Theory and Applications of Mathematics & Computer Science

(ISSN 2067-2764)
http://www.uav.ro/applications/se/journal/index.php/tamcs

Theory and Applications of Mathematics & Computer Science 11(1) (2026) 27–38

TrueVision: A Human-Aligned Explainable Visual Completion Dataset for Incomplete Artworks

Roxana Buzuriu[a,*], Darian M. Onchiş[a], Eduard Hogea[a]

[a]*West University of Timişoara, Faculty of Mathematics and Computer Science, Timişoara, Romania*

Abstract

Large vision-language models are increasingly used in domains where the input image is incomplete, including damaged paintings, cropped artworks, and restoration scenarios. In these settings, a plausible answer is not sufficient: the model should distinguish between visible evidence and inferred content, communicate uncertainty, and abstain when the available cues are too weak. This paper presents the current repository-backed stage of TrueVision, a bilingual English-Romanian dataset and annotation platform for explainable visual completion on artworks and other incomplete images. The current reportable experimental bundle contains a shared 100-task overlap annotated by three production annotators, yielding 300 bilingual annotation instances and deterministic 76/10/14 train-validation-test splits. The live platform already supports a richer structured workflow that separates visible cues from inferred hidden content in both languages, but the committed gold overlap still follows the earlier single-prediction annotation format and is therefore the basis of the quantitative results reported here. The paper includes four reproducible local baselines: zero-shot and few-shot Qwen2-VL 2B prompting, a metadata-aware retrieval baseline, and a lightweight supervised visual baseline trainable on Apple Silicon. On the held-out test split, retrieval remains the strongest non-learned reference in English similarity (0.4820), while the trained baseline provides the strongest learned result, with 0.4746 Romanian similarity, 0.4507 English token F1, 0.3223 Romanian token F1, 0.5093 confidence macro-F1, and 0.4853 difficulty macro-F1. These results show that the TrueVision pipeline is operational end to end and already supports controlled comparison of explainable visual completion methods in a low-resource bilingual setting.

Keywords: vision-language models, explainable AI, cultural heritage, dataset construction, multimodal reasoning, uncertainty estimation, Romanian resources

1. Introduction

General-purpose vision-language models such as Flamingo, BLIP-2, and LLaVA have demonstrated strong multimodal capabilities, but they are typically evaluated on tasks where the relevant evidence is present in the image (Alayrac *et al.*, 2022; Li *et al.*, 2023*a*; Liu *et al.*, 2023). Cultural heritage analysis and restoration introduce a different setting: the image may be partially destroyed, occluded, or cropped, while the system is still expected to produce meaningful, evidence-aware explanations. In this context, a visually plausible completion is only one part of the problem. The more important question is whether the model

*Corresponding author
E-mail address: roxana.buzuriu00@e-uvt.ro (Roxana Buzuriu)

can justify what is visible, state what is inferred, and signal uncertainty when the scene does not support a reliable conclusion.

Existing work already shows two complementary limitations. On the one hand, modern inpainting systems produce realistic completions, yet they do not explain why a specific completion is plausible (Lugmayr *et al.*, 2022). On the other hand, large vision-language models may hallucinate unsupported objects or claims when the visual signal is weak or incomplete (Li *et al.*, 2023*b*). This gap becomes even more problematic in art analysis, where historically or stylistically inappropriate completions can mislead human experts. Recent work on artwork explanation and damage analysis confirms that explanation quality, uncertainty, and domain specificity remain open challenges (Hayashi *et al.*, 2024; Ivanova *et al.*, 2025).

TrueVision is designed as a practical response to this gap. This research targets explainable visual completion on incomplete artworks, with bilingual annotations in English and Romanian and a workflow that explicitly captures abstention and confidence. The long-term goal is to fine-tune and evaluate models that can reason about missing content without collapsing into confident hallucination. The goal of the present paper is narrower and more realistic, to document the research problem, the dataset design, the implemented annotation tool, and the evaluation plan that will support the later experimental phase.

2. Related Work and Research Gap

Multimodal foundation models have expanded the scope of image understanding, but their benchmarks rarely require explicit reasoning under missing evidence (Alayrac *et al.*, 2022; Li *et al.*, 2023*a*; Liu *et al.*, 2023). Work on object hallucination shows that models can output linguistically fluent yet visually unsupported claims, which is especially risky when parts of the image are absent (Li *et al.*, 2023*b*). When the relevant region is hidden or damaged, these models may fall back to dataset priors rather than grounded visual reasoning. Cultural heritage and analogue media datasets show that damage patterns are diverse and difficult to generalize across, even for specialized segmentation and restoration tasks (Ivanova *et al.*, 2025).

TrueVision treats this gap as a benchmark design problem. Instead of asking only for a caption or an answer, it asks for a structured bilingual explanation of the likely hidden content together with uncertainty metadata. This shifts the focus from generic multimodal fluency to responsible reasoning under partial observability.

A further motivation is linguistic coverage. Romanian remains under-represented in multimodal tools and datasets, which limits both evaluation and downstream model adaptation in local contexts (Mitrofan *et al.*, 2025). TrueVision addresses this by treating bilingual annotation as a core design constraint, not a post-processing step. This gives the project value beyond restoration. It can also support explainability research in low-resource multimodal settings.

