How Physics Fidelity Shapes Perceived Plausibility in Mixed Reality: A Case Study of Bounce Behavior

Hyunsoo Kim, Jingyu Lee, Rajesh Balan and Youngki Lee

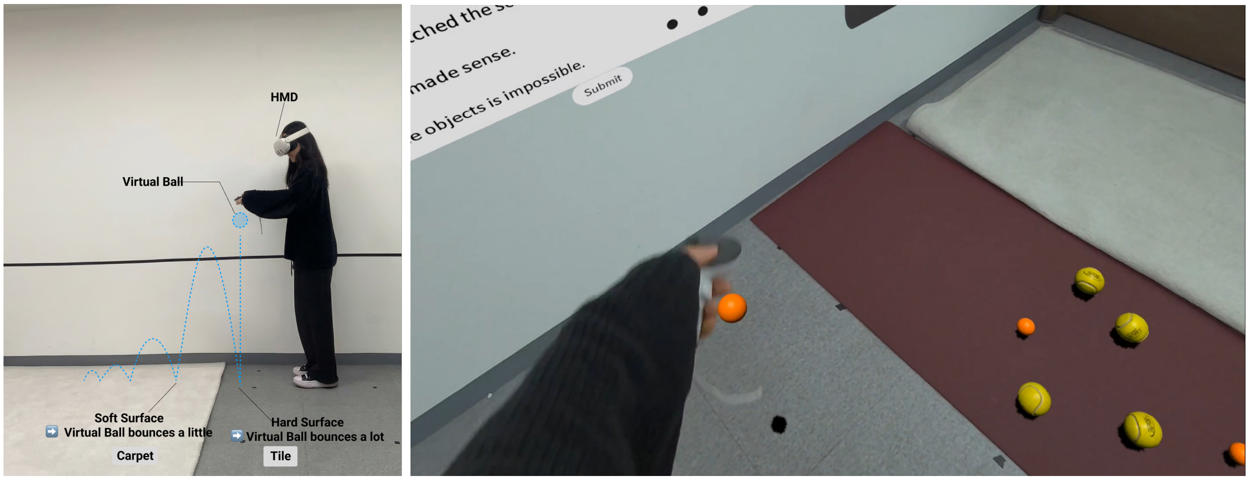

Figure 1: Left: Conceptual image of our MR ball-bouncing task across real-world surfaces. A participant drops a virtual ball onto visible real surfaces; the illustrated trajectories contrast the lower rebound on the soft surface (carpet) with the higher rebound on the hard surface (tile). Right: Experimental setup and in-headset view of the MR ball-bouncing task. The in-headset view shows the three surfaces used in the experiment (tile, yoga mat, and carpet) with the virtual balls (tennis and pingpong balls) being dropped by the user.

Abstract—Mixed Reality (MR) systems increasingly place virtual objects on visible real-world surfaces, yet it remains unclear how users judge the physical plausibility of these interactions when rebound behavior departs from reality. We study this question through ball bouncing, a simple but expressive MR interaction in which plausibility depends on both object properties and surface-dependent response. In a 4 × 2 within-subjects study, 20 participants interacted with virtual tennis and pingpong balls on three real surfaces under four rebound models, and evaluated each model before and after direct familiarization with the corresponding real balls and surfaces. The Rebound Model significantly affected perceived plausibility ($p < 0.001$), and this effect depended on the familiarization phase ($p < 0.001$), whereas phase alone showed no overall effect ($p = 0.399$). Before familiarization, participants showed only weak differences across models, but after familiarization they clearly distinguished the calibrated model from the alternatives. These findings suggest that, in familiar MR interactions, preserving the expected surface-dependent structure of motion may matter more than strict physical accuracy, offering practical guidance for designing future MR experiences that balance physical plausibility and system complexity.

Index Terms—Mixed reality (MR), Augmented Reality (AR), plausibility, experience, physics fidelity, familiarity.

1 Introduction

Mixed Reality (MR) systems increasingly place virtual objects on visible real-world surfaces [14, 28]. Yet, visually coherent placement alone does not guarantee a believable interaction. A virtual ball may appear well anchored to the floor, but still feels wrong if it rebounds in the same way on tile, carpet, and a yoga mat. As scene reconstruction and geometric alignment continue to improve, a key remaining challenge is not only where virtual objects should appear, but also how they should *behave* when they contact the physical environment.

This limitation is clear in current vision-centric MR platforms [2, 15, 16]. Commodity systems can now reconstruct scene geometry with sufficient accuracy for placement, occlusion, and coarse collision, but they usually treat real surfaces as geometric support rather than as physically distinct substrates. As a result, different surfaces can yield valid collisions while producing the same or overly generic response. This creates MR experiences that are visually grounded but physically shallow. That gap matters because many target MR applications involve everyday interactions for which users already have strong expectations about how objects should move.

In this paper, we study how physics fidelity shapes perceived plausibility in MR through a controlled case of ball bouncing on real surfaces. Ball–surface interaction is simple, observable, and easy to compare across conditions, yet it captures a core aspect of physical groundedness in MR. It also maps well to capabilities already available on commodity headsets, including passthrough, surface detection, and rigid-body simulation. Our goal is not to argue that full physical simulation is always necessary. Instead, we ask which deviations from real-world behavior become noticeable, and under what user conditions they matter.

A key challenge is that perceived plausibility depends not only on the simulation itself, but also on the user's reference for judging it. Before direct exposure, users may rely on prior expectations about how different balls should rebound on different surfaces. Those expectations may be incomplete, object-specific, or inaccurate, yet they can still shape

• *Hyunsoo Kim is with Seoul National University (SNU). E-mail: hyunsookim.hsk@gmail.com*
• *Jingyu Lee is with SNU. E-mail: jingyu.lee@hcs.snu.ac.kr*
• *Rajesh Balan is with Singapore Management University. E-mail: rajesh@smu.edu.sg*
• *Youngki Lee is with SNU. E-mail: youngki.lee@gmail.com*

Manuscript received xx xxx. 202x; accepted xx xxx. 202x. Date of Publication xx xxx. 202x; date of current version xx xxx. 202x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.202x.xxxxxxx

plausibility judgments. After users interact with the corresponding real balls and surfaces, however, they gain a fresh external benchmark that may change what mismatches they notice. We thus examine two user-side factors together: *baseline mental-model accuracy*, measured by how closely participants' expected rebound heights matched real-world values, and *recent real-world familiarization*, provided through direct interaction with the corresponding physical balls and surfaces.

To study these questions, we conducted a controlled 4×2 within-subjects experiment. Twenty participants interacted with virtual tennis and pingpong balls on three visible real surfaces—tile, carpet, and yoga mat—under four rebound models that varied in how closely the simulated behavior matched a calibrated real-world reference: **(A)** a Calibrated model, **(B)** a Uniform surface-insensitive model, **(C)** a Damped but correctly ordered model, and **(D)** a Reversed surface-dependent model. Participants evaluated all four models both before and after directly interacting with the corresponding real balls and surfaces (Pre vs Post-familiarization). Before this familiarization process, we also measured each participant's baseline mental-model accuracy through a separate estimation task.

Our results show that, overall, perceived plausibility depended strongly on the Rebound Model. This dependency became more pronounced after the familiarization phase, that is, after the participants interacted with the corresponding real balls and surfaces. Before familiarization, participants showed only weak differences among the four models, whereas after familiarization they clearly found the Calibrated model more plausible than the other models. These findings suggest that perceived plausibility in MR depends not only on physics fidelity itself, but also on the experiential reference through which users sharpen their evaluation of the interaction.

Our qualitative findings further suggest that participants were not judging plausibility based on bounce magnitude alone. They used the relative bounce pattern across surfaces and ball types as one important diagnostic cue, alongside rebound magnitude and temporal behavior. Baseline mental-model seemed to also depend on the familiarization: before direct real-world interaction, participants' prior expectations shaped how they interpreted the virtual behavior; afterward, direct experience provided a reference point. Together, these findings suggest that perceived physical plausibility in MR depends not only on simulation fidelity itself, but also on the experiential reference through which users evaluate the interaction.

