

Perceptual Evaluation of Liquid Simulation Methods

KIWON UM, Technical University of Munich
XIANGYU HU, Technical University of Munich
NILS THUEREY, Technical University of Munich

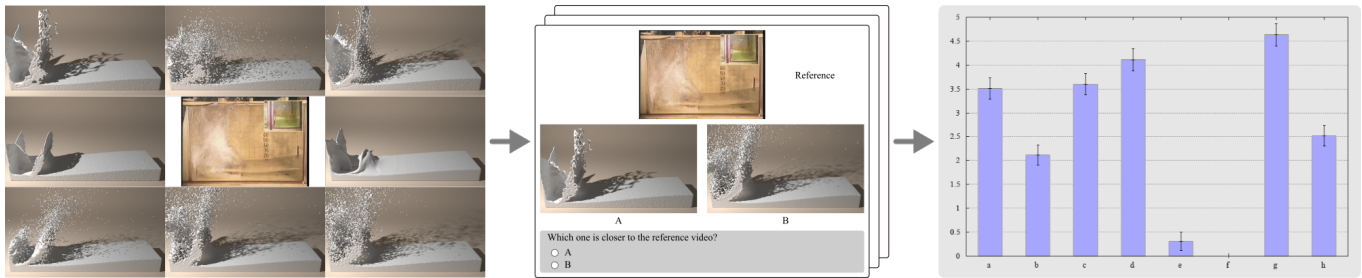


Fig. 1. We evaluate different simulation methods (left) with user study consisting of pair-wise comparisons with reference (middle). This allows us to robustly evaluate the different simulation methods (right).

This paper proposes a novel framework to evaluate fluid simulation methods based on crowd-sourced user studies in order to robustly gather large numbers of opinions. The key idea for a robust and reliable evaluation is to use a reference video from a carefully selected real-world setup in the user study. By conducting a series of controlled user studies and comparing their evaluation results, we observe various factors that affect the perceptual evaluation. Our data show that the availability of a reference video makes the evaluation consistent. We introduce this approach for computing scores of simulation methods as *visual accuracy* metric. As an application of the proposed framework, a variety of popular simulation methods are evaluated.

CCS Concepts: • **Computing methodologies** → **Physical simulation; Perception;**

Additional Key Words and Phrases: perceptual evaluation, liquid simulation, fluid-implicit-particle, smoothed particle hydrodynamics, crowd-sourcing

ACM Reference format:

Kiwon Um, Xiangyu Hu, and Nils Thuerey. 2017. Perceptual Evaluation of Liquid Simulation Methods. *ACM Trans. Graph.* 36, 4, Article 143 (July 2017), 12 pages.

DOI: <http://dx.doi.org/10.1145/3072959.3073633>

1 INTRODUCTION

In science, we constantly evaluate the results of our experiments. While some aspects can be proven by mathematical measures such as the complexity class of an algorithm, we resort to measurements for many practical purposes. When measuring a simulation, the metrics for evaluation could be the computation time of a novel optimization scheme or the order of accuracy of a new boundary condition. These evaluation metrics are crucial for scientists to

This work is supported by the *ERC Starting Grant* 637014.

Author's addresses: K. Um and N. Thuerey, Informatics 15, Technical University of Munich; X.Y. Hu, Chair of Aerodynamics and Fluid Mechanics, Technical University of Munich.

© 2017 Copyright held by the owner/author(s). Publication rights licensed to ACM. This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in *ACM Transactions on Graphics*, <https://doi.org/http://dx.doi.org/10.1145/3072959.3073633>.

demonstrate advances but also useful for users to select the most suitable one among various methods for a given task.

This paper targets numerical simulations of liquids; in this area, most methods strive to compute solutions to the established physical model, i.e., the *Navier-Stokes* (NS) equations, as accurately as possible. Thus, researchers often focus on demonstrating an improved order of convergence to show that a method leads to a more accurate solution [Batty et al. 2007; Enright et al. 2003; Kim et al. 2005]. However, for computer graphics, the overarching goal is typically to generate believable images from the simulations. It is an open question how algorithmic improvements such as the contribution of a certain computational component map to the opinion of viewers seeing a video generated with this method.

There are several challenges here. Due to the complexity of our brain, we can be sure that there is a very complex relationship between the output of a numerical simulation and a human opinion. So far, there exist no computational models that can approximate or model this opinion. A second difficulty is that the transfer of information through our visual system is clearly influenced not only by the simulation itself but also by all factors that are involved with showing an image such as materials chosen for rendering and the monitor setup of a user. Despite these challenges, the goal of this paper is to arrive at a reliable visual evaluation of fluid simulation methods. We will circumvent the former problem by directly gathering data from viewers with user studies, and we will design our user study setup to minimize the influence of image-level changes.

While there are interesting studies that investigate individual visual stimuli [Han and Keyser 2016] and the influence of different rendering methods for liquid simulations [Bojrab et al. 2013], our goal is to calculate the perceptual scores for fluid simulations on a high-level from animations produced with different simulation methods. We will demonstrate that a robust perceptual evaluation framework can be realized using crowd-sourced user studies that utilize carefully chosen simulation setups and a reference video. This

will allow us to retrieve reliable *visual accuracy* scores of different simulation methods evaluated in each study. In order to establish this framework, we ran an extensive series of user studies gathering more than 53,000 votes in total. The overview of our framework is illustrated in Figure 1.

In summary, we propose a novel perceptual evaluation framework for liquid simulations. To the best of our knowledge, the perceptual evaluation of physically-based liquid animations has previously not been studied, and we will use our framework to evaluate different simulation methods and parameterizations. From our evaluation results, we will draw useful observations for different simulation methods.

2 RELATED WORK

Fluid simulation methods typically compute solutions to the NS equations, which can be written as $\partial \mathbf{u} / \partial t + \mathbf{u} \cdot \nabla \mathbf{u} = \mathbf{g} - \nabla P / \rho + \nu \nabla^2 \mathbf{u}$ with the additional constraint to conserve volume: $\nabla \cdot \mathbf{u} = 0$, where \mathbf{u} is the velocity, \mathbf{g} is the gravity, P is the pressure, ρ is the density, and ν is the viscosity coefficient. Numerical solvers for these equations can be roughly categorized as Eulerian and Lagrangian methods. Fluid animations using Eulerian discretizations have been pioneered by Foster and Metaxas [1996], and the *stable fluids* solver [Stam 1999] has been widely used after its introduction. For liquids, the particle level set method has been demonstrated to yield accurate and smooth surface motions [Enright et al. 2002]. Currently, the fluid-implicit-particle (FLIP) approach, which combines Eulerian incompressibility with a particle-based advection scheme to represent small-scale details and splashes, is widely used for visual effects [Zhu and Bridson 2005]. The FLIP algorithm has been extended to many interesting applications such as the artistic control [Pan et al. 2013] and adaptivity [Ando et al. 2013]. In the following, we will focus on liquid simulations in simple domains without any adaptivity. We believe that this is a good starting point for our studies, but these extensions would of course be interesting for perceptual evaluations in the future.

The FLIP method was extended to incorporate position correction of the participating particles [Ando et al. 2012; Um et al. 2014] and to improve its efficiency by restricting particles to a narrow band around the surface [Ferstl et al. 2016]. Secondary effects generation has been a highly popular topic within the fluid simulation area in order to increase the apparent detail of the simulation [Ihmsen et al. 2012]. Many movies and interactive applications have incorporated hand-tuned parameters and heuristics to approximate where and how splashes, foam, and bubbles develop from an under-resolved simulation. Moreover, a unilateral pressure solver was proposed to enable large-scale splashes in FLIP [Gerszewski and Bargteil 2013]. Recently, several more FLIP variants were proposed to incorporate complex material effects that go beyond regular Newtonian fluids [Ram et al. 2015; Stomakhin et al. 2013]. We will later use the closely related affine particle-in-cell (APIC) variant [Jiang et al. 2015] as one of our candidates for simulation methods.

Lagrangian fluid simulation techniques in graphics are typically based on variants of the smoothed particle hydrodynamics (SPH) approach. After its first use for deformable objects [DeBunne et al. 1999], an SPH algorithm for liquids was introduced by Müller et

al. [2003], and then weakly-compressible SPH (WCSPH) was introduced by Becker and Teschner [2007]. The SPH algorithm was adopted and extended in a multitude of ways such as an adaptive discretization [Adams et al. 2007] and a predictor-corrector step that improves efficiency and stability [Solenthaler and Pajarola 2009]. Techniques for two-way coupling between rigid bodies and liquids have likewise been proposed [Akinci et al. 2012].