Work on multimodal explanations also supports the idea that natural-language justification should be treated as a direct modeling target rather than a by-product of captioning (Park *et al.*, 2018). This is especially relevant when the model must communicate what is supported by the visible image and what remains speculative.

The research gap can therefore be stated precisely: existing pipelines either prioritize visual completion without explanation, or textual explanation without explicit modeling of missing evidence. TrueVision positions itself between these directions by coupling incomplete-image inputs with structured human explanations, uncertainty signals, and a bilingual annotation workflow.

2.1. Image completion, amodal reasoning, and explanation

Inpainting methods such as Context Encoders, LaMa, RePaint, and Palette show that missing pixels can be reconstructed with high visual realism (Pathak *et al.*, 2016; Suvorov *et al.*, 2022; Lugmayr *et al.*, 2022;

Saharia *et al.*, 2022). Yet these methods do not explain why a completion is plausible or how certain the model is about the missing content. In parallel, recent amodal reasoning work, including Amodal Ground Truth and Completion in the Wild, pix2gestalt, AURA, and Open-World Amodal Appearance Completion, moves closer to the challenge addressed here by reasoning about occluded content rather than only visible regions (Zhan *et al.*, 2024; Ozguroglu *et al.*, 2024; Li *et al.*, 2025; Ao *et al.*, 2025). Complementary work on multimodal explanations shows that grounded language explanations are themselves a distinct modeling target rather than a by-product of captioning (Park *et al.*, 2018).

Still, these methods primarily optimize visual outputs such as masks, reconstructions, or amodal shapes. TrueVision differs because its target output is communicative: what cues are visible, what hidden content is plausible, and when the evidence is insufficient for a reliable answer.

2.2. Reliability, hallucination, and multilingual explanation

The reliability literature provides the final motivation. Large vision-language models are known to hallucinate unsupported content, especially when the input is ambiguous or weakly constrained (Li *et al.*, 2023*b*). This becomes more severe when the hidden region is semantically decisive. At the same time, multilingual studies show that explanation quality can degrade outside English. Recent artwork-explanation and cross-lingual artwork benchmarks are therefore particularly relevant to the bilingual design of TrueVision (Hayashi *et al.*, 2024; Ozaki *et al.*, 2025), while Romanian resource surveys highlight the practical need for evaluation in a low-resource multimodal setting (Mitrofan *et al.*, 2025). Calibration and abstention research further motivates explicit confidence and reject-option modeling when evidence is insufficient (Guo *et al.*, 2017; Geifman & El-Yaniv, 2019).

The research gap can be stated precisely. Current pipelines either emphasize visual completion without explanation, or textual explanation without explicit modeling of missing evidence. TrueVision positions itself between these directions by coupling incomplete-image inputs with human-authored explanations, abstention, confidence, and bilingual supervision.

3. Methodology and Benchmark Design

3.1. Task definition

For each masked image x and corresponding mask m, the benchmark target is a structured bilingual output that contains visible-cue text, hidden-content text, and uncertainty metadata. The output tuple is defined as follows:

$$y = (t_{\text{obs,en}}, t_{\text{obs,ro}}, t_{\text{inf,en}}, t_{\text{inf,ro}}, a, c, d) \tag{3.1}$$

In Equation (3.1), t_{obs} denotes observation grounded in the visible image, t_{inf} denotes the inferred hidden content, a is the abstention flag, c is a three-level confidence score, and d is the perceived difficulty label. Each annotation instance therefore records both content and epistemic status, allowing evaluation to measure not only semantic plausibility but also whether a method remains appropriately cautious under missing evidence.

3.2. Dataset construction pipeline

The benchmark pipeline begins with artwork collection, continues with synthetic mask generation and shared task assignment, and ends with annotation, gold aggregation, dataset preparation, and baseline evaluation. The pipeline is summarized in Figure 1. The dataset used in this paper includes a seeded pool of

1000 masked tasks, together with a shared evaluation subset of 100 tasks annotated by three production annotators. Synthetic masking is deliberately diverse: rectangular, elliptical, and irregular polygon masks are used so that the benchmark covers both semantically central and semantically peripheral missing regions.

Figure 1. TrueVision pipeline from artwork collection to evaluation.

3.3. *Annotation interface and masked-image-first workflow*

The annotation platform is implemented locally with Flask, SQLite, and Pillow. Each annotator logs in, receives a queued task, sees the masked image by default, and can optionally reveal the original image. This masked-image-first design is not cosmetic. It forces the primary judgment to begin from incomplete evidence rather than from direct inspection of the full painting. The combined interface views are shown in Figure 2.

