care, judgment, dexterity

CRAEFT

Scene and activity monitoring

Project Acronym	Craeft
Project Title	Craft Understanding, Education, Training, and Preservation for Posterity and Prosperity
Project Number	101094349
Deliverable Number	D3.3
Deliverable Title	Scene and activity monitoring
Work Package	3
Authors	Gavriela Senteri, Sotiris Manitsaris

This project has received funding from the European Commission, under the Horizon Europe research and innovation programme, Grant Agreement No 101094349.
http://www.craeft.eu/

Executive summary

This deliverable reports on the outcomes of Task T3.3 – Scene and Activity Monitoring, within WP3 (Digital Reenactment) of the project. The objective of this task is to facilitate the understanding of craft activities by jointly analyzing what is visible in the scene and what is implicitly expressed through human motion, with a particular focus on hand–tool–material interactions in real craft environments.

Traditional approaches to scene understanding often focus either on geometric reconstruction of the environment or on isolated recognition of actions. However, craft activities pose unique challenges in scene understanding, relying on fine-gained, skill-dependent movements that also include the collaboration of different operators. To address the involved challenges, T3.3 adopts a dual perspective on scene understanding, distinguished to explicit and implicit.

On the explicit side, this work introduces a dynamic scene representation pipeline that reconstructs craft scenes from egocentric video using a combination of a global motion scaffold and 3D Gaussian Splatting (3DGS). This representation enables the decomposition of scenes into hands, tools, materials, and background, and supports scene synthesis under occlusions (occlusion of the hand, the tool, the material). Quantitative evaluation metrics are applied to assess the quality of the synthesized scenes, including both pixel-level and perceptual metrics, as well as an evaluation of occlusion-revealed regions. Results across craft scenarios (marble carving, glass-blowing, silversmithing) demonstrate that the proposed approach produces visually consistent and structurally realistic reconstructions, even in partially or fully occluded regions.

Beyond geometric reconstruction, the deliverable focuses on explicit scene understanding, centered on the analysis of human movement complexity and hierarchical organization. Skills in crafts are not only expressed through visible transformation of objects, but through structured patterns of motion. For this, human activity is modelled hierarchically through movement primitives, actions, and activities, enabling a structured decomposition and interpretation of skills, as well as craft routines. Experimental results show that hierarchical modelling highlights the importance of implicit motion structure for reliable scene understanding.

Overall, the results of Task T3.3 demonstrate that combining explicit scene reconstruction with implicit hierarchical understanding of human motion provides a foundation for digital reenactment and skill analysis in craft contexts.

Document history

Date	Author	Affiliation	Comment
02/02/2026	Gavriela Senteri	ARMINES	First draft
02/02/2026	Sichen Su	ARMINES	First draft additions
08/02/2026	Gavriela Senteri	ARMINES	Second draft
09/02/2026	Sotiris Manitsaris	ARMINES	Draft internal review

Table of contents

Executive summary 2

Document history 3

Table of contents 4

1. Introduction 6

2. Motion capture of craft routines 8

3. Explicit Scene understanding 10

3.1. Explicit scene representation and synthesis 10

3.2. Dynamic Scene decomposition 11

3.3. Quantitative Evaluation of Scene Synthesis in Craft Environments 12

3.4. Quantitative results 13

3.4.1. Marble carving 13

3.4.2. Glass blowing with blowtorch 14

3.4.3. Silversmithing 15

3.4.4. Tapestry 17

3.4.5. Wood carving 18

3.4.6. Porcelain making 19

3.4.7. Glassblowing with pipe 20

3.5. Quantitative evaluation across craft scenarios 22

3.5.1. Marble carving 22

3.5.2. Glass blowing 23

3.5.3. Silversmithing 24

4. Implicit scene understanding 26

4.1 Hierarchical organization of craft movements 26

4.2.1 Hierarchical Multi-Task Backbone as Adaptation Prior 27

4.2.2 Forecast-Driven Adaptation Control 27

4.2.3 Stability and Knowledge Retention in crafts 28

5. Perspectives and future directions 30

5.1 Personalized Skill Assessment in Craft Training 30

5.2 Professional Reconversion and Skill Transfer 30

5.3 Future Directions 31

1. Introduction

Craft actions are characterized not only by the complexity of human movement itself, but also by the additional complexity that appears through the interactions between craftspeople, tools, and materials, shaped by the skills of each individual, their experience and the context of each craft. Understanding such actions is a challenge for digital reenactment, as it requires an interpretation of how these actions are performed, why certain movements are chosen, what are the reflexes behind them, as well as how the acquired expertise appears through movement, than a mere visual capture of them. This challenge is specifically addressed through Task T3.3, focusing on scene and activity monitoring for skill analysis and knowledge transmission.

Standard approaches for scene understanding typically emphasize explicit representations, such as object detection and localization, or visual feature extraction. While these methods are essential for describing the structure of a scene, they are not considered as sufficient when applied in isolation to the craft context. Craft activities are often recorded with cameras from exocentric viewpoints, where hands and body are frequently occluded by tools and materials, and where critical aspects of expertise are embedded in temporally extended motion patterns. As a result, an only explicit interpretation of the scene risks overlooking the elements that actually differentiate skilled performance from learner behavior.

To address this limitation, Task T3.3 deploys a dual interpretation of scene understanding, those of the explicit and implicit ones. Explicit scene understanding concerns the reconstruction and analysis of what is directly observable in the scene, such as hands, tools and materials, while implicit scene understanding, the interpretation of human motion semantics, capturing how movements are organized, combined, and repeated over time to form meaningful actions and activities for each one of the crafts. Within this framework, scene understanding is decomposed along two complementary analytical axes, the environment and the crafts person. The axis of the environment mainly focuses on the workspace, tools, and the materials, while the human axis one on skills, actions, and activities, and requires the modelling the hierarchical and temporal nature of human movement, together supporting the interpretation of craft activities.

The introduction of implicit scene understanding is particularly motivated by the complexity and variability of human movement in real-world craft settings. Skilled actions are not isolated events but are composed of recurring movement primitives that combine into actions and, ultimately, into complete activities or workflows. Ignoring this structure leads to systems that struggle to generalize across individuals, tools, or even craft domains. Through the modelling of a movement hierarchy, Task T3.3 aims to capture knowledge that supports both recognition and adaptation.

This deliverable presents the outcomes of Task T3.3 by introducing the framework for both explicit and implicit scene understanding. Quantitative and qualitative evaluations are provided to assess the robustness of scene reconstruction, followed by the analysis of hierarchical activity recognition and the adaptation through the captured knowledge to new, unseen data. Finally,

the deliverable demonstrates how the proposed methodology supports concrete CRAEFT applications, including personalized skill assessment and professional reconversion in craft domains.

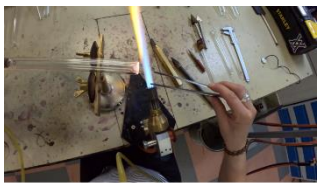
2. Motion capture of craft routines

As previously presented in the initial deliverable for this task, a dataset was created with a variety of craft professions, including glassblowing with pipe, glassblowing with blowtorch, marble carving, silversmithing, and porcelain pottery. The recordings took place in the first year of the project at the respective craft environment, in collaboration with experts in each craft, that are either responsible for teaching a craft, or for creating and promoting their craft through an association, or the family business.

Two more recordings were performed on the second year of the project, on the crafts of wood carving and tapestry. Concerning the wood carving profession, the recordings took place over two days in the town of Yecla, in Spain, in the workspace of an expert wood carver. According to the ethnographic protocol presented in WP1, the recordings started from an interview with the craftsperson, analyzing all the details of the process, the tools, the materials. After the interview the technical part of the recordings took place, the technical characteristics of which will be further analyzed below. The last step of the recording – also according to the established ethnographic protocol- was the part of the video elicitation, where the video recordings of the previous step where shown to the craftsperson, who in turn analyzed the craft routine, all the taken steps and decision processes involved, the used tools, as well as the behavior of the respective material. The same process was taken for the recordings of tapestry, where two days were also taken to follow the three aforementioned steps of the ethnographic protocol, in a tapestry workshop in the town of Aubusson, in France.

More specifically on the used sensors, two GoPro cameras were used, one in an egocentric (first-person) view, and one in an exocentric (third-person view). The egocentric view provided with the dexterous hand movements of the craft operators, along with the used tools, while the exocentric view captured the operator's ample body movements. Furthermore, two microphones, a contact one that captured the sound of the interaction between tools and materials, through their vibrations, along with a stereo microphone that recorded the sounds from the working environment, work objects, and materials, as well as the communication among different operators in collaboration, creating a multi-modal setup.

Table 1. The complete list with all the recordings performed during the first and second year of CRAEFT. Examples from the egocentric (top images) and the exocentric (bottom images) of each craft recording.