A different formulation using the position-based dynamics viewpoint was proposed for real-time simulations [Macklin and Müller 2013] while other researchers suggested an implicit method for better convergence rate [Ihmsen et al. 2014a]; this is known as implicit incompressible SPH (IISPH). From the Lagrangian field, we will restrict our visual accuracy study to a few selected methods: WCSPH and IISPH, which are typical and popular in graphics. Additionally, we also include an engineering SPH variant [Adami et al. 2012], from which we expect particularly accurate simulations; we denote this variant as SPH in our studies.

Naturally, researchers have been interested in combining aspects of the Lagrangian and Eulerian representations by bringing SPH and grid-based solving components together [Losasso et al. 2008; Raveendran et al. 2011]. We have not yet included these hybrid approaches in our studies, although FLIP arguably represents a hybrid particle-grid method. For a thorough overview of popular fluid simulations methods, refer to the book by Bridson [2015] and state-of-the-art report by Ihmsen et al. [2014b].

The human visual system and perception of image and video contents have received significant attention in computer graphics in order to study how algorithmic choices influence the final judgment of the created images. For example, in the area of rendering techniques, Cater et al. [2002] proposed to use selective and perceptually driven rendering approaches, and Dumont et al. [2003] introduced a theoretical framework to compute perceptual metrics. In photography, Masia et al. [2009] perceptually evaluated different techniques for tone-mapping HDR images with user studies. For videos, an approach for perceptually-driven up-scaling of 3D content was proposed [Didyk et al. 2010] while others investigated a computational model for the perceptual evaluation of videos [Aydin et al. 2010].

Beyond rendering and video, perceptual studies have also been used in the field of character animation. Especially, human characters have received attention. For instance, McDonnell et al. [2008] studied how to populate natural crowds for virtual environments. More recently, researchers also gathered data on the attractiveness of virtual characters [Hoyet et al. 2013]. In the area of deformable objects, Han and Keyser [2016] studied how visual details can influence the perceived stiffness of materials. Bojrab et al. [2013] studied how rendering styles of liquids influence user opinion. While this work also considers liquids, our goal is in a way orthogonal to theirs. We focus on simulation methods without being influenced by rendering styles.

3 VISUAL EVALUATION OF LIQUID SIMULATIONS

Despite the fact that most liquid simulation methods are physically-based and thus capable of approximating the NS equations in the limit, noticeable visual differences exist among animations created from the different methods. Being aware of these differences, we

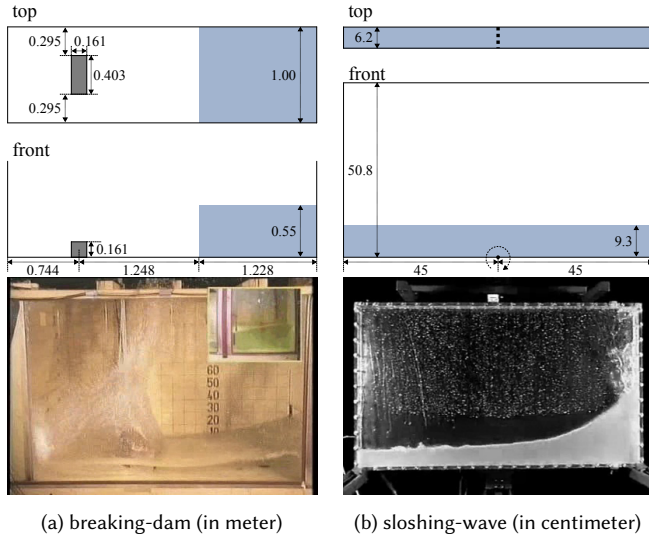


Fig. 2. Two simulation setups [Botia-Vera et al. 2010; Kleefsman et al. 2005] for evaluation of liquid simulation and example frames of real experiments.

propose a novel approach that employs user studies to evaluate the different methods in terms of how closely they match real phenomena. The goal of our approach is to robustly and reliably compare different liquid simulations such that the evaluation reflects a general opinion. Therefore, we employ a crowd-sourcing platform in order to recruit many participants to retrieve a reliable evaluation.

We focus on the perceptual evaluation of simulations in terms of what we call *visual accuracy*. We define this visual accuracy to be a score computed from user study data to compare different methods, and we will make sure that it can be computed in a robust and unbiased way. To collect data, we let users select a preferred video from pair-wise comparisons, and we found it crucial for robustness to provide participants with a visual reference. As we will outline below, this also makes the results very stable with respect to strongly differing rendering styles. These comparisons with a reference video are also our motivation to see the scores we compute as a form of *accuracy*.

Liquid simulations are commonly used tools in visual effects and applied for a vast range of phenomena from drops of blood to large scale ocean scenes. While it would be highly interesting to evaluate all of them, we focus on one particular regime of water-like liquids on human scales. This regime is highly challenging due to the low viscosity of water. The resulting flows typically feature high Reynolds numbers, complex waves, and large amounts of droplets and splashes. Although this naturally limits the regime of our study, we believe that it is particularly a representative for many effects and thus worth studying. Next, we will present two carefully chosen simulation setups that will also form the basis of our user studies in Section 3.2.

3.1 Simulation Setup

When selecting simulation setups, our requirements are that the setups are easy to realize in numerical simulations; thus, they do not

involve any specialized domain boundary conditions or any moving obstacles. Therefore, the setups should be easily reproducible. Nonetheless, the setups need to result in sufficiently complex dynamics such as overturning waves and splashes in order to be relevant for visual effects applications. Note that our setups stem from the engineering community. This has the additional benefit that detailed flow measurements are available as well as video data from real experiments. The latter is especially important for our user study later.

Our first setup is close to the popular breaking dam case often seen in graphics. Such a benchmark setup is also often used for validation in the engineering studies, which adds an obstacle in front of the breaking dam for additional complexity [Kleefsman et al. 2005]. This setup uses a tank of size $3.22\text{m} \times 1\text{m} \times 1\text{m}$ with an open roof, a static obstacle of $0.16\text{m} \times 0.16\text{m} \times 0.4\text{m}$, and an initial water volume of $1.23\text{m} \times 0.55\text{m} \times 1.0\text{m}$. As the tank is more than three meters in length and the initial column is considerably high, this breaking dam setup results in violent and turbulent splashes, which makes it tough but relevant for our purposes. We will denote this setup as *dam* in the following, and the details of its initial conditions are illustrated in Figure 2-(a).

Our second setup is a sloshing wave tank [Botia-Vera et al. 2010]; this is illustrated in Figure 2-(b). A rectangular tank partially filled with water experiences a periodic motion that continually injects energy into the system leading to waves and splash effects forming over time. The size of the tank is $0.9\text{m} \times 0.51\text{m} \times 0.062\text{m}$, and the rotation axis is located at the lower center of the tank. The initial water height is 0.093m . This setup has a significantly smaller overall water volume; it leads to interesting waves forming over time. These waves are more prominent here than in the dam setup. We will denote this setup as *wave* in the following. Additional documentation for both setups is available online [Issa et al. 2017].

For all simulations, we parameterize them according to the real-world dimensions given above using earth gravity as the only external force. Unless otherwise noted, we will not include any additional viscosity. In the following, we explain the user studies, which are based on one of these two setups.

3.2 User Study Design

The goal of the user studies is to reliably evaluate the visual accuracy across a set of m videos produced by different simulation methods. While many variants of user studies are imaginable [Leroy 2011], we opted for purely binary questions in order to reduce noise and inconsistencies in the answers. We also want to make the design as simple as possible to prevent misunderstandings. Thus, participants are shown two videos to consider in comparison to a reference video as illustrated in Figure 3. The videos are played repeatedly without time limit, and the participants are given the task to select one video which they consider to be closer to the reference video.

All participants have to give their vote for all possible pairs in a study. Thus, for m videos under consideration, we collect $m(m-1)/2$ responses per participant. In order to limit the workload per participant, we ensure that m is kept small, e.g., $m \leq 7$ for our studies. In order to identify untrustworthy participants, we duplicate the set of comparisons and randomize their order; then, we check the consistency of the answers. We reject participants

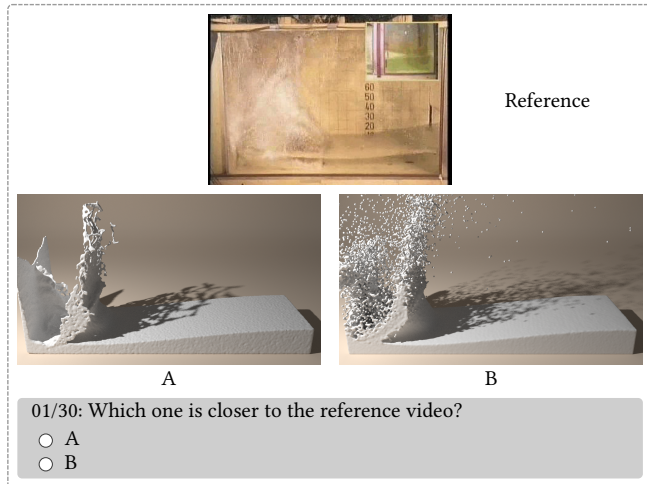


Fig. 3. Our user study design.

with a consistency of less than 70% [Cole et al. 2009]. Note that we also randomize the positioning of both videos for each question (i.e., left and right side).