This paper makes three contributions.

1. It provides an empirical account of how users judge the plausibility of ball–surface interactions in MR when rebound behavior varies in its fidelity to real-world physics.
2. It shows that these judgments depend not only on the simulated behavior itself, but also on the user's experiential reference: before familiarization, plausibility judgments were shaped in part by baseline mental-model accuracy, whereas after direct real-world exposure, users more clearly differentiated the calibrated model from the alternatives.
3. It provides practical design guidance for familiar, directly comparable MR interactions.

2 Related Work

2.1 Visual Fidelity

Visual realism has long been a central goal in VR and MR research. Prior work has improved the visual integration of virtual content into the real world through geometry-aware rendering, including occlusion, shadows, scene reconstruction, and relighting. Early studies showed that shadow and occlusion cues can substantially improve visual coherence in optical see-through MR, and later systems used scene geometry and depth sensing to support more robust visually grounded interaction on commodity devices [4, 9, 21]. More recent work has extended this direction with increasingly photorealistic relighting and compositing pipelines for reconstructed scenes and radiance fields [29]. Together, these efforts have made visual fidelity—that is, how well virtual objects appear embedded in the visible environment—a mature and well-studied dimension of MR realism.

2.2 Physics Fidelity and Perceived Plausibility

Compared to visual fidelity, physics fidelity remains less systematically studied in MR. Prior MR systems have shown that geometry-aware collision and material-aware interaction can make real-virtual interactions more believable. For example, *Forked!* demonstrated that physics-engine-based interaction can improve the believability of real-virtual object manipulation in MR [3], while later work used measured coefficients of restitution to simulate more realistic collision responses [22]. Other systems further explored semantic and material-aware scene representations so that virtual objects could respond differently to different classes of real surfaces [7]. However, this line of work has primarily focused on enabling such interactions, rather than characterizing how users perceive deviations in physical behavior once those interactions occur.

Evidence from VR and MR perception suggests that this distinction matters. Prior work has shown that object dynamics such as falling, collision, and rebound can directly shape plausibility, and that users do not judge realism from physical accuracy alone. For example, Do and Lee found that perceived collision reliability depends on how users integrate visual and bodily cues, rather than only on the default response of a physics engine [8]. More broadly, plausibility studies have shown that violations of expected object behavior can reduce subjective realism even when the surrounding environment remains otherwise convincing [5, 20, 27]. Particularly relevant to our work, Pouke et al. found that physically accurate rigid-body dynamics can be judged as *less* realistic than "movie physics" under scale manipulations [18], and later showed that contextual cues such as virtual character identity can substantially alter judgments of physics coherence [19]. Together, these findings suggest that perceived plausibility is shaped not only by physics fidelity, but also by expectations, context, and experiential reference. Our work extends this question to MR by examining how users judge surface-dependent rebound behavior on visible real-world surfaces, and how those judgments change with prior expectations and direct real-world familiarization.

2.3 Prior Expectations and Recent Experience

Users often evaluate physical behavior through prior expectations formed from everyday perceptual experience [26]. For familiar object interactions, these expectations may include assumptions about material properties and how objects should behave on different surfaces [1]. For example, people may already have beliefs about how a tennis ball or a pingpong ball should rebound from hard or soft surfaces. These expectations can guide plausibility judgments before users encounter a specific virtual simulation, but they may also be incomplete, approximate, or inaccurate. In this sense, prior expectations provide one reference point against which users judge whether virtual behavior feels physically plausible.

Recent experience can also affect what people notice and how they interpret later stimuli [12, 23]. Repetition priming describes one such effect: after people encounter a stimulus, the same or a related stimulus can become easier to process, recognize, or distinguish later [23]. In visual search, for example, recently seen features such as color, location, or spatial frequency can guide attention in later trials, making related targets easier to detect or identify [12, 13]. This suggests that recent direct experience may also provide a concrete reference for evaluating later behavior, making relevant differences easier to notice.

These two sources of reference are particularly relevant to MR, where virtual behavior can be compared with familiar physical interactions. We therefore examine how plausibility judgments vary with baseline mental-model accuracy and recent real-world familiarization.

3 Research Methodology

3.1 Overview and Research Questions

While prior work has focused on either enabling physically grounded MR interactions or understanding the broader perceptual roles of expectations and recent experience, less is known about how perceived plausibility varies across deviations in simulated physical behavior and how user familiarity shapes these judgments in MR.

To address this gap in a controlled setting, we focus on one foundational and highly observable interaction: ball bouncing. Bounce behavior provides a useful probe because it is common, dynamic, and directly shaped by both object and surface properties. It therefore offers a simple but expressive setting for evaluating simulated physical behavior against visible real-world surfaces.

Accordingly, our study addresses the following questions:

- **RQ1.** Does perceived plausibility differ across physics-fidelity conditions?
- **RQ2.** Does user familiarity shape plausibility judgments?

To answer these questions, we conducted a controlled 4 × 2 within-subjects study with two factors: Rebound Model (to examine the effect of physics fidelity) and Familiarization Phase (to examine the effect of user familiarity). We also measured baseline mental-model accuracy through a pre-study estimation task, allowing us to examine user familiarity in terms of both recent exposure and prior expectations.

3.2 Study Design

Rebound model. The primary manipulated factor in the study was the rebound model. We define the manipulated physics-fidelity condition in terms of how closely the rebound dynamics of a virtual ball match measured real-world behavior across different surfaces. When designing the models, we considered two aspects of rebound behavior: (i) rebound magnitude, represented by the approximate height reached after impact; and (ii) the surface-dependent ordering, represented by whether relatively high- and low-rebound surfaces maintained the same relationship observed in the real world.

To examine how different types of physical mismatch affect perceived plausibility, we constructed four rebound models that systematically varied their relationship to measured real-world behavior. The models differed in whether they preserved the overall rebound magnitude and the surface-dependent response pattern observed in the real world.

- **(A) Calibrated.** This model served as the approximately realistic reference condition. It was calibrated to reproduce the measured real-world trajectories for each ball–surface combination as closely as possible, both in terms of rebound magnitude and the surface-dependent ordering.
- **(B) Uniform.** This model removed any surface-dependent variation: the virtual balls used the same moderate rebound response on tile, yoga mat, and carpet, regardless of the measured differences among these surfaces. Uniform therefore served as a surface-insensitive baseline.
- **(C) Damped.** This model preserved the same surface-dependent response pattern of Calibrated, but systematically reduced the rebound magnitude. Surfaces producing relatively higher rebounds in the reference condition Calibrated therefore remained bouncy in Damped, but the resulting trajectories had consistently lower rebound heights.
- **(D) Reversed.** This model deliberately assigned the surface responses of Calibrated to different surfaces. Consequently, surfaces that produced relatively high rebound in the Calibrated condition instead produced relatively low rebound, and vice versa. Reversed was therefore unrealistic not only in magnitude, but also in its material-dependent pattern.

Together, the four models represented distinct types of deviation from real-world rebound behavior: an approximately realistic reference (Calibrated), a surface-insensitive baseline (Uniform), a systematically damped but still correctly ordered response (Damped), and an inverted surface-dependent response (Reversed). This design allowed us to examine users' sensitivity to both overall rebound magnitude and expected relative bounce patterns across surfaces.

