OWNERFUSION3D: MASK-GUIDED LOCAL IMAGE-TO-3D FUSION

Anonymous authors
Paper under double-blind review

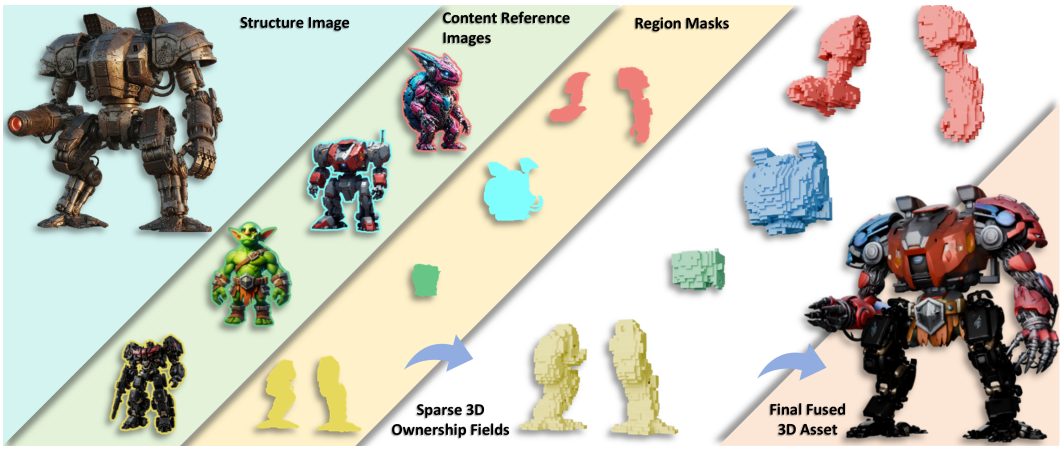

Figure 1: **Mask-specified local image-to-3D fusion.** Given a structure image, one or more image-plane masks, and corresponding content references, OwnerFusion3D discovers sparse 3D ownership and executes source-specific conditioning in a frozen sparse 3D generator. COR resolves overlapping ownership candidates in the multi-part setting.

ABSTRACT

Recent image-conditioned 3D generators can synthesize high-quality assets from single visual conditions. However, they cannot directly use a user-drawn 2D mask to assign an independent reference image to a chosen region of a 3D asset: treating multiple images as global conditions causes reference appearance to leak beyond the target region and weakens preservation elsewhere. We present **OwnerFusion3D**, a training-free framework for mask-guided local image-to-3D fusion. Given a structure image, an image-plane mask, and a content reference, OwnerFusion3D transfers reference appearance to the selected region while preserving the complementary structure. First, we introduce a **Mask-Conditioned Ownership Estimator (MOE)** that lifts the 2D mask to a sparse 3D owner field by contrasting frozen cross-attention responses to selected and competing foreground patches. Second, we propose **Owner-Exclusive Hierarchical Routing (OEHR)**, which uses this field to restrict every sparse query to its assigned condition source throughout the generator hierarchy. For multi-part fusion, **Categorical Ownership Resolution (COR)** resolves competing part assignments before the same routing process. OwnerFusion3D requires neither training nor 3D supervision and uses no post-hoc mesh or texture editing. Experiments across diverse object categories and visual styles demonstrate accurate local reference transfer together with strong preservation outside the selected mask. An anonymous implementation is available at https://anonymous.4open.science/r/anonymous-3d-release-7k2p.

1 INTRODUCTION

Image-conditioned 3D generators have made high-quality asset creation accessible from a small number of visual observations (Xiang et al., 2025; 2026; Hunyuan3D Team, 2025; Li et al., 2025).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The next practical challenge is not only producing a plausible asset, but controlling which visual evidence changes which part of it. A user may wish to retain the structure and pose of one image, transfer the appearance of an independent reference to a marked component, and preserve everything else.

Existing image-to-3D systems generally treat image conditions as global evidence. Concatenating two images therefore does not assign regional responsibility: the reference may leak into unrelated regions, while reducing its influence suppresses the requested transfer. We study this problem as **mask-specified local image-to-3D fusion**. Given a structure image I_s, an image-plane mask M, and an independent content reference I_c, the goal is to produce one coherent textured asset whose selected region receives appearance cues from I_c while its complement remains primarily governed by I_s. The images may differ in category and style; the mask specifies spatial responsibility, not semantic correspondence. The resulting modulation is bounded by the structure-derived sparse support and is not an unconstrained mechanism for creating topology absent from the structure image. Figure 1 illustrates the task.

Recent 2D generation and editing methods show that masks and image prompts can support localized appearance composition (Yang et al., 2023; Chen et al., 2024; Ye et al., 2023). Extending this capability to 3D is not a matter of copying a 2D mask into a rendered view. The mask is defined on pixels in one view, whereas a 3D generator operates on sparse latent elements that are not explicitly aligned with those pixels. Moreover, exposing multiple conditions does not tell each latent query which source it is allowed to read. A 2D-first composite can also introduce view-dependent seams and inconsistent depth cues that a subsequent 3D lifting stage must resolve without regional ownership. These challenges motivate treating local 3D fusion as an explicit ownership problem.

We introduce **OwnerFusion3D**, a training-free framework that separates local fusion into *ownership discovery* and *ownership execution*. MOE estimates a sparse 3D owner field from mask-conditioned attention, and OEHR enforces the resulting assignments as source-exclusive pre-softmax access through shape and texture generation. COR resolves competing assignments when multiple masks and references are provided. Thus, the mask is not a post-hoc image weight or a soft preference over globally available conditions: it is converted into ownership over generator-native sparse queries and retained as a source-access constraint across the hierarchy. The interface is compatible with generators that expose sparse queries, source-addressable attention, and cross-resolution correspondence. With a small generator-specific adaptation, the same ownership interface also transfers beyond TRELLIS.2, as illustrated with Direct3D-S2 (Wu et al., 2025) in Appendix D.3.

Our contributions are:

- We formulate mask-specified local image-to-3D fusion as sparse ownership discovery followed by source-specific execution in a frozen image-to-3D generator.
- MOE lifts a 2D mask to sparse 3D ownership through contrastive foreground attention and a fixed calibrated block-head readout, without 3D supervision.
- OEHR enforces source-exclusive pre-softmax access across the shape-to-texture hierarchy; COR resolves ownership conflicts in multi-part fusion.

2 Related Work

2.1 Image-Conditioned 3D Asset Generation

Image-conditioned 3D generation has evolved from feed-forward reconstruction models such as LRM and TripoSR to high-fidelity systems including InstantMesh, SF3D, Hunyuan3D 2.0, TripoSG, TRELLIS, and TRELLIS.2 (Hong et al., 2024; Tochilkin et al., 2024; Xu et al., 2024; Boss et al., 2025; Hunyuan3D Team, 2025; Li et al., 2025; Xiang et al., 2025; 2026). Multimodal reconstruction further shows the value of heterogeneous visual evidence (Cheng et al., 2023). TRELLIS.2 is most relevant to our work because its sparse representation and coarse-to-fine flow cascade expose a natural support on which regional responsibility can be defined.

Most image conditions, however, remain global: the generator internally decides how evidence is distributed over the 3D latent. OwnerFusion3D keeps the pretrained TRELLIS.2 support construction,

flow samplers, representation, and decoders, and adds an explicit mask-defined ownership interface inside conditional attention.

2.2 Reference-Guided 2D Editing and Fusion

2D generation has developed effective mechanisms for localized appearance composition. Paint by Example performs masked exemplar transfer, AnyDoor supports zero-shot object customization at a specified location, and IP-Adapter combines image-prompt and other conditions (Yang et al., 2023; Chen et al., 2024; Ye et al., 2023). Prompt-to-Prompt and MasaCtrl further expose attention correspondence as a control mechanism (Hertz et al., 2023; Cao et al., 2023). These methods motivate our 3D setting, but their target is a known image grid; we must instead identify and consistently control sparse 3D queries while the asset is generated. This motivates extending region-specific reference control from image editing to the latent stages of 3D asset generation.

2.3 Controllable 3D Generation

Controllable 3D generation spans text-driven synthesis (Poole et al., 2022; Lin et al., 2023), subject-driven generation and localized editing (Raj et al., 2023; Li et al., 2024; Cheng et al., 2024), and representation-level or image-guided editing (Haque et al., 2023; Zhuang et al., 2023; Sella et al., 2023; Zhuang et al., 2024; Gao et al., 2025; Bar-On et al., 2025; Gat et al., 2026; Hsiao et al., 2026; Xia et al., 2025; Preintner et al., 2026). Part-aware models expose semantic components or layouts as additional controls (Chen et al., 2025; Lin et al., 2025; Yang et al., 2025; Zhu et al., 2026). These methods differ in whether they optimize an existing representation, propagate edited views, or condition a feed-forward generator directly. Unlike supervised part-aware generators that commonly rely on curated 3D part annotations or a closed semantic vocabulary, OwnerFusion3D uses an inference- time, category-agnostic mask to specify spatial responsibility without 3D part supervision or fine-tuning.

Multi-image 3D fusion studies regional synthesis through multi-condition token fusion, 2D–3D alignment, and local attention control. Fuse3D (Jin et al., 2025) is a representative user-controllable approach: users can provide multiple images and specify regions ranging from global views to local details. Its fusion pipeline builds a shared multi-condition representation and enhances local attention to balance regional influence. OwnerFusion3D instead represents the control signal as sparse ownership and executes it through exclusive source access across the frozen hierarchy; COR extends this interface to multi-part conflicts.