Based on the pair-wise votes per study, we can now compute a set of scores for all m videos. For this purpose, we adopt the widely used Bradley-Terry model [Bradley and Terry 1952]. We review the model briefly here. Its goal is to compute scores s_i such that we can define the probability p_{ij} that a participant chooses video i over video j as:

$$p_{ij} = e^{s_i - s_j} / (1 + e^{s_i - s_j}). \quad (1)$$

Let w_{ij} denote the number of times where video i was preferred over video j in a user study. Assuming the observations are independent, w_{ij} follows a binomial distribution. Therefore, the log likelihood for all pairs among all videos can be calculated as follows:

$$L(\mathbf{s}) = \sum_{i=1}^m \sum_{j=1}^m (w_{ij} s_i - w_{ij} \ln(e^{s_i} + e^{s_j})) \quad (2)$$

where $\mathbf{s} = [s_1, s_2, \dots, s_m]$. The final scores of all videos are computed by solving for the \mathbf{s} that maximizes the likelihood function L in Equation (2) [Hunter 2004].

The vector of scores \mathbf{s} is what we use to evaluate the visual accuracy in the following. Note that these scores do not yield any “absolute” distances to the reference, and they cannot be used to make comparisons across different studies. However, we found that they yield a reliable scoring and probability (see Equation (1)) for all videos participating in a single study.

In order to prevent bias with respect to the participants, we ran a series of studies in three different crowd-sourcing platforms and found that differences were negligible. Details for these studies can be found in Appendix A. Across our studies, we also noticed that the consistency checks did not substantially influence the results, thus the large majority of participants was trustworthy. We gathered 50 answers per pair-wise comparison question from different participants in each user study. In total, we collected user study data for 53,300 questions from 612 participants in 65 countries.

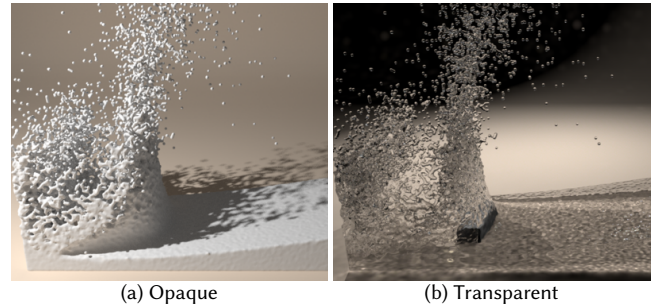


Fig. 4. Example frames of the opaque and transparent rendering styles.

Seeing the consistency of answers across different platforms, we believe that the user study design described above yields consistent answers. However, the existence of consistent scores by themselves does not yet mean that we can draw conclusions about the underlying simulation methods rather than about a certain style of visualization. In the next section, we will present a series of user studies to investigate whether we can specifically target simulation methods.

3.3 Visual Accuracy for Simulations

In order to show that there is a very high likelihood that our studies allow conclusions to be drawn about the simulation methods, we now turn to comparisons of studies. Thus, instead of considering individual visual accuracy scores s_i , we will consider multiple sets of score vectors \mathbf{s} to be compared with each other. Once we have demonstrated that our user studies allow us to draw conclusions with high confidence, we will discuss individual scores for specific simulation-related questions in Section 4.

In the following, we will analyze pairs of studies for which we make only a single change. For example, one study will have rendering style A, and a second study will have rendering style B while keeping all other conditions identical. We then perform a correlation analysis for these studies. If the studies turn out to be correlated, we can draw conclusions about the influence of the change on the outcome.

For the correlation analysis, we compute the Pearson correlation coefficient and statistical significance [Pearson 1920], which are widely used in statistics as a measure of the linear correlation between two variables $x, y \in \mathbb{R}^m$. This correlation coefficient r is the covariance of the two variables divided by the product of their standard deviations σ_x and σ_y , i.e., $r = \text{cov}(x, y) / \sigma_x \sigma_y$. A strong positive correlation, i.e., very similar score distributions, will result in values close to +1, while uncorrelated or inverted scores, hence very different user opinions, will result in correlations of 0 or even negative correlations of -1.

In order to investigate the robustness of our visual accuracy evaluation, we set up user studies with six videos. The six versions were chosen to broadly sample the space of typical resolutions and simulation methods. For the studies of this section, we are not particularly interested in the specific details of the simulation methods as long as they are representative for commonly used methods of graphics applications. With this goal in mind, we will use a popular Eulerian

Table 1. Correlation analysis for the sets of scores evaluated from different user studies using FLIP and SPH. Here, ref. denotes the reference video.

ID	Comparison (IDs in Table 7)	Constant parameters	r	p-value
C ₀	opaque (A) vs. transparent (B)	dam with ref.	0.97347	0.00105
C ₁	dam (A) vs. wave (C)	rendered in opaque with ref.	0.96557	0.00176
C ₂	opaque (A*) vs. transparent (B*)	dam w/o ref.	-0.01308	0.98039
C ₃	dam (A*) vs. wave (C*)	rendered in opaque w/o ref.	0.83895	0.03682
C ₄	with ref. (A) vs. w/o ref. (A*)	dam rendered in opaque	0.64540	0.16632
C ₅	with ref. (B) vs. w/o ref. (B*)	dam rendered in transparent	-0.60960	0.19887

method FLIP [Zhu and Bridson 2005] and Lagrangian method SPH [Adami et al. 2012] with three representative resolutions as shown in Table 2. Note that FLIP effectively is a hybrid Lagrangian-Eulerian method. However, we consider FLIP as Eulerian in our studies due to its Eulerian pressure solver, which is a key component of the algorithm. We put an emphasis on visual aspects with the studies described in the following section.

The space of possible visualization techniques for liquid animations is huge. Many freely available renderers exist to create realistic images. Real-time applications typically use specialized shaders for efficiency, and visual effects in movies employ very refined compositions of many layers to produce highly realistic visuals. Instead of trying to cover this whole space of possibilities, we focus on two extremes of the spectrum: a fully opaque rendering style and perfectly transparent surface. While the former employs a simple diffuse material similar to a preview rendering, the transparent rendering style exhibits complex lighting effects, such as refraction, reflection, and caustics. A consequence is that the surface is very clearly visible for the diffuse surface in contrast to the transparent rendering. Still images for an example of these two rendering styles can be found in Figure 4.

Comparisons of user studies: Assuming that our design for user studies is reliable, we expect to see a strong correlation when comparing two studies with these different rendering styles despite the differences in appearance. This hypothesis is confirmed with a correlation coefficient of more than 0.97 with a high confidence level ($p < 0.01$). The details for this correlation calculation C₀ as well as the following ones can be found in Table 1, and the full studies under

Table 2. Six simulation configurations for the experiments of Table 1 and 3. Here, S denotes the scaling factors of resolution, and M denotes the methods: Eulerian (Eu.) and Lagrangian (La.)

S	M	Resolutions for particle and grid	
		<i>dam</i>	<i>wave</i>
1x	Eu.	83k (80× 75× 25)	23k (75× 42× 5)
2x	Eu.	664k (160×150× 50)	186k (150× 84×10)
4x	Eu.	5,315k (320×300×100)	1,488k (300×168×20)
1x	La.	84k (80× 75× 25)	24k (75× 42× 5)
2x	La.	665k (160×150× 50)	186k (150× 84×10)
3x	La.	2,253k (240×225× 75)	634k (225×126×15)

consideration are given in Table 7. Considering the substantially different images resulting from these two rendering styles, we believe that the strong correlation is an encouraging result.

When removing the reference video from the user study design (C₂), i.e., only showing two videos of numerical simulations with the task to select the “preferred” version, the result changes drastically. Instead of a positive correlation, we now see a nearly no correlation (i.e., $r = -0.013$). Thus, without the availability of a reference video, the opinion between videos changes very strongly when switching from the opaque to the transparent rendering style.

While we used the *dam* setup for the study above, we now repeat this comparison keeping the rendering style constant (i.e., opaque) and comparing simulation setups (*dam* versus *wave*). When performing these studies with reference videos, we see a strong correlation of 0.97 (C₁) with a high confidence level ($p < 0.01$), whereas the correlation slightly drops to 0.84 when the reference video is removed (C₃). The absence of a reference video does not necessarily lead to inconsistent results for all cases, rather there is an increased chance of ambiguity and substantially different responses.