Glassblowing with pipe	Glassblowing with blowtorch	Marble carving
		
		

Silversmithing	Porcelain pottery	Tapestry	Wood carving
			
			

3. Explicit Scene understanding

Explicit scene understanding focuses on the analysis and the reconstruction of the different components of craft activities, such as the tools, the various materials, the environment and aims in providing a coherent and dynamic representation of the physical scene, despite the challenges introduced by egocentric viewpoints, continuous motion, and occlusions.

Unlike static reconstruction approaches, craft activities require scene representations that evolve over time and preserve interaction dynamics. Actions such as carving, shaping, or blowing glass involve continuous hand–tool–material interactions, where the relative motion between entities is often more informative than their absolute geometry. Explicit scene understanding in Task T3.3 therefore emphasizes dynamic representations that capture both spatial structure and temporal evolution.

This explicit representation is not intended to model skill or intent directly. Instead, it provides a physically grounded, interpretable description of the scene, enabling subsequent analysis steps such as decomposition, synthesis under occlusions, and the integration with implicit scene understanding. By separating observable scene structure from motion semantics, the methodology ensures modularity and interpretability, while remaining compatible with adaptive and hierarchical recognition strategies introduced later in this deliverable.

3.1. Explicit scene representation and synthesis

The first component of explicit scene understanding addresses the problem of representing dynamic craft scenes from egocentric video input. Egocentric viewpoints are challenging due to the continuous camera motion (mounted on the head of the human operator), and persistent occlusions caused by the hands and utilized tools. To address these challenges, Task T3.3 introduces a two-level scene representation that separates the global motion dynamics from the fine-grained parts of the craft scene.

Main part of this approach is the concept of a global motion scaffold, that provides a sparse, structured representation that captures the dominant motion patterns present in the scene. Rather than attempting to reconstruct all scene details directly, the scaffold acts as a stabilizer that encodes overall dynamics, (i.e. hand trajectories), and provides with a reference frame for subsequent reconstruction. This abstraction is particularly important in egocentric settings, where raw visual data is affected by rapid viewpoint changes and motion-induced noise. Building on this scaffold, the scene is reconstructed using 3D Gaussian Splatting (3DGS). In this representation, the scene is modelled as a collection of 3D Gaussian primitives, allowing for efficient rendering, smooth handling of occlusions, and high-fidelity visual synthesis. Compared to mesh-based or voxel-based representations, 3DGS offers a flexible compromise between visual quality and computational efficiency, making it well suited for dynamic and partially observable craft environments.

The use of the global motion scaffold alongside with 3D Gaussian Splatting provides a dynamic scene representation that preserves global coherence, as well as local detail. The scaffold constrains the overall motion and structure of the scene, while the Gaussian part captures fine-grained appearance information for hands, tools, materials, and background cues. This combination enables the reconstruction of complex hand–tool–material interactions over time, even when individual components are temporarily occluded.

Apart from supporting providing the reconstruction of the craft scene, this representation also supports scene synthesis. As hands, tools and material move, regions that were previously occluded in earlier frames may become visible in later frames. By leveraging this temporal consistency and the accumulated knowledge concerning the scene, our system can synthesize previously seen parts of occluded regions, providing a continuous representation of the craft scene. This forms the basis for steps in explicit scene understanding, including scene decomposition into semantic components (hands, tools, materials) and the quantitative evaluation of synthesis quality under occlusions.

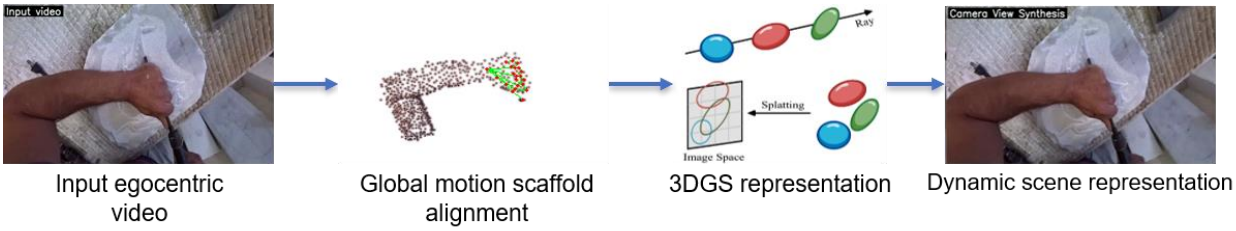


Figure 1. Overview of the dynamic scene reconstruction pipeline from egocentric craft video.

3.2. Dynamic Scene decomposition

Once a coherent dynamic representation of the craft scene has been established, the next step in explicit scene understanding consists in decomposing the scene into semantically meaningful components. In craft activities the tools, the materials, the background elements, as well as the hands of the crafts person and their collaborators show different motion characteristics, thus handling the scene as a single entity limits the ability of analyzing the interactions among the above and unveil their contribution to the overall craft activity. As such, dynamic scene decomposition addresses this limitation by separating the reconstructed craft scene into its elements, enabling the fine-grained analysis of both motion and the interaction patterns. In Task T3.3, decomposition is performed along three primary components of the craft scene: hands, tools, and materials, with the background treated as a supporting contextual layer.

Hands represent the primary source of motion and the main carrier of skill-related information. Their trajectories, coordination, and temporal consistency are critical indicators of expertise and are therefore isolated to support detailed motion analysis. On the other side, tools acts as the intermediaries between the craftsperson and the respective material (and its transformation). Materials are the elements of the scene that capture the outcome of this interaction and provide visual cues related to the task progression and quality. Apart from this, the performed decomposition improves robustness to occlusions, as when one component is hidden, its state can be inferred through its relationship with the other components of the scene. More specifically, through the observations of the craft scene over time, the system progressively reconstructs a complete view of the craft scene. In this process, the initial

During this process, an initial frame at time T=t is considered, where part of the scene is occluded by the hand. As the action unfolds and the hand, the tool and the material move, new frames at T=t+n reveal previously hidden regions. These later observed areas can work as a replacement for the ground truth, enabling the synthesis and evaluation of part of the craft scene that were not visible earlier. Through the reconstruction of occluded regions as they are revealed through the progress of a craft, the system can effectively preserve both visual plausibility and task relevance.

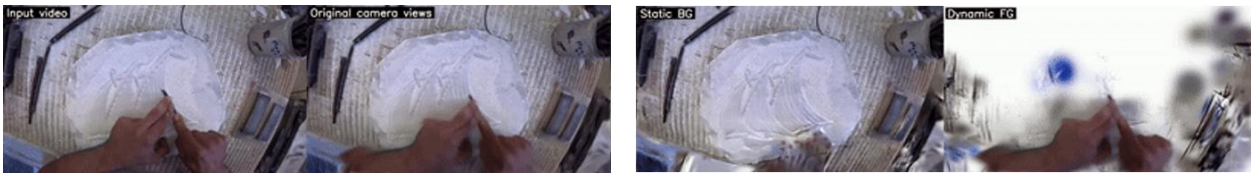


Figure 2. Illustration of occlusion handling through temporal observation. Regions occluded by the hand in an initial frame are progressively revealed in subsequent frames as the craft action unfolds, enabling synthesis and evaluation of previously hidden scene content

3.3. Quantitative Evaluation of Scene Synthesis in Craft Environments

Following the temporal reconstruction of occluded regions described in the previous section, it becomes necessary to establish a quantitative evaluation framework capable of assessing how faithfully the synthesized scene represents real craft interactions. In the context of craft analysis, evaluation is not limited to the visual perception of the quality of the image, but needs to be supported by metrics that prove the aforementioned logic of the interactions of the scene elements and the handling of occlusions. Thus, T3.3 deploys a multi-metric evaluation approach that combines both the numerical fidelity of the results, with perceptual and region-specific reliability measures.

Three complementing image quality measurements are utilized:

- Peak Signal-to-Noise Ratio (PSNR) assesses the pixel-level similarity between synthetic and ground-truth pictures. Higher PSNR values signify a more precise reconstruction of intricate visual features, which is especially pertinent in craft contexts for assessing the accuracy of material surfaces and tool edges.
- The Structural Similarity Index (SSIM) assesses perceptual and structural consistency by evaluating local texture and contrast patterns. Elevated SSIM values indicate that the generated image retains structural information crucial for discerning surface characteristics, such as carving marks in wood or marble, as well as material texture.
- The Number of Valid Pixels (Npix) indicates the dimensions of the assessed area. This metric offers statistical context, as assessment over broader areas enhances confidence that reported performance represents significant interaction zones rather than isolated segments.

Alongside these standard measurements, the Occlusion-disclosed Pixel Ratio (ORPR) quantifies the ratio of pixels that were initially occluded but subsequently disclosed and reconstructed with an error below a specified threshold. ORPR evaluates how effectively the system recovers hidden interaction regions that are important for analyzing manipulation quality and workflow continuity. The reference ranges shown in the evaluation table (Table 2) classify PSNR and SSIM values into qualitative categories (excellent, good, fair, poor). These ranges provide a scale for assessing reconstruction quality in practical craft scenarios. For example, PSNR values above 30 dB and SSIM values above 0.95 correspond to reconstructions that are almost identical to the observed reality, while lower ranges indicate increasing deviation and reduced reliability for skill analysis. Together, these metrics establish an evaluation protocol specifically tailored to craft environments.