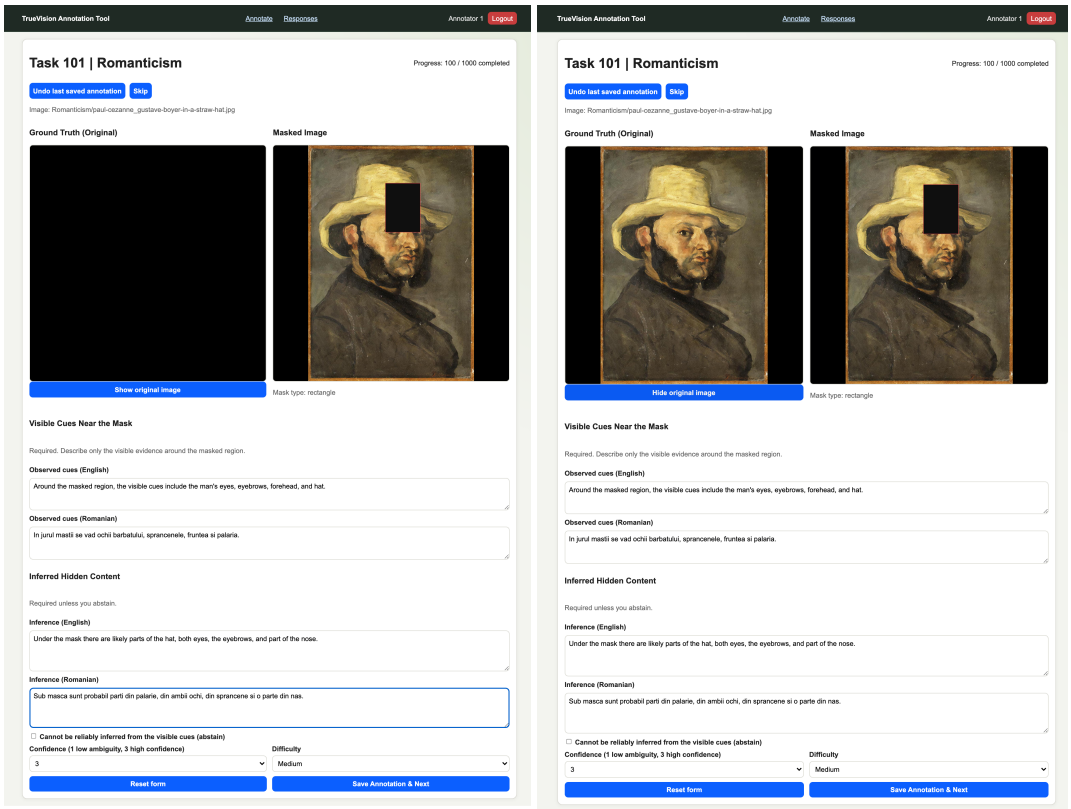

Figure 2. Annotation interface before and after revealing the original image.

The application supports bilingual text fields, confidence, difficulty, queue management, skip behavior, and export operations. The annotation controls used in the benchmark are summarized in Table 1, while the schema used for structured prediction is visualized in Figure 3.

Table 1. Annotation fields and controls used in the benchmark workflow.

Element	Description
Bilingual text entry	English and Romanian responses for the masked region, separated into visible cues and inferred hidden content.
Abstention	Boolean field used when the hidden content cannot be inferred reliably from the visible evidence.
Confidence	Three-level scale with values 1, 2, and 3.
Difficulty	Ordinal label with easy, medium, and hard.
Mask families	Rectangle, ellipse, and irregular polygon masks.
Queue actions	Save-and-next, skip, progress tracking, and undo-last-save behavior.

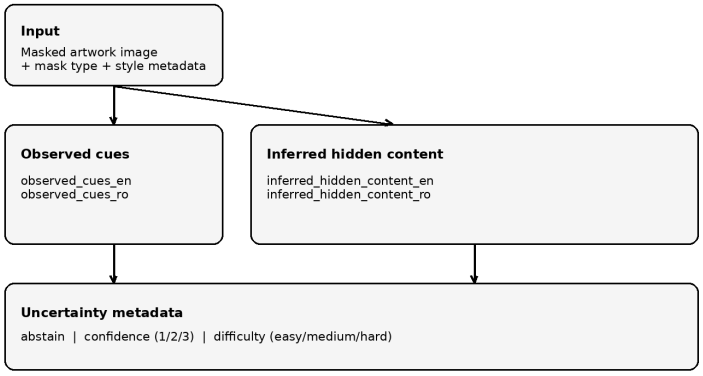

Figure 3. Annotation schema for visible cues, inferred hidden content, and uncertainty metadata.

Figure 4 restates the benchmark pipeline in a compact schematic form: data collection feeds masking and curation; curated tasks enter the bilingual annotation interface; structured exports support both baseline construction and evaluation; and the final release packages the dataset, code, and documentation needed for reproducible reuse.

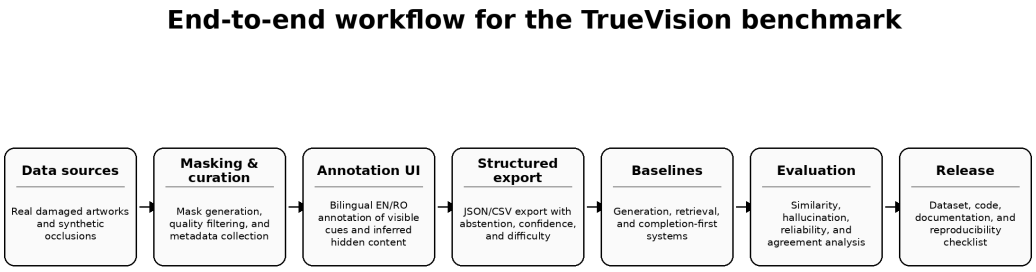

Figure 4. End-to-end workflow diagram for data sourcing, annotation, export, baselines, evaluation, and release.