Familiarization phase. The second manipulated factor in the study was the familiarization phase, which examined how recent direct

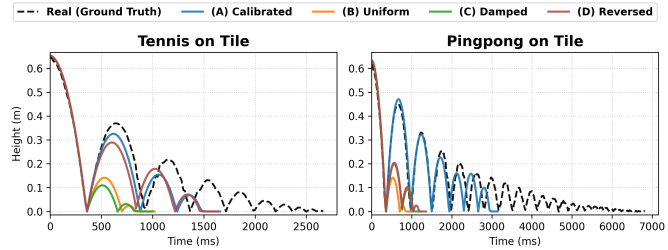

Figure 2: Measured and simulated bounce trajectories for tennis and pingpong balls on tile. Dashed black lines show the real-ball trajectories, and the colored lines show the four rebound models.

exposure to the corresponding real-world interaction affected plausibility judgments. We operationalized this factor through a familiarization phase, in which participants performed the real-world counterpart of the virtual interaction by dropping the corresponding physical balls onto the real surfaces and observing their rebound behavior. Participants evaluated the virtual ball-surface interactions both before (*Pre*) and after (*Post*) this familiarization phase. The Pre phase therefore captured judgments made without an immediate real-world reference, whereas the Post phase captured judgments made after recent direct exposure to the real-world interaction. This allowed us to examine whether real-world familiarization changed how participants evaluated virtual bounce behavior.

Baseline mental-model accuracy. Baseline mental-model accuracy was a participant-level covariate reflecting how closely participants' prior expectations of ball–surface rebound behavior aligned with real-world behavior. It complemented Familiarization Phase by capturing pre-existing expectations rather than recent direct exposure.

Together, Familiarization Phase and baseline mental-model accuracy captured two aspects of participants' reference state: recent direct exposure and pre-existing expectations, respectively.

3.3 Rebound Model Implementation and Calibration

We implemented the four rebound models in Unity with NVIDIA PhysX [17]. We manipulated the `PhysicMaterial.bounciness` values of virtual balls and real-surface colliders. This parameter controls how much velocity is retained in the opposite direction along the collision normal after impact: 0 indicates no bounce, whereas 1 indicates the maximum supported bounce response [25].

To obtain the values used in the *Calibrated* model, we first recorded real balls being dropped on each surface using a ZED Mini camera. We tracked the ball centroid in each frame using a DETR ResNet-50 detector [6] and reconstructed its 3D trajectory using the camera's depth information.

We then reproduced the same drop conditions in Unity and used Particle Swarm Optimization (PSO) [11] to identify the balls' and surfaces' bounciness values that minimized the difference between the simulated and measured real-world trajectories.

The other models were derived from these Calibrated values. Uniform assigned a bounciness value of 0.5 to all balls and surfaces. Damped was generated by subtracting 0.3 from each Calibrated value and limiting values below 0 to 0. Reversed reassigned the Calibrated surface values within each ball type: the carpet value was assigned to tile, the tile value to yoga mat, and the yoga mat value to carpet.

The resulting bounciness values for all ball-surface combinations are reported in Supplementary Material Table 1. Figure 2 shows representative measured and simulated trajectories for the tennis and pingpong balls on tile. Tile is presented as a representative example; the experiment included all three surfaces.

3.4 Controlled Variables

To isolate differences in physics fidelity, other physical properties and sensory cues were held constant across all models. Physical parameters such as object size, mass, linear drag, and friction were fixed in Unity

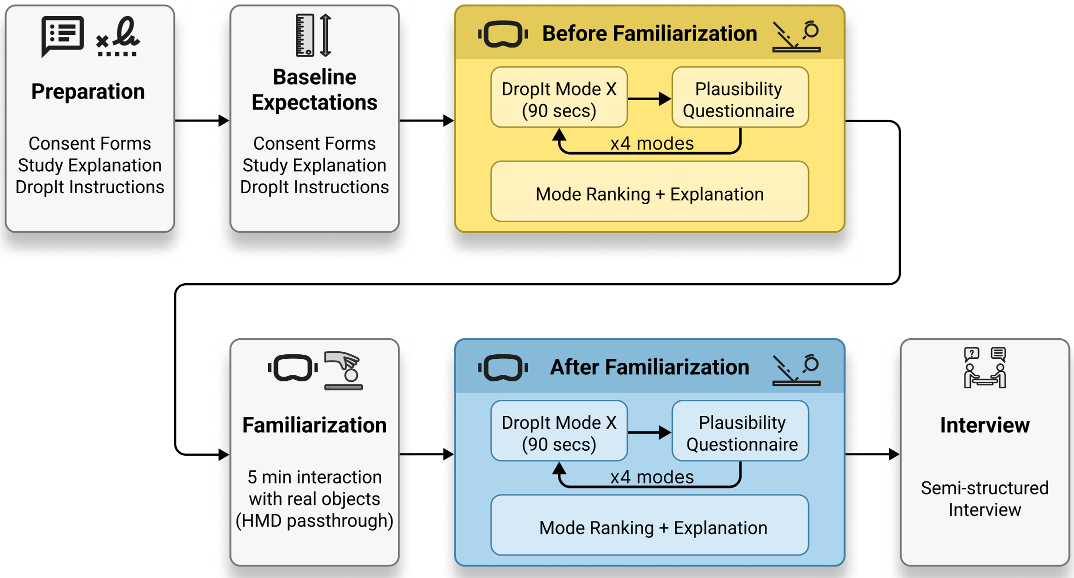

Figure 3: Study procedure overview. After baseline bounce-height estimation, participants completed the *Before Familiarization* and *After Familiarization* phases, interacting with each rebound model for 90 seconds and completing a questionnaire after each one. They then ranked the four models and explained their ranking. Between phases, a five-minute *Familiarization* stage involved interaction with the real balls and surfaces. The session ended with a semi-structured interview.

settings unless explicitly required by the rebound-model manipulation. Each virtual object was assigned the corresponding size and mass of its real-world counterpart; for example, the virtual tennis ball used a tennis-ball asset and was assigned the size and mass of a real tennis ball. Mass does not change the free-fall behavior of the virtual ball (as all virtual objects are set to be under the same gravitational acceleration), but it affects momentum exchange during ball-ball and ball-surface collisions. Therefore, assigning the corresponding real-world masses helped preserve plausible collision responses, such as the lighter pingpong ball being propelled farther when colliding with the heavier tennis ball.

Likewise, non-physics aspects of the experience were kept consistent across conditions. Visual assets, interaction flow, controller inputs, and task structure were identical across all models. Audio feedback was also held constant across models, surfaces, and ball types, with only playback volume varying as a function of impact force. By keeping these factors fixed, we aimed to ensure that any systematic differences in plausibility judgments were primarily attributable to the manipulated differences in physics fidelity rather than to unrelated changes in appearance, interaction, or sound.

4 User Study

4.1 Task and Procedure

Task The experimental task, which we call *DropIt*, asked participants to move freely within the PLAY area, drop virtual balls, and observe how they bounced on the different visible real surfaces. Each rebound-model block lasted 90 seconds, during which participants were instructed to perform at least five drops with each ball on each surface to ensure that they experienced every ball–surface combination. Five drops was a minimum engagement requirement; participants could perform additional drops freely within the available interaction time. Participants could also switch between the ball types at any time during the block. After each block, participants completed a plausibility questionnaire, and after experiencing all four rebound models, participants ranked the models from most to least realistic. They were allowed to revisit any model before finalizing their ranking.

We used two virtual ball types: a tennis ball and a pingpong ball. Their visual assets were selected to resemble their real-world counterparts as closely as possible, and the corresponding real balls were shown to participants at the beginning of the study, but participants were not allowed to physically handle the balls or observe them being dropped until the familiarization phase.

Procedure The overall study procedure consisted of six stages, as illustrated in Figure 3: preparation, baseline estimation, pre-familiarization MR evaluation, real-world familiarization, post-familiarization MR evaluation, and interview. Each session lasted approximately 65 to 90 minutes depending mainly on questionnaire completion time and interview length. Although the full session could approach 90 minutes, the main MR task was divided into two shorter evaluation phases of approximately 15–20 minutes each rather than one continuous session. Participants were monitored for discomfort throughout, and breaks were provided when needed; one participant requested an additional 10-minute break.

During preparation, participants were briefed on the study purpose, the overall structure of the session, and how to use the controllers to manipulate balls and respond to questionnaires.