From the first two pairs of comparisons, we draw the conclusion that the availability of a visual reference is crucial for a consistent evaluation of the liquid motion. Having a reference video even stabilizes results from strongly differing visualization styles as illustrated with the studies of C₀. The reference video is also the reason why we believe our results do not contradict previous work that found substantial influence of rendering styles on perception for animated water [Bojrab et al. 2013]. Regarding liquid motions in the human-scale regime, our results indicate that the influence of rendering can be made negligible by providing a visual reference. Note that our reference does not need to closely match the rendering style used for the simulation videos. The results are consistent even for considerably stylized and different rendering styles such as our opaque and transparent styles; both are very different from the reference video. The different correlation scores are summarized visually at the figure in Table 1. This figure again highlights that the low and even negative correlations are stabilized by the availability of a reference video.

To shed further light on this topic, we compute correlations between the studies with and without reference video. These correspondences can be found in C_{4,5} in Table 1. In both cases, the visual accuracy scores of the methods under consideration change substantially when the reference video is removed. This results in



Fig. 5. Example frames of our alternate reference video for *dam*.

the correlations that are not statistically significant ($p > 0.05$). Besides, the results with the transparent rendering style show a drastic change of user opinions. Thus, without a reference video, visual appearance can strongly influence the scores.

Reference videos: The video we used as reference for the *dam* example has a visual appearance that is clearly different from our renderings. We note that visual accuracy can be evaluated even when the simulated phenomena bear only rough resemblance to the reference video. Figure 5 shows the example frames of a video recorded in nature at a seashore. In an additional user study with the *dam* simulations, we use this seashore video as reference instead of the one shown in Figure 2. The resulting scores (A^+ in Table 7) are highly correlated with the results of the original study with the video of the *dam* experiment (A in Table 7). Here, the correlation is 0.93 with a high confidence level ($p < 0.01$).

On the other hand, when we use a video that differs more strongly in another additional user study, the results start to change. The correlation between a study using the *wave* video with the *dam* simulations (A^- in Table 7) and the original study (A in Table 7) is not statistically significant ($p > 0.05$). To summarize, our results show that a reliable visual accuracy can be established even if no reference to the exact simulation setup is available. The human visual system is powerful enough to correlate the visual inputs despite different appearance. However, the stability of the results drops when the physics differ substantially.

Representative methods: At this point, we also want to confirm our assumption that the two initially chosen simulation methods are representative for commonly used Eulerian and Lagrangian methods. We choose two different methods from the Eulerian and Lagrangian classes: APIC [Jiang et al. 2015] and IISPH [Ihmsen et al. 2014a]. With these two methods, we performed new user studies keeping the remainder of the user study and simulation setups constant; i.e., the simulations use the same resolutions of particle and grid as before (Table 2). The strong positive correlation for this pair of studies confirms our initial assumption (C_6 in Table 3). Note that our two sets of simulation methods are also correlated in studies without a reference video (C_7). Presumably, this indicates that the participants' tendency in preference among the two classes of methods is fairly consistent. In this case, the individual scores of each method change substantially between the FLIP&SPH and APIC&IISPH sets. Thus, this makes it difficult to draw a conclusion among the different methods of each class. However, the correlation between the two sets of methods confirms our assumption that these methods cover the space of Eulerian and Lagrangian classes well. In addition, we find that the availability of a reference video

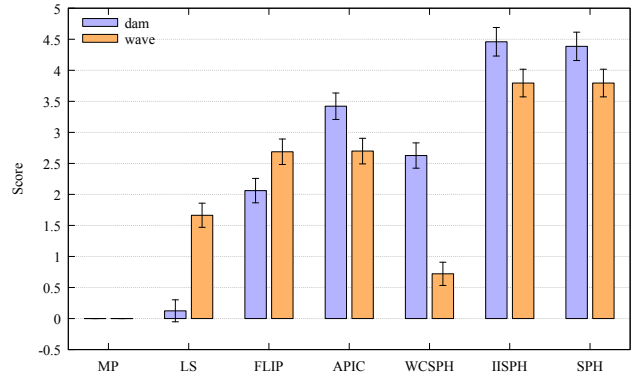


Fig. 6. Visual accuracy scores of the seven simulation methods under consideration (J and K in Table 7).

affects the stability also in these methods (C_8). This is consistent with the aforementioned results indicating again that the absence of reference video results in a chance of ambiguity.

4 APPLICATIONS AND RESULTS

In this section, we use our approach to evaluate the visual accuracy of various simulation methods (Section 4.1 and 4.2). We also demonstrate that our evaluation allows us to redeem heuristic approaches, such as the grid resolution for particle skinning (Section 4.3), or algorithmic modifications, such as a splash model for FLIP simulations (Section 4.4).

4.1 Liquid Simulation Methods

When establishing our evaluation framework, a central goal was to compare simulation methods. In the following, we evaluate seven simulation methods from the Eulerian and Lagrangian classes: marker-particles (MP) [Foster and Metaxas 1996], a solver with level set surface tracking (LS) [Foster and Fedkiw 2001], FLIP [Zhu and Bridson 2005], and APIC [Jiang et al. 2015] as representatives of Eulerian methods; WCSPH [Becker and Teschner 2007], IISPH [Ihmsen et al. 2014a], and a so-called wall-boundary SPH method [Adami et al. 2012] as representatives of Lagrangian ones. Note that this classification is primarily based on whether the method uses a grid in the pressure solver. Using these seven methods, we simulate our two simulation setups, i.e., *dam* and *wave* from Section 3.1.

The evaluation results are summarized in Figure 6. Interestingly, the Lagrangian methods (particularly, IISPH and SPH) consistently receive higher visual accuracy scores than the other methods. Among the Eulerian methods, the FLIP variants (i.e., APIC and FLIP) receive higher scores than MP and LS. Our guess for the latter results is that the MP and LS versions exhibit a very small amount of droplets. Note that the score of WCSPH is also noticeably low in the *wave* example; we observe that the amount of splashes is likewise very small, and the surface motion is highly viscous due to its artificial viscosity. Here, the level set method receives a higher score than WCSPH. We presume that this is caused by the artificial viscosity of the WCSPH solve, which often results in a stronger damping of

Table 3. Additional correlation analysis for two sets of simulation methods. Here, the *dam* example is used with opaque style.

ID	Comparison (IDs in Table 7)	Constant parameters	r	p-value
C ₆	FLIP&SPH (A) vs. APIC&IISPH (D)	with ref.	0.96057	0.00230
C ₇	FLIP&SPH (A*) vs. APIC&IISPH (D*)	w/o ref.	0.96932	0.00140
C ₈	with ref. (D) vs. w/o ref. (D*)	APIC&IISPH	0.72139	0.10562

its surface motion in comparison to the LS solve. Figure 7 shows several still frames of all the methods.

For the implementation of each method, we followed the original work without any considerable modifications. The Eulerian methods (i.e., MP, LS, FLIP, and PIC) used grid resolutions of $160 \times 150 \times 50$ for *dam* and $150 \times 84 \times 10$ for *wave*. All methods except LS used 665K particles for *dam* and 186K particles for *wave*. Although the Lagrangian methods did not use any grid in their solve, we used the same underlying grid for initializing the particles and sampled each cell with eight particles. While the Lagrangian methods used a uniform sampling, the Eulerian methods randomly jittered the particles to avoid aliasing. Note that this resulted in slightly different numbers of particles ($\sim 1k$) between the two classes of methods; the particles were not reseeded during the simulation. In order to ensure a comparable resolution for surface tracking in all methods, we used a doubled resolution for tracking the level set in LS. For the pressure solver of the Eulerian methods, we used a standard conjugate gradient method with modified incomplete Cholesky preconditioning [Bridson 2015]. All implementations and setups can be found online.

4.2 Limited Computational Budget

This experiment focuses on the four methods that ranked highest from the previous evaluation and re-evaluates them with the constraint of a limited computational budget per frame. While the previous study kept resolution and particle count constant, we have adjusted them to yield comparable runtimes for this study. We simulated the *dam* example using APIC, FLIP, IISPH and SPH such that they all required approximately 55 seconds per frame of animation. Here, we do not include the computational costs for non-simulation steps such as surface generation and rendering. We are aware that absolute comparisons of performance are difficult in general, but we have made our best efforts to treat all methods fairly and to bring all implementations up to a similar level of optimization (e.g., all implementations employ shared-memory parallelism with OpenMP for most of their steps).

The time restriction leads to significant reduction in resolution for the SPH-based methods. Both FLIP and APIC use a $320 \times 300 \times 100$ grid and 5,315k particles; IISPH uses 143k particles sampled from a $96 \times 90 \times 30$ grid, while SPH uses 84k particles sampled from a $80 \times 75 \times 25$ grid. Example frames for these simulation configurations are shown in Figure 8.

In contrast to the previous evaluation in Section 4.1, our participants gave the Eulerian methods higher visual accuracy scores. The results are shown in Figure 9. Thus, while the previous study suggests that Lagrangian methods capture large-scale splashes better at a given resolution, this study suggests that FLIP and APIC lead to improved results under a restriction in computation time.