3.4. Frozen splits and experimental material

The benchmark subset used in this paper consists of 100 distinct tasks completed by three production annotators, which yields 300 bilingual annotation rows. These rows are organized into frozen train, validation, and test bundles. The split statistics are given in Table 2. The reported dataset is intentionally fixed so that all baselines are compared under identical conditions and the same randomization of mask placement and annotation pairing.

Table 2. Dataset summary and frozen split composition.

Subset / component	Tasks	Annotation instances	Notes
Seeded task pool	1000	300	Masked-image pool used for task assignment and later expansion.
Shared evaluation subset	100	300	Three production annotators per task.
Training split	76	228	Used for retrieval indexing and supervised baseline training.
Validation split	10	30	Used for model selection and ablation analysis.
Test split	14	42	Held-out evaluation used only for final comparison.

To select the lightweight supervised baseline, the repository uses a validation composite score that prioritizes bilingual semantic quality while preserving structured-prediction behavior. The model-selection scalar is defined as:

$$S_{\text{comp}} = 0.3\,\text{Sim}_{EN} + 0.3\,\text{Sim}_{RO} + 0.2\,F1_{EN} + 0.2\,F1_{RO}. \tag{3.2}$$

Equation (3.2) is used for model selection, not for final evaluation. The final comparison relies on the full metric suite reported in the results section so that semantic quality, label prediction, and metadata reliability remain visible rather than being collapsed into a single scalar.

4. Experiments

4.1. Baselines

The baseline ladder contains four systems. The first two are prompt-only Qwen2-VL 2B variants: a zero-shot prompt and a few-shot prompt. The third is a metadata-aware retrieval baseline built on CLIP-style image-text representations (Radford *et al.*, 2021; Wang *et al.*, 2024) and training-split nearest-neighbor reuse. The fourth is a lightweight trained visual baseline that uses a frozen image encoder with trainable prediction heads for bilingual hidden-content labels and metadata. This final model is intentionally compact, but it is operationally important because it tests whether supervised learning is preferable to prompt-only behavior when the input is incomplete.

4.2. Metrics

The reported metrics combine free-text similarity and structured prediction. Sim_{EN} and Sim_{RO} measure English and Romanian semantic similarity against human reference annotations; these are normalized semantic-overlap scores for free-form generation and should not be interpreted as classification accuracy. $F1_{EN}$ and $F1_{RO}$ evaluate structured hidden-content label prediction derived from the textual outputs, while confidence and difficulty are evaluated with macro-F1. This combination is essential because a completion benchmark should reward not only plausible wording, but also stable label prediction and reliable uncertainty estimates.

4.3. Evaluation setting

Validation uses the 30 annotation instances associated with the shared 10-task validation slice. Test uses the 42 instances associated with the 14-task held-out slice. All experiments were run locally and reproduced under the same workflow so that each baseline is evaluated on the same benchmark split, with identical masks and paired human references.

5. Results

5.1. Validation comparison across baselines

The full validation comparison is shown in Figure 5 and summarized numerically in Table 3. Three conclusions stand out. First, zero-shot prompting is clearly inadequate: it yields low bilingual similarity and almost no useful structured prediction behavior. Second, the few-shot prompt improves lexical overlap, but it remains unreliable because it collapses to repeated answers, which indicates imitation of the prompt examples rather than grounded reasoning. Third, the retrieval baseline and the trained visual baseline form the first genuinely useful pair of baselines. Retrieval is competitive on English similarity and remains a strong non-parametric reference, but the trained visual baseline is stronger on label F1 and confidence prediction, which is more important for an explainable benchmark.

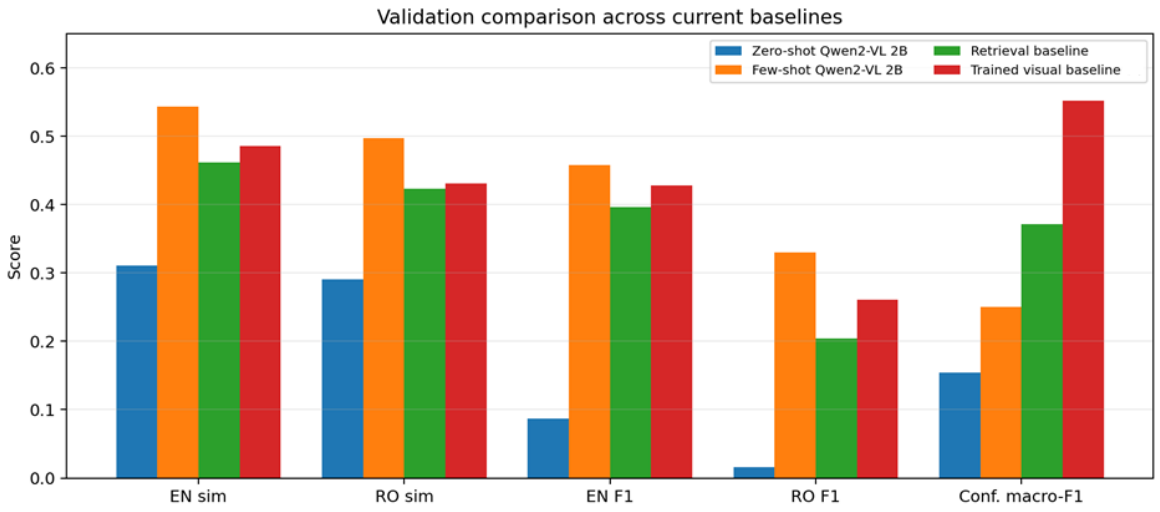

Figure 5. Validation comparison across the benchmark baseline ladder.