Next, before physically handling any real balls, participants completed a baseline bounce-height estimation task. They were shown the tennis ball, pingpong ball, three surfaces, and a 75-cm reference height marked with black tape, but were not allowed to interact with the balls. For each ball–surface combination, participants indicated by hand the height to which they expected the ball to rebound, and the experimenter recorded it using a tape measure.

Participants then wore the Meta Quest 3 and completed the first MR evaluation phase (*Pre*). In this phase, they experienced the four rebound models, completed plausibility questionnaires, and then ranked the models from most to least realistic while explaining their reasoning. Figure 4 shows the questionnaire interface used in the MR application. Within each phase, the order of the four rebound models was randomly assigned for each participant to minimize any potential order effects.

After the pre-familiarization phase, participants directly interacted with the corresponding real balls in the PLAY area for five minutes while continuing to see the physical surfaces through passthrough. This familiarization stage provided immediate real-world reference experience with ball–surface behavior.

Participants then completed the second MR evaluation phase (*Post*), again experiencing and evaluating the same four rebound models. Finally, a semi-structured interview was conducted to gather qualitative reflections on the criteria participants used when judging realism and how their judgments may have changed after familiarization.

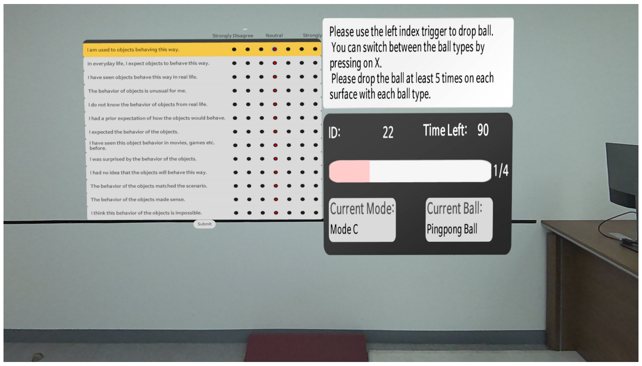

Figure 4: In-headset view during the *DropIt* task. The left panel presents the plausibility questionnaire. The right panel shows task instructions, remaining time, progress, current rebound model, and selected ball type.

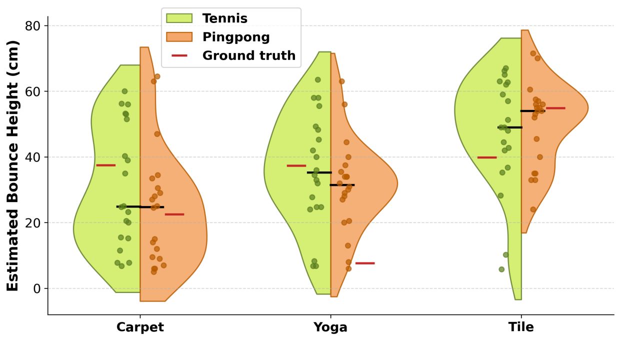

Figure 5: Baseline rebound-height estimates used to establish participants' baseline mental-model accuracy. Points show individual estimates, and split violins show distributions by ball type. Black and red bars indicate median estimates and measured real-world heights, respectively.

4.2 Measures

We collected three evaluation outcomes: plausibility ratings, model rankings, and qualitative explanations. Baseline mental-model estimates were analyzed separately as a participant-level covariate.

Baseline mental-model accuracy. For each participant, we calculated the mean absolute relative error between the estimated and measured rebound heights across the six ball–surface combinations; lower error indicated greater baseline mental-model accuracy. Figure 5 summarizes these measurements.

Plausibility questionnaire. We use *perceived plausibility* as the primary subjective outcome. Participants completed the 13-item plausibility questionnaire adopted from Brubach et al. [5], using a 7-point Likert scale ranging from 1 (*I do not agree at all*) to 7 (*I fully agree*). Responses were averaged to compute a composite plausibility score for each condition.

Ranking. At each familiarization phase, participants ranked the four rebound models from most to least realistic. These rankings captured participants' relative realism judgments across models.

Qualitative data. To contextualize the questionnaire and ranking results, we collected qualitative data from participants' explanations for their rankings and from a final semi-structured interview. The qualitative corpus consisted primarily of detailed contemporaneous notes taken by the researcher during the ranking explanation sessions and the 15–30 minute interview, supplemented by audio recordings when clarification was needed.

The qualitative analysis followed an iterative inductive content analysis process. One researcher reviewed the notes and coded meaning units related to judgment criteria, perceived physical mismatches, prior expectations, and changes in evaluation after real-world familiarization.

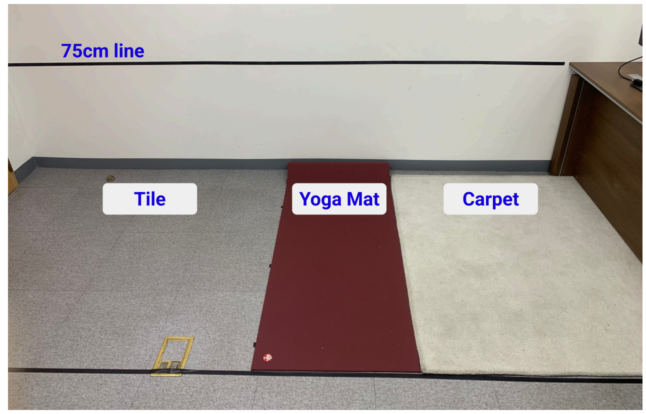

Figure 6: Study setup with tile, yoga mat, and carpet in the marked PLAY area. Participants dropped balls from the marked 75-cm reference height onto the real surfaces, which remained visible through headset passthrough during the MR task.

These codes were then grouped into higher-level categories to identify recurring patterns across participants.

4.3 Participants

An a priori power analysis using G*Power [10] determined the required sample size. Assuming a medium effect size (Cohen's $f = 0.25$), an alpha level of 0.05, a conservative repeated-measures correlation of 0.5, and a nonsphericity correction epsilon of 1, the analysis indicated a minimum of 16 participants to achieve 80% statistical power. To account for potential data loss or technical anomalies, we recruited 22 participants. Data from 20 participants were included in the final analysis, with two excluded due to technical issues. Demographic information for the final sample is reported in Table 1.

The study was approved by the university's ethics committee. Participants were recruited primarily through the university's online community forums and received monetary compensation equivalent to approximately 20 USD. Before the experiment, participants were informed about the study procedures, audio and screen recording, and their right to withdraw at any time. All participants provided written consent.

4.4 Apparatus and Study Setup

Apparatus Participants used a Meta Quest 3 and its controllers for the MR task, including dropping balls, switching ball types and rebound models, and completing questionnaires. The application was implemented in Unity Engine (v2022.3.21f) [24], using Meta XR SDK [15] and MRUK [16] for interaction and scene interpretation. We used the headset's passthrough capability so that participants could continuously see the real-world environment while interacting with virtual content.

Study setup Figure 6 shows the study setup. The study took place in a room containing three surface types within a designated PLAY area: tile, carpet, and yoga mat. The carpet occupied one side of the area, the yoga mat was placed adjacent to it, and the remaining floor area consisted of tile. This setup allowed participants to observe the same virtual balls bouncing on visibly different real substrates while remaining in an MR experience rather than a purely virtual one.

5 Results

5.1 Statistical analysis

To evaluate perceived plausibility, we analyzed the composite questionnaire score using linear mixed-effects models(LMMs). All models included a random intercept for each participant to account for repeated responses from the same person. We conducted three main analyses.

First, using the full dataset, we examined the effects of Rebound Model (A–D), Familiarization Phase (Pre vs. Post), and their interaction as fixed effects. This analysis tested whether perceived plausibility

Demographics	# of participants
Gender	14 male, 6 female
Age	Min: 22 Max: 67 $M = 28.35$ ($SD = 10.25$)
VR/MR experience	Never used: 3 Used 1–3 times: 12 Used 4–10 times: 2 Used 11 times or more: 3

Table 1: Demographic and media usage of all participants whose data was used (20 participants).