4.3 Particle Skinning

Our evaluation approach is also useful to redeem heuristic approaches, where parameters are typically chosen by intuition. One example is the grid resolution for generating a surface mesh from particle data, i.e., *particle skinning*. The commonly used heuristic for this is to use a two times higher resolution of the simulation grid, but there has been little motivation for this particular setting.

As the base simulation for this experiment, we use FLIP with a $160 \times 150 \times 50$ grid and 664k particles. After simulation, a signed distance field is computed from the particles [Zhu and Bridson 2005], which we triangulate with marching cubes. Since the particles are sampled at a 2^3 sub-grid, the cell size of the base resolution (1x) is $2h$, where h denotes the particle spacing. We perform the particle skinning using different resolutions with seven scaling factors relative to h : 0.5x, 0.75x, 1x, 1.5x, 2x, 3x, and 4x. In order to avoid missing particles in the grids that are more than h apart, the particle diameter is adjusted to the larger of either the grid spacing or the particle spacing. The example frames are shown in Figure 10.

As Figure 11 shows, the evaluation result indicates that the heuristic of 2x [Zhu and Bridson 2005] is a good one. The higher resolutions do not yield results that can reliably be considered better than the 2x factor, which thus represents the best performance.

4.4 Visual Impact of Splash Modeling

This section inspects a specific FLIP extension that claims to yield an increased amount of visual detail with secondary effects. It employs a neural-networks approach to model the sub-grid scale dynamics that lead to splashes [Um et al. 2017], and we will denote it as *MLFLIP* in the following. A visual comparison of example frames from both FLIP and MLFLIP can be seen in Figure 12.

In order to see whether this splash model indeed results in better visual accuracy scores, we evaluate both FLIP and MLFLIP with two additional methods for reference (i.e., MP and SPH). Figure 13 shows the resulting visual accuracy scores. For the *dam* setup, we observe that the MLFLIP approach yields a notable improvement in score from 2.28 for regular FLIP to 4.18 for MLFLIP. The gain for the *wave* setup is lower, from 1.83 to 2.66, but we can still find a statistically relevant improvement. These results indicate that splashes are an important visual cue for large-scale liquid phenomena.

5 DISCUSSION OF RENDERING STYLES

As our core method of evaluation, we propose to use measurements of visual accuracy scores from user studies with a reference video. However, seeing the strong variability in the previous results, especially for the transparent rendering style, we believe that it is important to discuss additional studies that we conducted to investigate the influence of rendering on the scores of simulation methods

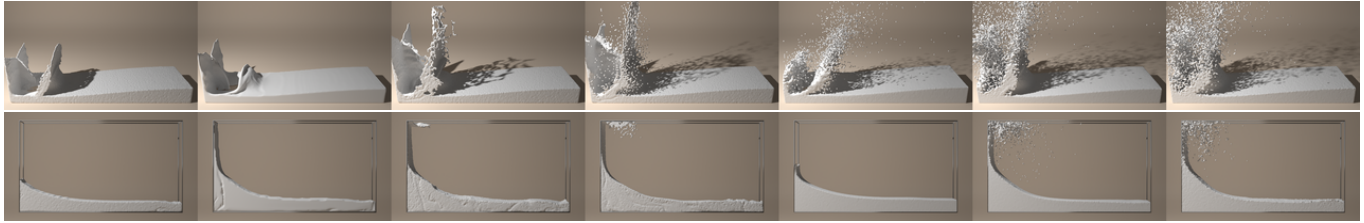


Fig. 7. Example frames of seven simulations in two examples: (top) *dam* and (bottom) *wave*. From left to right, MP, LS, FLIP, APIC, WCSPH, IISPH and SPH are shown.

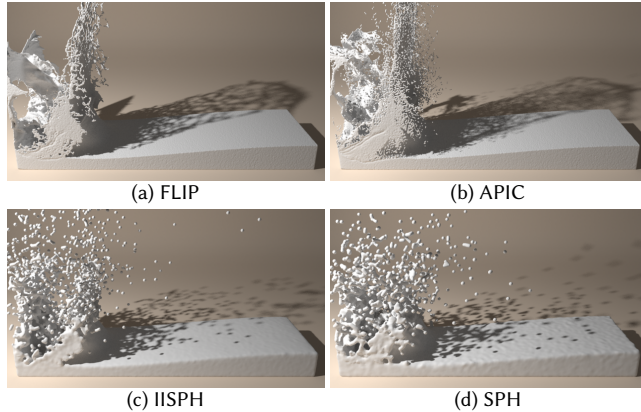


Fig. 8. Example frames of four simulations with a similar computation time.

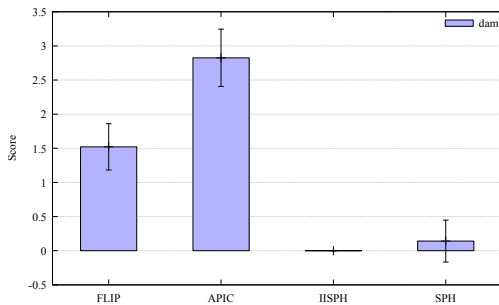


Fig. 9. Visual accuracy scores of the four videos simulated in a similar computation time (L in Table 7).

when no reference video is available. However, we found this area to be highly complex; thus, the following results are far from a complete mapping of rendering space.

In a first series of studies, we investigate the behavior of the transition between opaque and transparent rendering styles. We generated a sequence of three in-between versions by linearly blending the two styles in image space as shown in Figure 14 and performed user studies. Interestingly, the correlations between this series of studies change smoothly, albeit not linearly, when moving from opaque towards transparent. The data are shown in Figure 15. Due to the strong difference in initial results (C_2 from Table 1), we found

Table 4. Correlation analysis for the additional rendering styles.

Comparison (IDs in Table 7)	r	p-value
Opaque (A^*) vs. Glossy (H^*)	0.94329	0.00473
Opaque (A^*) vs. Translucent (I^*)	0.93170	0.00684
Transparent (B^*) vs. Glossy (H^*)	0.55867	0.24918
Transparent (B^*) vs. Translucent (I^*)	0.59764	0.21027

it surprising that the space between these two extremes behaves smoothly.

We also performed the user studies with the same setup using two additional rendering styles, which we selected to be different from both opaque and transparent styles. The first additional style is a dark-green *glossy* surface, while the second one is a *translucent* glossy volume with attenuation effects. These two rendering styles are shown in Figure 16. The correlation coefficients for these two styles with respect to our two initial styles indicate that both the *glossy* and *translucent* styles are strongly correlated with the opaque one as shown in Table 4. Note that all studies discussed in this section were performed without the reference video. The results indicate that the opaque style covers a broader range of other rendering styles by showing the strong correlations even when no reference video is given. Presumably, the transparent rendering style with its complex light effects triggers a very different “mental image” for the participants when no reference video is given. This leads to a substantially different evaluation of the videos with transparent rendering. However, note that all studies in Section 4 are conducted with the opaque rendering style and a reference since our goal is to reliably assess different methods.

6 CONCLUSIONS AND OUTLOOK

We have presented the first framework to perceptually evaluate liquid simulation methods by employing crowd-sourced user studies. By analyzing the evaluation results from controlled studies, we have demonstrated that our framework can reliably measure user opinions in the form of a visual accuracy score. Our key finding here is that the availability of a reference video makes stable evaluations possible. Most importantly, the scores are not influenced by a certain choice of rendering method.

The findings from our studies have led to several insights. For our chosen settings, the studies suggest that

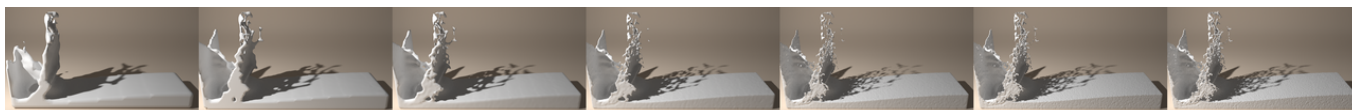


Fig. 10. Example frames of seven resolutions for particle skinning. From left to right, 0.5x, 0.75x, 1x, 1.5x, 2x, 3x, and 4x are shown.

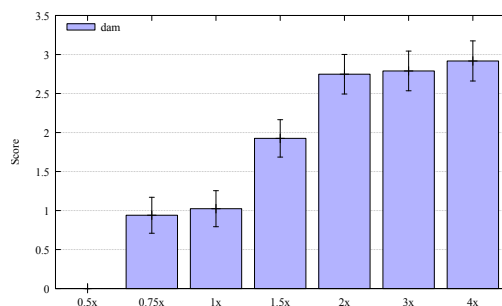


Fig. 11. Visual accuracy scores of the seven resolutions for particle skinning (M in Table 7).

