HGCalVis: Web-Based Importance-Ordered Visualization of CERN's HGCal Particle Showers

Supplementary Material: Detailed Description and Complete Results of the Preliminary User Study

This document accompanies the main paper and expands Section 7 (the preliminary user study) with the full methodology, the complete set of descriptive statistics and significance tests, three figures, the per-participant data, and the verbatim open-ended responses. It is preliminary: N=19, mixed expertise, and intended to motivate a larger replication rather than to settle the questions definitively.

1 Study Design

Goal. The study tests whether each of the three main visual contributions of HGCalVismeasurably helps on a task it is designed to support: (i) depth cues from cylinder shading and screen-space ambient occlusion (SSAO), (ii) Importance-Ordered Rendering (IOR), and (iii) the linked 2D schematic view with multi-attribute filtering. It is a within-subjects design: every participant performs every task under every condition. The study is not an evaluation of the prototype's interface, which was deliberately left unpolished.

Tasks and conditions. Three tasks, each isolating one feature group (Table 1). T_1 and T_3 contrast a control condition (feature disabled) with a treatment condition (feature enabled). T_2 uses three conditions, so that the contribution of the energy *colour map* can be separated from that of the importance-ordered *compositing*: (a) no cue, (b) trajectories coloured by energy cluster but rendered with ordinary depth compositing (occlusion intact), and (c) full IOR(energy clusters plus importance-ordered compositing that keeps high-energy tracks unoccluded).

Table 1: Tasks, the feature each isolates, and the conditions tested.

Task	Prompt	Conditions
T_1	"Of the two highlighted trajectories, which one is closer to the camera?"	depth cues off / on
T_2	"Click on the highest-energy trajectory in this shower."	no cue / colour only / IOR
T_3	"Find all direct daughter trajectories of the highlighted parent track."	schematic view + filtering off / on

Datasets and counterbalancing. Trials were drawn from the four production HGCal datasets used in the main paper, assigned deterministically so that no participant saw the same dataset twice and each dataset was used roughly equally across the cohort. The camera was reset to a fixed viewpoint at the start of every trial. Within each task, the order of conditions was counterbalanced across participants; task order was fixed ($T_1 \rightarrow T_2 \rightarrow T_3$).

Ground truth and metrics. Correct answers were pre-computed offline from the shower data and never shown to participants. For T_1, the two highlighted trajectories were chosen so their screen-space projections overlap while their depth differs by at least one detector unit; the correct answer is the nearer one (binary). For T_2, the correct answer is the unique trajectory with the largest end-point energy (binary; a click counts as correct if it falls within the bounding cylinder of that trajectory). For T_3, the reference set is every trajectory whose parent is the highlighted track; we report both exact set recovery (binary) and the F1 score over the set. Every task also offered an explicit "I don't know" response. Completion time was measured from prompt display to answer submission.

Procedure. The study ran entirely in the browser as a self-administered session of about 20 minutes: consent and demographics (2 min); an interactive tutorial on a small non-HGCal dataset (3 min); seven measured trials (T_1 and T_3 with two conditions, T_2 with three), each preceded by a 3-second "get ready" countdown ($\sim$12 min); and a short questionnaire (2 min). No experimenter was present.

Exclusions. Following the pre-registered rule, trials longer than five minutes are flagged as possible timeouts. Two T_3 treatment trials exceeded this threshold (552 s and 443 s), both on low-end hardware. They are retained in all results below; excluding them does not change the direction or significance of any reported effect (it only widens the T_3 accuracy and time gaps).

Participants. Nineteen volunteers completed the study (Table 2), unpaid, recruited through the authors' institutional and collaborator networks. Six were physicists or physics students; the rest had technical but non-physics backgrounds.

Table 2: Participant demographics (N=19). Familiarity is self-rated on a 1–7 scale.

Attribute	Categories	Count / value
Role	Engineer / computer scientist	11
	Physicist / physics student	6
	Other technical	2
Age	25–34 / 35–44 / 45+	5 / 6 / 8
Gender	Male / Female	12 / 7
Familiarity	Particle physics (mean±sd)	4.2 ± 2.0
	3D scientific vis. (mean±sd)	4.9 ± 1.5

2 Complete Results

2.1 Descriptive statistics

Table 3 reports, for every task and condition, the accuracy and the distribution of completion time. Figure 1 visualises the accuracies and Figure 2 the time distributions.