Data	Effect	LR_stat	df	p	sig
All data	Model × Phase	19.939	3	<0.001	***
	Model	42.323	3	<0.001	***
	Phase	0.712	1	0.399	
Pre only	Model × Prior	12.625	3	0.006	**
	Model	7.228	3	0.065	
	Prior	0.001	1	0.975	
Post only	Model × Prior	6.413	3	0.093	
	Model	60.223	3	<0.001	***
	Prior	0.588	1	0.443	

Table 2: Likelihood-ratio test results for the linear mixed-effects models. Model denotes Rebound Model. *LR_stat* is the likelihood-ratio statistic, *df* is the degrees of freedom, and *sig* indicates significance level.

differed across Rebound Models (RQ1), whether it changed after familiarization (RQ2), and whether familiarization affected the models differently.

Second, we analyzed only the Pre-familiarization data to examine whether participants' baseline mental-model accuracy was related to their initial plausibility judgments. This model included Rebound Model, baseline mental-model accuracy, and their interaction as fixed effects.

Third, we applied the same fixed-effects model but to the Post-familiarization data. This analysis tested whether baseline mental-model accuracy remained related to plausibility judgments even after participants had experienced the corresponding real-world interactions.

Together, these analyses allowed us to separate two conceptually distinct sources of familiarity: recent direct experience, captured by familiarization phase, and prior expectation, captured by baseline mental-model accuracy.

All statistical analyses were conducted in Python, and the LMMs were fitted using `statsmodels` (v0.14). We evaluated fixed-effect significance using likelihood-ratio (LR) tests against nested models. Follow-up pairwise comparisons and simple-slopes analyses were conducted using Wald z-tests, with Holm correction applied within each family of comparisons. Ranking data were analyzed using Friedman tests followed by Wilcoxon signed-rank tests with Holm correction.

Order-effect verification. Rebound-model order was randomized independently for each participant. To verify that the plausibility judgments were not affected by when each model was presented, we first added the model presentation position as a fixed effect to the analysis containing Rebound Model, Familiarization Phase, and their interaction. The results show that the model presentation position had no significant main effect on plausibility ($LR = 1.88$, $df = 1$, $p = 0.17$). We then included Rebound Model × Position and Phase × Position as additional fixed effects to test whether the effects of Rebound Model and Familiarization Phase differed depending on presentation position. The Rebound Model × Phase interaction was even stronger than originally reported ($LR = 22.04$ vs. $LR = 19.94$, $df = 3$, both $p < 0.001$), while Rebound Model × Position ($LR = 10.95$, $df = 9$, $p = 0.28$) and Phase × Position ($LR = 4.64$, $df = 3$, $p = 0.20$) were both non-significant.

Thus, we found no evidence that the model presentation position influenced the reported plausibility effects.

5.1.1 Influence of Rebound Model on Perceived Plausibility

The first mixed-effects model revealed a significant main effect of Rebound Model on perceived plausibility ($LR = 42.32, p < 0.001$), indicating that the four models were not judged equally. Looking at the descriptive statistics (see Supplementary Material Table 2), Model A (the Calibrated model) consistently yielded the highest plausibility ratings (M = 4.46 in Pre-familiarization, M = 5.68 in Post-familiarization) across both familiarity conditions, and Model C (the Damped model) yielded high plausibility ratings in the Unfamiliar condition (M = 4.38) but decreased significantly in the Familiar condition (M = 3.48). The other models consistently received lower scores in both phases.

5.1.2 Influence of recent real-world familiarization on perceived plausibility

The linear mixed model showed no significant main effect of Familiarization Phase on perceived plausibility ($LR = 0.71, p = 0.399$), indicating that recent experience did not uniformly increase or decrease plausibility ratings across models.

The linear mixed model did reveal, however, a significant interaction between Rebound Model and Familiarization Phase, with $LR = 19.94, p < 0.001$. This indicates that the effect of Rebound Model on perceived plausibility differed between the Pre and Post phases.

Specifically (see Supplementary Material Table 2), before familiarization, the mean plausibility ratings were relatively close to one another across models (SD of model means = 0.46), with Model A scoring higher than the other models (M = 4.46). After familiarization (Supplementary Material Table 3), this gap became more distinct: Model A increased from (M = 4.46) to (M = 5.68), whereas Model B changed from (M = 3.53) to (M = 3.32), Model C from (M = 4.38) to (M = 3.48), and Model D from (M = 3.75) to (M = 2.99).

To examine this interaction effect in more detail, we conducted follow-up simple-effect analyses. Results are shown in Table 3. Specifically, we tested the effect of Rebound Model separately within each phase, and the effect of Familiarization Phase separately within each model. We also conducted Holm-corrected pairwise post-hoc comparisons to identify which models differed from one another within each phase. Pairwise comparisons further showed that, in the Post phase, Model A received significantly higher plausibility ratings than Model B ($z = 7.45, p_{\text{Holm}} < 0.001$), Model C ($z = 6.94, p_{\text{Holm}} < 0.001$), and Model D ($z = 8.46, p_{\text{Holm}} < 0.001$). No pairwise comparisons survived correction in the Pre phase, and no other pairwise differences were significant in the Post phase.

The phase-within-model simple-effects analyses showed that Model A increased significantly from Pre to Post ($z = 3.56, p_{\text{Holm}} = 0.001$), whereas Model C decreased significantly ($z = -3.42, p_{\text{Holm}} = 0.002$) and Model D also decreased significantly ($z = -2.30, p_{\text{Holm}} = 0.043$). Model B did not significantly change across phases ($z = -0.88, p_{\text{Holm}} = 0.379$).

Taken together, these results indicate that recent real-world familiarization did not produce a uniform shift in plausibility ratings, but instead increased the differentiation among models. Before familiarization, participants did not significantly distinguish among the four Rebound Models. After familiarization, however, Model A was clearly preferred over all other models.

5.1.3 Influence of prior knowledge on perceived plausibility

To examine whether plausibility judgments depended on participants' baseline mental models of bounce physics, we conducted separate analyses for the Pre and Post phases using the prior knowledge error (Prior) as an index of prior knowledge.

For the Pre phase, the model revealed a significant interaction between Rebound Model and prior knowledge error, $LR = 12.63, p = 0.006$. However, neither the main effect of Rebound Model ($LR = 7.23, p = 0.065$), nor the main effect of Prior ($LR = 0.001, p = .975$), was significant. This indicates that before direct real-world familiarization, plausibility judgments depended on the combination of Rebound Model and participants' baseline mental-model accuracy, rather than on either factor alone.

To further interpret this interaction, we performed follow-up simple-slopes analyses to test whether the association between prior knowledge error and plausibility was significant within each model separately. These analyses showed that the interaction in the Pre phase was primarily driven by Model A: higher prior knowledge error was associated with higher plausibility ratings for Model A ($\beta = 2.44, z = 3.00, p_{\text{Holm}} = .011$), whereas the slopes for Model B ($\beta = -0.43, p_{\text{Holm}} = .642$), Model C ($\beta = -1.14, p_{\text{Holm}} = .476$), and Model D ($\beta = -0.81, p_{\text{Holm}} = .642$) were not significant.

For the Post phase, the interaction between Rebound Model and prior knowledge error (Prior) was not significant ($LR = 6.41, p = 0.093$). In contrast, the main effect of Rebound Model was significant ($LR = 60.22, p < 0.001$), while the main effect of Prior was not ($LR = 0.59, p = 0.443$). This suggests that after real-world familiarization, plausibility judgments were driven primarily by the Rebound Model, rather than by individual differences in baseline mental-model accuracy.

Taken together, these findings indicate that the role of prior knowledge was limited and phase-dependent. Before familiarization, prior knowledge error moderated plausibility judgments in a model-specific manner, with the effect primarily observed for Model A. After familiarization, this moderation pattern was no longer statistically reliable.