Table 2. Overview of the evaluation framework used to assess synthesized craft scenes, including PSNR and SSIM reference quality ranges and the formulation of the Occlusion-Revealed Pixel Ratio (ORPR).

Metric	Excellent	Good	Fair	Poor
PSNR (dB)	>30	25–30	20–25	<20
SSIM (0–1)	>0.95	0.90–0.95	0.80–0.90	<0.80

3.4. Quantitative results

3.4.1. Marble carving

To evaluate the effectiveness of the proposed scene synthesis framework in real craft scenarios, quantitative experiments were conducted on all the captured crafts during the CRAEFT project, starting from marble carving. Marble carving is specifically characterized by continuous hand-tool contact and frequent occlusions of the working surface, thus the material and its transformation. The evaluation focuses on three regions of the scene, the region of the material, thus the visible marble surface, the fully occluded region and the entire synthesized scene. Each one of the three regions is assesses with the used of the aforementioned metrics, the PSNR, the SSIM and the Npix, allowing a differentiated analysis of reconstruction quality.

Figure 3. Qualitative evaluation of scene synthesis for marble carving

Table 3. Object-wise and occlusion-region metrics evaluating reconstruction quality in a marble carving task. The table reports PSNR, SSIM, and evaluated pixel counts (Npix) for visible object regions, fully occluded regions, and the full synthesized scene.

	Region	PSNR	SSIM	Npix
Marble carving	Marble object region	25.93	0.963	19508
	Fully occluded region	12.24	0.868	24078
	Full-scene synthetic quality	29.72	0.980	409920

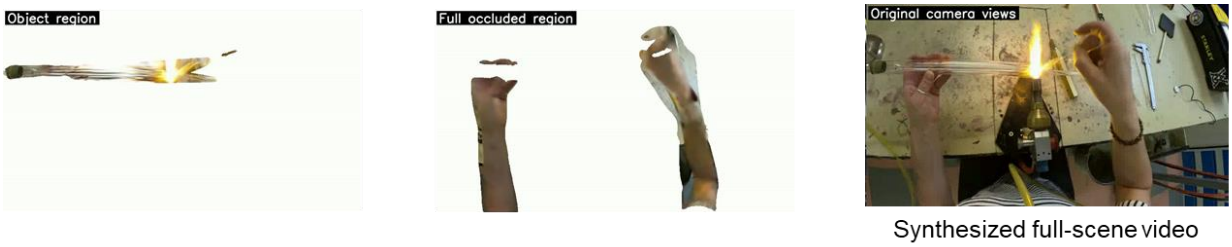
According to the results presented in Table 3, for the visible marble region, the reconstruction algorithm achieves a PSNR of 25.93 dB and an SSIM of 0.963 over 19.508 pixels. The high SSIM value, indicates that structural features of the marble surface, such as the carving traces, are preserved with high fidelity. Although the PSNR falls within the "good" range rather than the highest category, the result remains sufficient to support reliable interpretation of surface geometry in a craft context. This means that the reconstructed visible marble surface retains the visual cues necessary to analyze carving precision and progression. The fully occluded marble region seems to be more challenging, with the PSNR decreasing to 12.24 dB, showing the difficulty of reconstructing areas that were initially hidden from view. However, the SSIM remains at 0.868 across 24,078 pixels, indicating that while exact pixel values differ, the reconstructed region preserves meaningful structural consistency. For craft analysis, this suggests that the synthesized occluded surface captures the general shape and continuity of the material, even if fine color details are less accurate. The Occlusion-Revealed Pixel Ratio (ORPR) further explains this behavior. With a threshold $\tau = 20$, the ORPR value is 3.57%, meaning that a subset of previously occluded pixels is reconstructed with error below the defined tolerance. This percentage demonstrates that temporally revealed regions provide reliable context for validating occlusion recovery. In craft terms, this confirms that the system can recover at least part of the hidden working surface with quantifiable accuracy. For the full synthesized scene**,** the reconstruction achieves a PSNR of 29.72 dB and an SSIM of 0.980 over a large evaluation area of 409,920 pixels, indicating strong global consistency. The high SSIM suggests that the integrated scene, including hands, tools, and background, remains visually coherent and stable, which is essential for maintaining interpretability across extended craft sequences.

The above results highlight a property of the proposed framework, where while reconstruction of fully occluded regions remains challenging, the system has the ability to preserves structural continuity and at levels sufficient for analyzing craft routines. Visible object regions retain high fidelity, and even occluded areas maintain structural cues, ensuring that synthesized marble carving scenes remain suitable for tasks such as gesture analysis, skill assessment, digital reenactment.

3.4.2. Glass blowing with blowtorch

The second evaluation scenario considers glass blowing sequences, that are characterized by rapid hand movements, luminous and reflective material states, and complex tool–material interactions, due to the nature of the glass that can be in both liquid and stable form and the fire. Unlike marble carving, glass

Figure 4. Qualitative evaluation of scene synthesis for glass blowing with blowtorch.



blowing introduces additional visual challenges such as reflections, heat-induced color variation, and highly dynamic motion.

The evaluation again distinguishes between the object region, the fully occluded region**,** and the fully synthesized scene, enabling a detailed examination of reconstruction quality across interaction zones.

Table 4. Quantitative evaluation of scene synthesis for glass blowing with blowtorch.

	Region	PSNR	SSIM	Npix
Glass Blowing	Glass object region	24.69	0.978	12415
	Fully occluded region (Table)	14.81	0.97	16595
	Full-scene synthetic quality	29.2	0.89	409920

For the glass object region, the reconstruction achieves a PSNR of 24.69 dB and an SSIM of 0.978 over 12,415 pixels. The very high SSIM indicates strong preservation of the structural features, even though the specific craft is characterized by the visual complexity of the glass and the reflections that its transformations caused. This result suggests that the synthesized representation maintains the necessary visual structure for interpreting the position of the tool and the shaping of the glass as the used material. The slightly lower PSNR compared to marble carving reflects the difficulty of reconstructing reflective surfaces rather than a failure of structural representation. The fully occluded region, corresponding to the worktable area, achieves a PSNR of 14.81 dB and an SSIM of 0.970 across 16,595 pixels. Although the PSNR indicates noticeable pixel-level deviation, the SSIM remains high, showing that the reconstructed region preserves structural coherence and visual similarity. For craft analysis, this is particularly significant, as even when exact pixel values differ, the recovered workspace retains spatial consistency, allowing the interpretation of hand trajectories and tool placement relative to the environment.

The Occlusion-Revealed Pixel Ratio (ORPR) with threshold $\tau = 20$ reaches 13.68%, which is much higher than in the marble carving scenario. This suggests that a larger proportion of previously hidden pixels are reconstructed within acceptable error bounds. From a craft perspective, this shows that the dynamic motions inherent in glass blowing expose occluded regions more effectively, enabling stronger validation of reconstructed surfaces through temporal observation. For the full synthesized scene, the system achieves a PSNR of 29.2 dB and an SSIM of 0.89 over 409,920 pixels. While the SSIM is lower than in the marble carving case, it still indicates acceptable consistency given the visual complexity of the environment. Thus, structural fidelity is preserved in both visible and occluded regions, and the relatively high ORPR confirms that temporal synthesis can reliably recover hidden workspace areas.

3.4.3. Silversmithing

The third evaluation scenario examines a silversmithing sequence, a craft characterized by fine manual manipulations. Unlike marble carving or glass blowing, silverware making involves small-scale objects and extremely precise hand movements performed usually over a work table. This environment presents a difficult test for scene synthesis, as accurate reconstruction must preserve both fine object details and the spatial organization of the workspace.

As in the previous scenarios, evaluation is performed separately on the object region, the fully occluded region, and the entire synthesized scene, allowing a region-specific interpretation of reconstruction performance (Figure 5).

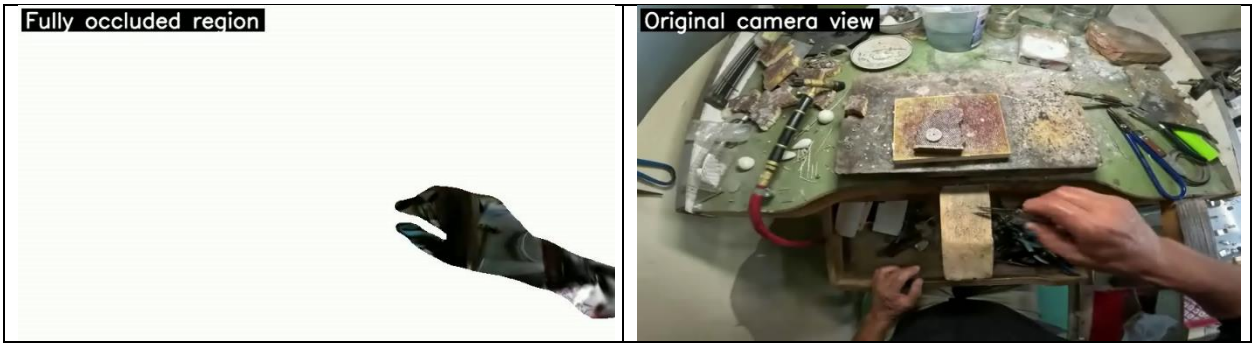


Figure 5. Qualitative evaluation of scene synthesis for silversmithing.