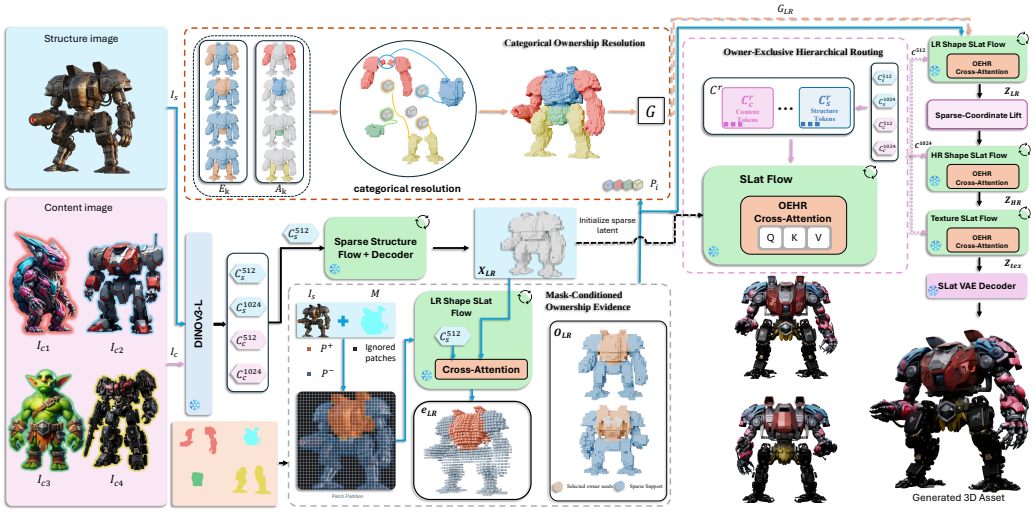

Figure 2: **OwnerFusion3D overview.** The blue path constructs structure-derived sparse support; the lower gray region uses MOE to obtain O_{LR}; the magenta region applies OEHR throughout the shape–texture hierarchy. The orange branch is the COR extension for multi-part fusion.

3 METHOD

OwnerFusion3D separates local fusion into ownership discovery and execution: MOE estimates sparse 3D ownership from a 2D request, OEHR enforces it during generation, and COR resolves multi-part conflicts (Figure 2).

3.1 SPARSE OWNERSHIP INTERFACE

Task. Let I_s be a structure image, M an image-plane mask, and I_c an independent content reference. We seek a textured asset $\widehat{\mathcal{G}}$ whose selected region receives appearance cues from I_c while the complement remains primarily governed by I_s. Since the two images may differ in category or style, the mask specifies spatial source responsibility rather than semantic correspondence. We interpret the mask as ownership of the corresponding structural component across its visible 3D extent rather than as a calibrated depth map; unresolved line-of-sight ambiguity remains structure-owned under the conservative support regularization. We represent this responsibility on active sparse queries:

$$O_{\mathrm{LR}} \in \{0,1\}^N, \tag{1}$$

where query $i \in \{1,\dots,N\}$ at coordinate x_i is content-owned when $O_{\mathrm{LR},i} = 1$ and structure-owned when $O_{\mathrm{LR},i} = 0$, with $X_{\mathrm{LR}} = \{x_i\}_{i=1}^N$ denoting the structure-derived support. MOE estimates this field and OEHR turns it into a source-access constraint.

Frozen hierarchical sparse generator. We instantiate OwnerFusion3D on the pretrained TRELLIS.2-4B generator (Xiang et al., 2026). Its weights, sparse-support construction, flow integration, and decoders remain unchanged; the method adds only a structure-only probe and source access control to conditional attention. The interface requires active sparse queries, source-addressable cross-attention, and cross-resolution correspondence. Image conditions use the frozen DINOv3 ViT-L/16 conditioner (Xiang et al., 2026; Siméoni et al., 2026). For generation resolution r,

$$C_s^r = \Phi^r_{\mathrm{DINOv3}}(I_s), \qquad C_c^r = \Phi^r_{\mathrm{DINOv3}}(I_c), \qquad X_{\mathrm{LR}} = \mathcal{S}_\Theta(C_s^{\mathrm{LR}}), \tag{2}$$

where C_s^r and C_c^r are the DINOv3 condition-token sequences, Φ^r_{DINOv3} is the frozen conditioner, and $\mathcal{S}_\Theta$ constructs the structure-derived support with the native parameters Θ. The content image is excluded from support construction, so ownership is assigned on a structure-defined support.

3.2 MASK-CONDITIONED OWNERSHIP ESTIMATION

The 2D mask and sparse support live in different spaces. MOE bridges them by reading the query-to-image correspondence produced by the frozen low-resolution shape flow. The structure-only probe prevents the content reference from affecting support construction or ownership. Figure 3 summarizes the conversion from mask-conditioned patches to an executable owner field.

Foreground-aware patch partition. We first apply the base generator's foreground crop to the structure image and the mask. The mask and the resulting structure foreground are then projected onto the spatial patch grid of C_s^{LR}. For patch j, let $u_j \in [0,1]$ denote its mask coverage and $v_j \in [0,1]$ its structure-foreground coverage. We retain two non-overlapping evidence roles:

$$\begin{aligned} \mathcal{P}^+ &= \{j \mid v_j \geq \eta_{\mathrm{fg}},\ u_j \geq \eta_+\}, \\ \mathcal{P}^- &= \{j \mid v_j \geq \eta_{\mathrm{fg}},\ u_j \leq \eta_-\}. \end{aligned} \tag{3}$$

Here j indexes a spatial DINOv3 patch, and η_{fg}, η_+, and η_- are fixed coverage thresholds with $\eta_+ > \eta_-$. The sets $\mathcal{P}^+$ and $\mathcal{P}^-$ contain selected and unselected foreground patches; background and ambiguous boundary patches are excluded. Comparing against unselected foreground, rather than background, makes the evidence part-specific instead of merely foreground-sensitive.

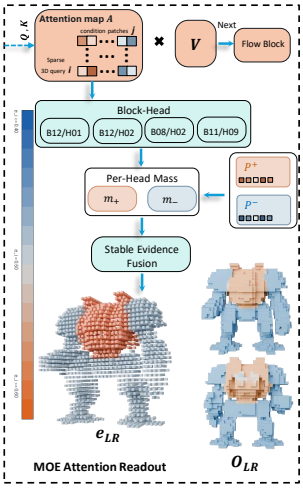

Figure 3: **MOE.** 2D mask evidence is lifted to sparse 3D ownership.

The implemented patch weights are the corresponding coverage weights,

$$w_j^+ = \mathrm{clip}_{[\epsilon_+,1]}(u_j), \qquad w_j^- = \mathrm{clip}_{[\epsilon_-,1]}\big(v_j(1-u_j)\big). \tag{4}$$

Here ϵ_+ and ϵ_- are numerical floors for the selected and competing weights, respectively. The ϵ used in Eq. 5 is the denominator floor. Their exact values and the coverage thresholds are listed in Appendix C.2.

Contrastive attention evidence. During a short structure-only probe, an observation $o = (t, \ell, h)$ indexes a sampling step, a cross-attention block, and an attention head. At coordinate x_i, the sparse latent supplies the query and C_s^{LR} supplies image keys and values. Let $A^o_{i,\iota(j)}$ be post-softmax attention to the condition token for patch j, where ι offsets auxiliary tokens; softmax ranges over all condition tokens. We compute selected and competing foreground mass, normalized by the total weight of each set:

$$d^o_{i,+} = \frac{\sum_{j\in\mathcal{P}^+} w_j^+ A^o_{i,\iota(j)}}{\sum_{j\in\mathcal{P}^+} w_j^+ + \epsilon}, \qquad d^o_{i,-} = \frac{\sum_{j\in\mathcal{P}^-} w_j^- A^o_{i,\iota(j)}}{\sum_{j\in\mathcal{P}^-} w_j^- + \epsilon},$$
$$e^o_i = \frac{d^o_{i,+}}{d^o_{i,+} + d^o_{i,-} + \epsilon}. \tag{5}$$

The normalized masses measure selected and unselected foreground response, and $e^o_i \in [0,1]$ is the relative selected-ownership preference.

Fixed block–head readout. Different block-head members provide heterogeneous correspondence signals. We therefore fix a four-member set $\mathcal{H}_*$ using an independent diagnostic split rather than selecting a new subset for each input. Each fixed member receives an input-dependent reliability weight, and the evidence fields are fused as

$$e_i = \sum_{m\in\mathcal{H}_*} \omega_m \bar{e}_i^m, \qquad \sum_{m\in\mathcal{H}_*} \omega_m = 1. \tag{6}$$

Here m indexes a block–head member, $\bar{e}_i^m$ is its probe-averaged evidence, and ω_m is its normalized reliability weight. Membership is fixed; only weights adapt (Appendix B).

Support-constrained owner construction. We convert continuous evidence into a binary field only on the original support:

$$\widetilde{O}_i = \mathbf{1}[e_i \geq \tau], \qquad O_{\mathrm{LR}} = \mathcal{R}_{X_{\mathrm{LR}}}(\widetilde{O}, e). \tag{7}$$

Here τ is the fixed threshold, $\widetilde{O}$ is the preliminary binary field, and $\mathcal{R}_{X_{\mathrm{LR}}}$ is sparse-support regularization: a local majority operation removes isolated assignments, followed by a score-qualified neighborhood expansion over coordinates that already belong to X_{LR}. It cannot create an owner outside the structure-derived support. The continuous evidence vector e and discrete O_{LR} shown in Figure 3 are therefore different objects: the former is evidence used for decision making, while the latter is the source-assignment interface consumed by OEHR.

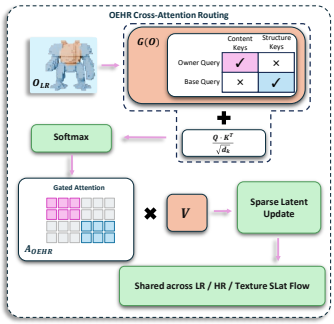

Figure 4: **OEHR.** Each query reads one condition source.

3.3 Owner-Exclusive Hierarchical Routing

MOE identifies candidate queries but does not itself prevent the two sources from competing. Ordinary concatenated conditioning gives every query access to both sources, and post-attention reweighting cannot undo that interaction. OEHR therefore converts ownership into a source-exclusive constraint at the cross-attention logits; Figure 4 shows its reuse across flows.