5.2 Ranking analysis

Participants ranked the four rebound models from most realistic to least realistic for each familiarity condition. We analyzed these ordinal data using Friedman tests followed by Wilcoxon signed-rank post-hoc comparisons with Holm correction. The ranking data are visualized in Figure 7. Before familiarization (Pre), the Friedman test indicated a significant but small effect of Rebound Model on ranking ($\chi^2(3) = 9.24$, $p = .026$, Kendall's $W = 0.154$). Model A obtained the best (lowest) mean rank ($M = 1.75$), while the other models showed similar ranks (Model B = 2.85, Model C = 2.65, Model D = 2.75). However, none of the pairwise comparisons remained significant after Holm correction (all $p > .05$). After familiarization (Post), the Friedman test indicated a significant and large effect of Rebound Model on ranking ($\chi^2(3) = 32.22$, $p < 0.001$, Kendall's $W = 0.537$). Pairwise comparisons after Holm correction showed that Model A ($M = 1.10$) significantly outperformed Model B, Model C, and Model D (all $p < .001$, Holm corrected). However, no significant differences were observed between the remaining models (all $p > .05$). These results align with the questionnaire findings: before familiarization, participants showed only weak preference differences across models, whereas after familiarization they more clearly distinguished the Calibrated model (Model A) from the alternatives.

5.3 Qualitative Findings

Participants relied on relative bounce structure and other motion cues rather than exact height alone. A central theme in the interviews was that participants considered whether the *relative pattern* of bounce behavior made sense across surfaces and ball types, rather than by exact numerical height alone. Participants such as P3, P12, P14, P21, and P22 explicitly explained their rankings in terms of whether the hard floor, yoga mat, and carpet produced the expected ordering. For example, participants described Model B as overly uniform because "the balls all bounced the same everywhere" (P2), making it feel flattened or insensitive to the scene structure (P7). Model D was also rejected because it contradicted the expected material logic of the environment. As P19 put it, "It was the most unrealistic; it was the opposite of what should happen... it made no sense." One participant (P21) explicitly distinguished between these two error types: the uniform response was dismissed as a "technological limitation", whereas the incorrect surface mapping was "much harder to excuse".

Some participants also attended to damping, rebound count, and the rate at which energy dissipated over time. P3 explicitly counted the number of bounces and compared those ratios in the second phase, while P18 focused on the settling motion near the end.

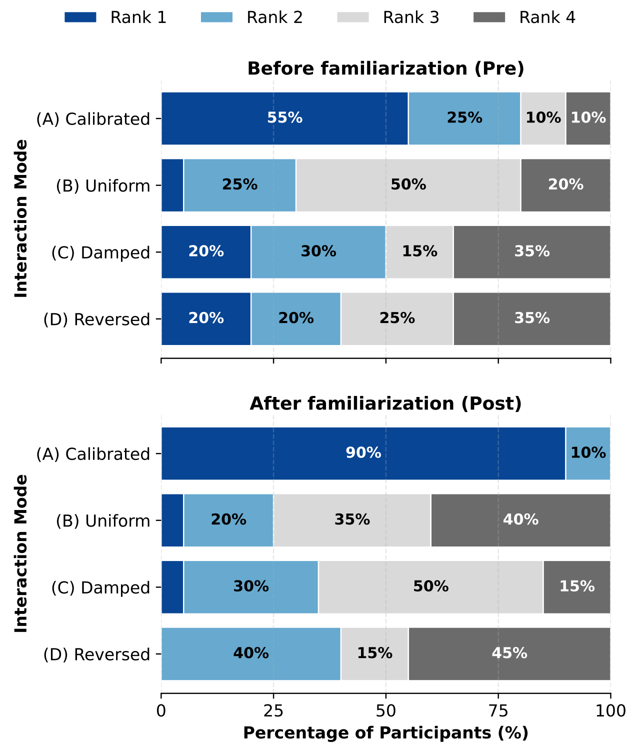

Figure 7: Rank distributions before and after familiarization, where Rank 1 is most realistic and Rank 4 is least realistic. Each stacked bar shows the percentage of participants assigning a model to each rank.

Together, these comments show that participants evaluated both the relationships across surfaces and the way each rebound unfolded over time, and not only on the overall rebound height.

Familiarization often made judgments stricter, clearer, and more confident. Participants described changes in their evaluation criteria after the real-world interaction. Multiple participants ($N = 10/20$) explicitly stated that their post-familiarization judgments became stricter, clearer, or more confidence-based. P19 described this shift clearly: "The absolute standard was established, so instead of just comparing modes, they had to satisfy that specific standard." Related comments from P8, P11, and P14 suggest a similar process: after physically interacting with the real balls and surfaces, participants felt they had a more concrete benchmark and could differentiate the virtual behaviors more sharply.

Prior expectations were object-specific, uneven, and sometimes unstable. Participants did not appear to rely on a single abstract sense of "physics." Instead, their prior expectations were often anchored to whichever ball they felt they understood best. At least seven participants explicitly said that they relied more heavily on the ball they were more familiar with, most often the pingpong ball. For example, P20 noted that she "took in consideration the pingpong ball more, because I was more familiar with it," and P21 said that she judged "based on the pingpong ball a lot." P7 and P22 reported a similar reliance on pingpong behavior. P6 and P11 showed the same mechanism in the opposite direction, relying more on tennis-ball expectations. These accounts suggest that prior knowledge mattered unevenly and in an object-specific way. In a few cases (P14 and P17), implausible motion did not immediately lead to rejection of the simulation, but instead caused participants to question their own memory. P14 offered a clear example: "In the beginning, when I saw C, I thought, 'Oh, did I know this [ball movement] incorrectly?' I even doubted myself." Only after later experiencing the calibrated reference did she reinterpret the earlier

Panel	Comparison	Means / Difference	Statistic	p
A. Simple effects of Rebound Model within phase	Rebound Model within Pre	A=4.46, B=3.53, C=4.38, D=3.75	$LR(3) = 7.23$	0.065
	Rebound Model within Post	A=5.68, B=3.31, C=3.48, D=2.99	$LR(3) = 59.89$	<0.001
B. Holm-corrected pairwise model comparisons	Pre: no significant pairwise differences			
	Model A vs. Model B (Post)	$\Delta = 2.37$ [1.74, 2.99]	$z = 7.45$	<0.001
	Model A vs. Model C (Post)	$\Delta = 2.20$ [1.58, 2.83]	$z = 6.94$	<0.001
	Model A vs. Model D (Post)	$\Delta = 2.69$ [2.07, 3.31]	$z = 8.46$	<0.001
C. Phase comparisons within model	Post vs. Pre within Model A	Pre=4.46, Post=5.68, $\Delta = 1.22$	$z = 3.56$	0.001
	Post vs. Pre within Model B	Pre=3.53, Post=3.31, $\Delta = -0.21$	$z = -0.88$	0.379
	Post vs. Pre within Model C	Pre=4.38, Post=3.48, $\Delta = -0.91$	$z = -3.42$	0.002
	Post vs. Pre within Model D	Pre=3.75, Post=2.99, $\Delta = -0.75$	$z = -2.30$	0.043

Table 3: Follow-up simple-effects and Holm-corrected pairwise comparisons for the Rebound Model × Familiarization Phase interaction on perceived plausibility. Reported significant results only.

model as manipulated: After later encountering model A, she stated: "After seeing A, I felt that my tendency was not wrong. My height estimation might have been off, but the directional trend was right." The interviews further show how participants interpreted motion mismatches. When bounce behavior felt implausible, several participants treated it not as an isolated animation error but as evidence that the ball, surface, or even the surrounding environment had different physical properties than expected. P17, for example, stated, "I didn't think the ball was strange, but rather that the material of the floor [yoga mat] was different or harder than I expected." In contrast, P20 concluded that the ball itself no longer felt like the advertised material: "[Because the ball didn't move like a tennis ball,] I thought the ball wasn't actually a tennis ball. It looked like one, but effectively it wasn't." One participant (P3) even questioned the gravity setting of the MR environment.