Table 3. Validation performance across baselines using semantic similarity and macro-F1.

Baseline	EN sim.	RO sim.	EN F1	RO F1	Conf. macro-F1
Zero-shot Qwen2-VL 2B	0.310	0.291	0.087	0.015	0.153
Few-shot Qwen2-VL 2B	**0.544**	**0.497**	**0.456**	**0.329**	0.249
Retrieval baseline	0.462	0.423	0.397	0.204	0.372
Trained visual baseline	0.485	0.430	0.428	0.261	**0.551**

The few-shot prompt deserves special caution. Its higher similarity scores are accompanied by near-total output collapse. In practice, it behaves like a template copier rather than a robust reasoner over masked

images. For this reason, the paper treats it as an exploratory ablation rather than as the preferred strong baseline. The comparison that matters more is retrieval versus trained supervision. The retrieval baseline shows that training-set reuse is already informative when the hidden region is constrained by painterly context, while the trained visual baseline shows that modest supervised learning can convert that signal into better metadata prediction and better label-level stability.

5.2. *Held-out test results*

The held-out comparison between the retrieval and trained visual baselines is shown in Figure 6 and reported in Table 4. On the test set, retrieval remains slightly stronger on English similarity, but the trained visual baseline is better on Romanian similarity, both label F1 scores, and both metadata macro-F1 scores. This pattern is encouraging because it suggests that supervision is helping precisely where the benchmark is structurally more difficult: cross-lingual consistency and uncertainty-related prediction.

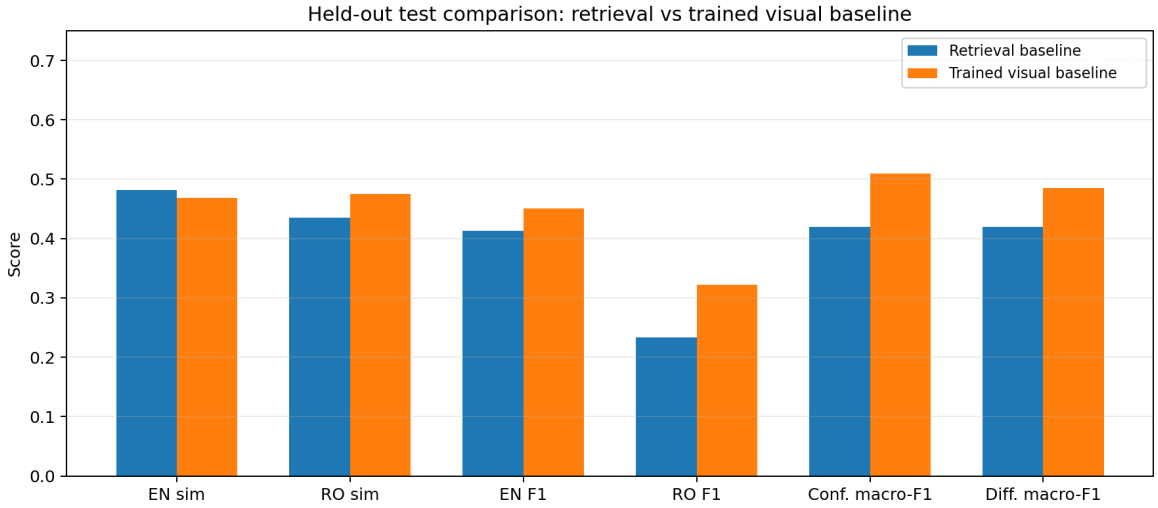

Figure 6. Held-out test comparison between retrieval and trained visual baselines.

Table 4. Held-out test results for the two strongest benchmark baselines.

Baseline	EN sim.	RO sim.	EN F1	RO F1	Conf. macro-F1	Diff. macro-F1
Retrieval baseline	**0.482**	0.436	0.412	0.235	0.420	0.420
Trained visual baseline	0.468	**0.476**	**0.450**	**0.322**	**0.509**	**0.486**

The test-set pattern highlights the complementary strengths of the two strongest methods. Retrieval remains competitive on English semantic similarity because it directly reuses stylistically related training examples. However, the trained supervised baseline is stronger on Romanian similarity, both hidden-content F1 scores, and both metadata macro-F1 scores. This indicates that supervision helps precisely on the dimensions that matter most for explainable visual completion: cross-lingual consistency, label stability, and uncertainty prediction.