Source-exclusive cross-attention. At resolution r, we concatenate content and structure conditions as $C^r = [C^r_c \mid C^r_s]$. For sparse latent $H_t^{(\ell)}$ at flow step t and block ℓ, frozen native projections produce query Q_i for sparse query i and key/value K_j, V_j for condition token j; d is the attention-head dimension. OEHR introduces no new projection or learned parameter; it changes only which

source tokens may contribute to each query's attention output. Let $\mathcal{K}_c^r$ and $\mathcal{K}_s^r$ be the content and structure key/value index ranges. Let O_i^r be the owner at resolution r (O_{LR} at LR or O_{HR} at HR), and write $\mathcal{K}_{O_i^r}^r$ for its permitted range: $\mathcal{K}_c^r$ if $O_i^r = 1$ and $\mathcal{K}_s^r$ otherwise. The gate is applied before normalization:

$$G_{ij} = \begin{cases} 0, & j \in \mathcal{K}_{O_i^r}^r, \\ -\infty, & j \notin \mathcal{K}_{O_i^r}^r, \end{cases} \qquad A_{ij}^{\mathrm{OEHR}} = \mathrm{Softmax}_j\left(\frac{Q_i K_j^\top}{\sqrt{d}} + G_{ij}\right), \\ Y_i = \sum_j A_{ij}^{\mathrm{OEHR}} V_j. \tag{8}$$

The output Y_i is the routed value aggregation. The $-\infty$ entries express hard exclusion before softmax; the gate is not learned and leaves native projections unchanged. It preserves relative attention within the allowed source while removing cross-source competition. The implementation's finite sentinel is specified in Appendix C.

The routing patch is applied only to cross-attention. Sparse self-attention, AdaLN/MLP updates, flow integration, and the native decoder remain untouched. Consequently, OEHR does not replace the base generator or impose a new latent representation; it changes the conditional access contract under which the frozen flow evolves each sparse query.

Propagation through the generator hierarchy. The LR owner field first routes the low-resolution shape flow. The generator then derives higher-resolution sparse support from the sampled LR shape latent. For an HR coordinate x, a deterministic quantized parent lookup $\pi(x)$ identifies its LR ancestor, and the ownership is lifted by

$$O_{\mathrm{HR}}(x) = O_{\mathrm{LR}}(\pi(x)). \tag{9}$$

Here $\pi(x)$ is the generator-native quantized parent lookup. The lifted assignment constrains higher-resolution shape and texture cross-attention through the available sparse-coordinate correspondence. If an exact LR parent is absent from the active LR map, the implementation uses the nearest valid nonzero LR owner only for the HR ownership readout; the LR field itself is not modified. This is a generator-native sparse correspondence, not a general 3D projection or registration algorithm. The final mesh and texture are decoded directly from the routed latents, without mesh-level or texture-level post-hoc editing.

3.4 Categorical Ownership Resolution for Multiple Parts

With multiple masks, independently estimated owners may claim one sparse query. COR resolves these categorical conflicts on the shared support; it is not a third synthesis mechanism. As a controlled w/o-COR baseline, *score-only* assigns each query to the highest-scoring admissible part independently. This selection rule cannot enforce spatial consistency when parts produce overlapping or weakly supported candidates. Full COR instead applies evidence-based resolution, bounded repair, conflict correction, and monotone completion only to owner-zero queries, never clearing a nonzero owner. It returns a mutually exclusive part owner $p_i \in \{0, 1, \ldots, P\}$, with zero denoting structure; parts sharing a content image are merged into the group owner $g_i \in \{0, 1, \ldots, K\}$ consumed by OEHR.

For reference groups indexed by $1, \ldots, K$, let γ_j denote the source group of condition token j, with $\gamma_j = 0$ for structure tokens and $\gamma_j = k$ for tokens from reference group k. The same exclusive execution rule then becomes

$$G_{ij}^{\mathrm{COR}} = \begin{cases} 0, & \gamma_j = g_i, \\ -\infty, & \gamma_j \neq g_i. \end{cases} \tag{10}$$

The final group field is exactly the field passed to OEHR. It is lifted as in Eq. 9 and used for shape and texture, leaving OEHR as the shared execution mechanism for single-mask and multi-part fusion. This separation lets COR resolve competing part assignments once, before generation, while OEHR applies the resulting source ownership consistently through the hierarchical shape and texture stages. COR therefore coordinates which reference may influence each supported query; it does not infer a complete 3D semantic partition beyond the generator's sparse support.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Tasks and baselines. Each single-mask input contains a structure image, a binary image-plane mask, and an independent content reference. The multi-part input contains several mask–reference pairs on the same structure image. We evaluate 50 single-mask 3D assets and 20 multi-part 3D assets, where each multi-part asset contains multiple mask-specified fusion regions. The assets cover diverse object categories and visual styles, including cross-category structure–reference pairs. Below, T1 denotes the original TRELLIS backbone and T2 denotes TRELLIS.2. The comparison uses Fuse3D (Jin et al., 2025) as the direct multi-image 3D fusion baseline, together with two 2D-first baselines: the hosted, closed-source Qwen-Image-2.0-Pro (Zhao et al., 2026) and the publicly available FLUX.2 (Black Forest Labs, 2025), based on the KLEIN 4B model. The latter two first produce a mask-conditioned 2D composite from the same structured fusion description and then lift that composite with the corresponding image-to-3D pipeline; their prompts and input ordering are fixed across methods and documented in Appendix C. OwnerFusion3D directly consumes the structure image, mask, and reference conditions with the frozen TRELLIS.2 generator, without training or fine-tuning. We use the released Fuse3D route with its native backbone. An attempted TRELLIS.2 migration did not yield a valid reproducible run; therefore, we retain the released native TRELLIS (T1) pipeline rather than report an unverified port. Fuse3D is consequently evaluated with its T1 reference pipeline, whereas OwnerFusion3D and the two 2D-first baselines use T2.

Backbone-paired evaluation. To anchor preservation to each method's own reconstruction backbone, Fuse3D is paired with a structure-only reference generated by T1, while OwnerFusion3D and the two 2D-first baselines are paired with T2 structure-only references. We manually align each structure-only reference once to the input camera, then reuse the saved camera and object transform for all associated method outputs; generated assets are not independently hand-aligned. At the aligned view, the selected region is compared with the content reference and the complementary region with the paired structure-only render. The full alignment, crop, and multi-part protocol is given in Appendix C.8. This within-backbone pairing reduces alignment and preservation bias. The same backbone-paired protocol is used for the VLM panels: each candidate is shown against its own backbone's aligned structure-only reference, so Fuse3D is not judged against a T2 reference. The protocol retains the native-backbone boundary described above.

Metrics. Our main measures are Joint LPIPS, Joint CLIP-I, and Joint DINO-I, based on LPIPS (Zhang et al., 2018), CLIP (Radford et al., 2021), and DINOv2 (Oquab et al., 2024). Each metric combines two equally weighted terms: an in-mask content term and an out-of-mask structure term, with the native direction of each metric retained (lower is better for LPIPS and higher is better for CLIP-I and DINO-I). This design prevents either objective from dominating: a conservative method that makes little or no target-region change may preserve the complement but will score poorly on in-mask transfer, whereas global fusion may match the reference more strongly but will be penalized for changing unselected regions. LPIPS is reported for complementary low-level fidelity, while CLIP-I and DINO-I provide more tolerant semantic comparison when the reference and generated crops differ in pose or layout. Joint scores are the primary summary because the task requires both transfer and preservation; the fixed component protocol is given in Appendix C.8.

4.2 QUANTITATIVE COMPARISON

Table 1: **Backbone-paired regional fusion.** Mean joint scores over the single-mask and multi-part benchmarks. Fuse3D uses the T1 reference pipeline; all other methods use T2. Best values are bold and second-best values are underlined.

Method	Single-mask			Multi-part		
	Joint LPIPS ↓	Joint CLIP-I ↑	Joint DINO-I ↑	Joint LPIPS ↓	Joint CLIP-I ↑	Joint DINO-I ↑
Fuse3D	**0.3476**	0.7867	0.5912	<u>0.3380</u>	<u>0.7907</u>	0.6166
Qwen-Image-2.0-Pro	0.3654	<u>0.8090</u>	<u>0.6511</u>	0.3445	0.7891	0.5970
FLUX.2	0.3698	0.8056	0.6382	0.3525	0.7798	<u>0.6191</u>
Ours	<u>0.3579</u>	**0.8389**	**0.7061**	**0.3355**	**0.8285**	**0.6973**

The joint objective measures transfer and preservation simultaneously, rather than either content similarity or structure retention in isolation. The results show that the ownership interface maintains this balance when moving from one selected region to multiple reference assignments, while the comparison pipelines trade regional transfer against preservation to different degrees.

VLM evaluation. We additionally use GPTEval3D (Wu et al., 2024) with GPT-5 (OpenAI, 2025) as an automated pairwise evaluator on the single-mask benchmark. The evaluator sees the conditions and aligned multi-view results, with method identities hidden and the candidates randomly assigned to A and B. It judges region fidelity, visual seamlessness, edit controllability, multi-view coherence, and overall preference. Table 2 reports the percentage of binary comparisons in which OwnerFusion3D is selected. For each criterion, the record contains one binary choice, A or B; no Tie decision is reported, so 50% is the pairwise parity point. The complete prompt and audit protocol are provided in Appendix C.9.

Table 2: **Mask-aligned GPTEval3D.** Percentage of pairwise comparisons in which OwnerFusion3D is preferred to each baseline on evaluated 3D assets. RC: region consistency; VS: visual seamlessness; EC: edit controllability; MV: multi-view coherence; OP: overall controlled-fusion preference.

Baseline	RC	VS	EC	MV	OP
Fuse3D	88	84	74	84	86
Qwen-Image-2.0-Pro	60	58	56	62	58
FLUX.2	66	62	62	56	70

4.3 Qualitative Results



Figure 5: **Single-mask qualitative comparison.**

Figure 5 gives concrete evidence of local control. In the elephant–robot row, the mask targets the elephant body while the chair and supporting structure remain unselected: our result transfers mechanical body cues without overwriting them, whereas the other methods preserve stronger elephant appearance or alter the surrounding body. In Figure 6, distinct references remain associated with the marked ornamental and winged parts, while competing results mix attributes or collapse them into one style. These examples show that the mask specifies spatial ownership rather than a semantic class match: unmarked structure remains anchored while MOE resolves assignments and OEHR enforces them during generation. More fusion results appear in Appendix D.5.