2.2 Paired significance tests

Accuracy differences were tested with the exact McNemar test (appropriate for paired binary outcomes at this sample size); completion-time differences with the Wilcoxon signed-rank test (Table 4). IORsignificantly improved T_2 accuracy over unordered rendering; the schematic view significantly improved T_3 accuracy. Depth cues left T_1 accuracy at ceiling but significantly reduced time. The T_2 colour-only versus IORcontrast trends in the same direction (McNemar p=0.07), while no cue versus colour-only shows no difference (p=1.0).

Table 3: Per-condition accuracy and completion-time statistics (N=19 trials per cell). For T_3, "Acc." is exact recovery of the daughter set and "F1" is the mean F1 over that set; times in seconds.

Task	Cond.	Acc.	mean	sd	median	IQR	range
T_1	off	0.84	10.2	7.1	7.7	5.2–13.5	1.0–26.7
	on	0.95	6.5	4.1	6.9	2.9–9.4	1.0–15.5
T_2	no cue	0.11	54.6	57.0	37.9	15.8–61.5	7.9–195.2
	colour	0.11	23.1	19.3	20.1	11.0–28.5	3.9–81.3
	IOR	0.42	49.3	49.8	33.7	19.7–53.5	2.3–222.0
T_3	off	0.37	62.8	80.1	26.9	12.9–72.0	1.0–255.1
	on	0.79	99.7	146.8	50.3	21.7–102.6	5.9–552.4
T_3 F1		off: 0.62	on: 0.83				

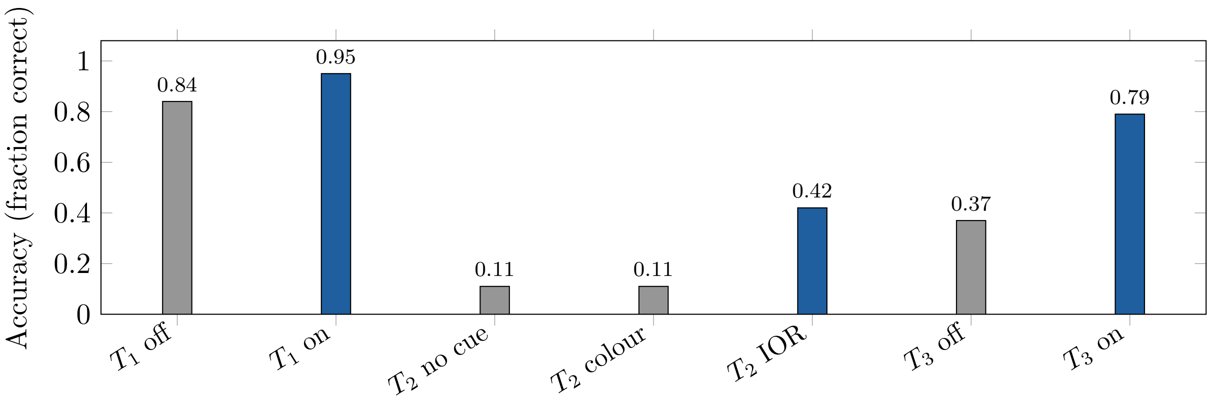

Figure 1: Accuracy by task and condition. Grey = control / intermediate conditions, blue = treatment. The T_2 panel is the key ablation: energy colour-coding alone leaves accuracy at the no-cue level (0.11), whereas full IORraises it to 0.42, locating the effect in the importance-ordered compositing rather than the colour map.

2.3 Effect across expertise

The IORbenefit on T_2 appears in both expertise groups (Table 5). Non-physicists could not identify the highest-energy trajectory without the cue (0.00) and reached 0.31 with IOR; physicists improved from 0.33 to 0.67. Subgroup sizes are small (n=6 and n=13) and these figures are indicative only.

2.4 Subjective questionnaire

The three seven-point Likert items, answered after the trials, were: **S1** "the 3D shading and ambient occlusion made the spatial layout easier to read" (depth cues); **S2** "the importance-ordered rendering helped me identify the high-energy trajectories" (IOR); and **S3** "the hierarchical schematic view and filtering helped me navigate the data" (schematic view). Figure 3 shows the full response distributions and Table 6 the summary statistics. Agreement is highest for the IORitem (S2: mean 5.8, 84% of responses $\geq$ 5).