Table 5. Quantitative evaluation of scene synthesis for silversmithing.

	Region	PSNR	SSIM	Npix
Silverware Making	**Silverware object region**	**24.31**	**0.981**	**423**
	Fully occluded region (Table)	**14.65**	**0.958**	**26164**
	Full-scene synthetic quality	**30.21**	**0.92**	**409920**

For the silversmithing object region, the reconstruction achieves a PSNR of 24.31 dB and an SSIM of 0.981 over 423 pixels. Even though the evaluated area is relatively small, the very high SSIM indicates an "excellent" preservation of the scene's features, suggesting that the system accurately captures the appearance of small metallic objects in the scene, while the small Npix value reflects the small size of the focal object rather than a reduced evaluation reliability. Despite the limited area, structural integrity is preserved, illustrating the framework's capacity to accommodate delicate craft elements. The entirely obscured area, primarily representing the table surface, attains a PSNR of 14.65 dB and an SSIM of 0.958 over 26,164 pixels. Despite observable pixel-level discrepancies, the elevated SSIM indicates that the reconstructed workspace maintains its overall structural integrity. The Occlusion-Revealed Pixel Ratio (ORPR) with a threshold of $\tau = 20$ attains 14.67%, the highest among the assessed situations. This signifies that a significant percentage of the previously occluded pixels is reconstructed within acceptable error limits, indicating that repetitive hand movements and changing views successfully expose occluded workspace areas, facilitating robust temporal validation of reconstructed surfaces. The complete synthesized picture attains a PSNR of 30.21 dB and an SSIM of 0.92 across 409,920 pixels, demonstrating robust consistency throughout the entire workspace.

All in all, the silversmithing results show that the proposed framework performs robustly in fine-grained, cluttered craft settings. Structural fidelity is maintained even for small objects and densely populated workspaces, and the high ORPR confirms effective recovery of occluded regions, properties essential for analyzing precision gestures and for preserving the spatial context of professional craft benches.

3.4.4. Tapestry

The fourth evaluation scenario considers the craft of tapestry, an activity characterized by repetitive, fine hand motions, and rapid localized interactions with the loom. Unlike heavy material shaping or luminous manipulation tasks, tapestry making emphasizes precision coordination and rapid micro-gestures within a visually textured environment. This makes it an important test use-case for evaluating reconstruction performance under conditions of fast hand motion and high structural detail. As in the previous evaluations, reconstruction quality is analyzed separately for the object region, the fully occluded region, and the full synthesized scene, enabling targeted interpretation of performance in interaction-critical areas.

Figure 6. Comparison between synthesized foreground reconstruction, occlusion handling, and full-scene synthesis during a tapestry making task

Table 6. Object-wise and occlusion-region reconstruction metrics for a tapestry making task, reporting PSNR, SSIM, and evaluated pixel counts (Npix) for visible object regions, fully occluded regions, and the full synthesized scene.

	Region	PSNR	SSIM	Npix
Tapestry	Tool object region	20.11	0.954	1895
	Fully occluded region (Table)	13.49	0.964	42117
	Full-scene synthetic quality	20.06	0.69	409920

The reconstruction for the tool/object region attains a PSNR of 20.11 dB and an SSIM of 0.954 across 1,895 pixels. Although the PSNR is inferior to those of previously assessed crafts, the SSIM demonstrates robust preservation of structural similarities. This indicates that the synthesized representation preserves the geometric arrangement of threads and tools, despite the motion blur caused by fast hand movements. The recreated area is visually discernible for assessing weaving accuracy and hand placement. The completely occluded area attains a PSNR of 13.49 dB and an SSIM of 0.964 over 42,117 pixels. Notwithstanding reduced pixel-level fidelity, the elevated SSIM suggests that the reconstructed table and loom structure retain perceptual consistency with the original scene. In tapestry operations, maintaining the spatial configuration of the loom is crucial for comprehending gesture alignment and repetitive motion patterns.

The large evaluated region further strengthens the reliability of this structural reconstruction. The Occlusion-Revealed Pixel Ratio (ORPR) analysis shows a convergence quality of 20.06 dB, with approximately 11.4% of previously occluded pixels reconstructed within the acceptable error threshold. This indicates that temporal synthesis effectively recovers a meaningful subset of hidden workspace regions, supporting stable interpretation of hand–loom interactions.

For the full synthesized scene, the reconstruction maintains overall structural coherence despite the rapid motion inherent to weaving actions. The global SSIM values confirm that the integrated scene remains visually close to reality. Minor blurring effects occur due to rapid hand motions; nonetheless, they do not impede the continuity necessary for evaluating repetitive craft gestures. The findings of tapestry production indicate that the proposed framework is effective in high-frequency precision crafts marked by swift micro-movements and intricate visual textures. Structural integrity is maintained in both visible and concealed areas, guaranteeing that rebuilt sceneries are appropriate for examining weaving rhythm, alignment, and gesture consistency. These properties are essential for training evaluation and for documenting fine motor skills in textile crafts.

3.4.5. Wood carving

The next evaluation scenario examines a wood carving sequence, a craft process characterized by continuous tool–material contact, textured organic surfaces, and rapid localized hand movements. Wood carving presents a challenging reconstruction environment due to irregular material geometry and frequent occlusions caused by close hand positioning. These properties make it an important benchmark for evaluating the robustness of scene synthesis in material-shaping crafts.

As in the previous scenarios, evaluation is conducted separately for the tool/object region, the fully occluded region, and the full synthesized scene, allowing region-specific interpretation of reconstruction fidelity.

Figure 7. Quantitative evaluation of scene synthesis for wood carving.

Visual examples of synthesized object regions, fully occluded regions, and full-scene reconstruction (top), together with object-wise and occlusion-region evaluation metrics (bottom) reporting PSNR, SSIM, and evaluated pixel counts (Npix) for the wood carving task. As shown in Table 3.12, the tool/object region achieves a PSNR of 20.70 dB and an SSIM of 0.993 over 2,088 pixels. The extremely high SSIM indicates excellent preservation of structural features, meaning that the reconstructed carved surface maintains realistic geometric continuity. This degree of structural accuracy facilitates precise interpretation of carving strokes and tool placement, even in the presence of pixel-level variations. Table 3.12 indicates a PSNR of 14.65 dB and an SSIM of 0.915 for the completely occluded region, based on 2,021 pixels. Despite

a reduction in pixel-level fidelity caused by occlusion, the SSIM indicates that the restored workspace maintains its perceptual structure. The convergence analysis reveals an average reconstruction quality of 26.45 dB, reflecting a variation of roughly 5.17% from the reference threshold. This indicates that a significant amount of previously concealed pixels is restored within acceptable error margins, confirming the efficacy of temporal synthesis in restoring occluded interaction areas. The complete synthesized picture attains a PSNR of 26.45 dB and an SSIM of 0.857 across 409,920 pixels, as detailed in Table 3.12. These values demonstrate globally consistent reconstruction despite fast motion and material intricacy. Although rapid hand motions produce considerable blur, the overall scene remains structurally coherent and comprehensible.

Table 7. Object-wise and occlusion-region reconstruction metrics for wood carving, reporting PSNR, SSIM, and evaluated pixel counts (Npix) for visible object regions, fully occluded regions, and the full synthesized scene.

	Region	PSNR	SSIM	Npix
Wood Carving	**Tool object region**	**20.70**	**0.993**	**2088**
	Fully occluded region (Table)	**14.65**	**0.915**	**20021**
	Full-scene synthetic quality	**26.45**	**0.857**	**409920**

The wood carving results indicate that the framework performs effectively in material-removal techniques on uneven surfaces. The significant structural similarity in the object region guarantees that carving trajectories and surface evolution are discernible, while the stable restoration of occluded areas maintains the context of the workspace. This equilibrium is crucial for evaluating technique, tool mastery, and advancement in sculptural arts.

3.4.6. Porcelain making

The concluding supplementary evaluation scenario investigates a porcelain production process, a craft endeavor defined by deformable materials, rotational symmetry, and ongoing manual shaping. Porcelain craftsmanship presents unique reconstruction issues because to uniform surface textures and nuanced geometric differences that must be maintained to faithfully depict shaping techniques. These characteristics make porcelain making an important test case for evaluating scene synthesis in crafts involving fine surface deformation and continuous material manipulation. As in the previous scenarios, reconstruction performance is evaluated separately for the tools/object region, the fully occluded region, and the full synthesized scene, allowing targeted interpretation of fidelity across interaction zones.