- viewers prefer SPH-based methods when comparable particle counts are used,
- FLIP and especially APIC are preferred when the computational resources are limited,
- the commonly used factor of two for particle skinning is confirmed by our experiment,
- and the splash effects are an important visual component for large-scale liquids.

As the perception of physical phenomena such as liquids is highly complex, our work clearly represents only a first step. We have not investigated the demographics of our participants in more detail. Moreover, we currently focus on a specific regime of liquid flows, and it is not clear how applicable our results are for other regimes. Likewise, we have only tested a small selection of simulation methods with our studies. There are many interesting variants that could be evaluated in addition to our current selection. In the future, we are

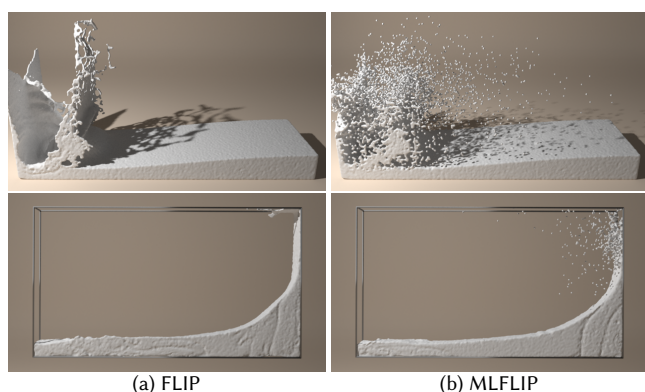


Fig. 12. Visual comparison of MLFLIP with FLIP in two examples: (top) dam and (bottom) wave.

also highly interested in extending our studies to smoke flows and other types of materials such as objects undergoing elasto-plastic deformations. As we have proposed a first perceptual evaluation framework for liquid simulation methods, we believe these directions are very interesting avenues for future work.

ACKNOWLEDGMENTS

We would like to thank all members of the graphics labs of TUM and IST Austria for the thorough discussions and the SPHERIC community for providing the experimental videos.

REFERENCES

- S. Adami, X. Y. Hu, and N. A. Adams. 2012. A generalized wall boundary condition for smoothed particle hydrodynamics. *J. Comput. Phys.* 231, 21 (Aug. 2012), 7057–7075. DOI: <https://doi.org/10.1016/j.jcp.2012.05.005>
- Bart Adams, Mark Pauly, Richard Keiser, and Leonidas J. Guibas. 2007. Adaptively Sampled Particle Fluids. *ACM Trans. Graph.* 26, 3, Article 48 (July 2007), 7 pages. DOI: <https://doi.org/10.1145/1276377.1276437>
- Nadir Akinci, Markus Ihmsen, Gizem Akinci, Barbara Solenthaler, and Matthias Teschner. 2012. Versatile Rigid-Fluid Coupling for Incompressible SPH. *ACM Trans. Graph.* 31, 4 (July 2012), 62:1–62:8. DOI: <https://doi.org/10.1145/2185520.2185558>
- Ryoichi Ando, Nils Thürey, and Reiji Tsuruno. 2012. Preserving Fluid Sheets with Adaptively Sampled Anisotropic Particles. *IEEE Transactions on Visualization and Computer Graphics* 18, 8 (2012), 1202–1214. DOI: <https://doi.org/10.1109/TVCG.2012.87>
- Ryoichi Ando, Nils Thürey, and Chris Wojtan. 2013. Highly Adaptive Liquid Simulations on Tetrahedral Meshes. *ACM Trans. Graph.* 32, 4 (July 2013), 103:1–103:10. DOI: <https://doi.org/10.1145/2461912.2461982>
- Tunç Ozan Aydın, Martin Čadik, Karol Myszkowski, and Hans-Peter Seidel. 2010. Video Quality Assessment for Computer Graphics Applications. *ACM Trans. Graph.* 29, 6 (Dec. 2010), 161:1–161:12. DOI: <https://doi.org/10.1145/1866158.1866187>
- Christopher Batty, Florence Bertails, and Robert Bridson. 2007. A Fast Variational Framework for Accurate Solid-fluid Coupling. *ACM Trans. Graph.* 26, 3, Article 100 (July 2007), 7 pages. DOI: <https://doi.org/10.1145/1276377.1276502>
- Markus Becker and Matthias Teschner. 2007. Weakly compressible SPH for free surface flows. In *Proceedings of the 2007 ACM SIGGRAPH/Eurographics symposium on*

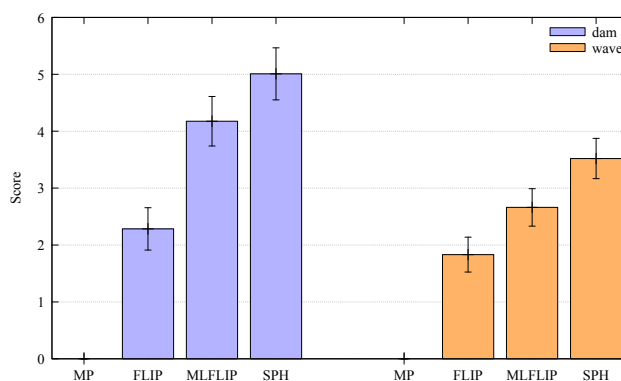


Fig. 13. Notable improvements of MLFLIP in visual accuracy in two examples (N and O in Table 7).



Fig. 14. Examples from our series of rendering styles transitioning from opaque to transparent.



Fig. 15. Correlation among the five sets of overall scores evaluated from the user studies with different rendering styles.

- Computer animation (SCA '07)*. Eurographics Association, Aire-la-Ville, Switzerland, Switzerland, 209–217. <http://dl.acm.org/citation.cfm?id=1272690.1272719>
- Micah Bojrab, Michel Abdul-Massih, and Bedrich Benes. 2013. Perceptual Importance of Lighting Phenomena in Rendering of Animated Water. *ACM Trans. Appl. Percept.* 10, 1 (March 2013), 2:1–2:18. DOI: <https://doi.org/10.1145/2422105.2422107>
- Elkin Botia-Vera, Antonio Souto-Iglesias, Gabriele Bulian, and L. Lobovský. 2010. Three SPH Novel Benchmark Test Cases for free surface flows. In *Proceedings of the 5th ERCOFTAC SPHERIC workshop on SPH applications*. Manchester, UK.
- Ralph Allan Bradley and Milton E. Terry. 1952. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika* 39, 3/4 (1952), 324–345. DOI: <https://doi.org/10.2307/2334029>
- Robert Bridson. 2015. *Fluid Simulation for Computer Graphics*. CRC Press.
- Kirsten Cater, Alan Chalmers, and Patrick Ledda. 2002. Selective Quality Rendering by Exploiting Human Inattentional Blindness: Looking but Not Seeing. In *Proceedings of the ACM Symposium on Virtual Reality Software and Technology (VRST '02)*. ACM, New York, NY, USA, 17–24. DOI: <https://doi.org/10.1145/585740.585744>
- Forrester Cole, Kevin Sanik, Doug DeCarlo, Adam Finkelstein, Thomas Funkhouser, Szymon Rusinkiewicz, and Manish Singh. 2009. How Well Do Line Drawings Depict Shape? *ACM Trans. Graph.* 28, 3, Article 28 (July 2009), 9 pages. DOI: <https://doi.org/10.1145/1531326.1531334>

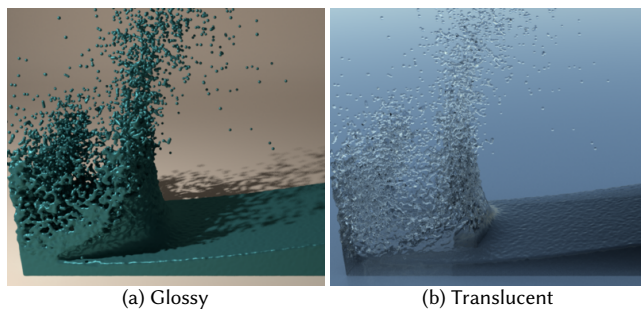


Fig. 16. Example frames of the rendering styles.

