RESEARCH READING GUIDE

HuanYu research reading guide and experimental design notes

IDENIFE · Methodological note · Version 1.0 · 2026-09-23

This guide connects cross-modal representation, domain semantics, conditional generation and evaluation through original papers and practical experiment proposals. It is an original IDENIFE methodological note, not a published research paper or a report of completed HuanYu experiments, measured performance or implemented architecture. The references are external research; citing them does not imply collaboration with or endorsement by their authors.

1. A reading route from representations to testable results

Research question	Suggested reading	Distinction to examine
How can images and language be compared?	CLIP [1]	Global similarity versus correct local relationships
How can a generator organise visual sequences?	DiT [2]	Architecture, latent compression and compute
How can a model learn a path from noise to data?	Flow Matching [3]	Training objective, probability path and inference solver
How can image and text information interact during generation?	MMDiT [4]	Modality-specific weights and joint information exchange
How can spatial conditions guide generation?	ControlNet [5]	Condition adherence and visual coherence
How can text-image faithfulness become explicit questions?	TIFA [6]	Question quality, visual answers and human review
Does confidence correspond to correctness?	Calibration [7]	Confidence, accuracy and distribution shifts
How can model selection account for quality and cost?	RouteLLM [8]	Routing policy, task distribution and quality constraints

Years refer to the first arXiv submission. DiT and Flow Matching first appeared in 2022; conference publication years may differ. The accompanying BibTeX library (huanyu-references.bib) cites the same arXiv records and includes complete author lists.

METHOD NOTE / 02

2. Cross-modal representation: retrieve, then check relationships

CLIP learns comparable image and text representations from paired data, providing a basis for retrieval and visual classification specified through language. Original paper 1

An independent industrial-design experiment could ask whether domain descriptions distinguish visually similar concepts with different functional relationships. This is a proposed experiment, not an industrial-design result established by the CLIP paper.

01 Split data by project, product family and design lineage. Keep recolours, crops and successive versions of the same concept in the same partition.

02 For each query, collect relevant examples and hard negatives. Change one important relationship in a negative, such as moving a port from the top to the side while preserving the product class and general style.

03 Compare the original brief, a normalised domain description and a description with relationship terms removed. Fix the candidate collection, encoder version and retrieval settings.

04 Report Recall@K, errors on relationship-focused negatives and sample counts for each product category. If several candidates satisfy the brief, retain all relevant labels rather than treating valid alternatives as negatives.

Use similarity to find candidates. Verify part counts, spatial relationships and unknown material fields through separate checks.

METHOD NOTE / 03

3. Domain semantics: preserve facts, requirements and unknowns

The domain layer below is a data-organisation proposal in this guide, not an industrial-design standard supplied by the cited papers. A record should retain provenance and usage rights, project and version identifiers, product class, visible parts, relationships, brief requirements, material and process descriptions, unknown fields, and annotation and review history.

This illustrative example is fictional and does not represent a company project:

Field	Example value	Evidence type
Product class	Portable inspection terminal	Specified by the brief
Operating condition	Gloved use	A requirement; a rendering alone cannot validate it
Port location	Top	Record visual evidence and brief requirements separately
Housing material	Unspecified	Unknown; do not infer metal
Internal assembly	Unknown	No structural drawing or physical evidence

Have at least two annotators independently label an initial sample, then reconcile reasons for disagreement and revise the instructions. Separate what is visible, what is required and what needs professional inference. Retain conflicts and unknowns instead of optimising field completion alone.

METHOD NOTE / 04

4. Conditional generation: change one factor at a time

DiT studies transformers operating on latent image patches. Flow Matching studies vector-field regression for generative modelling. MMDiT uses modality-specific weights with image-text information exchange. ControlNet introduces spatial conditioning for pretrained diffusion models. Architecture, training objective and condition inputs address related but distinct questions. Paper 2, paper 3, paper 4, paper 5

Start with a defined task and a model that supports the intended inputs:

Experiment	Input change	Main checks
A	Text condition	Objects, counts and semantic relationships
B	Same text plus contours	Layout adherence and overall form
C	Same text plus depth	Relative depth and occlusion
D	Local edit with an explicit preservation region	Intended change and drift outside the edit

Record an unsupported condition as unsupported. If a different model is required, treat model identity as an additional experimental factor; differences cannot then be attributed to conditioning alone.