Figure 6: **Multi-part qualitative comparison.**

4.4 Ablation Study

Table 3 isolates the MOE readout, MOE, and OEHR. *All-block-head* averages all diagnostic candidates; *w/o MOE* bypasses ownership discovery and routing; *w/o OEHR* retains the MOE owner field but leaves both sources accessible during cross-attention; *Full (Ours)* uses the fixed calibrated readout with exclusive routing.

Table 3: **Ablation study.** Mean joint scores on the 50 3D assets shared by all variants. Best values are bold and second-best values are underlined.

Variant	Joint LPIPS ↓	Joint CLIP-I ↑	Joint DINO-I ↑
w/o MOE	0.3919	0.7542	0.5173
w/o OEHR	0.3631	0.7888	0.5957
All-block-head	0.3616	0.8182	0.6804
Full (Ours)	**0.3579**	**0.8389**	**0.7061**

The results support both ownership discovery and exclusive execution. Fixed calibration also reduces the negative contributions of heterogeneous block–head members.

5 Conclusion

We introduced OwnerFusion3D, a training-free framework for mask-guided local image-to-3D fusion. MOE maps a 2D request to sparse 3D ownership, OEHR enforces exclusive access to the corresponding image sources during hierarchical generation, and COR resolves competing assignments for multi-part inputs before routing. Across our evaluations, this decomposition improves the balance between transferring reference appearance and preserving unselected structure; the results on a compatible alternative generator further indicate that the ownership interface can extend beyond the TRELLIS family. Future work can strengthen ownership inference with multi-view or geometric evidence and extend the routing interface to a broader range of sparse 3D generators. The current method assigns sources to supported generator queries, so its spatial scope remains bounded by the input structure and the available view evidence.

AI USE STATEMENT

In accordance with ICLR's policy on the use of Large Language Models (LLMs) during paper preparation, we declare the extent and nature of LLM involvement in this work. We used LLM-based tools during manuscript preparation and software development for language refinement, code debugging, and figure-layout iteration. The authors made the final decisions regarding the method formulation, implementation, experiments, evaluation, and conclusions.

ETHICS STATEMENT

This work does not involve human subjects, human-subject experiments, personally identifiable information, or biometric data. Our experiments use structure images, reference images, masks, and generated 3D assets for research evaluation. Region-controlled generation could be misused for unauthorized content manipulation or misleading asset creation. We intend the method and released assets for research purposes and will respect the licenses and usage restrictions of external models, images, and datasets.

REPRODUCIBILITY STATEMENT

OwnerFusion3D is training-free and uses the frozen TRELLIS.2-4B generator without additional 3D supervision or fine-tuning. The main paper describes the MOE, OEHR, and COR interfaces, while the appendix provides implementation details, inference configuration, block-head readout specification, baseline protocols, camera-alignment procedure, metric definitions, and additional results. An anonymous supplementary package will provide the code and accompanying materials required for reproducibility.

REFERENCES

Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18208–18218, 2022.

Roi Bar-On, Dana Cohen-Bar, and Daniel Cohen-Or. Editp23: 3d editing via propagation of image prompts to multi-view. *arXiv preprint arXiv:2506.20652*, 2025.

Black Forest Labs. FLUX.2: Frontier visual intelligence, 2025. URL https://bfl.ai/blog/flux-2. Model release.

Mark Boss, Zixuan Huang, Aaryaman Vasishta, and Varun Jampani. SF3D: Stable fast 3d mesh reconstruction with uv-unwrapping and illumination disentanglement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16240–16250, 2025.

Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22503–22513, 2023.

Minghao Chen, Roman Shapovalov, Iro Laina, Tom Monnier, Jianyuan Wang, David Novotny, and Andrea Vedaldi. Partgen: Part-level 3d generation and reconstruction with multi-view diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5881–5892, 2025.

Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6593–6602, 2024.

Xinhua Cheng, Tianyu Yang, Jianan Wang, Yu Li, Lei Zhang, Jian Zhang, and Li Yuan. Progressive3d: Progressively local editing for text-to-3d content creation with complex semantic prompts. *arXiv preprint arXiv:2310.11784*, 2024.

Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, Alexander Schwing, and Liangyan Gui. Sdfusion: Multimodal 3d shape completion, reconstruction, and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4456–4465, 2023.

Will Gao, Dilin Wang, Yuchen Fan, Aljaz Bozic, Tuur Stuyck, Zhengqin Li, Zhao Dong, Rakesh Ranjan, and Nikolaos Sarafianos. 3d mesh editing using masked lrms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7154–7165, 2025.

Inbar Gat, Dana Cohen-Bar, Guy Levy, Elad Richardson, and Daniel Cohen-Or. Shapeup: Scalable image-conditioned 3d editing. *ACM Transactions on Graphics*, 2026.

Ayaan Haque, Matthew Tancik, Alexei A. Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-nerf2nerf: Editing 3d scenes with instructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19683–19693, 2023.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. In *International Conference on Learning Representations (ICLR)*, 2023.

Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. LRM: Large reconstruction model for single image to 3d. In *International Conference on Learning Representations (ICLR)*, 2024.

Teng-Fang Hsiao, Bo-Kai Ruan, Yu-Lun Liu, and Hong-Han Shuai. VecSet-Edit: Unleashing pre-trained LRM for mesh editing from single image. In *Proceedings of the ACM SIGGRAPH Conference Papers (SIGGRAPH)*, pp. 1–12, 2026.

Hunyuan3D Team. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. *arXiv preprint arXiv:2501.12202*, 2025.

Xuancheng Jin, Rengan Xie, Wenting Zheng, Rui Wang, Hujun Bao, and Yuchi Huo. Fuse3d: Generating 3d assets controlled by multi-image fusion. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers (SIGGRAPH Asia)*, pp. 1–12, 2025.

Dong-Yang Li, Wang Zhao, Yuxin Chen, Wenbo Hu, Meng-Hao Guo, Fang-Lue Zhang, Ying Shan, and Shi-Min Hu. Pixal3d: Pixel-aligned 3d generation from images. *arXiv preprint arXiv:2605.10922*, 2026.

Yangguang Li, Zi-Xin Zou, Zexiang Liu, Dehu Wang, Yuan Liang, Zhipeng Yu, Xingchao Liu, Yuan-Chen Guo, Ding Liang, Wanli Ouyang, and Yan-Pei Cao. Triposg: High-fidelity 3d shape synthesis using large-scale rectified flow models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–18, 2025.

Yuhan Li, Yishun Dou, Yue Shi, Yu Lei, Xuanhong Chen, Yi Zhang, Peng Zhou, and Bingbing Ni. Focaldreamer: Text-driven 3d editing via focal-fusion assembly. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pp. 3279–3287, 2024.

Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 300–309, 2023.

Kuan Heng Lin, Sicheng Mo, Ben Klingher, Fangzhou Mu, and Bolei Zhou. Ctrl-X: Controlling structure and appearance for text-to-image generation without guidance. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 128911–128939, 2024.

Yuchen Lin, Chenguo Lin, Panwang Pan, Honglei Yan, Yiqiang Feng, Yadong Mu, and Katerina Fragkiadaki. Partcrafter: Structured 3d mesh generation via compositional latent diffusion transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 39823–39851, 2025.

OpenAI. GPT-5, 2025. URL https://openai.com/index/introducing-gpt-5/. Model release.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

Tobias Preintner, Yunfei Deng, Phillip Müller, Sebastian Illing, Adrian König, Thomas Bäck, Elena Raponi, and Niki van Stein. 3DMorph: Single-image-guided local 3d shape editing and morphing. *arXiv preprint arXiv:2606.07115*, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

Amit Raj, Srinivas Kaza, Ben Poole, Michael Niemeyer, Nataniel Ruiz, Ben Mildenhall, Shiran Zada, Kfir Aberman, Michael Rubinstein, Jonathan Barron, Yuanzhen Li, and Varun Jampani. Dreambooth3d: Subject-driven text-to-3d generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2349–2359, 2023.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Hyon Khedr, Roman Rädle, Chloé Rolland, Laura Gustafson, et al. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Etai Sella, Gal Fiebelman, Peter Hedman, and Hadar Averbuch-Elor. Vox-e: Text-guided voxel editing of 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 430–440, 2023.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=2NlGyqNjns. Also available as arXiv:2508.10104.

Dmitry Tochilkin, David Pankratz, Zexiang Liu, Zixuan Huang, Adam Letts, Yangguang Li, Ding Liang, Christian Laforte, Varun Jampani, and Yan-Pei Cao. Triposr: Fast 3d object reconstruction from a single image. *arXiv preprint arXiv:2403.02151*, 2024.

Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Yikang Yang, Yajie Bao, Jiachen Qian, Siyu Zhu, Xun Cao, Philip Torr, et al. Direct3d-s2: Gigascale 3d generation made easy with spatial sparse attention. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*, pp. 189214–189240, 2025. doi: 10.52202/085713-5688.

Tong Wu, Gang Yang, Zhen Li, Kai Zhang, Ziwei Liu, Leonidas J. Guibas, Dahua Lin, and Gordon Wetzstein. GPT-4V(ision) is a human-aligned evaluator for text-to-3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22227–22238, 2024.

Ruihao Xia, Yang Tang, and Pan Zhou. Towards scalable and consistent 3d editing. *arXiv preprint arXiv:2510.02994*, 2025.

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21469–21480, 2025.

Jianfeng Xiang, Xiaoxue Chen, Sicheng Xu, Ruicheng Wang, Zelong Lv, Yu Deng, Hongyuan Zhu, Yue Dong, Hao Zhao, Nicholas Jing Yuan, and Jiaolong Yang. Native and compact structured latents for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14419–14429, 2026.

Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024.

Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18381–18391, 2023.