5.3. *Learning dynamics of the trained visual baseline*

The learning dynamics in Figure 7 show that the trained visual baseline converges quickly, with the strongest validation composite appearing early and remaining stable afterward. This behavior is typical for

a compact supervised head on top of a frozen image encoder: training loss continues to decrease, but validation gains plateau once the model has extracted the most informative cross-task visual cues. Operationally, this is useful because it shows that the benchmark can support reproducible training even on modest local hardware.

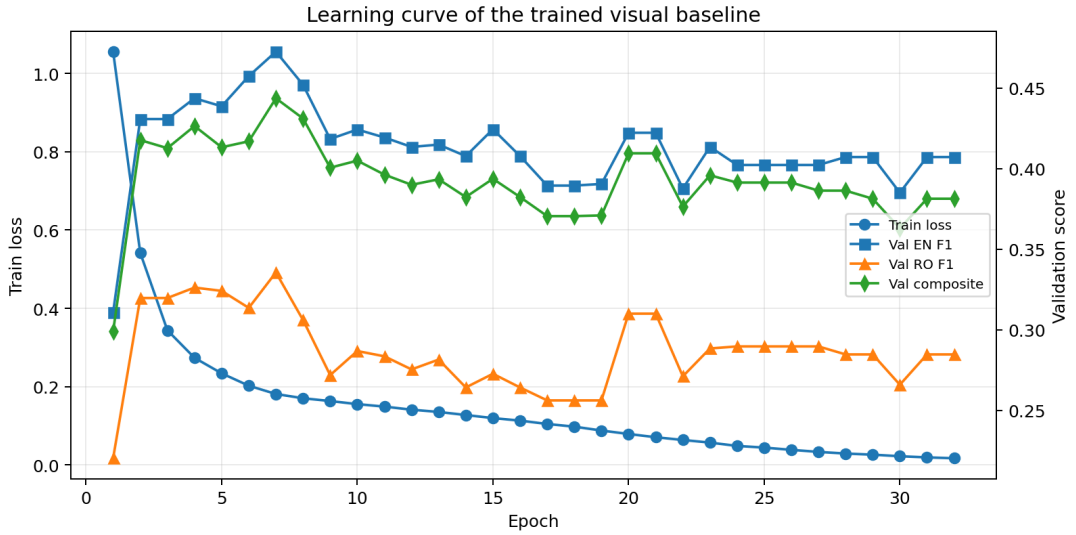

Figure 7. Learning curve of the trained visual baseline used for model selection.

5.4. *Agreement buckets and benchmark difficulty*

A useful property of TrueVision is that it does not collapse disagreement into annotation noise. The shared three-annotator design makes it possible to define agreement buckets that separate clearer cases from genuinely ambiguous ones. The current repository scaffold produces the frozen bucket distribution shown in Figure 8: 40 high-agreement tasks, 20 neutral tasks, and 40 ambiguous tasks. This near-symmetric structure is valuable because it prevents the benchmark from becoming dominated by trivially easy examples while still preserving a core of clearer cases for stable comparison.

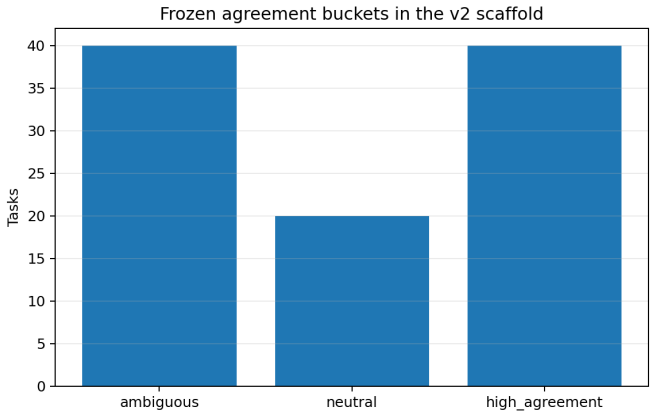

Figure 8. Agreement bucket distribution used in the benchmark split.

The existence of a large ambiguous bucket is not a weakness of the dataset; it is part of the phenomenon being studied. In incomplete-image reasoning, several completions may remain compatible with the visible evidence. A trustworthy model should therefore reflect this ambiguity through lower confidence, abstention, or less specific hidden-content claims instead of committing to a single overconfident hallucination.

5.5. Qualitative analysis

Qualitative inspection supports the same interpretation as the quantitative results. The representative examples in Figure 9 show a partial success and a clear failure of retrieval baseline v1. In the left example, the system retrieves another harbor scene and predicts boats and open water, which is compatible with the painterly context and mask placement. In the right example, retrieval confuses a face-and-hand sketch with a landscape scene, which demonstrates how brittle nearest-neighbor matching can become when the missing region is semantically central but the visible context is visually sparse or stylistically misleading.