2.5 Per-participant data

Table 7 gives the complete per-participant record: demographics, the correctness of every trial, the T_3 F1 scores, and the three Likert ratings.

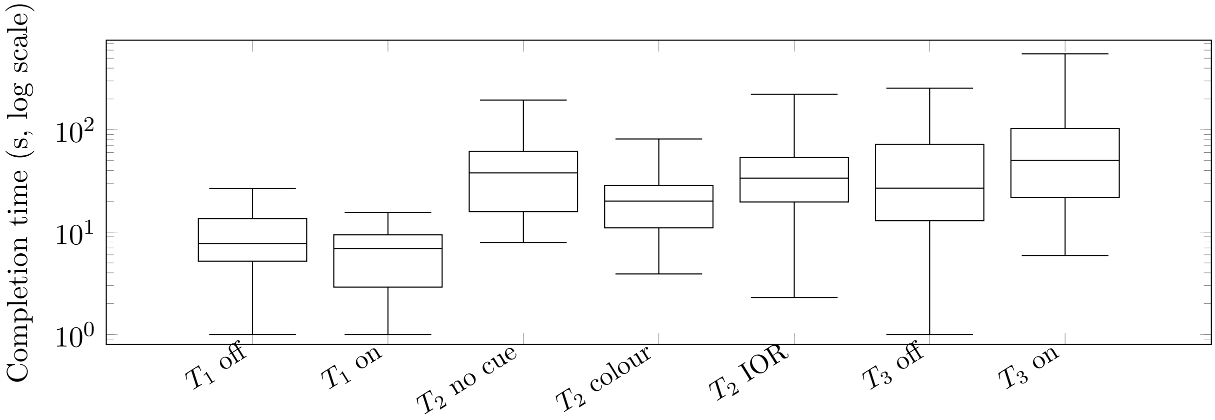

Figure 2: Completion-time distributions (whiskers span the full min–max range; log scale). Depth cues (T_1) shift times downward; the schematic view (T_3, on) increases time, consistent with deliberate tracing of the hierarchy rather than guessing.

Table 4: Paired tests, control (A) vs treatment (B). McNemar discordant pairs (b, c): b favouring treatment, c favouring control.

Task	Accuracy (A→B)	McNemar (b, c), p	Time Wilcoxon z, p
T_1	$0.84 \to 0.95$	$(3, 1)$, p=0.63	z=−2.45, p=0.01
T_2	$0.11 \to 0.42$	$(6, 0)$, p=0.03	z=−0.60, p=0.55
T_3	$0.37 \to 0.79$	$(8, 0)$, p<0.01	z=−1.97, p=0.05

2.6 Open-ended responses

The optional final question asked, "In one sentence: what did you find most useful?" Eight participants responded; the responses are reproduced verbatim below (participant identifier and role in brackets).

> **[P16, physicist]** "The visual aids helped interact with the collision data and better understand how trajectories move through 3D space."

> **[P10, physicist]** "Good stuff :)"

> **[P02, engineer]** "My low confidence is mostly due to my little understanding on the topic, I've found the UI pretty user friendly in any case."

> **[P11, engineer]** "Transparency would help."

> **[P06, engineer]** "The colour coding was misleading at the last task; I had to go back and forth twice to make sure that the highlighted parent in the 3D rendering (red) is the same as the one in the 2D view (blue)."

> **[P03, physicist]** "Not suitable for slow pc."

> **[P13, engineer]** "No comment."

> **[P05, other]** (abridged) ". . . The tree graph is not explained, but it is sort of intuitive. . . What I like: graphics are nice, although slow on my computer. . . I could not see the beginnings of the trajectories because they overlap like knots. Thank you, it was cool!"

The substantive responses are consistent with the quantitative results: a physicist linked the visual aids to improved 3D understanding, while the critical comments target the prototype interface (colour legend consistency, absence of an in-app tutorial, performance on low-end hardware) rather than the visual techniques themselves. The remark that trajectories "overlap like knots" is an unprompted description of the occlusion problem that IORaddresses.

Table 5: T_2 accuracy by participant group.