Yunhan Yang, Yufan Zhou, Yuan-Chen Guo, Zi-Xin Zou, Yukun Huang, Ying-Tian Liu, Hao Xu, Ding Liang, Yan-Pei Cao, and Xihui Liu. Omnipart: Part-aware 3d generation with semantic decoupling and structural cohesion. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers (SIGGRAPH Asia)*, pp. 1–12, 2025.

Chongjie Ye, Yushuang Wu, Ziteng Lu, Jiahao Chang, Xiaoyang Guo, Jiaqing Zhou, Hao Zhao, and Xiaoguang Han. Hi3DGen: High-fidelity 3d geometry generation from images via normal bridging. *arXiv preprint arXiv:2503.22236*, 2025.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018.

Bing Zhao, Chenfei Wu, Deqing Li, Hao Meng, et al. Qwen-image-2.0 technical report. *arXiv preprint arXiv:2605.10730*, 2026.

Yiheng Zhu, Kangle Deng, Jean-Philippe Fauconnier, Iñaki Navarro, Daiqing Li, Ava Pun, Yinan Zhang, Peiye Zhuang, Xiaoxia Sun, Maneesh Agrawala, Kiran Bhat, and Tinghui Zhou. Cubepart: An open-vocabulary part-controllable 3d generator. *ACM Transactions on Graphics*, 2026.

Jingyu Zhuang, Chen Wang, Lingjie Liu, Liang Lin, and Guanbin Li. Dreameditor: Text-driven 3d scene editing with neural fields. In *Proceedings of the SIGGRAPH Asia 2023 Conference Papers (SIGGRAPH Asia)*, pp. 1–10, 2023.

Jingyu Zhuang, Di Kang, Yan-Pei Cao, Guanbin Li, Liang Lin, and Ying Shan. TIP-Editor: An accurate 3d editor following both text-prompts and image-prompts. *ACM Transactions on Graphics*, 2024.

A CATEGORICAL OWNERSHIP RESOLUTION

For multi-mask fusion, Categorical Ownership Resolution (COR) converts part candidates into one mutually exclusive sparse ownership field before OEHR. It resolves ownership conflicts; it creates no coordinates or features. Here, a sparse support is the set of active 3D queries retained by the generator. For comparison, *score-only* assignment is the controlled w/o-COR rule that independently assigns each query to the candidate part with the largest MOE evidence, without core, neighborhood, conflict, or completion constraints.

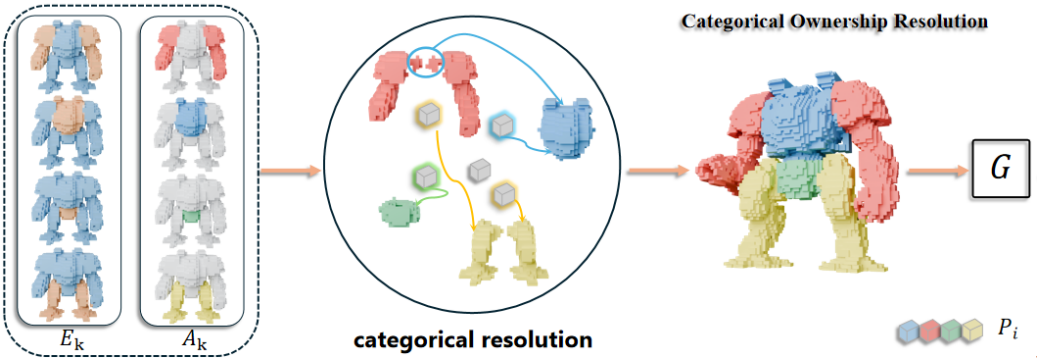

Figure 7: **Categorical Ownership Resolution (COR).** Independent part candidates are converted into one mutually exclusive ownership field before source routing; unlike score-only assignment, COR also applies the candidate, neighborhood, conflict, and completion constraints.

A.1 INPUT AND OUTPUT

Let $X_{\mathrm{LR}} = \{x_i\}_{i=1}^{N}$ be the active sparse support. For part k, MOE provides evidence $e_{ik} \in [0, 1]$ and candidate indicator $a_{ik} \in \{0, 1\}$ (the E_k, A_k fields in Figure 7). COR returns one part owner $p_i \in \{0, 1, \ldots, P\}$; zero denotes structure or uncertainty. Parts sharing a reference are merged into group owner $g_i \in \{0, 1, \ldots, K\}$ for OEHR. Completion fills zero owners.

A.2 RESOLUTION RULE

For each part, COR extracts a high-confidence core and measures candidate-restricted distance on the six-connected sparse graph. Unique candidates and core overlaps are assigned first; remaining conflicts use evidence, inverse core distance, core-neighbor support, and assigned-owner support:

$$s_{ik} = e_{ik} + \lambda_d \rho_{ik} + \lambda_c n_{ik}^{\mathrm{core}} + \lambda_a n_{ik}^{\mathrm{owner}}, \qquad a_{ik} = 1, \tag{11}$$

where $\rho_{ik} = 1/(1 + d_{ik})$ and d_{ik} is the candidate-restricted distance to the core. Small-margin conflicts remain structure-owned. Neighbor fill, residual completion, bounded conflict correction, direct-neighbor closure, and monotone zero-owner completion then refine the field; only bounded correction may replace an existing owner, under its fixed margin and support rule. The direct-neighbor closure is a single local pass over immediate sparse neighbors; newly assigned queries are not recursively propagated. In Eq. 11, e_{ik} is the MOE evidence for part k at query x_i; ρ_{ik} is proximity to that part's core; n_{ik}^{core} and n_{ik}^{owner} are normalized support from core neighbors and already assigned owner neighbors, respectively. The weights are fixed to $\lambda_d = 0.20$, $\lambda_c = 0.08$, and $\lambda_a = 0.08$; $a_{ik} = 1$ restricts the comparison to admissible candidates. The core is formed from the top 10% of candidate evidence subject to a minimum evidence of 0.55 and at least 16 queries, and a best-versus-second score margin below 0.06 is treated as unresolved.

A.3 QUALITATIVE EFFECT OF COR

Figure 8 compares the independent score-only rule defined above with COR, which adds candidate, neighborhood, conflict, and completion constraints. The result concerns source ownership, not complete 3D semantic segmentation.

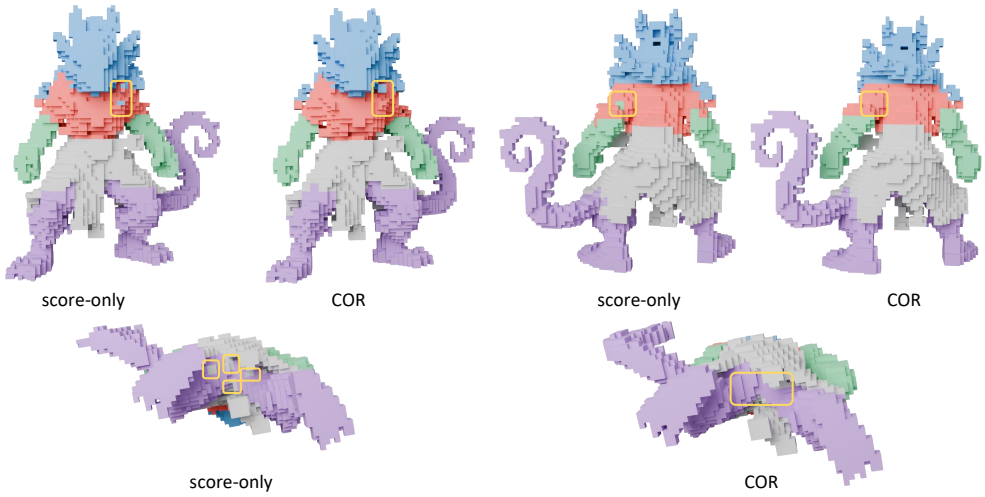

Figure 8: **Ownership resolution.** COR reduces local inconsistencies relative to score-only assignment by enforcing mutually exclusive, candidate-consistent part ownership.

Algorithm 1 Categorical Ownership Resolution (COR)

Require: LR support X_{LR}, MOE evidence $\{E_k\}$, candidates $\{A_k\}$, core and boundary thresholds
Ensure: Part owner $p \in \{0, 1, \ldots, P\}^N$

Build the sparse adjacency graph on X_{LR}.
for each part k **do**
    Extract a high-confidence core from E_k inside A_k.
    Compute candidate-restricted graph distances to the core.
end for
Initialize all part owners to $p_i \leftarrow 0$.
Assign unique-candidate queries and unique-core overlap queries.
for each remaining conflict query i **do**
    Rank admissible parts with Eq. 11.
    if the best score has a decisive margin **then**
        Assign p_i to the best part.
    else
        Keep $p_i = 0$ as an uncertain boundary.
    end if
end for
Apply one candidate-gated synchronous neighbor-fill pass.
Apply bounded local and component residual completion.
Apply constrained conflict correction and one non-propagating direct-neighbor closure.
Apply monotone owner-zero completion from existing owners.
Merge the resolved part owners into reference groups.
Return the group owner g for OEHR.

B Block–Head Diagnostic and Fixed Readout

MOE reads multiple frozen block–head members. An independent diagnostic split selects a fixed four-member readout; this is calibration, not training. Here, a block–head member denotes one transformer block and attention head whose frozen cross-attention map is read by MOE. Evidence is the resulting per-query response, and a selected query is one retained by the 2D mask projection; competing queries are the matched foreground queries outside that selected set.

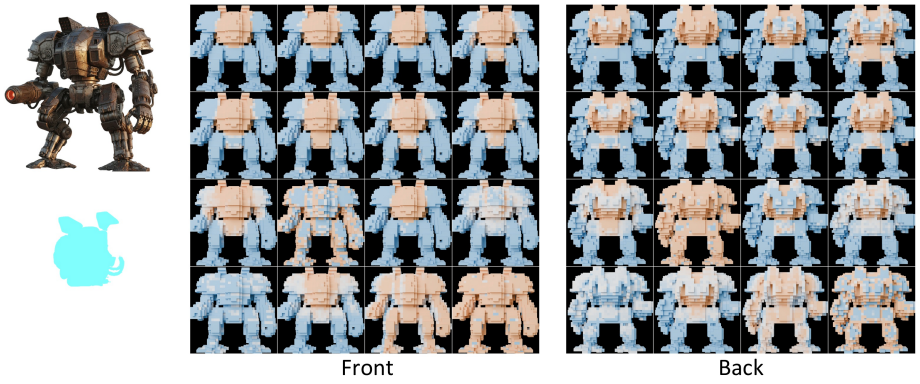

Figure 9: **Top-16 block–head diagnostic.** Front and back responses from the independent structure-only MOE probe; the diagnostic split is disjoint from all reported evaluation assets.