- Gilles Debunne, Mathieu Desbrun, Alan Barr, and Marie-Paule Cani. 1999. Interactive multiresolution animation of deformable models. In *Computer Animation and Simulation '99*. Springer, 133–144.
- Piotr Didyk, Elmar Eisemann, Tobias Ritschel, Karol Myszkowski, and Hans-Peter Seidel. 2010. Perceptually-motivated Real-time Temporal Upsampling of 3D Content for High-refresh-rate Displays. *Computer Graphics Forum* 29, 2 (2010), 713–722. DOI: <https://doi.org/10.1111/j.1467-8659.2009.01641.x>
- Reynald Dumont, Fabio Pellacini, and James A. Ferwerda. 2003. Perceptually-Driven Decision Theory for Interactive Realistic Rendering. *ACM Trans. Graph.* 22, 2 (April 2003), 152–181. DOI: <https://doi.org/10.1145/636886.636888>
- Douglas Enright, Ronald Fedkiw, Joel Ferziger, and Ian Mitchell. 2002. A Hybrid Particle Level Set Method for Improved Interface Capturing. *J. Comput. Phys.* 183, 1 (Nov. 2002), 83–116. DOI: <https://doi.org/10.1006/jcph.2002.7166>
- Doug Enright, Duc Nguyen, Frederic Gibou, and Ron Fedkiw. 2003. Using the Particle Level Set Method and a Second Order Accurate Pressure Boundary Condition for Free Surface Flows. In *Proceedings of 4th ASME-JSME Joint Fluids Summer Engineering Conference*, Vol. 2. 337–342. DOI: <https://doi.org/10.1115/FEDSM2003-45144>
- Florian Ferstl, Ryoichi Ando, Chris Wojtan, Rüdiger Westermann, and Nils Thuerey. 2016. Narrow band FLIP for liquid simulations. *Computer Graphics Forum* 35, 2 (2016), 225–232.
- Nick Foster and Ronald Fedkiw. 2001. Practical Animation of Liquids. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '01)*. ACM, New York, NY, USA, 23–30. DOI: <https://doi.org/10.1145/383259.383261>
- Nick Foster and Dimitri Metaxas. 1996. Realistic Animation of Liquids. *Graphical Models and Image Processing* 58, 5 (Sept. 1996), 471–483. DOI: <https://doi.org/10.1006/gmip.1996.0039>
- Dan Gerszewski and Adam W. Bargteil. 2013. Physics-Based Animation of Large-Scale Splashing Liquids. *ACM Trans. Graph.* 32, 6 (Nov. 2013), 185:1–185:6. DOI: <https://doi.org/10.1145/2508363.2508430>
- D. Han and J. Keyser. 2016. Effect of Low-Level Visual Details in Perception of Deformation. *Computer Graphics Forum* 35, 2 (May 2016), 375–383. DOI: <https://doi.org/10.1111/cgf.12839>
- Ludovic Hoyet, Kenneth Ryall, Katja Zibrek, Hwangpil Park, Jehhee Lee, Jessica Hodgins, and Carol O'Sullivan. 2013. Evaluating the Distinctiveness and Attractiveness of Human Motions on Realistic Virtual Bodies. *ACM Trans. Graph.* 32, 6 (Nov. 2013), 204:1–204:11. DOI: <https://doi.org/10.1145/2508363.2508367>
- David R. Hunter. 2004. MM algorithms for generalized Bradley-Terry models. *The Annals of Statistics* 32, 1 (Feb. 2004), 384–406. DOI: <https://doi.org/10.1214/aos/1079120141>
- Markus Ihmsen, Nadir Akinici, Gizem Akinici, and Matthias Teschner. 2012. Unified spray, foam and air bubbles for particle-based fluids. *The Visual Computer* 28, 6–8 (2012), 669–677.
- Markus Ihmsen, Jens Cornelis, Barbara Solenthaler, Christopher Horvath, and Matthias Teschner. 2014a. Implicit Incompressible SPH. *IEEE Transactions on Visualization and Computer Graphics* 20, 3 (March 2014), 426–435. DOI: <https://doi.org/10.1109/TVCG.2013.105>
- Markus Ihmsen, Jens Orthmann, Barbara Solenthaler, Andreas Kolb, and Matthias Teschner. 2014b. SPH Fluids in Computer Graphics. In *Eurographics 2014 - State of the Art Reports*. Eurographics Association, Strasbourg, France, 21–42. DOI: <https://doi.org/10.2312/egst.20141034>
- R. Issa, D. Violeau, Antonio Souto-Iglesias, and Elkin Botia-Vera. 2017. SPHERIC Validation Tests. <http://spheric-sph.org/validation-tests>. (2017).
- Chenfanfu Jiang, Craig Schroeder, Andrew Selle, Joseph Teran, and Alexey Stomakhin. 2015. The Affine Particle-in-cell Method. *ACM Trans. Graph.* 34, 4 (July 2015), 51:1–51:10. DOI: <https://doi.org/10.1145/2766996>
- ByungMoon Kim, Yingjie Liu, Ignacio Llamas, and Jarek Rossignac. 2005. FlowFixer: Using BFEC for Fluid Simulation. In *Eurographics Conference on Natural Phenomena*. Eurographics Association, Dublin, Ireland, 51–56. DOI: <https://doi.org/10.2312/NPH/NPH05/051-056>
- K. M. T. Kleefsman, G. Fekken, A. E. P. Veldman, B. Iwanowski, and B. Buchner. 2005. A Volume-of-Fluid Based Simulation Method for Wave Impact Problems. *J. Comput. Phys.* 206, 1 (June 2005), 363–393. DOI: <https://doi.org/10.1016/j.jcp.2004.12.007>

- Gondy Leroy. 2011. *Designing User Studies in Informatics*. Springer London. DOI: <https://doi.org/10.1007/978-0-85729-622-1>
- F. Losasso, J.O. Talton, N. Kwatra, and R. Fedkiw. 2008. Two-Way Coupled SPH and Particle Level Set Fluid Simulation. *IEEE Transactions on Visualization and Computer Graphics* 14, 4 (2008), 797–804. DOI: <https://doi.org/10.1109/TVCG.2008.37>
- Miles Macklin and Matthias Müller. 2013. Position Based Fluids. *ACM Trans. Graph.* 32, 4 (July 2013), 104:1–104:12. DOI: <https://doi.org/10.1145/2461912.2461984>
- Belen Masia, Sandra Agustin, Roland W. Fleming, Olga Sorkine, and Diego Gutierrez. 2009. Evaluation of Reverse Tone Mapping Through Varying Exposure Conditions. *ACM Trans. Graph.* 28, 5, Article 160 (Dec. 2009), 8 pages. DOI: <https://doi.org/10.1145/1618452.1618506>
- Rachel McDonnell, Michéal Larkin, Simon Dobbyn, Steven Collins, and Carol O’Sullivan. 2008. Clone Attack! Perception of Crowd Variety. *ACM Trans. Graph.* 27, 3, Article 26 (Aug. 2008), 8 pages. DOI: <https://doi.org/10.1145/1360612.1360625>
- Matthias Müller, David Charypar, and Markus Gross. 2003. Particle-Based Fluid Simulation for Interactive Applications. In *Proceedings of the 2003 ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA '03)*. Eurographics Association, Aire-la-Ville, Switzerland, Switzerland, 154–159.
- Zherong Pan, Jin Huang, Yiyi Tong, Changxi Zheng, and Hujun Bao. 2013. Interactive Localized Liquid Motion Editing. *ACM Trans. Graph.* 32, 6 (Nov. 2013), 184:1–184:10. DOI: <https://doi.org/10.1145/2508363.2508429>
- Karl Pearson. 1920. Notes on the History of Correlation. *Biometrika* 13, 1 (Jan. 1920), 25–45. DOI: <https://doi.org/10.1093/biomet/13.1.25>
- Daniel Ram, Theodore Gast, Chenfanfu Jiang, Craig Schroeder, Alexey Stomakhin, Joseph Teran, and Pirouz Kavehpour. 2015. A Material Point Method for Viscoelastic Fluids, Foams and Sponges. In *Proceedings of the 2015 ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA '15)*. ACM, New York, NY, USA, 157–163. DOI: <https://doi.org/10.1145/2786784.2786798>
- Karthik Raveendran, Chris Wojtan, and Greg Turk. 2011. Hybrid Smoothed Particle Hydrodynamics. In *Proceedings of the 2011 ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA '11)*. ACM, New York, NY, USA, 33–42. DOI: <https://doi.org/10.1145/2019406.2019411>
- B. Solenthaler and R. Pajarola. 2009. Predictive-corrective Incompressible SPH. *ACM Trans. Graph.* 28, 3, Article 40 (July 2009), 6 pages. DOI: <https://doi.org/10.1145/1531326.1531346>
- Jos Stam. 1999. Stable Fluids. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '99)*. ACM Press/Addison-Wesley Publishing Co., New York, NY, USA, 121–128. DOI: <https://doi.org/10.1145/311535.311548>
- Alexey Stomakhin, Craig Schroeder, Lawrence Chai, Joseph Teran, and Andrew Selle. 2013. A Material Point Method for Snow Simulation. *ACM Trans. Graph.* 32, 4 (July 2013), 102:1–102:10. DOI: <https://doi.org/10.1145/2461912.2461948>
- Kiwon Um, SeungHo Baek, and JungHyun Han. 2014. Advanced Hybrid Particle-Grid Method with Sub-Grid Particle Correction. *Computer Graphics Forum* 33, 7 (Oct. 2014), 209–218. DOI: <https://doi.org/10.1111/cgf.12489>
- Kiwon Um, Xiangyu Hu, and Nils Thuerey. 2017. Liquid Splash Modeling with Neural Networks. (2017). arXiv:1704.04456
- Yongning Zhu and Robert Bridson. 2005. Animating Sand As a Fluid. *ACM Trans. Graph.* 24, 3 (July 2005), 965–972. DOI: <https://doi.org/10.1145/1073204.1073298>

A CROWD-SOURCING PLATFORMS

There exist several crowd sourcing services that provide a web-based platform where the requester can launch user studies with a web-interface. This section compares three popular platforms: Amazon Mechanical Turk (MT), CrowdFlower (CF), and Microworkers (MW).