Table 8. Object-wise and occlusion-region reconstruction metrics for porcelain making, reporting PSNR, SSIM, and evaluated pixel counts (Npix) for visible object regions, fully occluded regions, and the full synthesized scene.

	Region	PSNR	SSIM	Npix
Porcelaine Making	**Tools object region**	**15.38**	**0.971**	**7944**
	Fully occluded region	**16.38**	**0.984**	**57327**

	Full-scene synthetic quality	**22.48**	**0.595**	**230400**

Figure 8. Visual examples of synthesized fully occluded regions, and full-scene reconstruction in porcelain making

As reported in Table 8, the tools/object region achieves a PSNR of 15.38 dB and an SSIM of 0.971 over 7,944 pixels. Although the PSNR is comparatively low due to the difficulty of reconstructing smooth, reflective porcelain surfaces, the high SSIM confirms strong preservation of structural geometry. In craft terms, this indicates that the reconstructed surface maintains the continuous curvature required to interpret shaping motions and rotational symmetry during the forming process. For the fully occluded region, Table 3.13 reports a PSNR of 16.38 dB and an SSIM of 0.984 across 57,327 pixels. The extremely high SSIM demonstrates that the recovered workspace structure is perceptually very close to reality. This is particularly important in porcelain making, where consistent spatial reference to the working surface is necessary for understanding hand positioning and shaping trajectories.

The convergence analysis indicates an average reconstruction quality of 22.48 dB, corresponding to approximately 9.43% deviation from the reference threshold. This suggests effective temporal recovery of previously hidden pixels, validating the synthesis mechanism under conditions of continuous hand occlusion. For the full synthesized scene, the reconstruction achieves a PSNR of 22.48 dB and an SSIM of 0.595 over 230,400 pixels, as shown in Table 8. While the SSIM is lower than in other evaluated crafts, the scene remains visually consistent and interpretable. The reduced perceptual similarity reflects the inherent difficulty of reconstructing smooth deformable materials undergoing rapid motion rather than instability in the reconstruction pipeline.

The porcelain making results demonstrate that the framework remains applicable to continuous surface-shaping crafts involving homogeneous materials. Structural preservation in both visible and occluded regions ensures that shaping gestures and material evolution remain observable. Despite increased visual ambiguity due to smooth textures, the reconstructed scenes retain sufficient coherence to support analysis of rotational symmetry, hand coordination, and forming technique.

3.4.7. Glassblowing with pipe

This evaluation scenario examines a glassblowing process utilizing a pipe sequence, a craft activity that incorporates elongated equipment, high-temperature substances, and significant arm actions. In contrast to the prior glass blowing scenario, the inclusion of an elongated pipe presents more issues concerning range of motion, dynamic balance, and occlusion patterns. The luminescent material, thus the liquid glass at the pipe tip generates strong illumination gradients that hinder visual reconstruction.

These factors make this scenario a demanding test case for scene synthesis in high-dynamic thermal craft environments. As in the previous evaluations, performance is assessed separately for the tool/object region, the fully occluded region, and the full synthesized scene, allowing region-specific analysis of reconstruction fidelity.

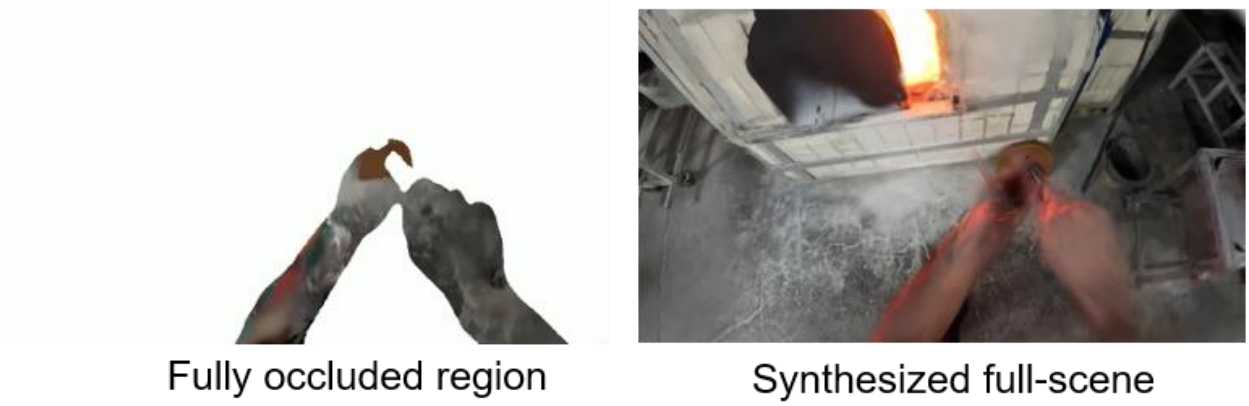


Figure 9. Visual examples of synthesized fully occluded regions, and full-scene reconstruction in glassblowing with pipe.

As shown in Table 9, the tool/object region achieves a PSNR of 15.90 dB and an SSIM of 0.983 over 3,791 pixels. The high SSIM indicates strong preservation of structural geometry despite intense illumination and rapid movement. From a craft perspective, this means that the reconstructed glowing glass tip and pipe alignment remain visually interpretable, supporting analysis of rotational control and shaping precision.

For the fully occluded region, Table 9 reports a PSNR of 14.10 dB and an SSIM of 0.943 across 49,658 pixels. Although pixel-level accuracy decreases due to occlusion and lighting variability, the SSIM confirms that the workspace remains structurally coherent. This ensures that hand trajectories and pipe positioning can be interpreted relative to a stable environment. The convergence analysis indicates an average reconstruction quality of 27.76 dB, corresponding to approximately 12.17% deviation from the reference threshold. This suggests effective temporal recovery of previously hidden regions even under extreme lighting and motion conditions. For the full synthesized scene, the reconstruction achieves a PSNR of 27.76 dB and an SSIM of 0.838 over 409,920 pixels. These values indicate globally consistent reconstruction despite the complexity of thermal illumination and extended motion. The integrated scene remains visually stable and suitable for analyzing large-scale gesture coordination.

The glass blowing with pipe results demonstrate that the framework performs robustly in high-dynamic thermal crafts involving extended tools and strong illumination contrasts. Structural preservation in both visible and occluded regions supports analysis of rotational symmetry, arm coordination, and tool balance.

Such capabilities are specifically important for the documentation and further study of the various glass shaping techniques.

Table 9. Object-wise and occlusion-region reconstruction metrics for glassblowing with pipe, reporting PSNR, SSIM, and evaluated pixel counts (Npix) for visible object regions, fully occluded regions, and the full synthesized scene.

	Region	PSNR	SSIM	Npix
Glass blowing with pipe	Tool object region	15.90	0.983	3791
	Fully occluded region	14.10	0.943	49658
	Full-scene synthetic quality	27.76	0.838	409920

3.5. Quantitative evaluation across craft scenarios

3.5.1. Marble carving

While quantitative metrics provide an objective assessment of reconstruction fidelity, qualitative evaluation is essential for interpreting how synthesized scenes support the visual analysis of craft actions. In marble carving, qualitative inspection focuses on whether the reconstructed scene preserves the perceptual cues required to understand hand positioning, interactions with the tool, and material transformation over time. The qualitative evaluation decomposes the synthesized scene into multiple visual layers that correspond to functional elements of the craft interaction. These layers allow direct inspection of how the dynamic representation separates and reconstructs hands, tools, and background components.

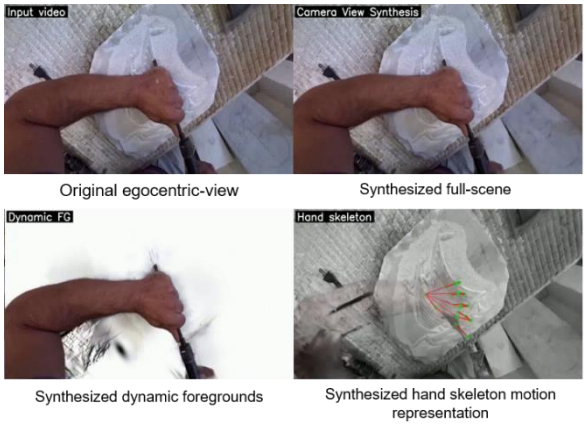


Figure 10. Visual comparison between original egocentric input and synthesized outputs, including full-scene reconstruction, dynamic foreground separation, hand skeleton motion representation, isolated hand and tool motion, and reconstructed static background.

The original egocentric-view video and the synthesized full-scene reconstruction demonstrate that the global visual structure of the marble carving task is preserved. The synthesis view preserves stable spatial relationships among the hand, tool, and marble surface, facilitating the comprehension of the carving

process, while the dynamic foreground visualization distinguishes moving items from the background. This division in marble carving emphasizes the trajectory and rhythm of the carving process, facilitating the observation of recurring movements and contact patterns that signify technique and proficiency. The skeletal depiction of hand motion abstracts hand movement into a structured kinematic framework. This depiction highlights collaborative coordination and the continuity of motion, serving as a link between explicit visual reconstruction and implicit motion analysis.

