall@5 of 84%, LPIPS of 0.082, and SUS score of 86.7. These findings indicate that semantic metadata, model-interface parameters,

quality metrics, and visual-output structures can be organized into a reusable technical framework for interoperable AI-assisted agricultural documentation. The work therefore contributes to the standardizationoriented development of multimodal AI in digital agriculture and provides a foundation for future ICT specifications, metadata schemas, and conformance-testing procedures.

Data Availability

The experimental validation is based on publicly available agricultural multimodal image-text datasets described in the manuscript. Processed evaluation sheets, prompt examples, representative visual outputs, and implementation details are available from the corresponding author upon reasonable request, subject to dataset licenses and institutional requirements. The proposed metadata fields and quality-metric definitions are reported in the manuscript to support reproducibility and future interoperability testing.

Conflicts of Interest

The authors declare no conflict of interest.

Ethics Statement

The experiments were primarily conducted on publicly available agricultural image-text datasets. The usability evaluation reported in Table 3 was conducted as an anonymous, minimal-risk comprehension and usability assessment of generated visual diaries. No personally identifiable information was collected, stored, or reported. Participants were informed of the academic purpose of the evaluation, the voluntary nature of participation, the anonymized use of responses, and their right to withdraw before submitting their responses. According to the applicable institutional guidelines for anonymous usability evaluation, formal ethics committee approval was not required. If required by the journal or institution, the authors will provide an institutional exemption statement as a peer-review-only file.

Acknowledgements

This work was supported by 2025 Shandong Provincial Social Science Planning Research Project: Construction of the Cultural Gene Atlas and Digital

Empowerment Research on New Year Picture Culture of the Shandong Section of the Grand Canal (25CWYJ25).

References

Aryal, A., Sapkota, Y., Lamichhane, B., and Shrestha, J. (2026). Applications of Internet of Things, remote sensing, and AI for precision agriculture and its adoption status in Nepal. Journal of Applied Agricultural Science and Technology, 10(1), 170–196.

Barbu, T. (2023). Edge detection techniques using nonlinear diffusion-based models. Bulletin of the Transylvania University of Brasov. Mathematics, Informatics, Physics. Series III, 3(1), 199–210.

Basir, M. S., Buckmaster, D., Raturi, A., and Zhang, Y. (2024). From pen and paper to digital precision: A comprehensive review of on-farm recordkeeping. Precision Agriculture, 25(5), 2643–2682.

Bowen, R., and Morris, W. (2024). Digital entrepreneurship in agrifood business: A resource bricolage perspective. International Journal of Entrepreneurial Behavior and Research, 30(2–3), 482–497.

Croitoru, F. A., Hondru, V., Ionescu, R. T., and Shah, M. (2023). Diffusion models in vision: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(9), 10850–10869.

Dhal, S., Wyatt, B. M., Mahanta, S., Bhattarai, N., Sharma, S., Rout, T., and Acharya, B. S. (2024). Internet of Things (IoT) in digital agriculture: An overview. Agronomy Journal, 116(3), 1144–1163.

Dzreke, S. S. (2025). Seeds of food sovereignty: AI, drones, and the fight against innovation apartheid in Africa's climate-smart agriculture. Frontiers in Research, 2(1), 1–19.

Ferreira, B. (2026). When the spotlight is shone on farmers: The transformative power of collaborative video in tourism research. Consumer Behavior in Tourism and Hospitality, 21(1), 157–170.

Garg, D., and Alam, M. (2023). Smart agriculture: A literature review. Journal of Management Analytics, 10(2), 359–415.

Liu, F. L., Fu, H., Lai, Y. K., and Gao, L. (2024). SketchDream: Sketch-based text-to-3D generation and editing. ACM Transactions on Graphics, 43(4), 1–13.

Mulungu, K., Kassie, M., and Tschopp, M. (2025). The role of information and communication technologies-based extension in agriculture: Application, opportunities and challenges. Information Technology for Development, 31(4), 1117–1146.

Namyenya, A., Daum, T., Rwamigisa, P. B., and Birner, R. (2022). E-diary: A digital tool for strengthening accountability in agricultural extension. Information Technology for Development, 28(2), 319–345.

Namyenya, A., Rwamigisa, P. B., and Birner, R. (2024). Strengthening the accountability of agricultural field agents: A principalagent perspective. The Journal of Agricultural Education and Extension, 30(3), 363–386.

Nismi Mol, E. A., and Santosh Kumar, M. B. (2023). Review on knowledge extraction from text and scope in agriculture domain. Artificial Intelligence Review, 56(5), 4403–4445.

Pearce, L. M. (2021). Photos in the field: Reflecting on environmental change through photo-elicited oral histories. Studies in Oral History: The Journal of Oral History Australia, 43, 166–193.

Peng, Y. (2025). A CLIP-based cross-modal matching model for image-text retrieval. Information Technology and Control, 54(3), 1030–1048.