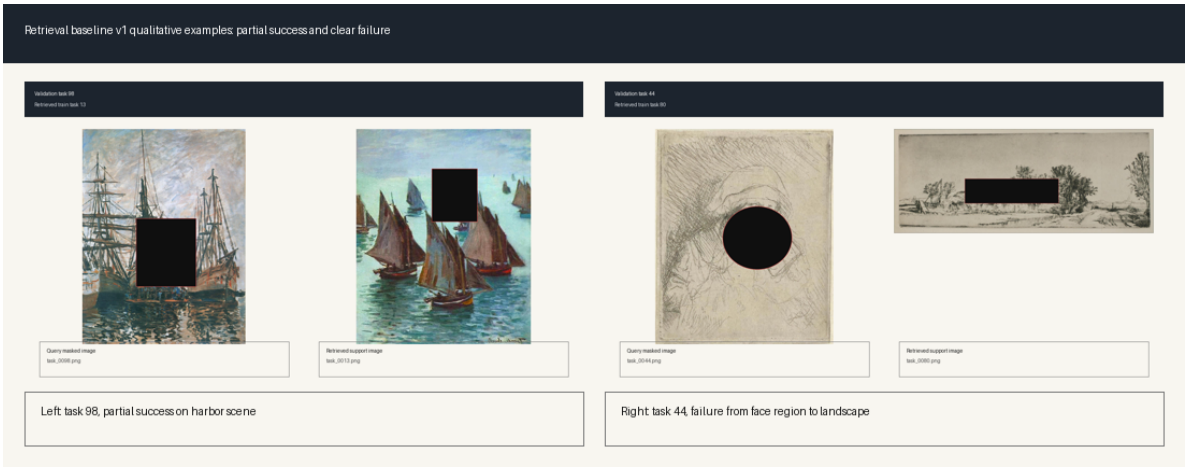

Figure 9. Qualitative retrieval examples: partial success on task 98 and clear failure on task 44.

This type of error is precisely why the benchmark should not be reduced to a retrieval-style similarity task. Strong retrieval can look convincing when painterly themes align, but it can also import the wrong object class, scene type, or narrative element when the visible evidence is weak. The trained visual baseline, even though still modest, is already more reliable on structured prediction because it does not depend on a single nearest-neighbor match.

6. Discussion

The reported semantic-similarity values should be interpreted in relation to the benchmark definition. Sim_{EN} and Sim_{RO} are free-form semantic-overlap scores, not exact-match accuracy values. Because the task asks for bilingual natural-language explanations of masked content, often with several plausible completions, scores in the 0.45–0.48 range already reflect meaningful alignment between model output and human reference explanations. In this setting, an apparently modest numeric difference often corresponds to a visible qualitative change in grounding and hallucination behavior.

Across baselines, the main lesson is that stronger prompting alone is not sufficient for reliable hidden-region reasoning. Prompt-only methods are prone to generic completions and mode collapse, while retrieval and supervised learning preserve more task-specific structure. The strongest overall result in this

paper is therefore not a single headline number, but the pattern across metrics: retrieval offers a robust non-parametric reference, whereas the trained visual baseline provides the most balanced performance on bilingual content prediction and uncertainty metadata. For cultural-heritage applications, this balance is more important than visually plausible but weakly supported completions.

7. Conclusion

TrueVision formulates incomplete-image reasoning as a benchmark for explainable visual completion rather than as a variant of generic captioning. The benchmark couples masked artworks, bilingual human explanations, and explicit uncertainty signals in order to evaluate whether a model can remain grounded when decisive visual evidence is missing.

The paper contributes a complete account of the dataset design, annotation workflow, benchmark splits, quantitative evaluation, and qualitative analysis. The included figures and tables document the full experimental pipeline, from masking and annotation to held-out comparison across baseline families.

Empirically, the benchmark shows a clear hierarchy of methods. Prompt-only inference is the weakest strategy because it collapses toward generic answers; metadata-aware retrieval is a strong reference baseline; and lightweight supervised learning delivers the best balanced results on Romanian semantic similarity, structured hidden-content prediction, and uncertainty metadata. These findings demonstrate that explainable visual completion is both measurable and experimentally tractable, and they establish TrueVision as a strong foundation for future multimodal research on incomplete artworks.

References

Alayrac, J.-B., J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson et al. (2022). Flamingo: A visual language model for few-shot learning. In: *Advances in Neural Information Processing Systems*. Vol. 35. https://papers.nips.cc/paper_files/paper/2022/hash/960a172bc7fbf0177ccccbb411a7d800-Abstract-Conference.html [Retrieved: 29.03.2026].

Ao, J., Y. Jiang, Q. Ke and K. A. Ehinger (2025). Open-world amodal appearance completion. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 6490–6499. https://openaccess.thecvf.com/content/CVPR2025/html/Ao_Open-World_Amodal_Appearance_Completion_CVPR_2025_paper.html [Retrieved: 29.03.2026].

Geifman, Y. and R. El-Yaniv (2019). Selectivenet: A deep neural network with an integrated reject option. In: *Proceedings of the 36th International Conference on Machine Learning*. PMLR. pp. 2151–2159. https://proceedings.mlr.press/v97/geifman19a.html [Retrieved: 29.03.2026].

Guo, C., G. Pleiss, Y. Sun and K. Q. Weinberger (2017). On calibration of modern neural networks. In: *Proceedings of the 34th International Conference on Machine Learning*. PMLR. pp. 1321–1330. https://proceedings.mlr.press/v70/guo17a.html [Retrieved: 29.03.2026].