For craft interpretation, this abstraction allows evaluators to examine posture stability and gesture consistency independently of surface texture. The isolated hand movement and tool motion visualizations further disentangle interaction components. By isolating these components, the method facilitates a focused examination of the tool's manipulation by the hand and its interaction with the marble surface. This analysis is especially beneficial for discerning nuanced distinctions in technique, such variations in pressure, angle, or fluidity of movement. The rebuilt static background verifies that the workspace maintains visual stability amid dynamic interactions. A consistent background is important for maintaining spatial reference and ensuring that motion analysis is not confounded by reconstruction artifacts.

Taken together, these qualitative visualizations demonstrate that the synthesized marble carving scene supports multi-level interpretation of craft actions. The capacity to examine full-scene coherence, separate dynamic elements, and abstract motion structure offers a thorough visual framework for examining carving processes. This stratified representation directly facilitates applications like training evaluation and gesture analysis, where comprehending both overarching workflows and detailed movements is crucial.

3.5.2. Glass blowing

The qualitative evaluation of the glass blowing with blowtorch sequence examines whether the synthesized scene manages to preserve the cues required to interpret coordinated craft actions involving heated and deformable materials. As in the previous scenario, the qualitative analysis decomposes the reconstructed scene into multiple visual layers corresponding to functional components of the interaction.

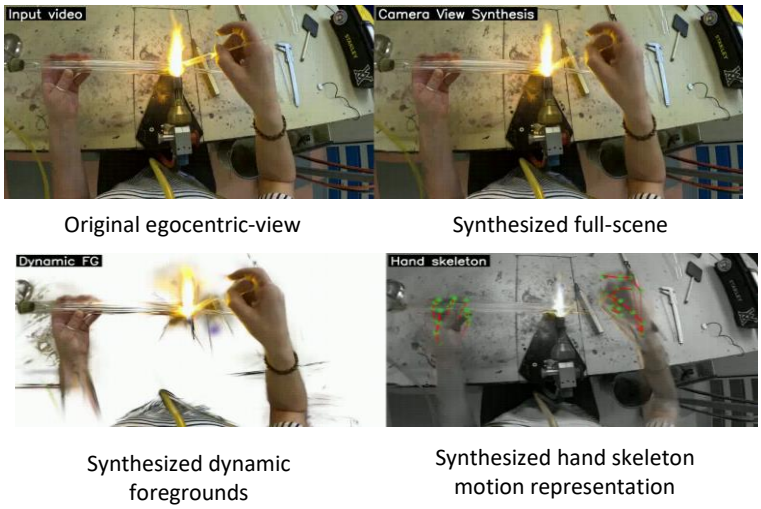


Figure 11. Visual comparison between original egocentric input and synthesized outputs, including full-scene reconstruction, dynamic foreground separation, hand skeleton motion representation, isolated hand and glass motion, and reconstructed static background.

The comparison between the original egocentric video and the synthesized full-scene reconstruction shows that the global dynamics of the glass blowing process are preserved. The reconstructed scene manages to maintain stable spatial relationships despite the changing visual conditions due to reflections, the changing state of the material from solid to liquid and to solid again. The dynamic foreground visualization highlights the motion of the hands and the glass object against the static workspace. In glass blowing, this separation makes it possible to observe the rhythm and fluidity of movement that characterize skilled handling of molten material. The hand skeleton motion representation provides a kinematic abstraction of gesture execution. In the context of glass blowing, this representation emphasizes coordination between both hands and the stability of posture during continuous shaping. This abstraction facilitates the examination of gesture smoothness and synchronization, which are important indications of skill development. The visualizations of the hand and glass movements explain the relationship between the artist and the material. The technology facilitates direct examination of the evolution of deformation and placement over time by isolating the glass object from the background. This is especially significant in glassblowing, because slight alterations in movement directly affect the object's ultimate shape. The reconstructed static background indicates that the workspace maintains visual consistency throughout the sequence. A constant background offers a consistent reference frame, facilitating precise interpretation of motion trajectories and minimizing ambiguity in gesture analysis. The qualitative findings affirm that the combined glassblowing scenes maintain the visual coherence essential for interpreting highly dynamic craft operations. The layered visualization facilitates the examination of both overarching workflows and detailed gesture coordination, providing a thorough representation appropriate for training analysis and digital reenactment. The capacity to uphold structural integrity under adverse visual conditions underscores the resilience of the suggested scene interpretation framework.

3.5.3. Silversmithing

The qualitative evaluation of silversmithing focuses on the system's ability to preserve perceptual clarity and structural coherence in fine-grained, precision-oriented craft workspaces. As in the previous qualitative evaluations, the reconstructed scene is analyzed through multiple visual layers corresponding to functional components of the craft interaction.

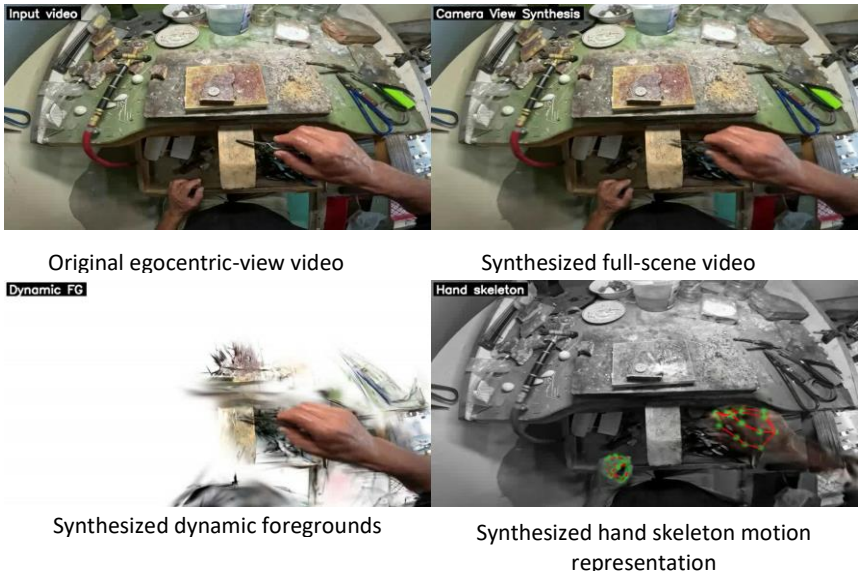


Figure 12. Visual comparison between original egocentric input and synthesized outputs, including full-scene reconstruction, dynamic foreground separation, hand skeleton motion representation, isolated hand and tool motion, and reconstructed static background in silversmithing

The comparison between the original egocentric video and the synthesized full-scene reconstruction demonstrates that the overall structure of the silverware workspace is preserved. Despite the presence of numerous small objects and tools, the reconstructed scene maintains stable spatial organization. This stability is important for interpreting the way that craftspeople navigate complex workbenches during precision tasks, while the dynamic foreground visualization isolates moving elements from the static environment. This separation, specifically in silversmithing, highlights the fine motor gestures involved in shaping and adjusting small components. The ability to track these movements without merging them into background clutter confirms that the dynamic representation preserves motion saliency in dense environments. The hand skeleton motion representation abstracts subtle hand articulations into a structured kinematic model. For precision crafts, this abstraction is particularly valuable, as it enables the inspection of micro-adjustments in posture and coordination that are difficult to perceive directly in raw video. The isolated hand and tool motion visualizations clarify how tools are manipulated within the workspace. By separating tools from surrounding objects, the system allows the targeted examination of interaction patterns, including grip stability and trajectory smoothness. This separation is critical for assessing skills in crafts where small variations in motion can significantly affect outcomes. The reconstructed static background confirms that the workspace remains visually coherent despite ongoing manipulation. A stable background reference ensures that motion analysis remains interpretable and that tool placement relative to the bench can be consistently evaluated. These qualitative results demonstrate that the proposed framework supports the visualization of craft activities, even in cluttered environments, with the layered decomposition allowing the simultaneous inspection of the global workspace organization and the execution of gestures.

4. Implicit scene understanding

While explicit scene understanding reconstructs the visible structure of craft environments, it does not by itself explain how actions are organized, how skills are expressed, or why certain motion patterns correspond to expertise. Craft activities are inherently hierarchical, meaning that complex workflows emerge from the coordinated composition of simpler movement units. Capturing this hierarchy is essential for interpreting the semantics of human motion and for linking visual reconstruction to skill analysis. Thus, implicit scene understanding focuses on modelling the structure of human movement rather than its visual appearance, which is particularly important in crafts since expertise and skills are often encoded in subtle temporal patterns, such as rhythm and hands coordination, elements that are not directly observable from the visual cues.

4.1 Hierarchical organization of craft movements

Human craft actions can be described through a three-level hierarchy:

- **Movement primitives**, representing motion units such as basic postures or joint-level motions.
- **Actions**, formed by various combination of movement primitives corresponding to meaningful task steps, without being merely groups of primitives, but also involve a level of expertise, different for every craft and also craftsperson.
- **Activities**, representing complete workflows composed of multiple actions.