Rabbi, M. F., Liza, Z. A., and Rahman, M. S. (2023). Capability enhancement of agricultural extension personnel through ICT interventions and its economic impact in Bangladesh. Economic Affairs, 68(4), 1937–1945.

Rizvi, S. K. J., Azad, M. A., and Fraz, M. M. (2021). Spectrum of advancements and developments in multidisciplinary domains for generative adversarial networks (GANs). Archives of Computational Methods in Engineering, 28(7), 4503–4521.

Rodríguez, T., Bonatti, M., Löhr, K., and Sieber, S. (2025). Rethinking knowledge systems for agroforestry: Insights from the mental models of cacao farmers in Colombia. Ambio, 54(11), 1852–1866.

Sarku, R., and Ayamga, M. (2025). Is the right going wrong? Analysing digital platformization, data extractivism and surveillance practices in smallholder farming in Ghana. Information Technology for Development, 31(4), 1039–1065.

Shin, A., Ishii, M., and Narihira, T. (2022). Perspectives and prospects on transformer architecture for cross-modal tasks with language and vision. International Journal of Computer Vision, 130(2), 435–454.

Singh, A. K., Junior, F. N. F., Mainsah, N. L., and Abdoul-Rahmane, B. (2024). Enabling data collection and analysis for precision agriculture in smart farms. IEEE Transactions on AgriFood Electronics, 3(1), 69–85.

Sivabalan, A., Rajan, M. A., and Balamuralidhar, P. (2013). Towards a lightweight Internet of Things platform architecture. Journal of ICT Standardization, 1(2), 241–252. https://doi.org/10.13052/jicts2245-800X.12a8.

Tessier, L., Bijttebier, J., Marchand, F., and Baret, P. V. (2021). Cognitive mapping, Flemish beef farmers' perspectives and farm functioning:

A critical methodological reflection. Agriculture and Human Values, 38(4), 1003–1019.

Utomo, A., McGregor, A., Som, B., Finlayson, C., Korn, C., McKinnon, K., and Cook, B. R. (2025). Team photo-diaries: Making places, people, and power more visible. Geographical Research, 63(4), 606–621.

Walshe, R., Casey, K., Kernan, J., and Fitzpatrick, D. (2020). AI and Big Data Standardization: Contributing to United Nations Sustainable Development Goals. Journal of ICT Standardization, 8(2), 77–106. https://doi.org/10.13052/jicts2245-800X.821.

Wang, W., and Wang, Y. (2022). Research on image capture technology of intelligent terminal and multi-exposure fusion to improve the resilience of agriculture production systems. Journal of Commercial Biotechnology, 27(2), 46–56.

Wang, Y., Chen, X., Ma, X., Zhou, S., Huang, Z., Wang, Y., and Liu, Z. (2025). LaVie: High-quality video generation with cascaded latent diffusion models. International Journal of Computer Vision, 133(5), 3059–3078.

Wulandari, D., Prayitno, P. H., Prasetyo, A. R., Gunawan, A., Syafruddin, A. B., and Indriati, W. N. (2025). Greenhouse fruit and vegetable picking tourism management model based on circular economy with productive waqf approach to support food independence of Ngajum Malang Prison. Jurnal Syntax Imperatif: Jurnal Ilmu Sosial dan Pendidikan, 6(5), 1276–1298.

Xing, F., Li, M., Wang, Y. G., Zhu, G., and Cao, X. (2024). CLIPVQA: Video quality assessment via CLIP. IEEE Transactions on Broadcasting, 71(1), 291–306.

Xu, P., Hospedales, T. M., Yin, Q., Song, Y. Z., Xiang, T., and Wang, L. (2022). Deep learning for free-hand sketch: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(1), 285–312.

Zeng, Y., Zhang, X., Li, H., Wang, J., Zhang, J., and Zhou, W. (2023). X2-VLM: All-in-one pre-trained model for vision-language tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(5), 3156–3168.

Zhang, D. J., Wu, J. Z., Liu, J. W., Zhao, R., Ran, L., Gu, Y., and Shou, M. Z. (2025). Show-1: Marrying pixel and latent diffusion models for text-to-video generation. International Journal of Computer Vision, 133(4), 1879–1893.

Zhao, Q. (2024). Research on the path of "new farmers" short video assisting the revitalization of rural cultural industry. International Journal of Frontiers in Sociology, 6(1), 23–27.

Biographies



Hui Li was born in Weihai, Shandong, P.R. China, in 1975. He received the master's degree from Shandong University. He works in the School of Art, Shandong Management University. His main research interests lie in traditional Chinese cultural arts and modern design.



Yu Chen was born in Jinan, P.R. China, in 1973. She received the master's degree from Hebei Normal University. She works in the School of Art and Design, Shandong Women's University. Her primary research interests include intangible cultural heritage and modern design.