B.1 Diagnostic Protocol

Each asset uses the structure-only MOE probe. Members are ranked by consistency, concentration, evidence separation, and support focus (Figure 9); the 50 diagnostic assets are disjoint from evaluation and no evaluation score selects the readout.

B.2 Runtime Reliability Calibration

For each member, we average evidence and selected/competing support over probe observations. Consistency is mean pairwise Pearson correlation of centered nonconstant evidence fields across probe observations; concentration is the mean evidence above the fixed threshold minus the mean evidence below it; support focus is selected-query support divided by total active-query support. Thus, these statistics measure repeatability, separation, and spatial selectivity rather than semantic-label accuracy. Degenerate cases use implementation fallbacks.

The three statistics are converted to higher-is-better ranks and combined with weights 0.40, 0.35, and 0.25. The four members are fused with softmax temperature 0.15: B08/H02, B11/H09, B12/H01, and B12/H02. The temperature controls how sharply the reliability weights favor higher-ranked members; it is fixed for all assets.

B.3 Cross-Asset Stability

The four members occur in per-asset Top-5 rankings at 76%, 76%, 76%, and 80%, respectively; all four are in Top-10 for 38/50 assets. Figures 10 and 11 visualize these cross-asset recurrence and stability patterns. These diagnostics measure recurring spatial evidence rather than semantic labels; *All-block-head* is an ablation that averages all available members.

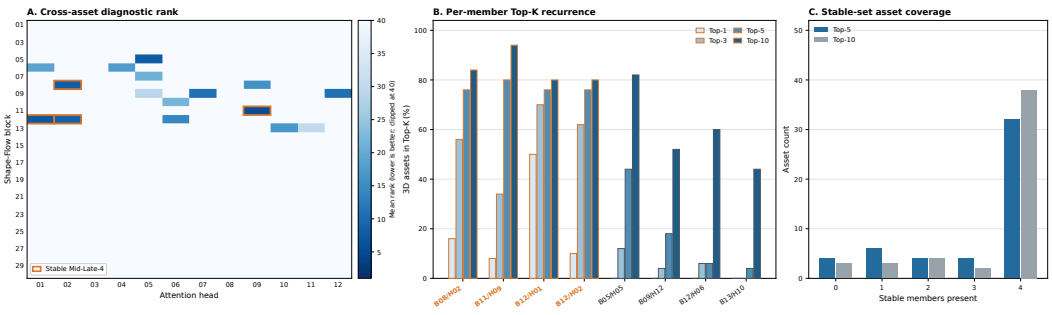

Figure 10: **Cross-asset block–head diagnostics.** Recurrence and reliability statistics for the candidate readout on the independent structure-only diagnostic split.

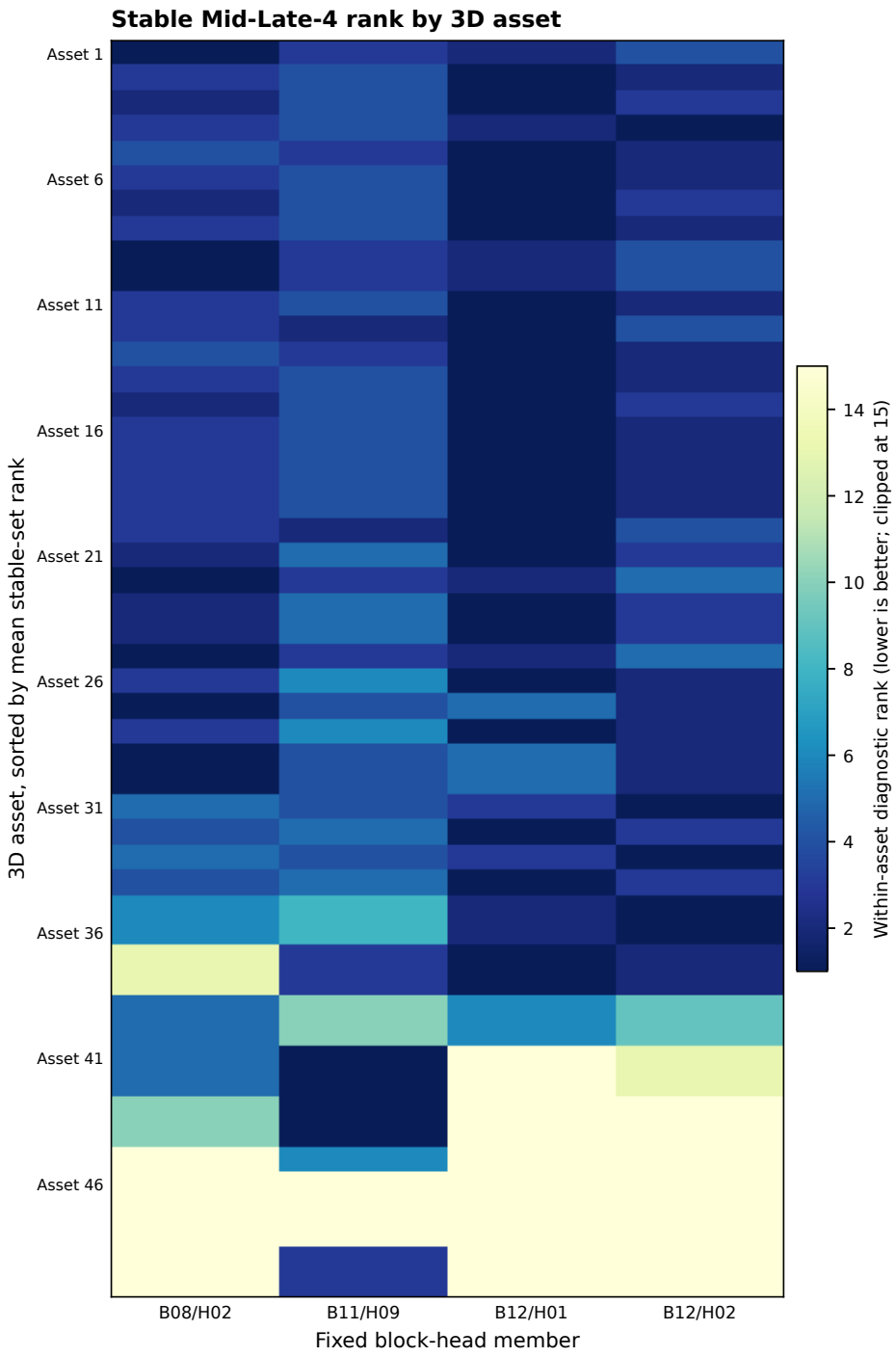

Figure 11: **Per-asset block–head stability.** Recurrence of the fixed readout members across diagnostic assets, rather than selection using the evaluation scores.

C Experiment Details

This appendix lists fixed inference, baseline, and evaluation settings.

C.1 Benchmark Asset Sources and Masks

Structure/reference images are from public assets released with Fuse3D (Jin et al., 2025), TRELLIS.2 (Xiang et al., 2026), Pixal3D (Li et al., 2026), and other listed open-source projects. Diagnostic and evaluation assets are disjoint. Masks are initialized with SAM 2 (Ravi et al., 2024), manually refined, fixed, and reused across methods.

C.2 Inference Configuration

OwnerFusion3D uses frozen TRELLIS.2-4B (Xiang et al., 2026); no generator parameter is trained. The seed is 2026; LR probing uses four steps and threshold 0.58; Stable Mid-Late-4 uses B08/H02, B11/H09, B12/H01, and B12/H02 with temperature 0.15.

The resolved field is reused for LR shape, HR shape, and texture. HR/texture sizes are 1024/4096, texture uses 12 steps, and the OEHR blocked-logit sentinel is $-10{,}000$. Masks are not dilated. Direct 1024 probes native HR support; the cascade probes 512/LR once, transfers ownership by native sparse correspondence, and independently samples HR shape and texture. Here LR and HR denote the low- and high-resolution sparse shape stages.

The DINO-I evaluator is frozen DINOv2-L `facebook/dinov2-large`; it is separate from the DINOv3 conditioner used by TRELLIS.2. Key settings are summarized in Table 4.

Table 4: **Key inference parameters.** Fixed defaults for reported runs.

Component	Parameter	Value
Global	seed; HR resolution	2026; 1024
Global	texture steps; texture size	12; 4096
MOE	probe steps; threshold τ	4; 0.58
MOE	readout; temperature	Stable Mid-Late-4; 0.15
OEHR	backend; routing scope	owner_grouped; LR/HR/texture
COR	resolution mode	full; monotone owner-zero completion

C.3 Runtime and Peak GPU Memory

Times exclude model loading, post-processing, and GLB export; memory is peak CUDA allocation in decimal GB. Logs use seed 2026, full-VRAM mode, and chunk size 2048. The table reports 50 single-mask and 20 multi-part assets. The stage-wise breakdown is reported in Table 5.

Table 5: **Generation time and stage costs.** Mean seconds per asset; `Gen.` excludes model loading, post-processing, and GLB export.