Fix splits, prompts, resolution, inference budget, condition preprocessing and evaluation rules before running. Use recorded seed lists where supported and retain the actual outputs. Identical seeds do not guarantee comparable random processes across systems. Alongside adherence, inspect small parts, text, boundaries and untouched regions. Appearance alone cannot establish dimensions, assembly correctness or manufacturability.

METHOD NOTE / 05

5. Evaluation: a question and evidence for each judgement

TIFA generates question-answer items from text and uses visual question answering to assess whether a generated image satisfies the description. It provides a way to inspect faithfulness at a finer level than a single similarity score. Original paper 6

Fix questions before examining outputs. Examples include “Is a port visible at the top?”, “How many ports are visible?” and “Are the display and grip regions separated?” A question about long-term comfort during gloved use needs other evidence and physical validation.

For each question, retain the expected answer, evidence region, model answer, expert review and failure category. Record reasons for removing ambiguous or visually unanswerable questions. Report slices for objects, attributes, counts, relationships, preservation constraints and insufficient evidence. An average should not hide a failed critical requirement. Blind human reviewers to model identity and randomise comparison positions; retain disagreement.

The evaluator can also misread images. Audit automated decisions against human review, distinguishing a correct generation rejected by the evaluator from an incorrect generation accepted by it. When estimating comparison intervals, resample by project or brief so that multiple outputs from one brief are not treated as fully independent tasks.

METHOD NOTE / 06

6. Confidence and routing: define correctness before thresholds

The calibration paper studies whether classifier probabilities match actual correctness and examines post-processing methods including temperature scaling. Applying temperature scaling requires appropriate classification scores and separate calibration data. A confident sentence is not a calibrated probability. Original paper 7

Define a reviewable event, such as passing the top-port check, and separate development, calibration and final test data. Record reliability plots, counts in each confidence bin and the binning rule used for calibration error. Also examine error rate against the fraction of tasks retained. Validate threshold behaviour again for unfamiliar categories or input quality.

RouteLLM learns quality-cost trade-offs between language models using preference data. Extending the idea to multimodal tasks requires a separate definition of task features, capability limits and preference labels; the original results cannot simply be inherited. Original paper 8

Compare always using a lower-cost model, always using a stronger model and routing by task. Define critical requirements in advance, then report pass rate, total call cost, latency distribution, human-review rate and abstention or clarification rate. Select thresholds on development data and reserve the final test set for an independent evaluation.

METHOD NOTE / 07

7. Minimum experiment record

Retain the following information so another reviewer can inspect the experiment or repeat it under comparable conditions:

Experiment identifier and date:
Question and hypothesis:
Data version, provenance scope and grouped split:
Model, weights or service version:
Baseline and the factor being changed:
Prompts, condition files and preprocessing:
Seeds, sampling settings and budget:
Predefined questions, thresholds and critical requirements:
Human-review rules, disagreements and unknowns:
Output files, failure categories and cost records:
Results and uncertainty: complete after running the experiment
Scope of conclusions and follow-up questions:

REFERENCES

References

These are external original papers. Years are first arXiv submission years; complete author lists and citation fields are in the accompanying BibTeX file.

1. Radford et al. (2021). Learning Transferable Visual Models From Natural Language Supervision. Paper record · Original PDF. Citation key: radford2021clip.

2. Peebles & Xie (2022). Scalable Diffusion Models with Transformers. Paper record · Original PDF. Citation key: peebles2022dit.

3. Lipman et al. (2022). Flow Matching for Generative Modeling. Paper record · Original PDF. Citation key: lipman2022flowmatching.

4. Esser et al. (2024). Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. Paper record · Original PDF. Citation key: esser2024mmdit.

5. Zhang, Rao & Agrawala (2023). Adding Conditional Control to Text-to-Image Diffusion Models. Paper record · Original PDF. Citation key: zhang2023controlnet.

6. Hu et al. (2023). TIFA: Accurate and Interpretable Text-to-Image Faithfulness Evaluation with Question Answering. Paper record · Original PDF. Citation key: hu2023tifa.

7. Guo et al. (2017). On Calibration of Modern Neural Networks. Paper record · Original PDF. Citation key: guo2017calibration.

8. Ong et al. (2024). RouteLLM: Learning to Route LLMs with Preference Data. Paper record · Original PDF. Citation key: ong2024routellm.

Bibliographic metadata checked on 2026-09-23. Unversioned arXiv links resolve to the latest revision. A reproducible experiment should additionally record the paper revision, code version and model version actually used.