In order to investigate consistency of all three platforms, we use our study setup for *dam* from Section 3.3 with six different versions. In addition, we included an additional seventh dummy video, which was synthesized by interleaving the six videos for each one second; we did not include reverse questions in these three studies.

Table 5 and Figure 17 show the evaluation results from the user study run on all three platforms, and Table 6 shows the resulting correlation coefficients. As all p-values (<0.01) in the table indicate, there is significant evidence with 99% confidence to conclude that the user studies obtained on the different platforms match. Thus, when only considering the results of a single study, all three platforms yield very similar results.

However, there are noticeable differences in the cost for each study. All platforms allow the requester to set a cost for each query

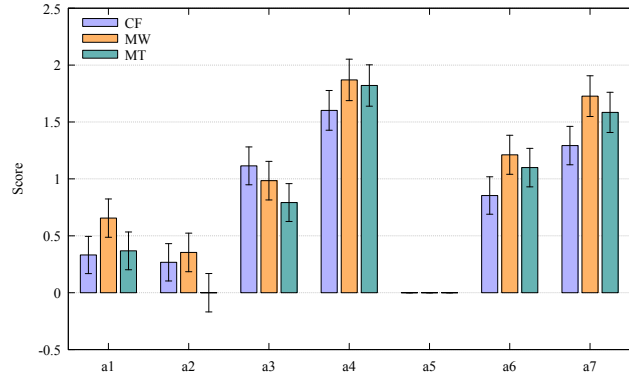


Fig. 17. Graph of seven scores evaluated from three platforms.

and the required number of participants. An additional service fee is typically charged on top of this. For our user study, we selected 50 participants for 21 queries and a per query payment of 0.01 USD, which resulted in costs of 21.00 USD for MT, 12.60 USD for CF, and 23.10 USD for MW. In addition, there were considerable differences in execution speed. With these settings, the MT platform took several weeks to complete the study, while the other two platforms yielded results in less than three days. Due to additional limitations with respect to the maximal number of queries in the CF platform, we chose the MW platform for all our studies.

Table 5. Three sets of scores evaluated from three platforms.

ID	Score (standard error)		
	CF	MW	MT
a1	0.3317 (0.1637)	0.6556 (0.1685)	0.3677 (0.1661)
a2	0.2673 (0.1640)	0.3539 (0.1693)	0.0000 (0.1687)
a3	1.1146 (0.1665)	0.9845 (0.1696)	0.7923 (0.1666)
a4	1.6024 (0.1744)	1.8701 (0.1820)	1.8208 (0.1820)
a5	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)
a6	0.8540 (0.1643)	1.2118 (0.1715)	1.0997 (0.1692)
a7	1.2931 (0.1688)	1.7273 (0.1790)	1.5849 (0.1766)

Table 6. Pearson’s correlations for the three platforms.

	CF, MW	MW, MT	MT, CF
r	0.95808	0.98596	0.95170
p-value	0.00068	0.00004	0.00096

Table 7. The visual accuracy scores (and standard errors).

Different examples and renderings (Section 3.3 and 5), where A*, B*, D*, and E*-I* denote the studies without the reference video.								
ID	Ex.	Rendering	FLIP, 1x	FLIP, 2x	FLIP, 4x	SPH, 1x	SPH, 2x	SPH, 3x
A	dam	opaque	0.0000 (0.0000)	3.1368 (0.4584)	4.6271 (0.4786)	4.9480 (0.4813)	6.5291 (0.4961)	6.7529 (0.4989)
A*	dam	opaque	0.0000 (0.0000)	1.0822 (0.1975)	1.6328 (0.2083)	0.0579 (0.1964)	0.9089 (0.1955)	1.0300 (0.1968)
B	dam	transparent	0.0000 (0.0000)	2.0498 (0.3272)	3.8288 (0.3572)	2.8715 (0.3428)	4.6016 (0.3700)	5.3260 (0.3864)
B*	dam	transparent	1.6860 (0.1785)	1.7125 (0.1789)	1.5685 (0.1765)	0.8198 (0.1694)	0.4223 (0.1695)	0.0000 (0.0000)
C	wave	opaque	0.0000 (0.0000)	3.2189 (0.4720)	3.6823 (0.4771)	3.0738 (0.4701)	5.2235 (0.4996)	5.0324 (0.4958)
			APIC, 1x	APIC, 2x	APIC, 4x	IISPH, 1x	IISPH, 2x	IISPH, 3x
D	dam	opaque	0.0000 (0.0000)	2.6095 (0.3411)	3.7208 (0.3541)	2.6466 (0.3416)	4.2966 (0.3618)	4.9892 (0.3751)
D*	dam	opaque	0.1480 (0.1816)	1.5857 (0.1857)	2.0321 (0.1933)	0.0000 (0.0000)	1.4117 (0.1835)	1.8044 (0.1890)
			FLIP, 1x	FLIP, 2x	FLIP, 4x	SPH, 1x	SPH, 2x	SPH, 3x
E*	dam	blend 25%	0.0000 (0.0000)	-0.0776 (0.1762)	0.0000 (0.1769)	-1.9924 (0.1972)	-1.4552 (0.1849)	-1.6837 (0.1895)
F*	dam	blend 50%	0.0000 (0.0000)	0.1456 (0.1629)	0.2132 (0.1636)	-1.2302 (0.1726)	-0.6418 (0.1629)	-0.7460 (0.1641)
G*	dam	blend 75%	0.0000 (0.0000)	0.8031 (0.1843)	1.1089 (0.1919)	-1.0177 (0.1914)	-0.2983 (0.1779)	-0.2034 (0.1772)
H*	dam	glossy	0.0000 (0.0000)	0.5613 (0.1489)	0.8232 (0.1524)	-0.7548 (0.1553)	0.1286 (0.1465)	0.0537 (0.1465)
I*	dam	translucent	0.0000 (0.0000)	0.8324 (0.1484)	0.8324 (0.1484)	-0.2321 (0.1456)	0.1135 (0.1437)	0.0723 (0.1438)
Different reference videos with <i>dam</i> simulations (Section 3.3).								
ID	Reference	FLIP, 1x	FLIP, 2x	FLIP, 4x	SPH, 1x	SPH, 2x	SPH, 3x	
A ⁺	seashore	0.0000 (0.0000)	1.2957 (0.1997)	2.1712 (0.2095)	1.1659 (0.1987)	2.6929 (0.2186)	2.9781 (0.2250)	
A ⁻	wave	0.0000 (0.0000)	1.3566 (0.1975)	2.0606 (0.2078)	0.3904 (0.1947)	1.7945 (0.2031)	2.2241 (0.2112)	
Seven simulation methods (Section 4.1).								
ID	Ex.	MP	LS	FLIP	APIC	WCSPH	IISPH	SPH
J	dam	0.0000 (0.0000)	0.1248 (0.1769)	2.0613 (0.1962)	3.4211 (0.2136)	2.6271 (0.2039)	4.4595 (0.2294)	4.3855 (0.2280)
K	wave	0.0000 (0.0000)	1.6646 (0.1943)	2.6871 (0.2058)	2.6987 (0.2060)	0.7209 (0.1876)	3.7943 (0.2229)	3.7943 (0.2229)
Four simulation methods for <i>dam</i> in similar computation time (Section 4.2).								
ID	FLIP	APIC	IISPH	SPH				
L	1.5215 (0.3387)	2.8256 (0.4205)	0.0000 (0.0000)	0.1410 (0.3070)				
Seven grid resolutions for particle skinning (Section 4.3).								
ID	0.5x	0.75x	1x	1.5x	2x	3x	4x	
M	0.0000 (0.0000)	0.9397 (0.2308)	1.0235 (0.2310)	1.9248 (0.2393)	2.7473 (0.2533)	2.7891 (0.2542)	2.9170 (0.2572)	
Four methods including MLFLIP (Section 4.4).								
ID	MP	FLIP	MLFLIP	SPH				
N	0.0000 (0.0000)	2.2833 (0.3723)	4.1758 (0.4353)	5.0077 (0.4569)				
O	0.0000 (0.0000)	1.8312 (0.3069)	2.6612 (0.3282)	3.5203 (0.3538)				