Hayashi, K., Y. Sakai, H. Kamigaito, K. Hayashi and T. Watanabe (2024). Towards artwork explanation in large-scale vision language models. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics. pp. 705–729. https://doi.org/10.18653/v1/2024.acl-short.65.

Ivanova, D., M. Aversa, P. Henderson and J. Williamson (2025). ARTeFACT: Benchmarking segmentation models on diverse analogue media damage. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. pp. 7439–7449. https://doi.org/10.1109/WACV61041.2025.00723.

Li, J., D. Li, S. Savarese and S. C. H. Hoi (2023*a*). BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: *Proceedings of the 40th International Conference on Machine Learning*. PMLR. pp. 19730–19742. https://proceedings.mlr.press/v202/li23q.html [Retrieved: 29.03.2026].

Li, Y., Y. Du, K. Zhou, J. Wang, W. X. Zhao and J.-R. Wen (2023*b*). Evaluating object hallucination in large vision-language models. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. pp. 292–305. https://aclanthology.org/2023.emnlp-main.20/ [Retrieved: 29.03.2026].

Li, Z., H. Yoon, S. Lee and W. Lin (2025). Unveiling the invisible: Reasoning complex occlusions amodally with AURA. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 21927–21937. https://openaccess.thecvf.com/content/ICCV2025/html/Li_Unveiling_the_Invisible_Reasoning_Complex_Occlusions_Amodally_with_AURA_ICCV_2025_paper.html [Retrieved: 29.03.2026].

Liu, H., C. Li, Q. Wu and Y. J. Lee (2023). Visual instruction tuning. In: *Advances in Neural Information Processing Systems*. Vol. 36. https://papers.nips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html [Retrieved: 29.03.2026].

Lugmayr, A., M. Danelljan, A. Romero, F. Yu, R. Timofte and L. Van Gool (2022). RePaint: Inpainting using denoising diffusion probabilistic models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 11461–11471. https://doi.org/10.1109/CVPR52688.2022.01116.

Mitrofan, M., E. Irimia and V. Păiș (2025). Multimodal romanian language resources and tools: Challenges and perspectives. *Discover Data* **3**(1), Article 26. https://doi.org/10.1007/s44248-025-00060-4.

Ozaki, S., K. Hayashi, Y. Sakai, H. Kamigaito, K. Hayashi and T. Watanabe (2025). Towards cross-lingual explanation of artwork in large-scale vision language models. In: *Findings of the Association for Computational Linguistics: NAACL 2025*. Association for Computational Linguistics. pp. 3773–3809. https://doi.org/10.18653/v1/2025.findings-naacl.209.

Ozguroglu, E., R. Liu, D. Suris, D. Chen, A. Dave, P. Tokmakov and C. Vondrick (2024). pix2gestalt: Amodal segmentation by synthesizing wholes. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 3931–3940. https://openaccess.thecvf.com/content/CVPR2024/html/Ozguroglu_pix2gestalt_Amodal_Segmentation_by_Synthesizing_Wholes_CVPR_2024_paper.html [Retrieved: 29.03.2026].

Park, D. H., L. A. Hendricks, Z. Akata, A. Rohrbach, B. Schiele, T. Darrell and M. Rohrbach (2018). Multimodal explanations: Justifying decisions and pointing to the evidence. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 8779–8788. https://openaccess.thecvf.com/content_cvpr_2018/html/Park_Multimodal_Explanations_Justifying_CVPR_2018_paper.html [Retrieved: 29.03.2026].

Pathak, D., P. Krähenbühl, J. Donahue, T. Darrell and A. A. Efros (2016). Context encoders: Feature learning by inpainting. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 2536–2544. https://openaccess.thecvf.com/content_cvpr_2016/html/Pathak_Context_Encoders_Feature_CVPR_2016_paper.html [Retrieved: 29.03.2026].

Radford, A., J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal et al. (2021). Learning transferable visual models from natural language supervision. In: *Proceedings of the 38th International Conference on Machine Learning*. PMLR. pp. 8748–8763. https://proceedings.mlr.press/v139/radford21a.html [Retrieved: 29.03.2026].

Saharia, C., W. Chan, H. Chang, C. A. Lee, J. Ho, T. Salimans, D. J. Fleet and M. Norouzi (2022). Palette: Image-to-image diffusion models. In: *ACM SIGGRAPH 2022 Conference Proceedings*. Vol. 15. pp. 1–10. https://doi.org/10.1145/3528233.3530757.

Suvorov, R., E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov et al. (2022). Resolution-robust large mask inpainting with fourier convolutions. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. pp. 2149–2159. https://openaccess.thecvf.com/content/WACV2022/html/Suvorov_Resolution-Robust_Large_Mask_Inpainting_With_Fourier_Convolutions_WACV_2022_paper.html [Retrieved: 29.03.2026].

Wang, P., S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai et al. (2024). Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*. https://arxiv.org/abs/2409.12191 [Retrieved: 29.03.2026].

Zhan, G., C. Zheng, W. Xie and A. Zisserman (2024). Amodal ground truth and completion in the wild. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 28003–28013. https://openaccess.thecvf.com/content/CVPR2024/html/Zhan_Amodal_Ground_Truth_and_Completion_in_the_Wild_CVPR_2024_paper.html [Retrieved: 29.03.2026].