Surrounding interactions and sensory cues contributed to plausibility judgments. Beyond the ball–surface rebound behavior, participants also discussed physical interactions, visuals, and audio when explaining their judgments about the interactions. Sixteen participants explicitly praised ball–ball interactions as realistic or surprisingly well implemented, often mentioning rebound direction, differential response across ball pairings, or the sense that basic physical laws were being respected. At the same time, eleven participants explicitly identified missing interactions with the body, furniture, or non-target real objects as a clear plausibility break. Audio and visual fidelity were usually described as secondary to bounce motion when participants ranked the models. Several participants reported that the visuals were generally acceptable or realistic enough not to dominate the judgment process (e.g., P10, P12, P14, P21), though some noted shadow artifacts, blur, or pass-through limitations. Audio was also often described as noticeable but not central; however, a few participants reported that unrealistic pingpong audio weakened realism (e.g., P2, P10), and P7 noted that after familiarization she became more certain that the same sound was being reused, making the mismatch more salient.

6 Insights and Implications

6.1 Discussion

Taken together, our findings show that plausibility differed across rebound models (RQ1) and that these differences depended on participants' reference state (RQ2). Recent real-world exposure did not shift ratings uniformly but rather sharpened differentiation among models, whereas baseline mental-model accuracy played a more limited, model-specific role. We next discuss these findings and their design implications.

6.1.1 Physics Fidelity Shaped Plausibility and Scene Interpretation

RQ1 asked whether perceived plausibility differed across physics-fidelity conditions. Our results showed that Rebound Model affected perceived plausibility. The Calibrated model received the highest mean rating in both phases, whereas the Damped model received lower ratings despite preserving the same surface-dependent ordering as Calibrated. This contrast suggests that plausibility depended on multiple aspects of motion rather than one rebound feature alone. Our qualitative findings help explain this pattern: participants considered multiple aspects of the motion—including rebound height, differences across surfaces, and the overall rebound pattern—when judging whether the virtual object, the visible real surface, and the resulting response formed a coherent physical scene. This interpretation is consistent with prior work [1,18,19,26] showing that perceptual experience and learned object–material associations shape expectations about how objects should behave [1,26]. In VR, physical accuracy alone does not guarantee perceived realism, and familiar contextual cues such as virtual character identity may influence judgments of physics coherence [18, 19]. Our findings extend these ideas to virtual objects interacting with visible real surfaces.

In a small number of cases, physical mismatch also appeared to shape how participants interpreted the scene. When the motion conflicted with their expectations, some participants questioned their own prior knowledge or reconsidered the properties of the object, surface, or environment.

To summarize, these findings suggest that physics fidelity matters in MR not only because it affects perceived plausibility, but also because physical mismatches may shape how users interpret the scene. Participants appeared to judge the overall physical coherence of the interaction—how well the object, surface, motion, and resulting response fit together—rather than the accuracy of any single motion parameter.

6.1.2 Prior Expectations and Recent Real-World Experience Shaped Plausibility Judgments

RQ2 asked how baseline mental-model accuracy and recent real-world familiarization related to plausibility judgments. These two reference sources showed different patterns in our results. Recent real-world exposure did not shift plausibility ratings across all models in the same direction. Instead, it made the Calibrated model easier to distinguish from the alternatives. The interviews similarly suggest that participants used more concrete and confident criteria after interacting with the real balls and surfaces. This sharper discrimination may have emerged through two processes.

First, direct real-world exposure provided a concrete external benchmark against which participants could recalibrate their evaluation criteria. Before familiarization, participants relied on uncertain individual expectations; afterward, they could compare the virtual behavior with a recently experienced physical reference.

Second, repetition priming [12, 13, 23] offers a possible account of how relevant motion cues became easier to notice. Although the physical and virtual interactions were not identical stimuli, recent exposure to rebound features—such as bounce height, surface-dependent differences, and energy dissipation—may have increased their salience during subsequent virtual evaluation. Recalibration may therefore have changed the standard against which the interaction was judged, while priming may have made relevant deviations easier to detect.

Together, these processes help explain why familiarization did not produce an overall increase or decrease in plausibility ratings, but instead made differences among the rebound models more pronounced.

Baseline mental-model accuracy played a more limited role. Before direct real-world exposure, its association with plausibility varied across Rebound Models, with a significant slope only for the Calibrated model. This does not necessarily mean that prior expectations mattered only for this model. Rather, our rebound-height error measure may have captured an aspect of prior knowledge particularly relevant to judging the Calibrated model. Because Models B–D introduced larger or structurally different deviations from real-world behavior, participants may have relied more on other cues, such as relative bounce patterns across surfaces and temporal characteristics, whereas the Calibrated model may have invited closer comparison with expected rebound heights. Interview comments are consistent with this multi-cue interpretation, as many participants described using relative bounce structure and other motion cues in addition to rebound height.

In summary, the results show that user familiarity shaped plausibility judgments in two ways. Recent real-world exposure mainly affected Post judgments by sharpening discrimination among the models, while baseline mental-model accuracy may have mattered mainly when the simulated behavior was realistic enough to invite comparison with prior expectations.

6.1.3 Design Implications for Physically Plausible MR

The above findings suggest that physics fidelity should be considered throughout the design and evaluation of familiar MR interactions. We translate them into three practical implications.

Evaluate plausibility before and after real-world exposure. When an MR interaction has a familiar and directly comparable real-world counterpart, we recommend evaluating it both before and after users experience that counterpart. The initial evaluation captures existing expectations, whereas the second serves as a stress test that can expose physical mismatches overlooked initially. For example, an MR tennis application could first be tested in its virtual form, followed by the same action with a real ball and surface, and then tested virtually again. This approach is most useful when users know the physical interaction or can directly compare the virtual event with reality.

Check both relative behavior and the full response. This two-part check is especially important when a virtual object interacts with a visible real object or surface. The real element provides a physical reference, allowing users to form expectations and directly compare the virtual response with familiar real-world behavior. Designers should first verify that objects and surfaces behave in the expected relation—for example, that a ball rebounds higher on a hard floor than on carpet. A practical first step is to create a matrix of object–surface combinations and check whether each produces the expected higher or lower response. This can reveal behavior that is uniform across conditions or assigned in the wrong direction. Designers should then inspect the response, including rebound height, damping, rebound count, and settling, because correct ordering alone does not guarantee plausibility. The same approach may extend to other physical behaviors. For rolling or sliding, designers could check both which surface slows an object more and how it comes to a stop. For collisions, they could check both the expected direction of motion and the resulting speed change. Because our study focused on bouncing, these extensions should be treated as design directions rather than directly tested effects.

Check surrounding interactions and sensory cues. Evaluation should extend beyond the main interaction/event to surrounding interactions and sensory feedback. In our interviews, participants identified clear plausibility breaks when virtual balls passed through their bodies, ignored nearby furniture, or produced implausible or repetitive audio. A practical coherence audit should therefore cover object–object collisions, contact with the user's body and furniture, sound timing and variation, shadows, and pass-through artifacts. The goal is not to maximize the fidelity of every cue, but to remove clear contradictions across the interaction. Prior work also shows that users combine visual and bodily cues when judging physical interactions [8]. Because we did not manipulate these cues independently, this recommendation should be understood as an audit for possible problems, not as evidence that every cue needs equal fidelity.

In conclusion, designers should evaluate plausibility against users' real-world references, assess both relative and full physical behavior, and eliminate contradictions in interactions from the surrounding area. Though derived from ball–surface interactions, we believe that these principles may guide other physically grounded MR experiences.

6.2 Limitations

6.2.1 Interaction scope

This study focuses on a deliberately controlled MR interaction: familiar balls bouncing on visible real-world surfaces. This setting lets us isolate how physics fidelity, prior expectations, and direct real-world reference shape perceived plausibility. At the same time, it represents only one class of physically grounded MR behavior. Many MR experiences involve different cues, such as rolling, sliding, grasping, deformable contact, or longer multi-step interactions. Our findings should therefore be read as a focused account of bounce-based plausibility, and as a starting point for studying richer classes of MR interaction.