This hierarchical decomposition of movement aligns with how craft knowledge is transmitted in practice, where craftspeople first learn the main gestures of the respective craft and later integrate them into full craft procedures. Modelling movement at these multiple scales allows the system to capture both local gesture structure and global workflow organization.

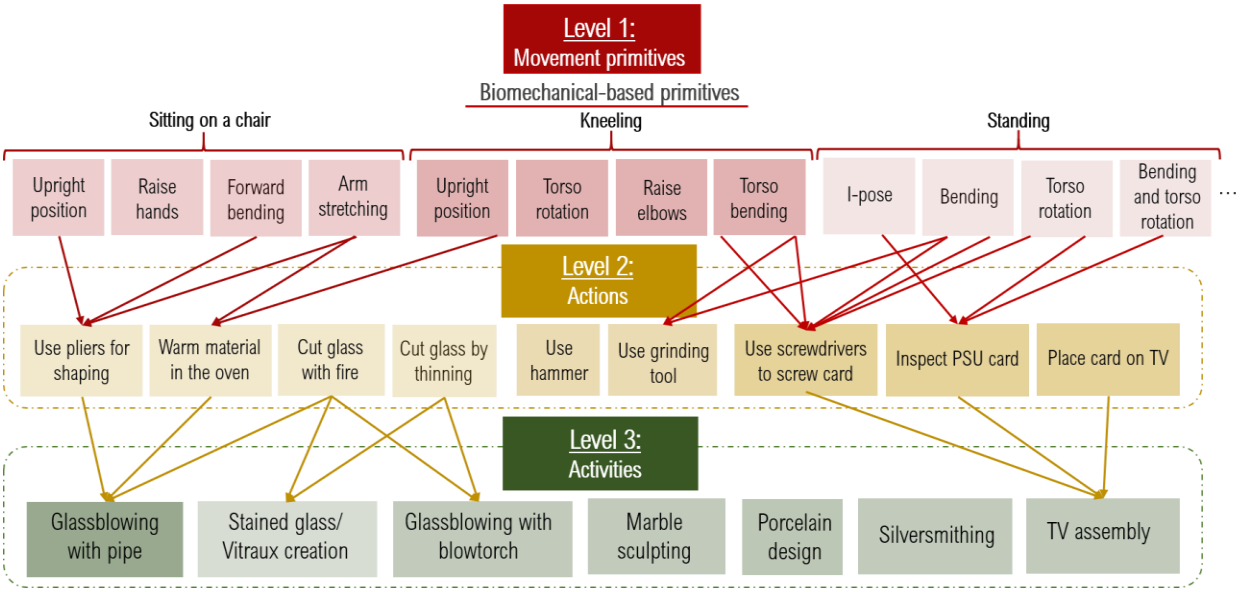


Figure 13. Illustration of the movement hierarchy linking primitives, actions, and activities, and its role in improving recognition performance through shared representations.

Experimental analysis demonstrates that incorporating the primitive level significantly improves recognition at higher hierarchical scales, as further shown in the previous version of this deliverable. In summary, during an ablation test, it was noticed that when movement primitives are removed from the learning process, action recognition accuracy decreases by approximately 3%, while activity recognition drops by 13%, indicating that low-level motion structure provides essential information for interpreting more complex craft movements.

From a craft perspective, this finding confirms that fine-grained gestures are foundational to understanding complete activities. Ignoring primitive movements reduces the system's ability to capture the building blocks of skilled performance, leading to weaker interpretation of complex tasks. The hierarchical approach to implicit scene understanding enables a transition from visual reconstruction to skill-aware interpretation. By modelling movement at multiple levels, the framework captures reusable patterns that generalize across individuals and craft domains. This capability is critical for applications such as training evaluation and professional adaptation, where understanding how a gesture is performed is as important as recognizing which task is being executed. While hierarchical implicit scene understanding provides a structured representation of craft movements, real-world deployment requires an additional capability, thus a controlled adaptation to new users, tasks, and craft environments under data scarcity. Data scarcity is specifically found in crafts and other manual professions, due to different reasons that include the availability of many different users to be captured and the confidentiality that some of those professions require. Focusing again on the adaptation to new users and environments, craft settings are inherently variable, with differences in execution style, tools, and working conditions. A recognition system must therefore adapt selectively, without destabilizing previously learned knowledge.

To address this challenge, Task T3.3 integrates a forecast-driven Meta-Learning adaptation mechanism built on top of the hierarchical Multi-Task Learning backbone presented in the initial version of this deliverable. In this case, rather than adapting continuously to every new observation of crafts, the system treats adaptation as a temporally controlled decision process, guided by predictions of its own learning dynamics.

4.2.1 Hierarchical Multi-Task Backbone as Adaptation Prior

The adaptive framework starts from a hierarchical Multi-Task Learning model that jointly learns movement primitives, craft actions, and activities and exchanges knowledge among them. By sharing representations across these levels, the backbone encodes an embedding of craft motions, reflecting how gestures compose into craft routines. During this Multi-Task Learning phase, the system also learns the temporal evolution of its loss dynamics through an autoregressive state-space formulation. These learned dynamics are stored as priors and reused during the adaptation process presented below. This means that the model does not only learn *what* to recognize, but also develops an internal expectation of *how the learning of a craft should normally evolve*.

4.2.2 Forecast-Driven Adaptation Control

During deployment, new craft data arrive sequentially in small chunks that simulate realistic acquisition events. For each chunk, the system predicts the expected evolution of its total loss using the stored autoregressive priors. It then compares this forecast with the observed loss on the incoming data. If the deviation between predicted and observed behavior exceeds a learned threshold, the system interprets the new data as incompatible with its existing knowledge and triggers adaptation to new craft data. Otherwise, it remains stable, avoiding unnecessary updates. This mechanism, referred to as Stop–Go control, ensuring that adaptation is event-driven rather than continuous.

This strategy is particularly suited to craft environments. Many incoming observations correspond to routine variations that should not destabilize the model, but just provide information, without affecting its current performance. Thus, selective adaptation allows the system to focus learning effort on novel or informative craft data, improving both stability and data efficiency.

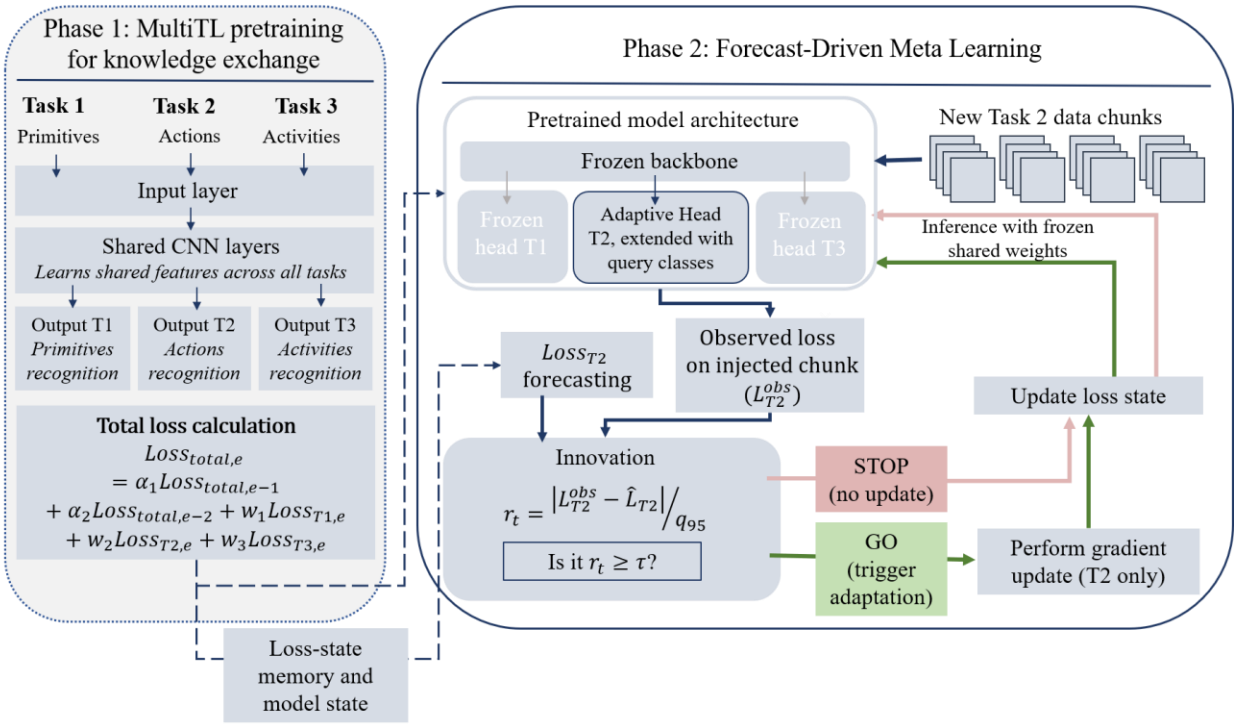


Figure 14. Integration of hierarchical Multi-Task Learning with a forecasting-based Meta-Learning controller. The system predicts expected loss evolution and triggers adaptation only when observed behaviour deviates significantly, enabling event-driven learning. The left part with the knowledge exchange among the three level of human movement hierarchy in crafts is analysed in the previous deliverable with T3.3.