Group	no cue	colour only	IOR
Physicists (n=6)	0.33	0.17	0.67
Non-physicists (n=13)	0.00	0.08	0.31

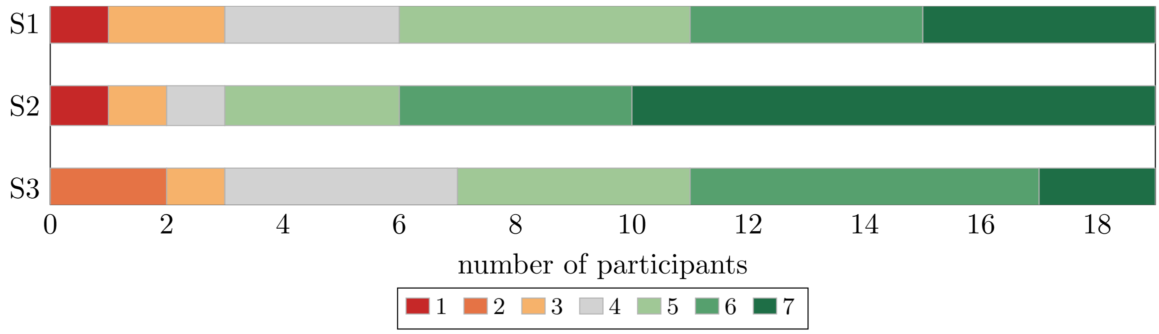

Figure 3: Distribution of Likert responses (1 = strongly disagree, 7 = strongly agree) for the three questionnaire items. S1 = depth cues, S2 = IOR, S3 = schematic view.

Table 6: Likert summary (1–7 scale, N=19).

Item	Topic	mean±sd	95% CI	median	≥ 5
S1	depth cues	5.05 ± 1.61	[4.27, 5.83]	5	68%
S2	IOR	5.79 ± 1.65	[4.99, 6.59]	6	84%
S3	schematic view	4.89 ± 1.49	[4.18, 5.61]	5	63%

Table 7: Per-participant data. Ph/Gx = self-rated particle-physics / 3D-graphics familiarity (1–7). ✓/× = correct/incorrect. T_2 columns: n = no cue, c = colour only, i = IOR. $\mathrm{F1}_a$/$\mathrm{F1}_b$ = T_3 F1 off/on.

P	Role	Ph	Gx	T_1a	T_1b	T_2n	T_2c	T_2i	T_3a	T_3b	$\mathrm{F1}_a$	$\mathrm{F1}_b$	S1	S2	S3
P01	Eng	4	6	✓	✓	×	×	✓	✓	✓	1.00	0.86	6	6	5
P02	Eng	4	6	✓	✓	×	✓	×	✓	✓	1.00	1.00	6	6	4
P03	Phys	3	2	✓	✓	✓	✓	✓	×	×	0.00	0.00	5	5	6
P04	Other	3	5	✓	✓	×	×	×	×	×	0.40	0.67	3	3	2
P05	Other	5	4	✓	✓	×	×	×	×	✓	0.08	0.86	3	5	5
P06	Eng	3	7	✓	✓	×	×	×	✓	✓	1.00	1.00	4	7	4
P07	Eng	5	7	✓	✓	×	×	✓	×	×	0.00	0.55	6	5	3
P08	Eng	2	6	✓	×	×	×	×	✓	✓	1.00	1.00	7	7	6
P09	Eng	1	5	✓	✓	×	×	×	×	✓	0.80	0.86	5	7	4
P10	Phys	7	6	✓	✓	×	×	×	×	✓	0.80	1.00	7	7	7
P11	Eng	6	6	✓	✓	×	×	✓	✓	✓	1.00	1.00	5	7	6
P12	Phys	5	3	×	✓	×	×	✓	×	✓	0.80	1.00	5	7	6
P13	Eng	1	5	×	✓	×	×	×	×	✓	0.50	1.00	1	1	2
P14	Eng	1	6	✓	✓	×	×	×	×	✓	0.80	1.00	6	7	6
P15	Eng	5	5	✓	✓	×	×	×	×	×	0.00	0.00	7	7	5
P16	Phys	6	3	×	✓	×	×	✓	×	✓	0.80	1.00	4	6	6
P17	Phys	7	3	✓	✓	✓	×	✓	✓	✓	0.86	1.00	5	6	5
P18	Phys	7	3	✓	✓	×	×	×	✓	✓	1.00	1.00	4	4	4
P19	Eng	5	5	✓	✓	×	×	✓	×	✓	0.00	1.00	7	7	7