6.2.2 Pre/post familiarization design

All participants completed the MR evaluation before and then after real-world familiarization. This design was effective for testing how judgments changed once participants had a direct physical reference, but it does not fully separate familiarization from other effects of repeated exposure, such as improved comparison skill or shifts in evaluation strategy. This matters because the main pattern was not a uniform phase effect, but sharper differentiation among rebound models after familiarization. Future work can test this pattern more cleanly through counterbalanced or between-subject designs that isolate the contribution of direct real-world reference.

6.2.3 Interaction ecology

Our analysis centers on rebound behavior, but participants' comments suggest that plausibility was shaped by more than bounce alone. They also noticed the absence or inconsistency of related cues involving the body, nearby furniture, surrounding objects, and other sensory feedback. This suggests that physical plausibility in MR emerges from the coherence of a broader interaction ecology rather than from any single motion primitive. Our prototype intentionally simplified many of these surrounding factors to isolate differences in rebound behavior, while richer multi-cue MR settings offer a next step for future work.

7 Conclusion

This work shows that physics fidelity strongly shapes the perceived plausibility of MR interactions. In our ball-bouncing study, users did not judge plausibility solely by whether motion looked approximately correct; they were especially sensitive to whether the interaction preserved the expected surface-dependent response structure across balls and surfaces. This sensitivity increased after real-world familiarization, leading participants to distinguish conditions more clearly and prefer physically calibrated behavior. These findings suggest that believable MR requires virtual dynamics that preserve the physical structure users expect, not just geometric alignment and visual coherence. In familiar interactions, maintaining correct relative response patterns may matter as much as matching exact numerical dynamics.

Acknowledgments

We sincerely thank our reviewers for their valuable feedback. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. RS-2022-NR070595 and No. RS-2024-00463802).

REFERENCES

[1] L. M. Alley, A. C. Schmid, and K. Doerschner. Expectations affect the perception of material properties. *Journal of Vision*, 20(12):1–1, 2020. 2, 8

[2] Apple. Implementing scene understanding and reconstruction in your realitykit app. `https://developer.apple.com/documentation/realitykit/realitykit-scene-understanding`. Apple Developer Documentation, accessed 2026-03-17. 1

[3] D. Beaney and B. M. Namee. Forked! a demonstration of physics realism in augmented reality. In *2009 8th IEEE International Symposium on Mixed and Augmented Reality*, pp. 171–172. IEEE, 2009. doi: 10.1109/ISMAR.2009.5336479 2

[4] O. Bimber and B. Frohlich. Occlusion shadows: Using projected light to generate realistic occlusion effects for view-dependent optical see-through displays. In *Proceedings. International Symposium on Mixed and Augmented Reality*, pp. 186–319. IEEE, 2002. 2

[5] L. Brübach, F. Westermeier, C. Wienrich, and M. E. Latoschik. Breaking plausibility without breaking presence-evidence for the multi-layer nature of plausibility. *IEEE Transactions on Visualization and Computer Graphics*, 28(5):2267–2276, 2022. 2, 5

[6] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. *CoRR*, abs/2005.12872, 2020. 3

[7] L. Chen, K. Francis, and W. Tang. Semantic augmented reality environment with material-aware physical interactions. In *2017 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pp. 135–136. IEEE, 2017. doi: 10.1109/ISMAR-Adjunct.2017.49 2

[8] S. Do and B. Lee. Improving reliability of virtual collision responses: A cue integration technique. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, art. no. 518, 12 pp. ACM, 2020. doi: 10.1145/3313831.3376819 2, 9

[9] R. Du, E. Turner, M. Dzitsiuk, L. Prasso, I. Duarte, J. Dourgarian et al. Depthlab: Real-time 3d interaction with depth maps for mobile augmented reality. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. ACM, 2020. doi: 10.1145/3379337.3415881 2

[10] H. Kang. Sample size determination and power analysis using the g* power software. *Journal of educational evaluation for health professions*, 18, 2021. 5

[11] J. Kennedy and R. Eberhart. Particle swarm optimization. In *Proceedings of ICNN'95-international conference on neural networks*, vol. 4, pp. 1942–1948. ieee, 1995. 3

[12] Á. Kristjánsson and G. Campana. Where perception meets memory: A review of repetition priming in visual search tasks. *Attention, Perception, & Psychophysics*, 72(1):5–18, 2010. 2, 8

[13] V. Maljkovic and K. Nakayama. Priming of pop-out: I. role of features. *Memory & cognition*, 22(6):657–672, 1994. 2, 8

[14] C. E. Mendoza-Ramírez, J. C. Tudon-Martinez, L. C. Félix-Herrán, J. d. J. Lozoya-Santos, and A. Vargas-Martínez. Augmented reality: survey. *Applied Sciences*, 13(18):10491, 2023. 1

[15] Meta Platforms, Inc. Meta XR SDKs for Unity. `https://developers.meta.com/horizon/documentation/unity/unity-sdks-overview/`, 2025. Meta Horizon OS Developers documentation, accessed 2026-03-14. 1, 5

[16] Meta Platforms, Inc. Mixed Reality Utility Kit - Overview. `https://developers.meta.com/horizon/documentation/unity/unity-mr-utility-kit-overview/`, 2026. Meta for Developers documentation, accessed 2026-03-14. 1, 5

[17] NVIDIA. Nvidia physx sdk. `https://www.nvidia.com/en-us/`, 2018. Accessed: 2026-03-14. 3

[18] M. Pouke, K. J. Mimnaugh, A. P. Chambers, T. Ojala, and S. M. LaValle. The plausibility paradox for resized users in virtual environments. *Frontiers in Virtual Reality*, 2:655744, 2021. doi: 10.3389/frvir.2021.655744 2, 8

[19] M. Pouke, K. J. Mimnaugh, T. Ojala, and S. M. LaValle. The cat is out of the bag: The effect of virtual characters and scale cues on physics coherence. *Frontiers in Virtual Reality*, 5:1250618, 2024. doi: 10.3389/frvir.2024.1250618 2, 8

[20] R. Skarbez, S. Neyret, F. P. Brooks, M. Slater, and M. C. Whitton. A psychophysical experiment regarding components of the plausibility illusion. *IEEE Transactions on Visualization and Computer Graphics*, 23(4):1369–1378, 2017. doi: 10.1109/TVCG.2017.2657158 2

[21] K. Someya and Y. Itoh. Blending shadows: Casting shadows in virtual and real using occlusion-capable augmented reality near-eye displays. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 422–430. IEEE, 2021. doi: 10.1109/ISMAR52148.2021.00059 2

[22] I. Takeuchi and T. Koike. Augmented reality system with collision response simulation using measured coefficient of restitution of real objects. In *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology*, art. no. 49, 2 pp. ACM, 2017. doi: 10.1145/3139131.3141211 2

[23] E. Tulving and D. L. Schacter. Priming and human memory systems. *Science*, 247(4940):301–306, 1990. 2, 8

[24] Unity Technologies. Unity, 2022. Game development platform. 5

[25] Unity Technologies. Collider surface bounciness. `https://docs.unity3d.com/6000.4/Documentation/Manual/collider-surface-bounce.html`, 2026. Accessed: 2026-07-12. 3

[26] M. Vicovaro. Grounding intuitive physics in perceptual experience. *Journal of Intelligence*, 11(10):187, 2023. 2, 8

[27] F. Westermeier, L. Brübach, M. E. Latoschik, and C. Wienrich. Exploring plausibility and presence in mixed reality experiences. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2680–2689, 2023. 2

[28] J. Xiong, E.-L. Hsiang, Z. He, T. Zhan, and S.-T. Wu. Augmented reality and virtual reality displays: emerging technologies and future perspectives. *Light: Science & Applications*, 10(1):216, 2021. 1

[29] Y. Xu, J. Bo, J. Wang, H. Tian, Y. Zhang, K. Li et al. Real-time physically-based relighting and composition of radiance fields with proxy meshes. In *2025 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 348–358. IEEE, 2025. doi: 10.1109/ISMAR67309.2025.00046 2