Pipeline	Setting	Prep.	MOE	OEHR-LR	OEHR-HR	COR	Texture	Gen.	Alloc.
OwnerFusion3D (single-mask)	512	1.34	0.59	2.28	–	–	1.30	5.96	19.52
OwnerFusion3D (single-mask)	1024	1.52	3.58	–	15.37	–	8.80	30.06	23.08
OwnerFusion3D (single-mask)	1024 cascade	1.54	0.41	2.49	23.65	–	13.68	42.94	23.04
OwnerFusion3D (multi-part)	512	1.82	2.34	2.43	–	0.04	1.39	10.19	19.68
OwnerFusion3D (multi-part)	1024	1.81	15.24	–	16.72	0.21	9.56	46.02	23.69
OwnerFusion3D (multi-part)	1024 cascade	2.46	2.30	2.44	34.84	0.04	9.70	55.35	23.84

Prep. covers encoding and sparse-support construction; MOE includes probing and candidate processing; COR is summed over parts. Direct 1024 probes HR support, whereas cascade probes LR once and still samples HR shape independently.

C.4 Probe-step Selection

We sweep 1–12 LR probe steps on one asset with other settings fixed. Four steps is the earliest stable visual setting; later steps add cost without improving the fixed-readout diagnostics. The numerical sweep is given in Table 6, with the corresponding visual outputs shown in Figure 12.

Table 6: Probe-length sensitivity on one 3D asset. Generation excludes GLB export.

LR probe steps	1	2	3	4	5	6	7	8	9	10	11	12
Observed layers	30	60	90	120	150	180	210	240	270	300	330	360
Selected ratio (%)	7.39	8.08	7.92	7.71	7.62	7.60	7.58	7.50	7.50	7.52	7.50	7.54
Consistency	1.0000	0.9713	0.9493	0.9322	0.9260	0.9152	0.9081	0.8920	0.8983	0.8953	0.8919	0.8793
Concentration	0.8859	0.8770	0.8599	0.8460	0.8573	0.8511	0.8472	0.8168	0.8380	0.8361	0.8324	0.8035
MOE probe (s)	0.27	0.37	0.54	0.73	0.92	1.10	1.30	1.52	1.76	1.98	2.20	2.44
Generation (s; no export)	63.59	60.21	60.40	60.88	60.76	61.39	61.08	61.45	61.73	62.34	62.01	62.25



Figure 12: **LR probe-length sensitivity.** Outputs for 1–12 probe steps with the fixed Stable Mid-Late-4 readout; four steps is the selected cost–quality operating point.

C.5 Stage-wise OEHR Visualization

The columns show LR shape-only, LR+HR shape, and LR+HR+texture outputs (Figure 13).



Figure 13: **Stage-wise OEHR visualization.** Routing at LR shape, LR+HR shape, and LR+HR+texture.

C.6 Unified Caption Construction

One caption record is built from the same structure, mask, and references for every asset. GPT-5 (OpenAI, 2025), the same model used for the pairwise VLM evaluation, generates the concise English JSON record. It describes the base structure, regional appearance, fusion intent, and global content; multi-part inputs store one description per part. The record is then converted deterministically to prompts and shared by all baselines. The single-mask and multi-part records are illustrated in Figures 14 and 15.

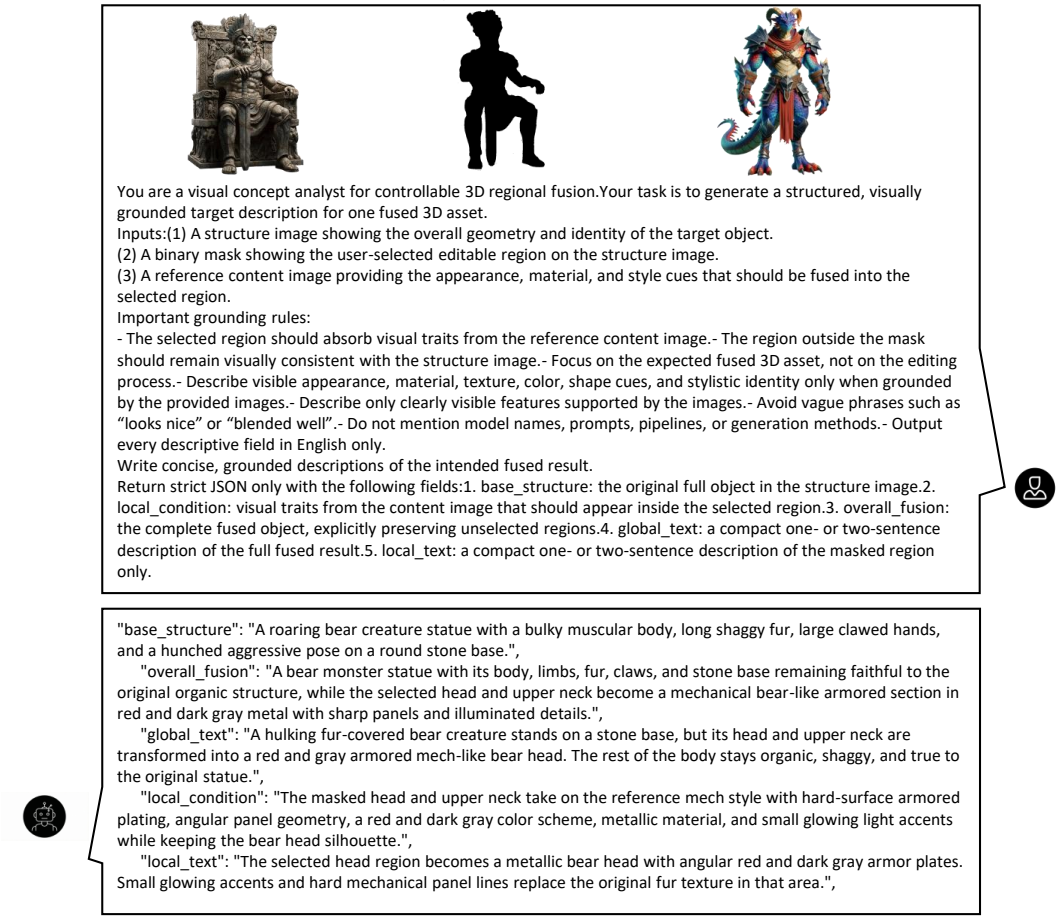

Figure 14: **Unified caption construction for single-mask fusion.** GPT-5 generates the structured caption from the shared visual inputs.

C.7 2D-First Baseline Protocol

Qwen-Image-2.0-Pro (Zhao et al., 2026) receives the structure, mask, reference, and shared fusion description; multi-part inputs are sequential. FLUX.2 (Black Forest Labs, 2025) uses the same inputs plus a color-coded part map and explicit mappings. Each composite uses the same frozen TRELLIS.2 pipeline as its comparison group; neither baseline uses MOE, OEHR, or COR.

Fuse3D (Jin et al., 2025) uses its released native interface on the original TRELLIS backbone (T1); the attempted TRELLIS.2 (T2) port was not reproducible. Fuse3D is paired with T1 structure-only references, whereas OwnerFusion3D and the 2D-first baselines use T2 references.

Additional 2D baselines include IP-Adapter (Ye et al., 2023), Blended Diffusion (Avrahami et al., 2022), and Ctrl-X (Lin et al., 2024), which provide image-prompt, masked-editing, and global structure–appearance controls. Their composites are lifted with TRELLIS.2; none has an explicit 3D ownership field. The resulting intermediate comparisons are shown in Figure 16.

C.8 Backbone-Paired Alignment and Metrics

Each structure image has one native-backbone structure-only reference, meaning the structure-only asset generated by the same backbone used for that comparison group; camera and object transforms are reused. Masks are not dilated. The selected region is compared with the reference foreground and the complement with the paired render. For each generated/reference pair, we independently



You are a visual concept analyst for controllable multi-part 3D fusion.Your task is to generate a structured, visually grounded target description for one fused 3D asset.
Inputs:(1) A structure image defining the base object identity and geometry anchor.(2) A part label image and per-group masks defining the editable parts.(3) One reference content image per editable part, each specifying the desired appearance cues for its own selected part.
Part groups:- 1. p01- 2. p02- 3. p03- 4. p04- 5. p05
Important grounding rules:- Each selected part should inherit traits only from its paired reference image.- Non-selected regions should stay visually consistent with the structure image.- Different edited parts should remain semantically separated rather than leaking into each other.- Describe visible appearance, material, texture, color, shape cues, and stylistic identity when grounded by the images.- Describe only clearly visible features supported by the provided images.- Do not mention model names, prompts, pipelines, or generation methods.- Output every descriptive field in English only.
Write concise, grounded descriptions of the intended fused result.Requirements:1. `base_structure` briefly describes the original full object in the structure image.2. `overall_fusion` describes the final fused object after all local edits.3. `global_text` is a compact 1-2 sentence caption for the full fused result.4. `groups` must contain exactly one object for every required group name, in this exact order: p01, p02, p03, p04, p05.5. Copy every `group_name` exactly as supplied above. Do not use generic labels such as `style_a`, `style_b`, `group_1`, or `part_1`.6. Each group object must include: - `local_condition`: what traits from that group's reference image should appear in the selected part. - `local_text`: a compact 1-2 sentence caption for only that selected part after fusion.7. Each local description must describe only its own selected part, not the whole object.
Return strict JSON only with this exact structural schema:
{"base_structure":"...","overall_fusion":"...","global_text":"...","groups":[{"group_name":"p01","local_condition":"...","local_text":"..."},{"group_name":"p02","local_condition":"...","local_text":"..."},{"group_name":"p03","local_condition":"...","local_text":"..."},{"group_name":"p04","local_condition":"...","local_text":"..."},{"group_name":"p05","local_condition":"...","local_text":"..."}]}