4.2.3 Stability and Knowledge Retention in crafts

A central concern in adaptive systems, such as this one is catastrophic forgetting, where learning new gestures degrades previously acquired knowledge of the deployed Machine Learning model. The proposed framework mitigates this risk through two mechanisms:

1. Selective adaptation, which limits updates to statistically justified events.

2. Distillation constraints, which preserve compatibility with previously learned gesture representations.

Together, these mechanisms allow the system to incorporate and learn new craft gestures while maintaining stable performance on the existing ones. Experiments on different craft datasets demonstrate that forecasting-guided adaptation achieves higher accuracy and smoother learning trajectories than continuous fine-tuning or random update policies, particularly under few-shot conditions. Also, from a craft perspective, this adaptive strategy enables recognition systems to evolve alongside artisans. The model can personalize itself to individual operators or new workflows without requiring large retraining datasets or external pretraining. This capability is essential for the practical deployment in workshops where data collection is limited, while variability is high. Within the broader CRAEFT framework, forecast-driven adaptation acts as a control layer that connects implicit movement understanding to real-world usability. The hierarchical backbone provides semantic structure, while the Meta-Learning controller regulates how this structure evolves over time. This combination supports scalable deployment in craft settings, where systems must remain both interpretable and adaptable.

5. Perspectives and future directions

The methodological framework developed in this deliverable establishes a foundation for explicit and implicit scene understanding in craft environments. Beyond its current experimental validation, the framework opens several forward-looking perspectives for craft training, workforce development, and adaptive human movement analysis. These perspectives are not presented as completed applications, but as envisioned directions that extend the present work toward practical deployment and long-term research evolution.

5.1 Personalized Skill Assessment in Craft Training

One envisioned application of the proposed scene and activity monitoring framework concerns personalized skill assessment in craft training environments. Traditional evaluation of craft expertise relies on subjective observation by experienced craftspeople. While this approach captures tacit knowledge, it is difficult to standardize, reproduce, or scale across institutions. The integration of explicit and implicit scene understanding offers a computational basis for supporting objective and fine-grained assessment of craft performance. Explicit scene reconstruction provides a temporally consistent representation of the physical workspace, enabling detailed visualization of hand–tool–material interactions. In parallel, implicit hierarchical modelling interprets movements in terms of primitives, actions, and activities, linking observable gestures to semantically meaningful units of skill. The combination of these two layers makes it possible to compare trainee motion patterns with reference executions encoded in the hierarchical model and to identify deviations in rhythm, coordination, and movement structure. Such deviations reflect not only what action is performed, but how it is executed. Variations in tool angle consistency, gesture smoothness, and temporal segmentation of primitives could be quantified and visualized to support structured feedback. The forecasting-driven Meta-Learning controller further enables selective adaptation to individual execution styles. Rather than imposing a fixed template, the system could adjust to each trainee's movement distribution while preserving shared structural knowledge learned from expert demonstrations. This would allow longitudinal tracking of skill progression relative to a personalized baseline.

From an operational perspective, this perspective includes gestural profiling through hierarchical metrics and adaptive benchmarking against expert references and prior personal performance. These capabilities suggest how scene monitoring could evolve into a tool for structured pedagogical feedback. Importantly, such tools are conceived as augmentations to human mentorship rather than replacements, preserving the central role of craft trainers while providing interpretable visualizations and reproducible measurements.

5.2 Professional Reconversion and Skill Transfer

A second perspective concerns professional reconversion and cross-domain skill transfer. Many craft professions share underlying movement structures even when tools and materials differ. Identifying these shared structures is essential for facilitating workforce mobility and preserving embodied knowledge across evolving industrial contexts.

The hierarchical representation of human movement provides a natural framework for revealing such correspondences. At the level of movement primitives, gestures are often biomechanically similar across professions. Actions in different crafts may reuse common primitive sequences despite serving distinct functional roles. By modelling gestures hierarchically, the framework can expose structural relationships that would remain invisible to flat action classification systems. In a reconversion scenario, a worker's existing gesture repertoire could be compared with the movement structure required by a target profession. Explicit scene understanding would reconstruct how tools and materials are manipulated, while implicit hierarchical modelling would highlight shared primitive and action patterns. The adaptive Meta-Learning controller could then specialize recognition to the new domain with limited additional data, mirroring how human learners transfer prior experience when acquiring new skills.

Practically, this perspective includes skill similarity mapping, targeted training recommendations focused on genuinely novel movement patterns, and data-efficient adaptation to new professions. Such capabilities align with broader European priorities related to workforce reskilling and sustainable employment. By formalizing embodied craft knowledge as structured motion hierarchies, the framework suggests a basis for evidence-informed reconversion strategies that respect both technical and human dimensions of skilled work.

5.3 Future Directions

In parallel with these application perspectives, methodological research directions are required to evolve the framework toward scalable adaptive craft intelligence.

A first direction concerns the scaling and enrichment of explicit scene representation. While the current pipeline shows a robust scene reconstruction across multiple craft domains, future work will investigate improvements mainly in the temporal stability. Incorporating richer spatial priors, and multimodal sensing, such as depth or inertial cues, could enhance reconstruction fidelity in cluttered environments. These developments would support more precise capture of hand–tool–material interactions and higher resolution analysis of the craft workspace. A second direction involves extending the hierarchical movement representation. Beyond movement primitives, actions, and activities, other movement notions could be further introduced, such as task intentio. Such an extension could strengthen transfer across professions and improve adaptation efficiency, particularly in reconversion scenarios.

Another research axis focuses on adaptation under extreme data scarcity. The forecasting-driven Meta-Learning controller demonstrates that event-driven adaptation can stabilize learning in low-data regimes. Future work will further examine how different task structures and hierarchical configurations influence adaptation dynamics and how hierarchical priors can further guide adaptation decisions.

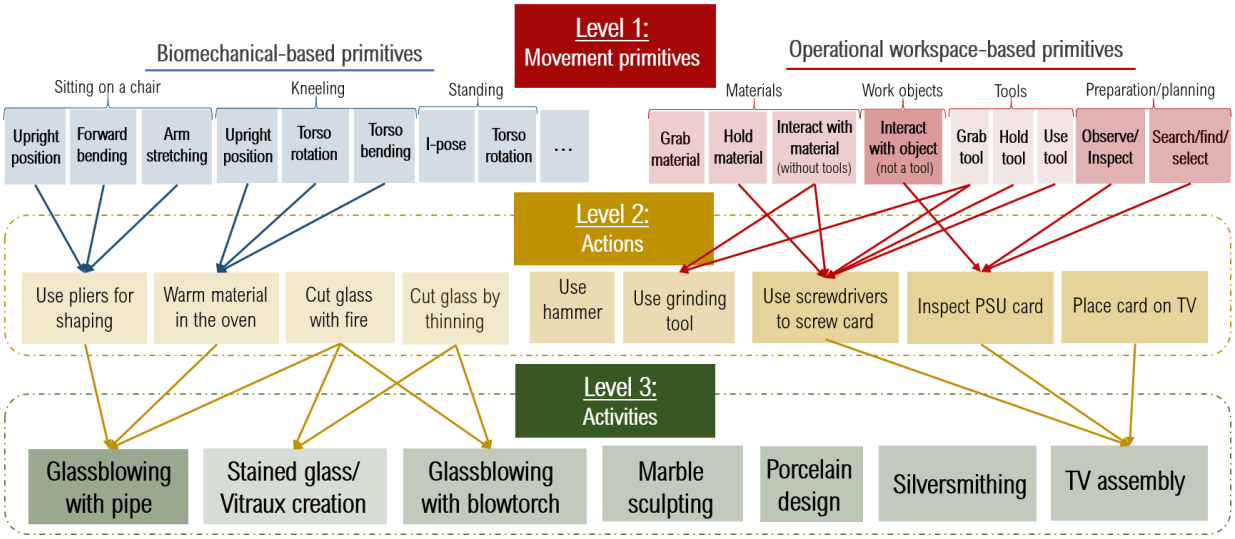


Figure 15. The extension of the current hierarchy toward a richer movement primitives' layers, identifying the interactions of craftspeople with tools and materials in the crafts space.

Together, these perspectives and methodological directions outline a path toward an integrated platform for adaptive craft intelligence. By jointly advancing explicit scene reconstruction and implicit hierarchical learning, future systems could continuously refine their understanding of human movement while preserving interpretability and alignment with embodied craft knowledge. Such integration is essential for long-term deployment in real craft environments, where accurate scene representation and adaptive motion understanding must operate in concert.