"base_structure": "A crouching, muscular reptilian monster with blue-gray scales, a horned and spiked head, glowing yellow eyes, a tan segmented belly, clawed limbs, and a long spined tail.",
"overall_fusion": "The fused creature retains its imposing reptilian silhouette while combining a rusty mechanical head, ivory-and-gold armored torso, turquoise industrial robot arms, moss-covered carved-stone legs, and a warm red-orange dragonlike tail.",
"global_text": "A heavily customized reptilian guardian with distinct mechanical, sacred armor, ancient stone, and dragon-inspired regions. Its original monster form remains clear beneath the separated material treatments.",
"groups": [
"group_name": "p01",
"local_condition": "Apply turquoise-blue industrial armor plating with orange-gold trim, dark mechanical seams, vents, joints, and layered hard-surface panels from the reference.",
"local_text": "Both arms become bulky turquoise mechanical limbs with segmented armor shells and orange-gold hardware accents. Dark joints and panel lines preserve articulated claws and forearms."
"group_name": "p02",
"local_condition": "Apply weathered gray stone, deeply carved ornamental patterns, blocky layered construction, and patches of vivid green moss from the reference.",
"local_text": "The legs become heavy ancient-stone limbs covered in carved geometric reliefs and moss. Their broad clawed stance remains intact beneath the rough monument-like surface."
"group_name": "p03",
"local_condition": "Apply the reference dragon's smooth warm red-orange skin, darker ridged spine details, and lighter orange underside accents where visible.",
"local_text": "The tail takes on a smooth red-orange dragon finish with dark raised dorsal spikes. It remains long, curled, and sharply segmented at the silhouette."
"group_name": "p04",
"local_condition": "Apply polished ivory-white armor, fine gold filigree, gold edging, small jewel-like accents, and restrained brown leather details from the reference.",
"local_text": "The torso is clad in refined ivory armor with ornate gold borders and delicate engraved decoration. The belly region reads as protective ceremonial plating rather than exposed scales."
"group_name": "p05",
"local_condition": "Apply rusty brown mechanical plating, rivets, layered metal components, exposed tubing, and warm amber glowing details from the reference.",
"local_text": "The head becomes a rugged rust-brown mechanical beast head with layered riveted plates and industrial detailing. Warm amber illumination enriches the eye area while the horned reptilian profile remains recognizable."]

Figure 15: **Unified caption construction for multi-part fusion.** GPT-5 generates one structured local description for each marked part.

compute a foreground bounding box, and resize the resulting canonical crop to 256×256. For multi-part assets, each part is scored separately and the part scores are averaged equally (macro-averaged). CLIP-I/DINO-I are primary semantic metrics, LPIPS auxiliary, and joint scores equally weight in-mask transfer and out-of-mask preservation. The canonical crop is computed independently from each image's foreground mask, so pixelwise alignment is not assumed across different object categories. CLIP-I and DINO-I are image-image semantic similarity scores, whereas LPIPS is an

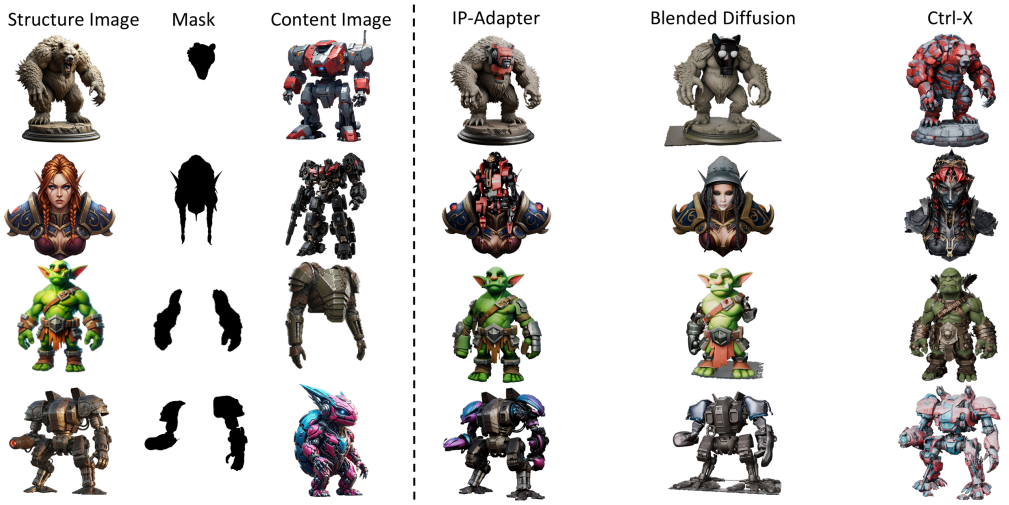

Figure 16: **Additional 2D baseline comparison.** Intermediate outputs from IP-Adapter, Blended Diffusion, and Ctrl-X.

auxiliary perceptual distance; joint scores are reported so transfer and preservation are considered together.

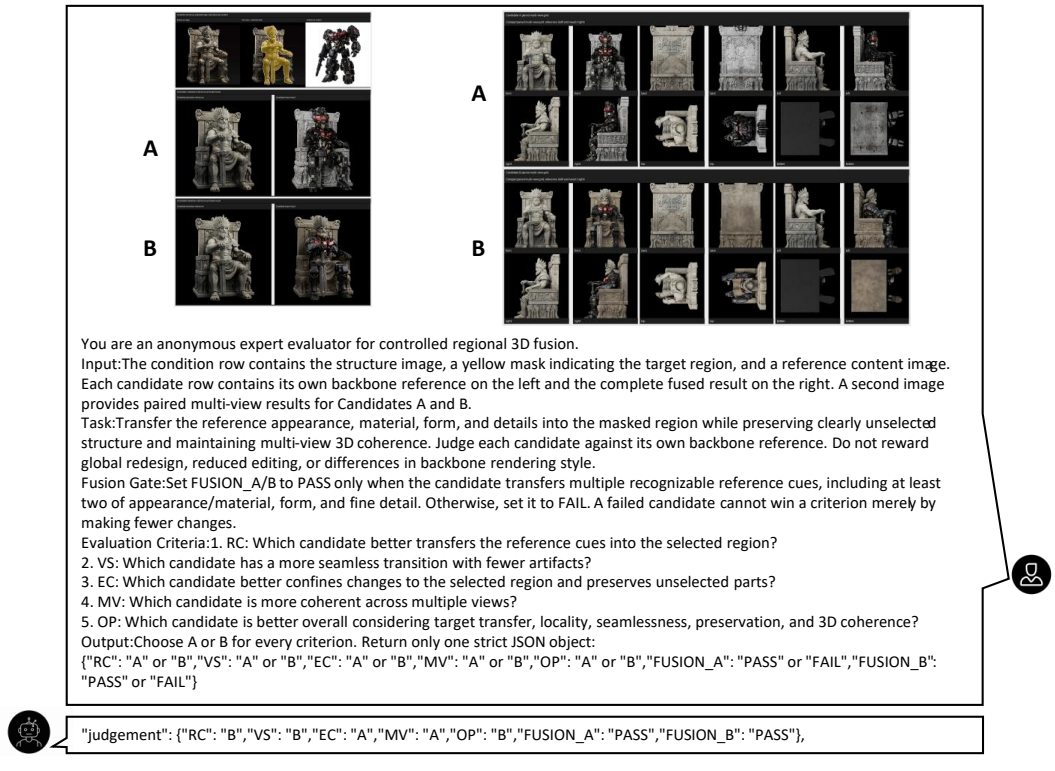

Figure 17: **Mask-aligned pairwise VLM evaluation.** Candidates are compared with their own backbone references under identical conditions.

C.9 GPTEval3D and Pairwise VLM Protocol

GPTEval3D (Wu et al., 2024) uses GPT-5 (OpenAI, 2025) for pairwise evaluation. Method names are hidden, candidates are randomized to A/B, and both are shown with native references, aligned results, masks, content references, and multi-view renders. Criteria cover part fidelity, boundary/cross-part control, unmasked preservation, multi-view coherence, and overall preference. The presentation and decision layout are illustrated in Figure 17. Each criterion returns one binary A/B choice; 50% is parity.

D Qualitative Ablation and Additional Fusion Results

This appendix provides qualitative ablations, ownership visualizations, transfer results, and additional mask-controlled assets. The notation "w/o X" denotes the controlled variant that removes module X while keeping the remaining pipeline and inputs fixed.

D.1 Qualitative Ablation Comparison

Figure 18 compares w/o MOE, w/o OEHR, All-block-head, and Full (Ours) under identical inputs. All-block-head averages the evidence from every available block–head member, whereas Full uses the fixed calibrated readout described in Appendix B.



Figure 18: **Qualitative ablation comparison.** Four variants under identical structure, reference, and mask.

D.2 Sparse Ownership Visualizations

Figure 19 visualizes source assignments on sparse 3D supports, where each colored query corresponds to an active generator coordinate and its assigned reference source.



Figure 19: **Sparse ownership visualizations.** Source assignments on sparse 3D supports.

D.3 Cross-Generator Transfer

We adapt the ownership interface to Hi3DGen (Ye et al., 2025), Direct3D-S2 (Wu et al., 2025), and Pixal3D (Li et al., 2026) through native sparse attention or projection interfaces. Figure 20 shows local transfer with outside-mask structure retained.



Figure 20: **Cross-generator transfer.** OwnerFusion3D instantiated on Direct3D-S2, Hi3DGen, and Pixal3D under the same inputs.

D.4 Failure Case and Scope of the Limitation

Figure 21 illustrates a failure case in which the transferred eagle talons do not preserve a convincing avian appearance. The mask selects a thin, articulated region, while the content reference contains an armored character holding a cup. The 2D mask establishes where the content source should be used, but it does not specify complete 3D correspondence, part topology, articulation, or a physically plausible attachment boundary. Consequently, the sparse ownership field can localize the transfer correctly yet still provide insufficient evidence to reconstruct fine talon geometry. The generator then completes the underconstrained region using its learned shape and texture priors, which may produce locally inconsistent claws or a part that reads as a generic mechanical appendage. This example separates local source control from guaranteed semantic or geometric correctness and motivates future work on explicit correspondence and fine-detail constraints.



Figure 21: **Failure case.** A thin eagle-talon region is selected for local fusion. Ownership localizes the transfer, but the generated talons do not retain a fully convincing avian shape.

D.5 More Fusion Assets

Figure 22 shows additional single-region results, including cross-category structure–reference pairs. Figure 23 shows multi-region results in which colored mask regions specify distinct part–reference assignments. Together, these examples illustrate local appearance transfer while the complementary structure and assignments remain separated.



Figure 22: **More fusion assets: single-region results.** Cross-category structure–reference pairs with a selected transfer region.



Figure 23: **More fusion assets: multi-region results.** Colored regions specify distinct part–reference assignments.